Application of Multivariate Gaussian Convolution and Mixture Models for Identifying Key Biomarkers Underlying Variability in Transcriptomic Profiles and the Diversity of Therapeutic Responses

Bastien Chassagnol

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

SORBONNE UNIVERSITÉ

Doctoral School **École Doctorale Sciences Mathématiques de Paris Centre**
University Department **Laboratoire de Probabilités, Statistique et Modélisation**

Thesis defended by **CHASSAGNOL BASTIEN**
Defended on **August 2, 2023**
In order to become Doctor from Sorbonne Université
Academic Field **Applied mathematics**
Speciality **Statistics and Modeling, Bioinformatics**

Application of Multivariate Gaussian Convolution and Mixture Models for Identifying Key Biomarkers Underlying Variability in Transcriptomic Profiles and the Diversity of Therapeutic Responses

Thesis supervised by	Grégory NUEL	Supervisor
	Etienne BECHT	Co-Supervisor
	Mickaël GUEDJ	Co-Monitor
	Pierre-Henri WUILLEMIN	Co-Monitor

Committee members

Referees	Julien CHIQUET	Senior Researcher at INRAE	
	Zlatko TRAJANOSKI	Senior Researcher at CNRS	
Examiners	Emeline PERTHAME	Research Engineer at Institut Pasteur	
	Stéphane ROBIN	Professor at UPMC, LPSM	Committee President
	Marie CHION	Research Associate at University of Cambridge	
Guest	Michael RERA	Senior Researcher at CRI	
Supervisors	Grégory NUEL	Professor at UPMC, LPSM	
	Etienne BECHT	Junior Researcher at Les Laboratoires Servier	
	Mickaël GUEDJ	Nanobiotix	
	Pierre-Henri WUILLEMIN	Associate Professor at UPMC, LIP6	

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

SORBONNE UNIVERSITÉ

École doctorale **École Doctorale Sciences Mathématiques de Paris Centre**
Unité de recherche **Laboratoire de Probabilités, Statistique et Modélisation**

Thèse présentée par **CHASSAGNOL BASTIEN**

Soutenue le **2 août 2023**

En vue de l'obtention du grade de docteur de l'Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Statistiques et modélisation, Bioinformatique**

Application de modèles de convolution et de mélange gaussiens pour l'identification des biomarqueurs clés sous-jacents à la variabilité des profils transcriptomiques et à la diversité des réponses thérapeutiques

Thèse dirigée par	Grégory NUEL	directeur
	Etienne BECHT	co-directeur
	Mickaël GUEDJ	co-encadrant
	Pierre-Henri WUILLEMIN	co-encadrant

Composition du jury

Rapporteurs	Julien CHIQUET	directeur de recherche à l'INRAE	
	Zlatko TRAJANOSKI	directeur de recherche au CNRS	
Examinateurs	Emeline PERTHAME	ingénieure de recherche à l'Institut Pasteur	
	Stéphane ROBIN	professeur à l'UPMC, LPSM	président du jury
	Marie CHION	chercheuse à l'University of Cambridge	
Invité	Michael RERA	directeur de recherche au CRI	
Directeurs de thèse	Grégory NUEL	professeur à l'UPMC, LPSM	
	Etienne BECHT	chargé de recherche au Les Laboratoires Servier	
	Mickaël GUEDJ	Nanobiotix	
	Pierre-Henri WUILLEMIN	MCF à l'UPMC, LIP6	

This thesis has been prepared at the following research units.

Laboratoire de Probabilités, Statistique et Modélisation

Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

☎ +33 1 57 27 93 16
Web Site https://www.lpsm.paris/

Laboratoire d' Informatique de Paris 6

Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

☎ +33 1 44 27 72 60
Web Site https://www.lip6.fr/

Les Laboratoires Servier

Institut de Recherche et Développement Servier Paris-Saclay
Rue Francis Perrin
91190 Gif-sur-Yvette
France

Web Site https://servier.com/

I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

Isaac Newton

Remerciements

En considérant un travail de thèse, on pourrait s'arrêter en première lecture au travail personnel conséquent, de longue haleine, aux courtes nuits pour de longues journées, sous la lumière du plafonnier, et aux instants de doute. Mais ce serait aussi oublier l'incroyable aventure humaine que ce travail représente, la richesse des échanges, qu'ils soient scientifiques, politiques ou plus grégaires, et les joies partagées face aux épreuves affrontées avec succès.

Il est temps de rendre la monnaie de leurs pièces, à toutes ces personnes ayant contribué au succès de cette thèse. Et tout d'abord, je dois remercier la pièce centrale de l'échiquier complexe des interactions requises par la thèse, en la personne d'Etienne BECHT. Mon encadrant industriel a fait plus, bien plus que superviser mon travail de recherche. Il a été présent dans mes moments de doute, il m'a permis d'accomplir de nombreux progrès, que ce soit certes académiques, mais aussi communicationnelles, artistiques et humains (dont, non des moindres, la découverte des œuvres de Mozart, et un intérêt renouvelé pour les chorales grégoriennes). Je me dois aussi de remercier Grégory NUEL, mon responsable académique, d'avoir relu mes différents travaux, et m'avoir sorti de l'ornière et éclairé les nombreux défis statistiques que j'ai pu rencontrer, en explorant des voies totalement innovantes sur l'étude de données biologiques. Une pensée enfin pour mes deux autres encadrants, Mickaël GUEDJ, pour avoir su déniché en moi une propension à me confronter à des problèmes complexes, et m'avoir guidé sur une période charnière de ma thèse, et Pierre-Henri WUILLEMIN, pour sa gentillesse et son accueil au bureau du LIP6 que je sais être ouvert pour mes condoléances, en espérant que je pourrai un jour revenir aux réseaux bayésiens [1].

Je souhaiterai aussi remercier tous les membres de mon jury de thèse, à commencer par les rapporteurs de mon long (trop long certainement) manuscrit de thèse, que ce soit Zlato Trajanoski, alors même que le sujet présentait une composante statistique conséquente, et Julien CHIQUET pour son travail considérable (et de façon infortune double) de relecture, présentant de nombreux commentaires précieux pour la suite de mon aventure dans le monde de la recherche, et même des pointes d'humour et de style. Un grand merci aussi à Emeline Perthame (à qui je dois la découverte du milieu souterrain de Paris, une aventure qui j'espère est loin de signer la première), à Marie CHION (pour avoir inlassablement traqué les erreurs d'assignation dans la page de garde de mon manuscrit), et à Stéphane ROBIN pour sa gentillesse et son sens du devoir, ainsi qu'à notre invité mystère du jour, Michael rera, et sa doctorante très enthousiaste, savandara.

Evidemment, il me faut aussi remercier les doctorants que j'ai pu côtoyés tout au long de cette thèse, en commençant par ceux de Sorbonne Université (puisqu'après tout, c'est pour eux que je me retrouve à rédiger cette partie des remerciements à l'arrache, temps que j'aurai mieux fait de consacrer à ma préparation de soutenance). Et tout d'abord, les vaillants du bureau du « froid » (merci à Lucas Brx par ailleurs, pour m'avoir fait découvrir cette hiérarchie de bureaux dans ta partie remerciements), avec une attention spéciale à Lucas D (préparez-vous, il y'en une

[1] Ce que je pense fort probable, puisque de la même manière que tous les chemins mènent à Rome, toutes les démarches statistiques intégrant une part d'explicabilité aboutissent aux graphes probabilistes

palanquée), mon co-doctorant académique et nos nombreuses discussions sur de multiples sujets, de la descente tragique de l'ASM au défi de maintenir un régime sportif, en passant par des discussions philosophiques profondes sur le sens de la vie, de l'amitié, des relations amoureuses et j'en passe, à Alexandra ma seconde co-doctorante partie trop vite et dont les accessoires (une tente Quechua entre autres, du moins sa partie extérieure) ont égayé notre intérieur tristounet, à Jérôme pour son talent de comédien caché, avec le sketch toujours renouvelé de la rénovation des cabinets, à Elias pour nos discussions géopolitiques approfondies, en langage des signes à défaut de comprendre l'auvergnant, à Maxence, cet ange tout droit tombé du ciel et dont le sourire permanent serait à même de briser toutes les carapaces, à Nadège, avec qui j'ai enfin eu l'honneur de déjeuner au CROUS, et à Vladimir, une guimauve dans un corps d'ours. Il me faut ensuite, par ordre de continuité, remercier ces gens bizarres, fort sympathiques, mais traitant d'objets abscons (je soupçonne d'ailleurs que tous ces symboles ésotériques cachent une secte aux motivations obscures), à savoir la caste des probabilistes. Tiens, se serait que l'intitulé de leur exposé de ce matin : une démarche, un peu balourde, d'un éléphant se déplaçant dans une grille en 2D, et atteint d'un Alzheimer sélectif. Tout mon soutien à Yohan (Johan?, pour diverses raisons il semblerait qu'il y'ait un manque de consensus sur l'écriture de ce prénom, appelons-le Tardy(grade), un arthropode connu pour son exceptionnelle capacité de résistance) pour ton rire facile (et nous donner l'impression d'être de grands humoristes dans ce labo), à Emilien pour nos discussions méridiennes d'une grande profondeur, sur l'art notamment de déposer délicatement les guirlandes sur l'arbre de Noël et ses récits épiques de traque au patou dans les Pyrénées, ou dans un tout autre registre, de la meilleure manière de prendre soin d'une haie de maison, à Nicolas au grand cœur, qui au-delà d'un aspect général raspoutinien, est un homme ayant de nombreuses cordes à son arc, digne descendant d'Henri Poincaré et curieux de tous les champs d'exploration mathématiques (désolé encore de t'avoir infiniment déçu avec mes crêpes au pâté), à Sonia la gentille succube, pour ces nombreuses sessions de psychothérapie (et m'avoir enfin détrôné du statut de plus gros beauf du labo), à Jean David pour ses histoires rocambolesques impliquant entre autres une nouvelle stratégie, à base de diffusion de tracts amoureux et à même de détrôner les applis de rencontre, et la découverte du produit diamant, à David tout court, qui a découvert mon secret, à savoir mes racines arabes (ou gitanes, ou..., compléter la nationalité qui vous sied le plus), à Guillaume pour m'avoir fait le plus beau compliment que l'on ne m'ait jamais adressé, à Loic le loustic (ou loubard) pour sa bonhommie permanente et son style rétro à la Freddy Mercury, à Antonio pour son vif échange sur l'intérêt respectif d'avoir un chat ou un chien dans sa vie et à Robin, pour avoir définitivement montré à Sonia, que le street fight, c'est une histoire de gros bras.

Et j'en oublierai encore, si je ne mentionnais les statisticiens, dont, apparemment je ferai partie. Encore merci donc à Ulysse et à son ancienne directrice Magalie pour tous ces verres partagés à Bruxelles et sa définition de la p-value, à Anna Bonnet pour m'avoir fait découvrir cette belle initiative de cours à destination des déshérités de notre système aveugle. Une pensée aussi aux petites mains de notre laboratoire, que l'on ne remercie sans doute jamais assez à hauteur de leur contribution. Et donc notamment aux différents membres du secrétariat pour leur implication dans le financement et l'organisation des événements de la vie du labo, et à Hughes Moretto, de son blase HM, pour sa patience toujours renouvelée face à notre lenteur de la stricte application des Maj de nos systèmes.

Un petit passage par le milieu industriel, je souhaite adresser mes remerciements, pour les conseils techniques et éclairés des équipes NITA et QP, dans des domaines aussi variés et passionnants que l'immuno-inflammation, la déconvolution cellulaire, la médecine personnalisée ou encore le repositionnement de médicaments, ... Une pensée toute particulière à Philippe Moingeon et à Laurence Laigle pour s'être efforcés de comprendre mes projets statistiques, et leur compréhension globale des enjeux du développement de méthodes innovantes pour accéder à

une compréhension holistique de maladies hétérogènes, ainsi qu'à Emiko et Sandra, pour m'avoir encore un peu gardé en vie au croisement d'un carrefour.

Au sein de l'équipe de médecine computationnelle, je souhaitais remercier en particulier Céline Lefebvre, qui a su intégrer notre équipe au sein de son groupe, une tâche fort périlleuse dans le milieu industriel, à Antoine BICHAT pour nos nombreux échanges de geek sur la meilleure utilisation des packages de R, et l'application des pratiques FAIR dans ce langage, à Xavier pour nos discussions sur l'intérêt de déployer des ontologies cellulaires, à Sophie pour le vif intérêt porté à mes projets de recherche, à Yufei pour son aide sur le déploiement d'un pipeline standardisé pour l'application de méthodes de déconvolution cellulaire, à Antoine pour nos conversations de commère sur les couples en formation à Servier, à Fabien pour sa positive-attitude toujours renouvelé, à Daniel, Sahar, Jérémy, Perrine...

Plus particulièrement, je souhaiterai exprimer ma gratitude à l'égard de ma seconde co-thésarde cœur, Cheima BOUDJENIBA, pour sa sincérité sans ambages et son sens de l'accueil et de l'hospitalité and my third (2.7 seemingly) co-PhD student, namely Nanna BARNKOP, for its renewed hospitality and welcomeness. Il me reste à exprimer ma reconnaissance au travail prodigieux accompli par la communauté ShineDocs des doctorants et post-docs, menée par leur éternelle figure de proue Selena BOUFFETTE, tant sur le plan de la production et de la vulgarisation scientifiques que pour leur organisation d'événements et échanges informels entre jeunes chercheurs. Je remercie notamment cette dernière pour avoir stimulé de façon inégalé les échanges entre des communautés qui ne se seraient jamais parlé sans son intervention, et de façon plus personnelle, pour nos échanges passionnels dans des domaines aussi variés que la littérature (merci encore de m'avoir fait découvrir tout un pan de la culture slave), la géopolitique ou la biologie. Une pensée aussi pour ces autres membres, à Agathe pour avoir fait découvrir au plouc du groupe les endroits huppés de la capitale, à (E)(stéphania) pour son petit accent péruvien trop mignon, Chloé la seconde vraie socialiste de notre groupe, à Zanou la gazelle du désert, Thomas le dieu du stade (enfin, tout du moins à l'échelle du labo des chimistes), et Amandine.

Enfin, je voudrais savoir gré de l'appétence scientifique sans cesse renouvelée et des interrogations pas si naïves des stagiaires et alternants de nos équipes, nous poussant à creuser sans cesse les failles de nos modèles existants pour mieux les dépasser.

Au-delà des frontières du monde professionnel, qui s'avèrent perméables comme vous avez pu le constater dans le corpus précédent, je souhaiterai remercier ma copine, mon bébé, ma petite Sarah au cœur fragile, mais à la tête de bois. Ta détermination, ta ténacité et ton courage m'ont toujours, et continueront de me guider, vers un choix non-carriériste ou vénale de ma vie professionnelle, mais bien plutôt vers une décision du cœur et humaniste. Et j'espère que tous ces échanges, toutes ces aventures qu'on a vécues et traversées ensemble, de la difficulté à communiquer avec nos encadrants à la traversée du Sancy en raquette, ne signent que le début d'une longue et douce histoire, avec ton patapouf préféré. Je souhaiterai aussi remercier mes amis, d'école d'ingé avec qui j'ai partagé les 400 coups, romain et amaury pour nos tête-à-tête au Ninkasi, Julie et sa passion rock and roll, Paul pour son flegme légendaire, digne de la trame d'un leader et Aurélie pour l'avoir enfin rendu présentable, contrairement au clown, Ubuesque Kaelan et son bagout légendaire, Alexis le poète et joueur de tam-tam, Pierre pour ses créations culinaires de grande qualité (et notre fournisseur alimentaire lors des randonnées), Jürgen le techno-solutionniste, le groupe des bretons, avec au premier plan le brave Nicolas (qui a compté une infinité de fois jusqu'à l'infini, alors que cheuck norris, ce gros nul, il ne l'a fait que quatre fois), Pierre pas C le tibo in shape de la bande, la maman Mylène, le brave et courage Samuel, si prompt à organiser des événements festifs, le filou et la future JK Rowling du roman de science-fiction Marine. Du coté auvergnat, encore un grand merci à Mehdi et Quentin pour leur art de la table (et notamment une discussion passionnante sur l'art de bien préparer des frites), Justin pour ses clips de rap d'une grande qualité, Lou-Mary pour ses anecdotes médicales, Baptiste pour ses

envolées lyriques et son partage de la vie de galérien d'un thésard, Anne pour ses gros biscottos, Marie Camille pour ses anecdotes sur la vie montluçonnaise,

Un petit mot pour mes équipes de rugby, que ce soit de touch (et notamment au coach Kewwin, pour avoir essayé de m'inculquer l'esprit intrinsèquement collaboratif de ce sport, et à Etienne pour avoir détricoté tout cela dès notre premier tournoi), et du côté des cocks of the tiger, un grand merci aux joueurs et notamment aux anciens pour tenter de m'inculquer dans ma tête de bois des rudiments du sport, encore merci notamment à Oui Oui pour la dance de la mouette, aux ronds de C et à Ali pour avoir pris sous votre aile (ou plutôt en guise de pilier) le gringalet moineau que je suis, Bizuth pour ses gros plats de viande et son super appart du 15ème et Paddy pour être le dernier socialiste à avoir voté pour Hidalgo à la présidentielle (ou au moins le prétendre). (bon là, en vrai, il a fallu que je parte imprimer mon manuscrit, la suite des remerciements et leur plus grande personnalisation sera dans le prochain DLC sur Zenodo).

Et pour terminer, je souhaiterai remercier ma famille, à commencer par mes parents qui m'ont soutenu tout le long de cette thèse, tout en s'efforçant de comprendre mon sujet fort éloigné de leurs domaines de compétence. Et pourtant, c'est une allégorie culinaire qui a rencontré le plus de succès, et m'a permis de vulgariser le processus complexe de la déconvolution, en une image parlante pour tous. Je souhaitais remercier aussi mon frère, pour ces parties de tennis endiablées (et essayer de s'adapter à mon niveau fort élémentaire), et pour qui je souhaite le même parcours universitaire que le mien. Toutes mes félicitations à ma sœur aussi, pour avoir su conjuguer ses motivations personnelles et son sens de l'engagement civique et écologique avec une mission marketing, un défi semblant à première vue insolvable. Une pensée aussi aux autres membres de ma famille, à mon cousin Florian, qui s'avère le plus sérieux et le plus engagé de la famille, à Manu et son sens du devoir maternel, à Cathy auquel je souhaite tout le succès du monde pour sa carrière d'ingénieure, à Marion l'artiste de la famille, à Agnès pour sa gentillesse et Cécile pour son bagout et son sens de l'entraide (et sans qui j'aurai dû dormir sous les ponts en fin de stage), à Papy Aimé pour avoir maintenu envers et contre tous nos grandes tablées familiales, et à Régis et nos conversations passionnantes sur les enjeux de l'IA, impactant tous les domaines, y compris agricoles. Finalement, vient le temps de conclure ce panégyrique sur une note plus mélancolique, en exprimant une dernière pensée pour ma mamie Annie, dont je suis certain que de là-haut, elle veille encore à la réussite de mes projets, et me soutient dans mes rêves lorsque je suis plongé dans l'obscurité. Je n'ai malheureusement pas pu/su trouvé le traitement qui aurait pu te guérir, mais saches que je prévois, par cette cicatrice, de trouver le chemin vers une meilleure compréhension, et une meilleure personnalisation des traitements thérapeutiques actuels, pour qu'un jour, plus personne n'ait à subir ton calvaire, ou celui de ta grande amie et confidente, la généreuse Marithée.

Publications

B. Chassagnol, G. Nuel, and E. Becht, "An updated State-of-the-Art Overview of transcriptomic Deconvolution Methods". Oct. 13, 2023. *aRxiv*, https://cnrs.hal.science/hal-04253032

B. Chassagnol, G. Nuel, and E. Becht, "DeCovarT, a multidimensional probalistic model for the deconvolution of heterogeneous transcriptomic samples". Feb. 1, 2024. *aRxiv*, https://cnrs.hal.science/hal-04208010

Boudjeniba, Cheïma, Perrine Soret, Diana Trutschel, Antoine Hamon, Valentin Baloche, Bastien Chassagnol, Emiko Desvaux, et al. 'Consensus Gene Modules Strategy Identifies Candidate Blood-Based Biomarkers for Primary Sjögren's Disease'. *medRxiv*, 6 July 2023. https://doi.org/10.1101/2023.07.05.23292036

Chassagnol, Bastien, Bichat Antoine, Boudjeniba Cheima, Wuillemin Pierre-Henri, Gohel David, Guedj Mickaël, Nuel Grégory, and Becht Etienne. 'Gaussian Mixtures in R'. *R Journal*, November, 2023 https://journal.r-project.org/articles/RJ-2023-043.

Desvaux, Emiko, Antoine Hamon, Sandra Hubert, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Audrey Aussy, et al. 'Network-Based Repurposing Identifies Anti-Alarmins as Drug Candidates to Control Severe Lung Inflammation in COVID-19'. *PLOS ONE* 16, no. 7 (juil 2021), https://doi.org/10.1371/journal.pone.0254374.

Soret, Perrine, Christelle Le Dantec, Emiko Desvaux, Nathan Foulquier, Bastien Chassagnol, Sandra Hubert, Christophe Jamin, et al. 'A New Molecular Classification to Drive Precision Treatment Strategies in Primary Sjögren's Syndrome'. *Nature Communications 12*, no. 1 (10 June 2021): 3523. https://doi.org/10.1038/s41467-021-23472-7.

Oral communications

Peer-reviewed communications

Bastien, Chassagnol, Nuel Gregory, and Becht Etienne. 'DeCovarT: Robust deconvolution of cell mixture in transcriptomic samples by leveraging cross-talk between genes'. Talk of 15 minutes presented at the Journées de Statistique 2023, *54ems Journées de Statistique de la SFdS*, Université libre de Bruxelles, Belgique, 7 July 2023. https://drive.google.com/file/d/15cWwt7daYMu6ppwnakWyDNTeaw80tm7I/view?pli=1.

Bastien, Chassagnol, and Becht Etienne. 'DeCovarT, a Holistic Deconvolution Algorithm to Explore Highly Heterogenous Tissues'. PhD in 3 minutes presented at the *Seedpods Day* 2023, Servier Paris-Saclay R&D Institute, 29 June 2023. https://servier.com/en/newsroom/seedpods-day-boosting-innovation-post-doc-PhD-researchers

Bastien, Chassagnol, Nuel Gregory, and Becht Etienne. 'DeCovarT: Robust deconvolution of cell mixture in transcriptomic samples by leveraging cross-talk between genes, Prix de la meilleure communication orale.' Talk of 15 minutes presented at the *JOBIM 2023*, Online, 27 June 2023. https://www.youtube.com/@sfbisocietefrancaisedebioi5270/videos

Bastien, Chassagnol, Nuel Gregory, and Becht Etienne. 'DeCovarT, a R Package for a Robust Deconvolution of Cell Mixture in Transcriptomic Samples Using a Multivariate Gaussian Generative Framework, Sciencesconf.Org:Rr2023:468014'. Lightning, short talk of 5 minutes presented at the *Rencontres R*, Avignon Université, 21 June 2023. https://github.com/Rencontres-R/Rencontres_R_2023/blob/main/Presentations/1_Mercredi/3_Lightning_I/1_rencontresR_bastien_chassagnol_short_talk.pdf

Bastien, Chassagnol, Bichat Antoine, Nuel Gregory, and Becht Etienne. 'Clustering-Based Models in R, with Application on Univariate Gaussian Mixtures: Review, Evaluation and Extensions'. Talk of 15 minutes presented at the *Rencontres R*, Paris, AgroParisTech, 13 July 2021. https://rr2021.sciencesconf.org/356189

Invited seminars

Bastien, Chassagnol, Nuel Gregory, and Becht Etienne. *DeCovarT: Robust deconvolution method to reconstitute the cellular composition of the tumoral micro environment (TME).* Talk of 10 minutes, presented at *Computational Systems Biology of Cancer*, Institut Curie, Paris, 29 September 2023. Reward for best PhD presentation. https://training.institut-curie.org/courses/sysbiocancer2022.

Bastien, Chassagnol, Becht Etienne, and Nuel Gregory. 'Présentation du package R DeCovarT, en accès libre, pour l'estimation robuste de la composition cellulaire à partir de données transcriptomiques'. Highlight of 20 minutes presented at Conference DOME (Données Omiques Médecine de précision et Empowerment), 20 September 2023, as part of 'Société Française de Médecine Prédictive et Personnalisée' congress, at Paris. https://9eme-congres-de-la-sfmpp.site.calypso-event.net/

Bastien, Chassagnol, Becht Etienne, and Nuel Gregory. 'Robust deconvolution of transcriptomic samples using the gene covariance structure.' Highlight of 20 minutes presented at the *StatOmique 2022*, Agro Paris-Saclay, 22 place de l'Agronomie à Palaiseau, 25 October 2022. https://statomique.netlify.app/media/StatOmique2022-presBChassagnol.pdf

Bastien, Chassagnol, Yufei, Luo, and Becht Etienne. 'Robust deconvolution of transcriptomic samples using the gene covariance structure'. Talk of 20 minutes presented at the Conférence *IA et santé*, Centre Henri Lebesgue, Nantes, 29 June 2022. https://www.lebesgue.fr/fr/node/4708

Poster sessions

Bastien, Chassagnol, Yufei, Luo, Nuel Gregory, and Becht Etienne. 'Overview of DeCovarT, a Holistic Method for the Deconvolution of Heterogeneous Transcriptomic Samples, Obtention d'une Bourse Du GDR BIM'. Poster presented at the *ISMB-ECCB 2023, The 31st Annual Intelligent Systems For Molecular Biology and the 22nd Annual European Conference on Computational Biology*, Online and palais des Congrès, Lyon, France, 24 July 2023. https://iscb.junolive.co/ismb2023/live/exhibitor/ismbeccb2023_poster_1420

Bastien, Chassagnol, Nuel Gregory, and Becht Etienne. 'Robust deconvolution of transcriptomic samples using the gene covariance structure'. Poster presented at the *JOBIM 2022*, Université de Rennes 1, France, 5 July 2022. https://jobim2022.sciencesconf.org/resource

Application of Multivariate Gaussian Convolution and Mixture Models for Identifying Key Biomarkers Underlying Variability in Transcriptomic Profiles and the Diversity of Therapeutic Responses

Abstract

The diversity of phenotypes and conditions observed within the human species is driven by multiple intertwined biological processes. However, in the context of personalized medicine and the treatment of increasingly complex, systemic, and heterogeneous diseases, it is crucial to develop approaches that comprehensively capture the complexity of the biological mechanisms underlying the variability in biological profiles. This spans from the individual level to the cellular level, encompassing tissues and organs. Such granularity and precision are essential for clinicians, biologists, and statisticians to understand the underlying causes of the diversity in responses to clinical treatments and predict potential adverse effects.

This manuscript primarily focuses on two biological entities of interest, namely transcriptome profiles and immune cell populations, for dissecting the diversity of disease outcomes and responses to treatment observed across individuals. The introductory section provides a comprehensive overview on the intertwined mechanisms controlling the activity and abundance of these inputs, and subsequently details standard physical methods for quantifying them in real-world conditions.
To comprehensively address the intricate multi-layered organization of biological systems, we considered two distinct resolution scopes in this manuscript. At the lowest level of granularity, referred to in this manuscript as an "endotype" we examine variations in the overall bulk expression profiles across individuals. To account for the unexplained variability observed among patients sharing the same disease, we introduce an underlying latent discrete factor. To identify the unobserved subgroups characterized by this hidden variable, we employ a mixture model-based approach, assuming that each individual transcriptomic profile is sampled from a multivariate Gaussian distribution.
Subsequently, we delve into a bigger layer of complexity, by integrating the cellular composition of heterogeneous tissues. Specifically, we discuss various deconvolution techniques designed to estimate the ratios of cellular populations, contributing in unknown proportions to the total observed bulk transcriptome. We then introduce an independent deconvolution algorithm, `DeCovarT`, which demonstrates improved accuracy in delineating highly correlated cell types by explicitly incorporating the co-expression network structures of each purified cell type.

Keywords: gaussian mixture models, cellular deconvolution, transcriptome pipeline, drug repositioning

Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France
Laboratoire d' Informatique de Paris 6 – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Application de modèles de convolution et de mélange gaussiens pour l'identification des biomarqueurs clés sous-jacents à la variabilité des profils transcriptomiques et à la diversité des réponses thérapeutiques

Abrégé

La diversité des phénotypes et des conditions observées au sein de l'espèce humaine est le résultat de multiples processus biologiques interdépendants. Cependant, dans le contexte de la médecine personnalisée et du traitement de maladies de plus en plus complexes, systématiques et hétérogènes, il est crucial de développer des approches qui capturent de manière exhaustive la complexité des mécanismes biologiques sous-jacents à la variabilité des profils biologiques. Cela s'étend du niveau individuel au niveau cellulaire, englobant les tissus et les organes. Une telle précision et une telle granularité sont essentielles pour que les cliniciens, les biologistes et les statisticiens comprennent les causes sous-jacentes de la diversité des réponses aux traitements cliniques et puissent prédire d'éventuels effets indésirables.

Afin d'aborder de manière exhaustive la complexité hiérarchique et stratifiée des systèmes biologiques, nous avons considéré deux niveaux d'étude dans ce manuscrit. Au niveau de granularité le plus bas, désigné dans ce manuscrit sous le terme "endotype", nous examinons les processus conduisant aux variations observées ans les profils d'expression transcriptomiques entre individus. Notamment, pour tenir compte de la variabilité non expliquée observée entre patients affectés par la même maladie, nous introduisons une variable latente discrète. Pour identifier les sous-groupes non observés, dépendant de cette variable cachée, nous utilisons des modèles de mélange probabilistes, en supposant que chaque profil transcriptomique individuel est échantillonné à partir d'une distribution gaussienne multivariée, dont les paramètres ne peuvent pas être directement estimés dans la population générale.
Ensuite, nous nous intéressons à un niveau de complexité supplémentaire, en passant en revue les méthodes canoniques permettant d'estimer la composition des tissus, souvent très hétérogènes, au sein d'un même individu. Plus précisément, nous discutons de diverses techniques de déconvolution conçues pour estimer les ratios de populations cellulaires, ces dernières contribuant en proportions inconnues au profil transcriptomique global mesuré. Nous présentons ensuite notre propre algorithme de déconvolution, nommé "DeCovarT", qui offre une précision améliorée de la délimitation de populations cellulaires fortement corrélées, en incorporant explicitement les réseaux de co-expression propres à chaque type cellulaire purifié.

Mots clés : modèles de mélange gaussiens, déconvolution cellulaire, filière de traitement de données transcriptomiques, repositionnement de médicaments

Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France
Laboratoire d' Informatique de Paris 6 – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Résumé long de la thèse

La diversité des réponses thérapeutiques à un même traitement, au sein de patients affectés par la même maladie, résulte d'interactions complexes entre de nombreuses entités biologiques. Or, dans le cadre du paradigme de la médecine personnalisée, associée à la prise en charge de maladies complexes et hétérogènes et avec l'explosion du volume de données biologiques fournies par des méthodes de pointe, l'utilisation d'outils statistiques dédiés est devenu indispensable pour embrasser une approche holistique et robuste des mécanismes sous-jacents de la variabilité inter-, and intra-individuelle, des systèmes biologiques.

Pour ce faire, nous allons décomposer la tâche intrinsèquement complexe de l'identification des acteurs causaux de la variabilité biologique, en considérant deux niveaux d'organisation biologiques. Soit, au niveau le plus général, l'étude de la diversité au sein d'une cohorte de patients, quantifiée par leurs profils d'expression transcriptomiques globaux. Puis, à une strate plus fine, nous allons nous intéresser aux méthodes pour caractériser l'hétérogénéité au sein d'un même échantillon biologique, telle que mesurée par ses composantes cellulaires (voir le résumé graphique, 1).

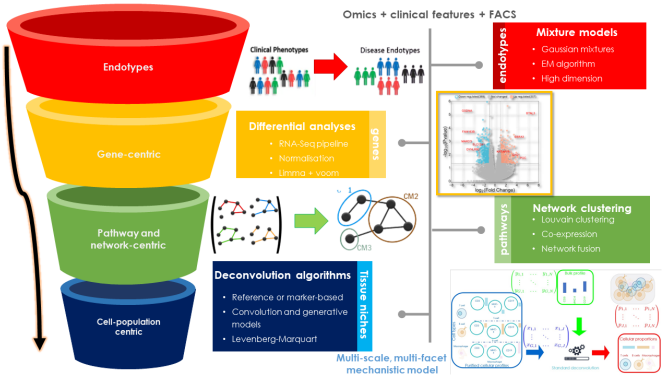


Figure 1: Résumé visuel du manuscrit de thèse. Nb: les deux strates intermédiaires du diagramme en entonnoir ne sont évoqués qu'en annexe.

Nous proposons en complément la figure 2 replaçant les acteurs biologiques présentés auparavant dans la hiérarchie communément admise de la “biologie des systèmes”. Cette dernière expose trois grandes sources biologiques distinctes contribuant à la diversité des profils transcriptomiques obervés entre les individus, voire au sein d’un même tissu, ordonnés par niveau hiérarchique: l’environnement ou la condition phénotypique (stade d’avancement d’une maladie donnée, localisation du tissu prélevé, etc.), les profils de mutations génétiques (haplotypes) et les altérations de la composition cellulaire.

Enfin, il convient de souligner que le signal biologique d’intérêt pour le biologiste est souvent dégradé par la présence d’artefacts introduits par des biais techniques (erreurs humaines de manipulations, variations des protocoles entre centres de séquençage, etc.), potentiellement inconnus.

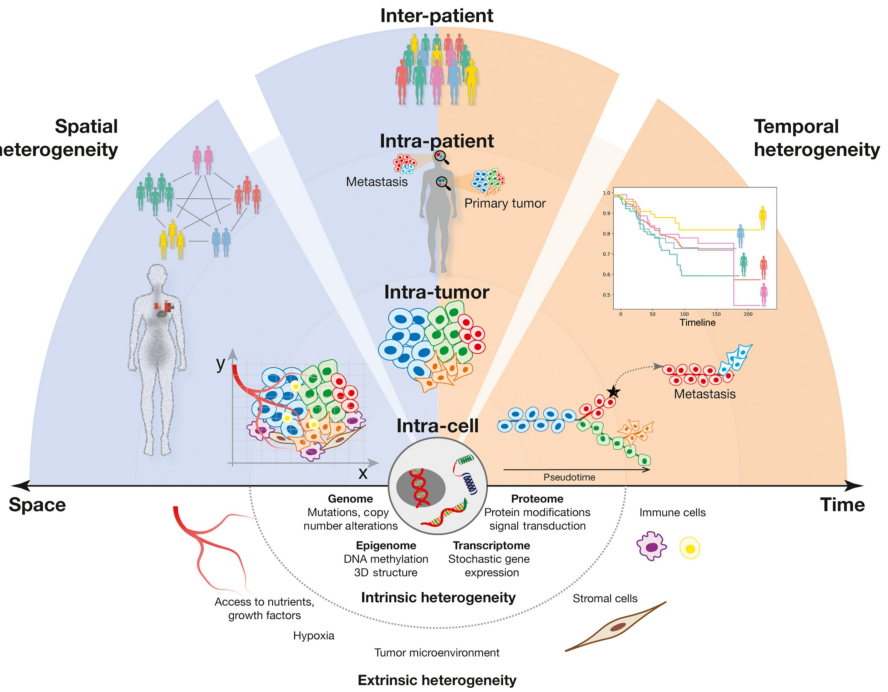


Figure 2: La diversité des profils moléculaires, entre individus, voire au sein d’un même tissu, d’une combinaisons de facteurs *intrinsèques*, et *extrinsèques*. Les sources intrinsèques incluent les mécanismes génétiques, transcriptionnels et protéomiques, généralement de nature stochastique. Les sources extrinsèques englobent entre autres les interactions entre les populations cellulaires résidant habituellement dans le tissu et l’environnement extérieur proche. Dans la figure, ces dernières sont illustrées par les échanges complexes tissés entre les populations immunitaires circulant dans le système lymphatique, et les cellules épithéliales ou mutées appartenant au clone tumoral. La figure est reproduite de [Kas+22b, Fig. 1].

Partie I: Introduction Biologique

Chapitre 1: Étude du transcriptome Le transcriptome est l'ensemble des transcrits, c'est à dire des portions d'ADN effectivement convertis en ARNm, mesurés dans un échantillon biologique donné. L'étude des variations des profils d'expression génique, au sein d'individus affectés par la même maladie permet entre autres d'identifier les principaux facteurs responsables de la défaillance des mécanismes de régulation de l'expression transcriptomique, souvent à l'origine de l'apparition d'états pathologiques.

Nous passons en revue dans cette section différents mécanismes de régulation du transcriptome, en soulignant que ces derniers agissent potentiellement sur l'ensemble du processus conduisant d'une matrice nucléaire ADN initiale à la synthèse d'une protéine fonctionnelle. Parmi ces derniers, nous soulignons notamment l'importance des réseaux de co-expression impliquant des facteurs transcriptomiques. Ces derniers agissent finement sur le niveau d'expression d'un gène, en contrôlant l'accessibilité du génome par l'ARN polymérase chargé de transcrire l'ADN en ARNm. Nous verrons ensuite dans les autres chapitres comment la prise en compte explicite de ces réseaux permet une identification plus fine et robuste des acteurs biologiques responsables de la diversité moléculaire des individus.

Nous présentons ensuite un ensemble de méthodes physiques dédiées à quantifier le transcriptome. Nous nous focalisons notamment sur les nouvelles technologies de séquençage (NGS). L'intérêt de cette nouvelle génération de méthodes réside principalement en leur flexibilité à répondre à un grand nombre de questions biologiques complexes, auparavent non résolues par les méthodes pionnières de puces à ARN. Nous montrerons toutefois que cette flexibilité vient avec un coût analytique complexe, requérant le développement de routines bio-informatiques automatisées de traitement de données, afin d'assurer la reproductibilité des analyses.

Chapitre 2: Introduction au système immunitaire Le système immunitaire joue un rôle central dans la défense de l'organisme contre les agents pathogènes et la régulation de l'homéostasie. Nous montrons notamment que les voies de signalisation transcriptomiques contrôlent finement les interactions entre les populations cellulaires, jouent un rôle pivot dans la flexibilité et l'efficacité du système immunitaire humain. Notamment, nous mettons en exergue l'importance de la coopération entre les acteurs de l'immunité inné, généralistes mais peu spécifiques, et les acteurs de l'immunité adaptative. Nous montrons toutefois qu'un accroc dans cette mécanique bien huilée peut entraîner une dérégulation du système immunitaire, et dans le pire scénario, conduire selon la situation à la prolifération incontrôlée de métastases ou au développement de maladies auto-immunes handicapantes.

Ensuite, nous exposons une série de techniques physiques pour estimer la composition immunitaire d'un échantillon biologique. Ces méthodes sont généralement catégorisées en deux groupes : les méthodes de cytométrie de flux, qui nécessitent la séparation physique des populations cellulaires, et les méthodes d'imagerie, qui permettent l'identification des types cellulaires *in-situ*, à l'aide de marqueurs fluorescents. Si les premières se caractérisent par leur moindre coût et leur capacité de traitement supérieure, nous montrons aussi que les secondes permettent de conserver l'organisation spatiale des sous-types cellulaires.

Après une introduction exhaustive aux concepts biologiques de ma thèse, le corps principal de mon manuscrit détaille un ensemble de méthodologies statistiques destinées à analyser les principales sources de variation des profils moléculaires des patients. Notamment, nous nous focalisons sur les modèles de mélange et les algorithmes de déconvolution. Pour chacune de ces approches, nous commençons par un état de l'art des solutions existantes, puis nous illustrons leur intérêt sur un cas biologique concret, tout en proposant des pistes d'amélioration pour améliorer leur robustesse, et leur performance.

Partie II: Transcriptome et Modèles de mélange

Chapitre 3: Article 1, modèles de mélange Gaussiens dans l'environnement R Les modèles de mélange sont des outils statistiques qui permettent de décrire des distributions multimodales, tirées de plusieurs lois statistiques. En médecine de précision, ces modèles sont utilisés pour regrouper des patients, présentant des profils en apparence distincts en sous-groupes homogènes, dénommés endotypes. Il est alors commun de supposer que chacun de de ces profils peut être modélisé par sa distribution propre, souvent une loi gaussienne car présentant de nombreuses propriétés statistiques utiles. L'objectif est alors de retrouver la classe, non observée, de chaque observation.

Dans cette perspective, nous avons comparé systématiquement les performances de 7 packages R: `bgmm`, `EMcluster`, `GMKMcharlie`, `flexmix`, `mclust`, `mixtools`, et `Rmixmod`. Ces derniers inférent les paramètres caractérisant les modèles de mélange gaussiens, en utilisant l'algorithme EM pour ce faire. En effet, [DLR77] a démontré la consistance, la convergence et l'efficacité asymptotique vers les paramètres recherchés de cette approche itérative. Notre étude comparative évalue notamment les performances statistiques et computationnelles de ces packages en fonction du choix de la méthode d'initialisation des paramètres et de la complexité du mélange à reconstituer.

Nous montrons une réduction significative de la précision et de la robustesse associées à l'estimation des mélanges gaussiens, lorsque le niveau de recouvrement et d'entropie des classes identifiées, ainsi que la dimensionalité du problème, augmentent. De façon imprévue, cette étude a aussi mis en évidence une nette dichotomie de performance entre les packages comparés. En particulier, nous montrons que de légères déviations à la méthode originelle pour gérer les débordements numériques entrainent un comportement atypique et non consistent des packages, notamment `bgmm` et `EMcluster` en grande dimension. Plus généralement, les packages `mixtools` et `Rmixmod` présentent une tendance de fonds à retourner des estimations moins biaisées mais aussi plus variables que les autres. Ce travail est actuellement sous presse au sein du `R Journal`, et devrait être officiellement publié et accessible en octobre 2023.

Chapitre 4: Article 2, une nouvelle classification moléculaire pour le syndrome primaire du Sjögren Nous avons appliqué cette comparative exhaustive de la modélisation par modèles de mélanges gaussiens pour la stratification de 304 patients atteints du syndrome de Sjögren primaire (pSD). Ce travail a notamment fait l'objet d'une publication dans le journal *Nature Communications*. La maladie de Sjögren est une affection auto-immune, caractérisée notamment par un asséchement des glandes salivaires.

Une classification non supervisée des patients en endotypes, reposant sur l'utilisation de modèles de mélange gaussiens appliquées sur des données transcriptomiques de sang, ont permis d'identifier quatre groupes distincts. Chacun d'eux est caractérisé par une signature biologique unique, analysée plus en détail à l'aide de données de cytométrie, métaboliques et sérologiques.

L'identification agnostique d'endotypes a notamment mis en relief des changements significatifs de populations cellulaires immunes entre les clusters identifiés, liés à une activation différentielle des voies de signalisions contrôlant le niveau d'activation et l'abondance de ces dernières. Plus généralement, cette étude, en révélant des mécanismes pathogéniques propres à chaque sous-groupe, suggère d'adapter les traitements en fonction de l'endotype prédit pour chaque patient. Par exemple, seuls les individus du groupe 3 présentent une sur-activation des modules de gènes liées à la maturation des lymphocytes B naifs en lymphocytes mémoires ou plasmocytes effecteurs, et donc pourraient bénéficier d'un traitement visant à diminuer la proportion de ces derniers.

Toutefois, des analyses de sensibilité supplémentaires ont révélé que les variations observées dans l'expression transcriptomique entre les patients étaient principalement dues à des altérations de la composition cellulaire immunitaire et non à des changements inhérents de l'activité de populations cellulaires nativement présentes.

Il s'agit d'un problème majeur rencontré par les méthodes statistiques utilisant des données RNA-Seq à une échelle globale. En effet, l'étude du transcriptome au niveau d'un tissue implique de moyenner les contributions individuelles de chaque population cellulaire. Ce faisant, elles ignorent l'impact d'altérations du pool cellulaire, provoquées par des variations de la motilité ou de la différenciation. Or, ces dernières jouent aussi un rôle clé dans la régulation des processus biologiques, qui se retrouvent ainsi ignorées dans les analyses de sensibilité et d'identifcaiton des facteurs clés menées ultérieurement.

A ce titre, dans la partie concluant ce manuscrit de thèse, nous détaillons des modèles statistiques alternatifs, dédiés à l'estimation de la composition cellulaire. En offrant une granularité améliorée, nous verrons que ces approches permettent de retrouver les composantes individuelles d'un mélange au niveau tissulaire, et donc d'analyser les causes des variations du transcriptome au niveau individuel.

Partie III: populations cellulaires et algorithmes de déconvolution

Chapitre 5: Article 3, état de l'art des méthodes de déconvolution cellulaires Dans cet article, nous présentons un état de l'art actualisé des algorithmes de déconvolution pour inférer automatiquement la composition cellulaire d'un échantillon biologique. Nous nous focalisons en particulier sur les approches numériques dites "partielles", qui retrouvent les abondances relatives des composants cellulaires d'un mélange hétérogène à partir des profils d'expression cellulaires purifiées et de l'expression transcriptomique totale, quantifiés par les nouvelles technologies de séquençage.

Si la majorité des modèles de déconvolution analysés supposent que l'expression transcriptomique totale du mélange peut être retrouvée en sommant les contributions individuelles de chaque sous-type cellulaire, pondérées par leur abondance, nous montrons toutefois que ces derniers se distinguent par la nature des objectifs poursuivis et des contraintes biologiques prises en compte dans le modèle. Par exemple, certains des algorithmes ont été développés pour estimer les caractéristiques d'une population cellulaire inconnue (souvent un clone tumoral), tandis que d'autres approches se focalisent sur la robustesse des résultats, en incluant des étapes de sélection supplémentaire des gènes utilisés dans la déconvolution.

Les méthodes actuellement développées présentent toutefois des performances limitées pour l'estimation de populations cellulaires rares, ou présentant un profil moléculaire proche de celui d'autres populations présentes dans l'échantillon. Ces limitations ont notamment été soulignées dans la revue comparative de [FT18], dans laquelle il suppose qu'une des causes de la faible

robustesse des algorithmes de déconvolution est leur hypothèse commune d'"absence d'interactions entre les gènes.

Pour pallier à ces défauts, nous proposons, dans le dernier chapitre de cette thèse, un algorithme de déconvolution cellulaire, `DeCovarT`, adoptant une approche multivariée et connectée de la résolution de la "soupe" transcriptomique.

Nous élargissons ensuite notre attention à l'ensemble du protocole de déconvolution, de la récupération des données jusqu'à l'évaluation statistique et l'interprétation biologique des résultats. Nous montrons notamment dans cette partie l'importance du choix, du prétraitement et du nettoyage des données transcriptomiques sur la qualité finale des résultats. Ainsi, l'obtention de profils d'expression transcriptomiques purifiées de qualité, en nombre suffisant et proches du contexte biologique étudiée, est certainement plus importante que le choix de l'algorithme lui-même ([Fin+19b]).

Nous terminons cet état de l'art par une discussion plus spécifique sur l'avenir que nous entrevoyons pour les méthodes de déconvolution cellulaire. Nous montrons notamment la complémentarité de ces dernières avec les technologies de single cell RNA-Sequencing pour améliorer la précision et la résolution des méthodes de transcriptomiques spatiales.

Chapitre 6: Article 4, `DeCovarT`, un algorithme de déconvolution robuste exploitant les réseaux transcriptomiques La principale contribution personnelle statistique de cette thèse est le développement d'une méthode de déconvolution innovante, `DeCovarT`, capable d'intégrer explicitement les réseaux de co-expression transcriptomiques. Contrairement aux méthodes de déconvolution classiques évoqués plus haut, nous relaxons en effet les hypothèses d'*exogénéité* et d'indépendance entre les expressions individuelles des gènes au sein d'une population cellulaire. Nous avons ainsi changé de paradigme, en replaçant les profils transcriptomiques de

chaque population cellulaire dans son réseau de co-expression. Précisément, nous postulons que le vecteur d'expression purifiée $\boldsymbol{x}_j$, caractérisant l'expression transcriptomique de chaque population cellulaire, suit une distribution gaussienne multivariée. La matrice de covariance, $\boldsymbol{\Sigma}_j$, de cette dernière, permet notamment d'encoder explicitement les interactions directes entre gènes. Nous supposons alors que le mélange transcriptomique global peut être reconstruit comme une convolution de variables statistiques indépendantes, les profils d'expression cellulaires, pondérées par des poids fixés inconnues, les ratios cellulaires. Nous montrons, par la propriété d'invariance par transformation affine des lois multivariées gaussiennes, que la distribution du profil d'expression totale, conditionnée par les profils d'expression purifiés individuels, est identifiable à une loi multivariée gaussienne. La représentation graphique de ce cadre de modélisation est reportée dans la figure 3.

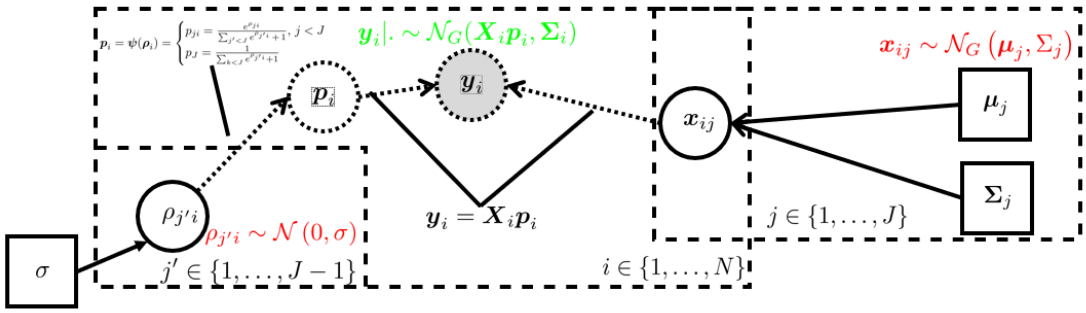


Figure 3: DAG décrivant le modèle génératif sous-tendant l'algorithme de déconvolution `DeCovarT`.

Puisque la loi de distribution du mélange est connue, nous pouvons alors aisément dériver la log-vraisemblance du modèle génératif associé, et retrouver par maximum de vraisemblance les paramètres d'intérêt, à savoir les ratios cellulaires ($\boldsymbol{p}$) et les variations transcriptomiques individuelles pour chaque population cellulaire ($\boldsymbol{X}$). En pratique, nous avons utilisé l'algorithme de minimisation itératif de Levenberg-Marquardt pour extraire les racines du gradient de la fonction de vraisemblance.

Sur une simulation numérique réduite, ne considérant que deux populations cellulaires caractérisées par deux gènes, nous avons démontré que le niveau de corrélation mutuel entre les deux gènes influence significativement l'estimation finale des ratios cellulaires. Nous avons toutefois montré que la baisse de précision imputable au niveau de corrélation entre les deux distributions bivariées décrivant chaque population cellulaire, était partiellement contrebalancé par la prise en compte des structures de co-expression avec l'algorithme `DeCovarT`. En particulier, nous avons montré que `DeCovarT` présentait des performances systématiquement supérieures par rapport à un algorithme standard, reposant sur une fonction d'optimisation quadratique du simplex unitaire.

De plus, comme le nouvel algorithme de déconvolution que nous proposons repose sur un modèle génératif, nous pouvons aisément dériver asymptotiquement des intervalles de confiance, et donc des tests statistiques. Ces derniers permettent d'évaluer théoriquement la significativité des variations de la composition cellulaire du milieu biologique.

Conclusion et perspectives Nous concluons cette thèse en passant en revue quelques pistes d'exploration, dont certaines en cours d'implémentation, visant à améliorer la pertinence biologique et la justesse statistique des modèles présentés dans cette thèse. Nous discutons notamment des opportunités offertes par l'intégration de sources biologiques diverses, tout en soulignant la difficulté de coupler des données pouvant aboutir à des conclusions contradictoires. Nous suggérons enfin brièvement une généralisation possible des modèles présentés, en remplaçant les distributions symétriques et standardisées des lois gaussiennes par des densités de Poisson log-normales ou de binomiales négatives; ces dernières sont en effet susceptibles de décrire plus justement la nature des processus biologiques.

Annexes Nous présentons dans ces annexes des tâches relevant plus de l'ingénierie, liées à mon statut de doctorant CIFRE. Plus précisément, le matériel supplémentaire se décompose ainsi; tout d'abord, les quatre premières annexes viennent compléter les chapitres principaux 1, 3, 5 et 6 de la thèse, respectivement. Les deux derniers chapitres de l'annexe présentent ensuite des collaborations en lien avec les objectifs thérapeutiques de Servier, ayant fait par ailleurs l'objet de publications respectivement publiés dans `Plos One Communications` (actuellement en phase de révision) et `Plos One`.

Annexe A du chapitre 1 de thèse Le pipeline développé englobe notamment le nettoyage des données et leur standardisation sous la forme d'un `ExpressionSet`, le prétraitement et la normalisation des matrices de compte pour corriger les biais techniques et se conformer aux hypothèses des modèles d'analyse, validées par différents contrôles visuels de qualité. Finalement, nous avons adapté en interne les principaux outils statistiques visant à comprendre les mécanismes sous-tendant la variabilité observée de l'expression transcriptomique, entre différentes conditions biologiques. Notamment, nous proposons des outils de visualisation "maisons" pour l'identification de modules, ou de gènes significativement variants. Nous concluons ce tutoriel technique en soulignant les l'utilité de la mise en place d'une filière standardisée pour l'analyse de données omiques pour la reproductibilité des résultats, et la réduction drastique des coûts d'analyse.

Annexe B du chapitre 3 et article 1 de la thèse Dans cette annexe, nous présentons des simulations numériques étendant la portée des résultats évoqués dans le chapitre 3 de la thèse. Tout d'abord, nous avons comparé les performances des modèles de mélange gaussien standards avec des modèles supprimant automatiquement les observations aberrantes, sur des jeux de données volontairement bruités. Les résultats montrent sans surprise que le package `otrimle` est significativement plus robuste que les packages canoniques, notamment `mclust`, en intégrant explicitement une composante additionnelle modélisée par une loi uniforme impropre. Dans un autre cadre de simulation, j'ai exposé théoriquement, puis illustré en pratique, l'intérêt des méthodes de projection et d'une paramétrisation parcimonieuse des modèles de mélange gaussiens pour visualiser et estimer des structures latentes en grande dimension.

Annexe C du chapitre 5 et article 3 de la thèse Dans cette partie, nous présentons plus en détail différentes techniques d'optimisation, très documentées, et utilisées en pratique par de nombreux algorithmes de déconvolution pour contrebalancer le manque de données de qualité et la complexité du calcul des paramètres minimisant la fonction de coût. Nous rappelons aussi le théorème de Gauss-Markov, étant donné que les hypothèses statistiques sous-tendant son application, sont rarement remises en question par la majorité des méthodes de déconvolution.

Annexe D du chapitre 5 et article 4 de la thèse Dans cette annexe, nous proposons deux pistes d'amélioration possibles pour améliorer la robustesse et la précision de `DeCovarT`. Premièrement, il serait intéressant d'intégrer l'approche réseau de notre algorithme de déconvolution, dès la sélection de l'ensemble minimal de gènes discriminants utilisés pour la construction de la matrice de signature cellulaire. Nous présentons notamment une stratégie holistique, qui prendrait en compte non seulement les variations de moyenne mais aussi de voisinage dans le réseau de co-expression, pour la sélection de marqueurs spécifiques d'une population cellulaire. Enfin, nous dérivons une méthode de Monte-Carlo, reposant sur une estimation des paramètres via une chaîne de Markov, pour échantillonner la distribution a posteriori des paramètres d'intérêt. L'intérêt majeur de cette méthodologie est la possibilité de prendre en compte les variations individuelles du transcriptome, ainsi que la dérivation naturelle d'intervalles de crédibilité pour les ratios cellulaires.

Annexe E: Article 5, classification non supervisée de gènes, appliqué au syndrome primaire de Sjögren Cet article présente une autre méthode de classification non supervisée, l'algorithme de Louvain, reposant sur l'optimisation de la modularité pour la séparation automatique de groupes de gènes. Elle favorise ainsi les sous-groupes de gènes fortement interconnectés entre eux. Cette approche modulaire facilite l'interprétation biologique des groupes de gènes identifiés, prélude à une compréhension causale et systémique de la diversité des réponses thérapeutiques observées à un même traitement.

En pratique, la majorité des 13 modules a pu être effectivement annotée, et reliée à une population cellulaire et/ou une voie de signalisation. D'autre part, des études cliniques sur des données réelles ont confirmé l'intérêt de ces groupes de gènes comme biomarqueurs prognostiques pour l'évolution de la sévérité des cas de Sjögren primaires ou l'évaluation de l'efficacité d'un traitement thérapeutique.

Annexe E: Article 6, repositionnement de médicaments appliqué au traitement de la COVID-19 basé sur une approche réseau Si nous avons évoqué différentes approches statistiques pour décrypter la diversité des profils biologiques observés chez l'homme, l'objectif poursuivi est la conception de thérapies ciblées et présentant moins d'effets secondaires, dans le cadre de la médecine de précision. Il s'agissait d'ailleurs de l'objectif initial de mon sujet de thèse, avant un changement orthogonal de direction, vers des approches orientées sur la biologie des systèmes.

Après avoir rappelé les difficultés et les forts taux d'attrition associés au développement d'une nouvelle thérapie, nous montrons dans cette annexe, en guise de préambule, l'intérêt général des méthodes de repositionnement de médicaments pour réduire les coûts et la durée de développement. Nous passons ensuite en revue les principales approches statistiques pour accélérer l'identification agnostique de biomarqueurs, via l'exploitation d'importants volumes de données.

J'introduis enfin la solution de repositionnement de médicaments historiques, propre aux laboratoires Servier et dénommé `Patrimony`. Cette dernière repose sur une approche systématique et holistique, dont le point de départ est la reconstruction d'un graphe de connaissances combinant de nombreux types de données omiques et cliniques.

Le projet `Patrimony` a été appliqué expérimentalement, et avec succès, pour l'identification de thérapies anti-inflammatoires visant à réduire la sévérité des formes les plus graves de la COVID-19. Des anticorps, contrecarrant la production d'interférons pro-inflammatoires, ont été notamment identifié pour traiter les cas de "tempête de cytokines". Leur intérêt thérapeutique a été ultérieurement validée par des essais cliniques.

Contents

Remerciements v

Publications ix

Oral communications x
Peer-reviewed communications . . . x
Invited seminars . . . xi
Poster sessions . . . xi

Abstract xii

Résumé long de la thèse xiv

Contents xxiii

I Biological introduction 1

1 Study of the transcriptome 2
1.1 Regulation of the Transcriptome . . . 2
1.1.1 Overview: Importance of meticulous Regulation of Gene Expression . . . 2
1.1.2 Epigenetics: Implications in Molecular Profile Diversity . . . 6
1.2 Tools for exploring the transcriptome . . . 8
1.2.1 Microarray technology to quantify gene expression . . . 8
1.2.2 RNASeq technology . . . 9
1.2.3 Perspectives: single cell and spatial transcriptomics . . . 15
1.2.4 Conclusion: The Significance of Transcriptomic Data in Computational Medicine . . . 18

2 Introduction to the Immune System 21
2.1 Key actors of the Immune System . . . 21
2.1.1 The innate system . . . 21
2.1.2 The adaptive system . . . 23
2.1.3 Exchange of goodwill between the two immune systems . . . 24
2.1.4 Immune dysregulation . . . 26
2.2 Physical methods for studying changes of cellular Composition . . . 28
2.2.1 Cytometry analyses . . . 28
2.2.2 Imaging methods . . . 29

II Transcriptome and mixture models **34**

3 Article 1: Gaussian Mixture Models in R **35**
3.1 Article 1 . . . 35
3.2 Main results . . . 58
3.3 Perspectives . . . 58

4 Article 2: A new molecular classification in primary Sjögren's syndrome **60**
4.1 Article 2 . . . 63
4.2 Main results . . . 82
4.3 Limitations and perspectives . . . 84

III Cell populations and deconvolution algorithms **86**

5 Article 3: review of cellular deconvolution methods **87**
5.1 Article 3 . . . 88
5.2 Conclusion: major Limitations of existing Deconvolution Algorithms Solutions 122

6 Article 4: DeCovarT, a deconvolution algorithm leveraging co-expression networks **123**
6.1 Article 4 . . . 123
6.2 Conclusion . . . 133
6.3 Publication Outline . . . 133

Thesis overview **135**

A Appendix of Chapter 1 **139**
A.1 Data import and cleaning . . . 144
A.1.1 Import relevant files . . . 144
A.1.2 Data wrangling with ExpressionSet . . . 145
A.2 Pre-processing of raw counts, using a Nextflow pipeline . . . 159
A.3 From raw counts to normalised gene expression and downstream analyses . . . 160
A.3.1 Sample filtering . . . 160
A.3.2 Gene filtering . . . 160
A.3.3 Normalization and transformation . . . 164
A.3.4 Quality control and data exploration . . . 169
A.3.5 Batch effect correction . . . 173
A.4 Downstream analyses . . . 178
A.4.1 Differential expression analysis . . . 178
A.4.2 Multi-level classification of cell populations . . . 190
A.5 Conclusions and perspectives . . . 193
A.6 Appendix A: Gene notations . . . 195
A.6.1 Gene terminologies . . . 195
A.6.2 Automated methods for gene annotation . . . 196

B Appendix of Article 1 **200**

C Appendix of Article 3 **265**
C.1 Appendix of Reference-Based Approaches . . . 265
C.1.1 Linear regression and Gauss-Markov theorem . . . 265
C.1.2 Robust regression approaches . . . 268
C.1.3 Regularised linear approaches . . . 271
C.1.4 Probabilistic approaches . . . 272
C.2 Statistical Appendix of Marker-Based Approaches . . . 275
C.2.1 Gene Set Enrichment Analysis . . . 275
C.2.2 Hypergeometric Distribution . . . 276
C.2.3 Limitations of Marker-Based Approaches . . . 276
C.3 Statistical Appendix of Reference-Free Approaches . . . 277
C.4 Appendix of Cellular Deconvolution Pipeline . . . 278
C.5 Biological Appendix to the Fate of Deconvolution Algorithms . . . 279

D Appendix of Article 4 **281**
D.1 Optimisation and calculus . . . 282
D.1.1 Multivariate distributions and basic algebra properties . . . 282
D.1.2 Matrix calculus . . . 284
D.1.3 First and second-order derivation of constrained DeCovarT . . . 285
D.2 A MCMC Algorithm for the Joint Distribution of Purified Profiles and Ratios . 286
D.2.1 An Introduction to Gibbs and Metropolis Hasting Samplers . . . 287
D.2.2 Pseudo-code Gibbs sampler . . . 288
D.2.3 Derivation of the Acceptance Probability Function . . . 290

E Article 5: Gene clustering applied to primary Sjögren's disease **291**
E.1 Conclusion . . . 313
E.2 Perspectives . . . 313

F Article 6: Network-based repurposing applied to COVID-19 **314**
F.1 Drug repurposing: a brief historical overview . . . 316
F.2 Introduction to the Patrimony initiative . . . 321
F.3 Repurposing applied to severe COVID-19 cases . . . 323
F.4 Conclusion . . . 342
F.5 Perspectives . . . 342

Contents **348**

Glossary **352**

Glossary **352**

Bibliography **364**

Part I

Biological introduction

Chapter 1

Study of the transcriptome

1.1 Regulation of the Transcriptome

Introduction to Omics Data and Their Role in Computational Medicine Omics data refers to commonly large datasets, generated using cutting-edge technologies; they notably encompass genomics, transcriptomics, proteomics, and metabolomics. These methodologies enable biologists to explore interactions between biological systems on a large scale, providing valuable insights into cellular functions and mechanisms.

This chapter primarily focuses on transcriptomic data. The transcriptome represents the set of transcripts, namely DNA segments effectively transcribed into mRNA, measured within a biological sample at a given time. Transcriptomic analyses, by retrieving differential gene expression patterns and co-expression networks, yield crucial insights into how genes are activated or silenced under various conditions.

Analysing such variations of the expression profiles facilitates the identification of the key drivers responsible for the breakdown of transcriptomic expression regulation mechanisms, one of the core onset of pathological conditions.

1.1.1 Overview: Importance of meticulous Regulation of Gene Expression

The central dogma is a structural cornerstone of molecular biology that states that genetic information is unidirectional, from Deoxyribonucleic Acid (DNA) to proteins. DNA stores the hereditary material of an organism while Ribonucleic Acid (RNA) serves as a template for the synthesis of proteins.

The process by which RNA is synthesised from DNA is called *transcription* while the synthesis of proteins from RNA is called *translation*. The central dogma states that DNA is never directly translated into proteins, implying that the flow of genetic information is uni-directional [Cri70]. The genetic information flow from DNA to proteins is usually referred to as *gene expression*.

Fined-tuned regulation of gene expression plays a crucial role in the development and adaptation of the human organism to diverse environmental conditions. Remarkably, every individual cell type contains an identical set of genetic material; yet, by selectively expressing specific subsets of genes in a precise temporal pattern, each cell type acquires the capability to fulfil highly specialized biological functions.

In the following subsections, we present an overview of the regulatory mechanisms that govern the activation or suppression of genes with precision.

A relevant analogy for visualizing these complex regulatory processes is the symphony concert illustration. As individual orchestra members tune their instruments, the resulting sound is initially a disordered cacophony. However, as soon as the conductor's baton is raised, all instruments harmonize, synchronizing their pitch with precision and right on time, turning the initial discord into a gleeful symphony.

Among the various regulatory processes involved, we particularly underscore the relevance of co-expression networks that involve transcriptomic factors. Additionally, we illuminate the pervasiveness of regulatory processes throughout the transformation from a DNA template to a functional protein. These mechanistic controls of the regulation of the transcriptomic expression encompass:

Regulation of the chromatin structure

Firstly, the level of transcription is modulated by the availability of the initiation sites and by extension to the level of compactness of the chromatin fibre. Indeed, the DNA of eukaryotic cells is packed with proteins in a "protein complex" known as chromatin, the basic unit of which is the *nucleosome*. When histones and DNA are tightly bound in the chromatin fibre, they limit the access of the RNA polymerase to promoters. This dynamic control of the chromatin architecture is termed *chromatin remodelling*.

The regulation of gene expression, in terms of chromatin compactness, is influenced by three primary factors: firstly, the gene's promoter location relative to initiation sites; secondly, the overall compactness of the chromatin structure, exemplified by genes within heterochromatin, a densely packed region, which typically remain unexpressed; and lastly, local chemical modifications.

Chemical modifications are catalysed by specific enzymes that modify the tri-dimensional configuration of histone tails protruding from nucleosomes or the DNA sequence. Reversible addition of methyl groups (-CH3) to amino acids in histone tails can promote the condensation of the chromatin, while addition of a phosphate group, namely *phosphorylation*, or *histone acytelation*, namely when acetyl groups attach to lysines, induce a looser structure of the chromatin.

The *histone code* hypothesis suggests that the specific combination and order of chemical modifications yields the final configuration of the chromatin, which in turn influences transcription.

While some enzymes interact chemically with the tails of histone proteins, *DNA methylation* refers to the process performed by a different set of enzymes, methylating directly certain bases in the DNA itself. It generally occurs at CpG (Cytosine-phosphate-Guanine) sites, the regions of the DNA sequence where a guanine follows a cytosine. DNA methylation mechanisms play a critical roles in gene silencing by inducing the condensation of the chromatin.

Interestingly, unlike the reversible changes induced by histone methylation, DNA methylation patterns, termed *genomic imprinting*, remain permanent, with highly-specialised cells keeping the methylation blueprint acquired during the developmental phase.

Co-expression networks of transcription factors

Chromatin-modifying enzymes provide the first initial regulation of the gene expression by controlling the access of a region of the DNA either more. Once the promoter site is reachable, the initiation of transcription involves the recruitment of proteins, building a *transcription initiation complex*.

Among them, the RNA polymerase II transcribes the gene and synthesises it into a primary RNA transcript (pre-mRNA). However, appropriate binding and efficient transcription requires the combination of regulatory proteins called Transcription Factors (TF) with their corresponding binding sites acting as control elements. Together, Transcription Factor interplay to precisely achieve the fine-tune regulation of gene expression seen, either activating or inhibiting the transcription process.

We set apart two types of transcription factors:

- *General transcription factors:* They are required for the initiation of the transcription of all genes, by either binding to the TATA box, a sequence present in most promoters, to other Transcription Factor or to RNA polymerase. However, their presence only leads to a low rate of RNA production.
- *Specific transcription factors*: Context-dependent production of a given gene involves another set of proteins, the so-called specific Transcription Factor. They recognise specific sequences of DNA, that might be located close to the promoter, hence termed *proximal control elements*, or may be distant up to thousands of nucleotides upstream or downstream, hence called the *distal control elements*. A group of control elements acting together to which a TF uniquely binds to is called an *enhancer*. Interestingly, while a gene may be associated to several enhancers, whose availability varies over time, each enhancer is only associated with a gene.

  Transcription Factor can either act as activators or repressors, and bind to other Transcription Factor or the DNA sequence itself. In the last case, they display the same two structural elements: a *DNA-binding domain* that binds to the corresponding DNA control elements and several *activation domains* that ease or inhibit Protein-Protein Interactions (PPI) networks and by extension the efficiency of the transcription.

Amazingly, while up to 20 000 genes must be regulated in a human cell, only a dozen distinct nucleotide sequences control the pairing between TF and the DNA sequence, each enhancer being composed on average of about ten control elements. It is the particular combination of control elements that yields the one-way matching between the enhancer and its corresponding TF.

In summary, interactions between transcription factors form complex co-expression networks that delicately modulate gene expression levels by controlling genome accessibility for the RNA polymerase.

Post-transcriptional regulation

The strand of mRNA resulting from the transcription is not yet mature, and undergoes a series of biochemical interactions modifying its properties. Among them, *alternative splicing* is the most fundamental mechanism of post-transcriptional regulation.

Indeed, while all the introns are removed, not all the exons, even though coding directly for a protein, are kept in the final mature mRNA. It enables the production of protein variants, *isoform proteins*, starting from the same pre-mRNA, as illustrated with the troponin gene (Figure 1.1). Thus, a single gene can achieve different functions in the cell by controlling the proportion of each isoform produced. In fact, alternative splicing is the most likely explanation until now (90% of human protein-coding genes undergo splicing) for the low number of human genes identified (around 20 000), similar to that of a soil worm (nematode) or a mustard plant.

Once the mature mRNA produced, it still must be translated into a protein. Conversion from a mRNA fragment to a functional protein highly depends on its average life span, which in return

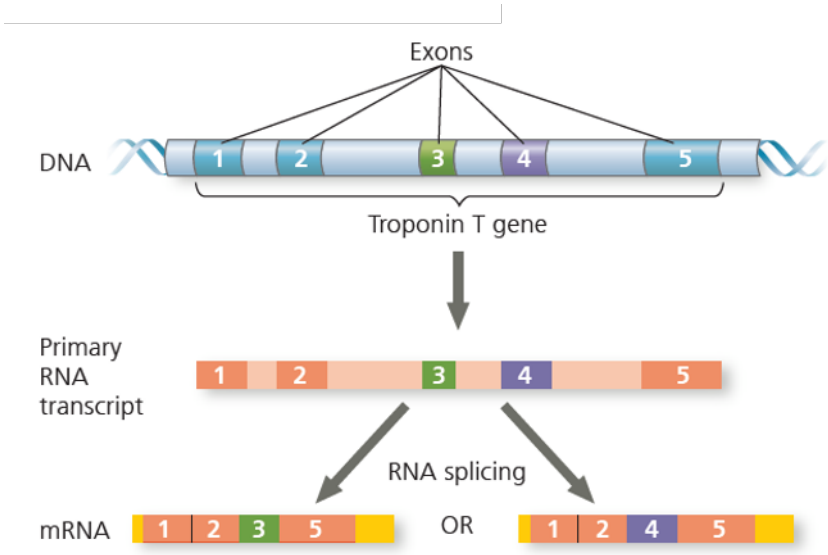


Figure 1.1: Alternative RNA splicing of the troponin T gene..This figure is reproduced from [Cam+20, Fig. 18.14, p. 378]. Colour background of the exons are dark and introns light. The primary transcript of this gene can be spliced in several mRNA strands: one ended up with exon 3 (in green) and the other with exon 4 (in purple). Both mRNAs are muscle proteins, differing slightly by their operational region.

is related to the presence of translation regulatory proteins that recognise sequences of the 3 or 5' UTR(specific inhibition control, by preventing the binding of both RNA ends to the ribosome) or general effect protein factors, notably involved in development and differential phase. We display the mechanisms involved in Figure 1.2.

Interestingly, recent researches highlight the combined role of small non-coding RNAs (ncRNAs) molecules, known as microRNAs or miRNAs, and regulatory proteins in modulating gene activity. By base pairing to mRNAs, microRNAs mediate translational repression or the degradation of mRNAs. Other non-coding RNAs, including long intergenic ncRNAs (lincRNAs) or small interfering RNAs (siRNAs), seems to be involved but the associated regulation mechanism still needs to be deciphered ([dSou12], [Eck+12], [Ger+07] and [Sny+20]).

The most complex biological functions require a coordinate set of chemical reactions, for instance the enzymes involved in a metabolic pathway or a transduction signal. Unlike bacteria, in which genes involved in the same biological function are associated to the same promoter, a structural organisation termed *operon*, genes requiring simultaneous expression are often scattered over the whole genome in human cells. In that case, coordinate control is triggered by cellular communication, that in turn promote the recruitment of Transcription Factor.

Post-translational regulation

Finally, post-translational modifications can occur at any time after the translation of the protein. By inducing chemical modifications to proteins, they alter its metabolic function or *targeting*, a part of the protein sequence itself that addresses the final polypeptide sequence to its appropriate localisation. The most common modifications involve covalent additions of one or more side groups, including phosphorylation, acetylation, alkylation, and glycosylation.

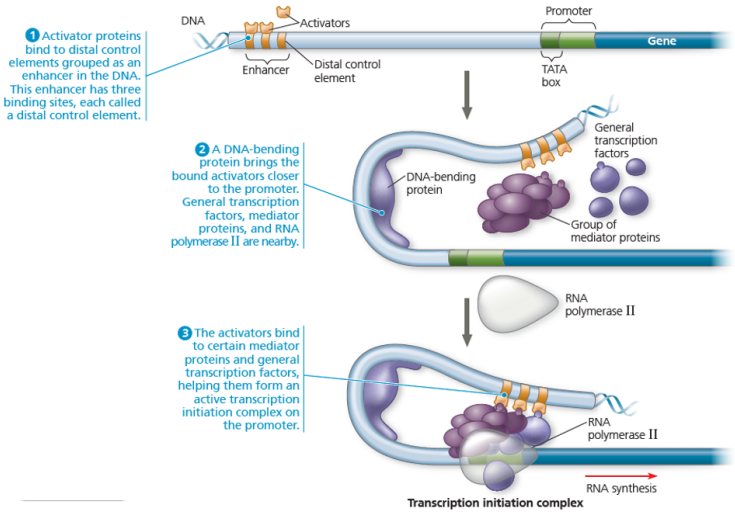


Figure 1.2: Model for the coordinate action of RNA polymerase and Transcription Factor..This figure is reproduced from [Cam+20, Fig. 18.11, p. 375]. The transcription initiation complex involves the coordinate binding of RNA polymerase II and general transcription factors, while fine-tune regulation is enabled by specific Transcription Factor that bind to the enhancers (here, represented with three control elements in gold) and to mediator proteins. It is DNA bending that enables enhancers to influence the regulation remotely.

Finally, the life span of each protein is strictly regulated by selective degradation. Giant protein complexes, the *proteasomes*, recognise the ubiquitin-tagged proteins and degrade them.

In section 1.1.2, we exemplify the substantial impact of epigenetic mechanisms on the multitude of cell types, each addressing a specific biological function, and the spectrum of molecular profiles. We subsequently detail how the variations in the regulation of the transcriptomic expression influence treatment responses and disease progression.

1.1.2 Epigenetics: Implications in Molecular Profile Diversity

The various modifications discussed earlier do not involve changes in the DNA sequence, yet they can still be passed on to future generations of cells. The transmitted patterns are referred to as *epigenetic inheritance* which, unlike DNA mutations, are generally reversible.

An increasing number of experiments confirm the significance of epigenetics in the development of genetically-based diseases or the initiation of pro-tumoral conditions that promote the onset of cancers. We gather all the processes involved in gene regulation in Figure 1.3.

It's worth noting that all these regulation processes are usually intertwined. For instance, Transcription Factor (TF) can additionally influence chromatin structure by recruiting proteins that acetylate histones to enhance transcription or directly by removing acetyl groups, resulting in transcriptional *silencing* [Wal+12].

In Appendix E, we detail unsupervised, data-driven methods, borrowed from the graph theory field coupled with high-dimensional clustering methods, to reconstruct from mere correlations across transcripts these set of intertwined regulation mechanisms.

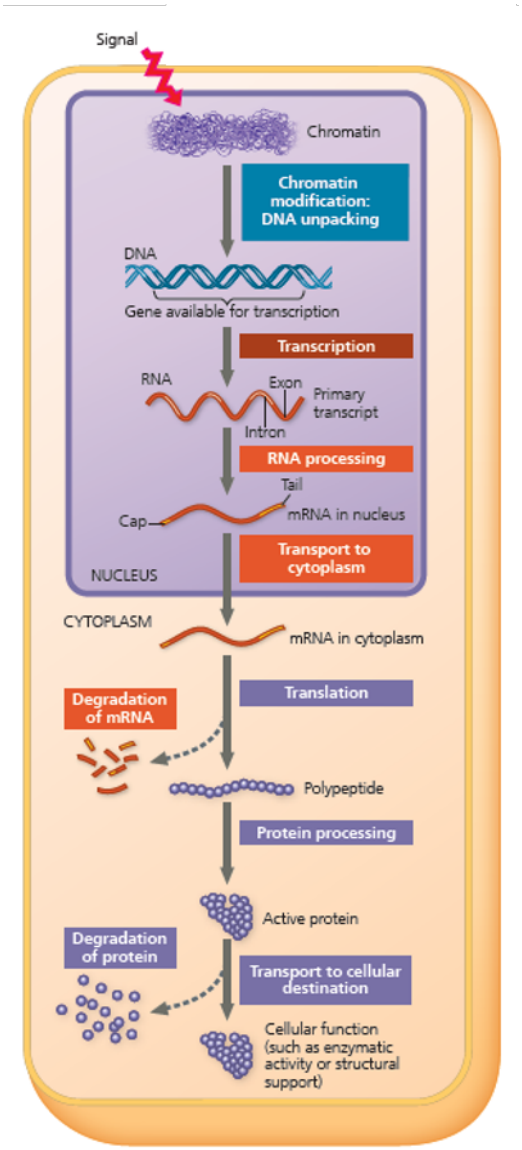


Figure 1.3: *The key stages of gene regulation.* This figure is reproduced from [Cam+20, Fig. 18.6, p. 370]. In this diagram, each colour indicates the type of molecule regulated (blue = DNA, red or orange = RNA, purple = protein).

Gene expression is quantified using advanced techniques, including microarrays and RNA sequencing (RNA-seq). These methods enable comprehensive profiling of the transcriptome and the identification of changes in gene expression levels under various disease states.

In the following section 1.2), we provide a detailed overview of the latest sequencing tools capable of simultaneously measuring transcripts at the tissue scale and with high-throughput analysis capacity.

1.2 Tools for exploring the transcriptome

Overview: historical outline Historically, tools for quantifying gene expression were primarily designed for single-gene analysis. The most commonly used methods included *in-situ hybridization* techniques like Northern Blots or Reverse Transcriptase Polymerase Chain Reaction (RT-PCR), as well as *sequence-based* approaches such as Serial Analysis of Gene Expression (SAGE) or Comparative Expressed Sequence Tag (EST) analysis.

All the technologies described hereafter harness the sequence of interest as a template (or primer) to synthesise a complementary and single-stranded RNA using nucleic acid hybridisation.

RT-PCR first involves synthesising a double strand, complementary to the DNA sequence and trimmed of introns, by the combined use of reverse transcriptase and DNA polymerase with a RNAse enzyme in charge of degrading the original mRNA material. Indeed, RNA is reverse transcribed to cDNA since DNA is more stable and enables more efficient amplification and the use of mature DNA sequencing technology. Then follows an *amplification stage* generating many copies of the RNA segment of interest: the PCR step, which notably requires *primers*, short chemically synthesised oligonucleotides that initiate the replication by attaching the DNA polymerase in place. Finally, newer methods such as qRT-PCR (q stands for *quantitative*) enable precise quantification while avoiding the burdensome and time-consuming electrophoresis step.

However, any of the methods described previously can only capture a limited number of mRNAs, while it is usually more interesting to pursue a systematic mRNA estimation. As seen in Section 1.1.1, different genes, scattered across the human genome, act in concert to perform complex cellular process.

To that end, we focus in this chapter on the technologies that truly enable to analyse systematically the *transcriptome*, with, by chronological order, microarray technologies Section 1.2.1, followed by more versatile Next-Generation Sequencing methods Section 1.2.2.

Of note, the Human Genome Project ([Col+98]), in early 21$^{\text{th}}$ century underlies the fast development of these high-throughput technologies. Indeed, most of them rely on a reference and universal human genome mapping to target specific mRNAs in the biological sample.

1.2.1 Microarray technology to quantify gene expression

Microarray plates are small chips onto which tens of thousands oligonucleotide probes, with sequences complementary to fragments of actual genes, are engraved. Generally, each gene is normally represented by a set of probes, so called a *probeset*, each representing a different but highly-specific gene region.

one of the most widely used technology being the Affymetrix suite [1], with a bunch of dedicated

[1] For instance, the Human Genome U133 Plus 2.0 Array (HGu133+, released by the Affymetrix company) gathers 54 000 probesets, spotted by 11 different probes, comprising overall more than 1 300 000 distinct oligonucleotide sites.

R packages and methods to aggregate probesets into an unique gene expression measure while discarding background noise or accounting for specific bias ([Gau+04]).

However, the versatility of microarray technologies is significantly hindered by the complex design of probes. The main challenge is to identify a set of nucleotides both specific to the gene targeted and unrelated to any other genome region, whereas technical limitations prevent the length of the complementary probe going beyond 60 base pairs. Without careful and customised choice of probes, the risk is otherwise of detecting a spurious pairing.

New promising microarray applications continue to emerge, ranging from SNPs detection to copy number variations through identification of methylated regions or protein binding sites. Due to their high miniaturisation and standardisation, they strike the balance between the requirement of high-throughput capacity analysis and cost-saving.

1.2.2 RNASeq technology

Historical development of RNA-Seq

Next-Generation Sequencing (NGS) methods are cutting-edge technologies, emerging from early 2000's, that bridge the gap between the high-throughput capacity analysis of microarray-based technologies and the versatility of traditional sequencing methods. We generally classify sequencing methods in three generations, differing by the nature of the reads generated (either short or long, direct RN or requiring prior conversion to cDNA) and the sequencing throughput.

First-generation RNASeq Sequencer platforms, like the ABI capillary technology, offer high accuracy and longer read lengths, but lack high-throughput sequencing.

Second-gen platforms (e.g., Illumina's MiSeq, HiSeq, NextSeq and Thermo Fisher Scientific's Ion Torrent) achieve both high throughput and accurate *base-calling* via parallel sequencing. However, their shorter read lengths pose challenges in assembling repetitive sequences.

In contrast, third-gen sequencers (e.g., Pacific Biosciences' PacBio, Oxford Nanopore Technologies' MinION, PromethION, SmidgION) generate long reads at high throughput by sequencing single-molecule, at the expense of a higher error rate.

In next section 1.2.2, we focus on the Illumina sequencing protocol, since all the datasets used for our analysis were sequenced through this platform [2].

However, the interested reader may report to glossary keys RNASeq Sequencer platforms for a comprehensive review of other sequencing techniques and RNA librarys for an overview of biological applications with respect to the nature of the generated reads.

Additionally, [SGA18] thoroughly elucidates the fundamental principles that underpin each technological advancement in the field, accompanied by a compilation of valuable papers and web resources for each sequencing platform. Furthermore, [Lam+12], [VD18] and [Cot18] benchmark comprehensively various RNA-Seq methodologies, while proposing future avenues for new biological applications.

Outline of RNA-Seq analysis

Library preparation The transcriptome of the sample or tissue of interest, after an isolation stage, is reverse-transcribed into complementary DNA (cDNA) for stability purposes. The next

[2] Going further,it is up to 90% of all DNA sequenced data that were generated with the Illumina platform, from estimations reported in Illumina leaflet, in 2015.

step consists of randomly fragmenting cDNA into strands of smaller sizes, called *inserts*, and ligating two distinct adaptors, one at each end, in a process called "tagmentation". The set of obtained cDNA fragments is termed the "library" (see also subfigure a, in Figure 1.4).

DNA amplification Compared to old-fashioned sequencing methods, NGS technologies mostly differ by their higher capacity of DNA amplification and sequencing. After an initial amplification stage, employing similar techniques to RT-PCR, comes the *clonal amplification* stage itself in order to increase by several orders of magnitude the total amount of DNA available for sequencing, compared to more traditional approaches [3].

Precisely, the clonal amplification stage mitigates the lower sequencing quality inherent to NGS in comparison to conventional methods, while achieving a significantly stronger genome coverage [4]. This is particularly valuable for sequencing challenging regions of the genome, including repetitive segments, GC-rich regions and *homopolymers*, which tend to pose accuracy hurdles during sequencing.

With respect to Illumina process, initial amplification relies on a lawn of oligonucleotides, complementary to the sequences of the adaptors, which are attached to a glass *flow cell*, composed of thousands of *tiles*. In parallel, the DNA library is *denatured* to form single-stranded DNA (ssDNA) fragments which subsequently tether to the surface-bound oligonucleotides. Subsequently, "bridge amplification" hybridised each tethered fragment and cloned it separately and simultaneously into ≈ 1000 copies. The set of copies is called a *cluster*. This process is enabled by the two types of oligonucleotides bound to the flow cell: they indeed allow alternated synthesis in both directions of the cDNA fragments, a process repeated until the desired amount of DNA has been reached (see also subfigure b, in Figure 1.4).

Raw sequencing Genome sequencing consist of determining the sequence of nucleotides of the whole set of fragments composing the DNA library and the technique is itself highly dependant on the platform used. The resulting data output is typically stored in FASTA or FASTQ files.

Again, we focus on the Illumina high-throughput and massively parallel sequencing protocol. Briefly, fast sequencing of fragments rely on a proprietary reversible "terminator" method that detects single addition of bases in the mean time they are incorporated.

Specifically, each sequencing cycle starts with the addition of a primer and fluorescent reversible-terminator nucleotide (rt-dNTPs), ensuring that only one nucleotide is added at a time. Any unbound nucleotide is then washed away, while lasers excite the fluorescent tags of each incorporated nucleotide, each being associated to its own wavelength. Finally, the tag is cleaved off, putting an end to the cycle.

This procedure, known as "sequencing-by-synthesis" is iteratively performed until the desired sequence length is achieved. The number of cycles employed dictates the resulting sequence length, a balanced trade-off between generating overly brief strands that might yield ambiguous mapping information in the subsequent assembly stage, and overly extended ones that carry an elevated risk of compromised synthesis quality at the 3' end.

The resulting images are then processed to return the *reads* themselves (the string sequence, each character coding for one of the four RNA nucleotides). To achieve this, the nucleotide identification process computes the signal intensity and noise for each of the identified cluster,

[3] Amplification is mostly required for methods that do not directly sequence RNASeq, but rather convert it first into c(omplementary)DNA (Figure 1.4)

[4] Up to 180 million reads are generated by the HiSeq 2000 platform

in every set of raw image files generated during the "sequencing-by-synthesis" phase. This multifaceted procedure is referred to as *base-calling* (see also [ill23] for an industrial overview or [RP18] and [MK10] for a didactic and unbiased academic point of view).

Better sequencing quality can be further achieved by synthesising fragments in both directions of each cDNA sequence, as illustrated in [ill23, Figure 4: Paired End Sequencing]. The resulting "paired-end" library has twice the number of reads for the same dedicated efforts, increasing the quality of the alignment and providing refined opportunities for the detection of "indels" (insertion deletion modifications), SNVs and SNPs mutations (see also subfigure c, in Figure 1.4).

Mapping From the billions of reads sequenced, computer software and bioinformatic tools rebuild the whole transcriptome, by locating each read in the genome. Two strategies are used, depending on the level of prior information available: *de-novo* strategy, as its name suggests, refers to construction of genomes when no annotations are available while the *genome-guided* strategy aligns and maps reads onto a genome of reference.

De-novo alignment is a highly-challenging task, especially for covering highly repetitive regions of the genome, or rare and abnormal genomic events, such as chromosomic rearrangements, Single-Nucleotide Polymorphism (SNPs) or even indels (for inserts and deletions).

However, such issues can be alleviated by coupling short-read assembly with long-read paired-end sequencing information. This strategy, known as Genomic scaffoldings involves generally an intermediate step which consists of generating larger individual Contigss before the final reconstruction of the whole genome. [Luo+21], [RG19] and [SN18] established that the optimal sequencing quality and coverage are achieved by combining reads of diverse sizes: shorter ones derived from NGS and longer inserts obtained through traditional sequencing methods, generally with lower read depth.

We now focus on the *reference-based strategy.* Historically, the methods developed focused on alignment methods for long reads generated through conventional sequencing, such as the renowned BLAST [Alt+90].

However, such methods are not tailored for aligning short reads returned by NGS technologies, nor mapping fragmented and discontinuous RNA fragments [5]. In addition, short reads with repetitive patterns have an equal chance of aligning to various distant regions of the genome, as discussed in [TS12].

Hence, numerous bioinformatic alignment tools have been developed to address the challenge of short-read sequencing (see a comprehensive review in [TS12]). These tools often use a strategy called "seed-and-extend" to align shorter portions of the reads, through dynamic programming, to find the best alignment with the most overlapping sequences. Some of the most commonly used mapping methods include BowTie2 [LS12], TopHat2 [Kim+13], STAR [Dob+13], and HISAT2 [Kim+19] algorithms.

In particular, in our custom industrial Nextflow pipeline (Appendix A.2), we use the STAR (Spliced Transcripts Alignment to a Reference) ([Dob+13]) mapping process, whose operational steps encompass:

1. Before mapping reads, STAR builds an index of the reference genome. This index facilitates fast alignment by pre-processing the genome into smaller segments, called "seeds" and storing them in a compact hash table, allowing for direct memory access.

[5] The alternative splicing phenomena, Section 1.1.1, involves that sections of the DNA sequence template are discarded in the transcription of a mature mRNA, including all *introns*

2. STAR then identifies similar seed patterns from the read sequences and matches them in the indexing hash table. The seed extension algorithm allows for robust mapping, even with mismatches and sequencing errors.

3. Interestingly, STAR is designed to handle spliced reads across exon-exon junctions, accounting for the presence of introns, thus enabling better accurate mapping of transcripts in eukaryotic organisms. To integrate this feature, STAR integrates a two-pass mapping strategy to improve alignment accuracy. The first pass identifies potential splice junctions which are further utilised in the second pass to align reads across spliced junctions.

4. STAR evaluates the quality of all potential alignments of a given read onto the reference genome, prioritising the mapping associated with the lower number of mismatches, indels, and mapping quality.

5. STAR returns SAM/BAM files that contain the aligned reads with their mapping positions on the reference genome and their overall quality alignment score.

The final output is a BAM file, which can be interpreted as a global map of the genome, in which each read is assigned to a unique pair of "genomic spatial coordinates".

[Sri+20] benchmarks a huge collection of alignment and mapping methods to determine their impact on the estimation of transcript abundance estimation. They observe significantly different and variable performance between lightweight-mapping and more traditional alignment-based methods. They notably observe that preprocessing spliced alignment to the genome and then projecting these alignments to transcriptome provides better mapping performance, compared to directly aligning against the transcriptome.

Quantification Once all reads are aligned to a reference genome, the final stage of the RNA-seq pre-processing workflow involves the estimation of transcript abundances, generally under the form of raw count measures at the gene or transcript isoform level. The final number of high-quality reads that could have been mapped unequivocally to the reference genome, is the *library size*, alternatively the *sequencing depth* (see also Panel d, in Figure 1.4).

Historically, only the reads overlapping perfectly a given exon of the original genome sequence were annotated, for instance consider HTSeq [APH15] and FeatureCounts [LSS14] tools. Since then, advanced methods that are particularly effective with limited annotation information involve a preliminary construction of Contigss, enabling to rebuild from scratch newly unobserved transcript variants by including junction reads and unannotated transcripts (see for instance Cufflinks [Tra+10] or StringTie [Per+15] tools).

Ultimately, Kallisto [Bra+16] and Salmon [Pat+17], an updated version that accounts for sample-specific biases, such as fragment GC content or positional bias, are both lightweight methods relying both on a probabilistic framework and the identification of k-merss regions, to make the alignment methods more scalable and accurate. Hence, they both stand out as cutting-edge methodologies by achieving similar accuracy to predecessors while being significantly faster.

We delve into details about the functional operations involved in FeatureCounts [LSS14], since it is the primary method to quantify reads for a given gene implemented in our industrial Nextflow RNA-Seq pipeline (refer to Appendix A.2):

1. Any quantifier algorithm usually accepts SAM/BAM files, storing the mapped reads and their position on the reference genome, as returned by any aligner tool, such as STAR or HISAT2 (refer to Section 1.2.2).

2. Quantifier algorithms additionally require a transcript feature annotation table, which capitalises on databases like GTF or GFF (General Transfer/Feature Format), to select a subset of reads mapping the regions of interest (usually exons, introns, or whole genes).

3. Each mapped read is then assigned to the most appropriate functional gene, matching the position of the read with the most universally known boundaries of the known genes on the genome. If a read overlaps with multiple features, genes are ranked by order of priority, receiving a larger weight if associated with a longer transcript or closer to the read's start position.

4. Once unambiguously associated to a known biological feature, the algorithm simply counts the number of reads assigned to a given gene, and this integer number is in turn considered as a proxy of the expression level or abundance of that gene.

Of note, raw counts from RNA-Seq data are in reality, inherently *compositional*, meaning they only reflect relative gene expression levels, due to variations in sequencing library depth across samples and batches.

This variability in sequencing depth can complicate direct inter-sample comparisons, but approaches like spike-in normalization (consider for instance the Bioconductor package BRGenomics ([DeB23]), or the use of normalization functions (Appendix A.3.3), can facilitate meaningful comparison between samples.

Experience-dependant and quality controls A variety of quality controls and optional operations can additionally be performed throughout the whole RNA-Seq workflow process, depending on the biological context and the objectives of the study.

Ribosomal RNA makes up a significant portion of total RNA and can even dominate nucleic RNA-Seq data. When the study focuses on protein-coding genes, bioinformaticians often remove rRNA reads using tools like `SortMeRNA` [KNT12] or `rRNASelector` [LYC11].

Adaptive trimming removes low-quality bases and trims adapter sequences on raw reads, using adaptive trimming tools like `Trimmomatic` [BLU14] or `Cutadapt`[Mar11].

`FastQC` [And10] is a tool designed to compute multi-quality control metrics, such as read length distribution, per-base sequence quality, GC content, and adapter contamination, helping to identify potential biases in the data at an early stage of the workflow pipeline.

`RSeQC` [WWL12] and `SAMtools` [Li+09] provides additional quality control metrics, including gene body coverage, read distribution, and strand specificity, performed on the BAM files outputs of aligner tools (refer to Section 1.2.2).

The main stages involved in any RNA-Seq sequencing workflow are summarised in Figure 1.4.

Microarray vs RNA-Seq

RNA-seq offers several advantages over microarrays.

Firstly, it doesn't require a labelled probe or a well-annotated genome, making it versatile for studying model organisms without reference transcriptome [WGS09], whereas microarray methodologies require a reference organism to build customised, complementary probe sites.

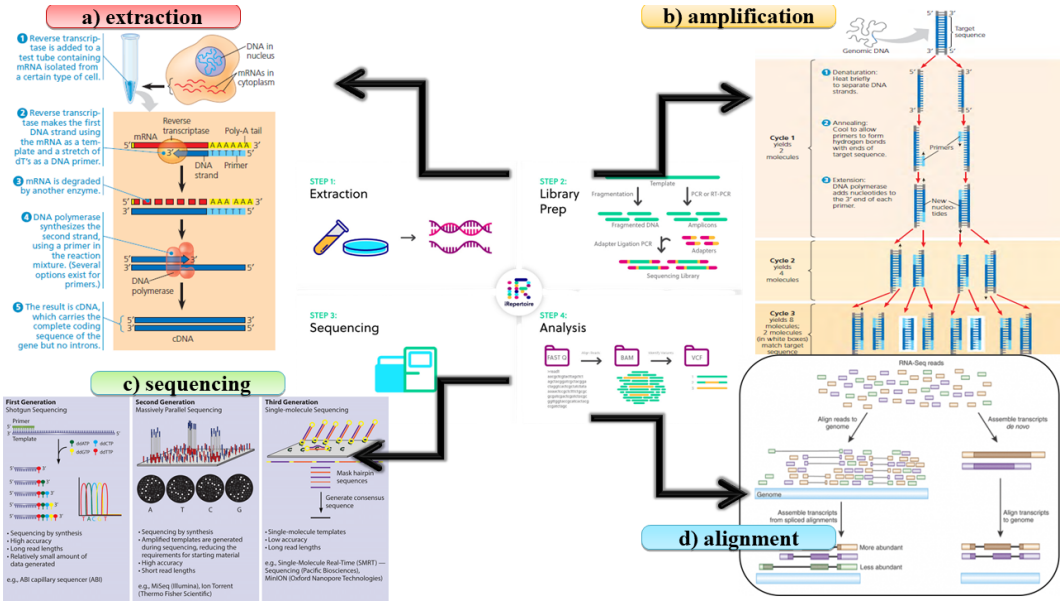


Figure 1.4: The entire NGS workflow can be broken down into four steps:

1. *Sample extraction:* When RNA is the starting template, an additional step, between the RNA extraction, and the library preparation itself, is required to convert the RNA into c(omplementary)DNA by reverse transcription (reproduced from [Cam+20, Fig 20.10]).
2. *Library preparation:* Preparation of a sequencing library typically involves two steps: 1) *RNA amplification* to increase the amount of appropriately sized target sequences and 2) the addition of *adapters* to uniquely identify them. PCR (for polymerase chain reaction) is one of these techniques to copy many times any target sequence within a test tube, which only requires double-stranded DNA containing the target sequence, a heat-resistant DNA polymerase, all four nucleotides, and two single DNA strands that serve as primers, one for each end of the target sequence (reproduced from [Cam+20, Fig 20.7]).
3. *Sequencing RNA library:* Currently, we generally classify all sequencing platforms into three generations, each coming with distinct strengths and weaknesses (subfigure c reproduced from [Ron+16, Fig .1]).
4. *RNASeq analysis:* when studying an organism with a reference genome, it is possible to directly map the reads onto the reference transcriptome. However, without a reference organism, the individual genome must be reconstructed from scratch: "de-novo assembly", see key Contigs and Genomic scaffolding for details (subfigure d reproduced from [Ron+16, Fig. 1]).

The general framework of this illustration is reproduced from [G20, Fig. 1].

Secondly, RNA-seq provides a broader dynamic range for expression levels, quantifying gene expression as integer counts, and reducing issues related to noise and saturation [WL09].

Thirdly, RNA-seq is highly versatile, enabling the detection of alternative splicing isoforms and rare genetic mutation disorders, including single nucleotide polymorphisms (SNPs) and complex transcript fusion events. These capabilities make RNA-seq a powerful tool for transcriptomic analysis [Man+14].

Comprehensive benchmarking of RNA-seq against microarray technologies, relying on concordance correlation with RT-qPCR expression data, is proposed in [Eve+17].

To mitigate these statements on the presumed better performance of RNASeq-based techniques over microarray, [Zha+15] demonstrates that while RNA-seq outperformed in quantifying the transcriptome of neuroblastoma, both methods performed similarly in predicting clinical endpoints (see also [Xu+13] for a benchmark between Illumina RNA-Seq and Affymetrix).

Confirming empirically these specific clinical examples, the meta-review [MO11] and opinion paper [Man+14] confirm the utility of microarrays, notably as being more cost-effective and reliable for gene expression profiling in model organisms. In addition, both approaches exhibit a strong congruence when asked the same biological question.

In the subsequent section 1.2.3, we review emerging RNA-Seq applications that enhance our understanding of biology, by going down to single-cell level, or by preserving the spatial and temporal context of tissue structural organization.

1.2.3 Perspectives: single cell and spatial transcriptomics

Limitations of RNA-Seq technologies Conventional RNA-Seq technologies, while performing accurate and high-throughput sequencing, are tailored to dissect bulk mixtures, aggregating different biological entities. Thus, they hinder the identification of key biological drivers involved in complex cellular processes.

Additionally, RNA-Seq outcomes are simply snapshots of the current transcriptome state, without temporal or spatial context. Thus, they can not be used to understand long-term interactions occurring between distinct biological compartments.

Single-cell RNASeq technologies

Outline of Single-cell RNASeq Single cell RNA-Seq is a promising technological advancement, that has notably been employed to investigate rare subpopulations at a single-cell resolution level, and which are typically silenced in bulk transcriptomics [GA18].

The isolation of individual cells, notably those strongly embedded within tissues, is one of the major challenge of scRNA-Seq ([FFF15]). Historical methods relied on manual isolation of individual cells, hereby limiting the analysis to a few dozen cells, including micro manipulation [Tan+09] and laser capture micro-dissection (LCM) techniques [GCS21].

In contrast, modern approaches involve high-throughput and automated dissection and sorting of tissues, encompassing microdroplets [BR19], microfluidics (10X Chromium,[Zhe+17] [Sar+19]) or microchips (Smart-seq2, [Pic+13].

scRNA-Seq additionally require amplification of minute amounts of RNA for each individually captured cell. The sequencing and amplification of RNA-Seq implies the same stages as standard RNA-Seq methods, namely (1) reverse transcription, (2) cDNA amplification with PCR, for

an exponential amplification, or in vitro hybridisation, for a linear amplification, and (3) the sequencing of the library itself (see Section 1.2.3, Section 1.2.2 and [Isl+14]).

Applications and limitations of single-cell The high resolution of scRNA-Seq technologies allows a better understanding of complex biological pathways, which were not accessible to standard methods of bulk cell population profiling, a snapshot of new avenues offered by this groundbreaking technology is unveiled in Section 1.2.3. Notably, they can be used to finely characterise subpopulations within a sample by identifying cellular cell markerss ([Seg+18]).

Pioneering and promising approaches go a step further, by attempting to reconstruct the continuum of cell states rather than definite stages [Ren+20]. Such models even allow the possibility to switch from one cell phenotype to another.

pseudo-time analysis hence consider a continuous paradigm, able to provide temporal context of the mechanisms involved in the regulation of cell differentiations. Notably, this approach aims at inferring "cell trajectories", namely ordering the developmental stages of single cells, based on the variations of the transcriptome. For instance, [Mys+21] reconstructs the branching trajectory of the cell state transitions conducing naive cells into differentiating into cytotoxic CD8+ T cells, with the DDRTree algorithm [Qiu+17]. Monocle [Tra+14] uses a minimal spanning tree approach to prove that CX3CR1+ macrophages and iNOS+ macrophages could not coexist in an early tumoral stage.

We should point out that scRNA-Seq is proned to technical biases, expensive and resource-consuming ([Pfi+21], [DC21]).

The major limitations of scRNA-Seq encompass *dropouts*, biased capture of cell types, detection of "doublets" or deal cells and identification of isolated cell types.

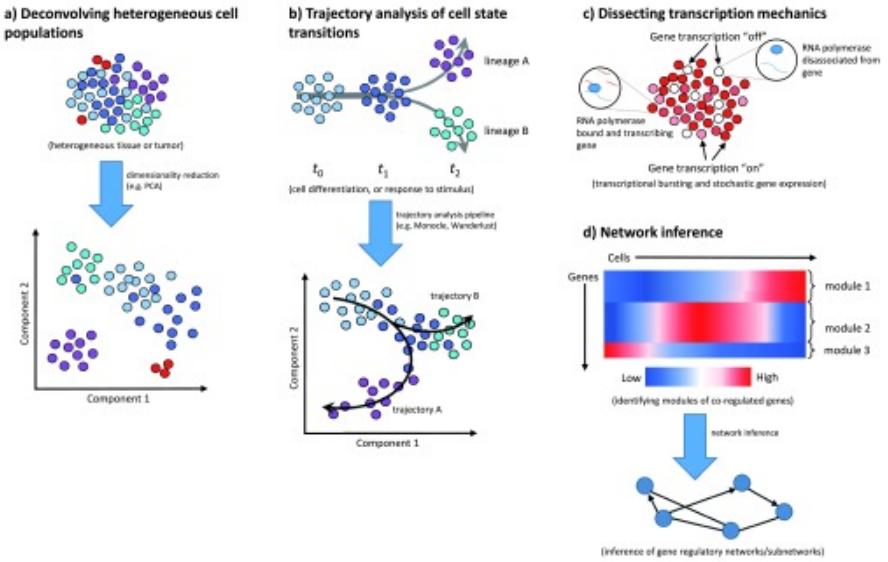
dropouts are technical artefacts, due to inefficient RNA capture, that result in highly variable and sparse transcriptomic expression matrices, exhibiting numerous null values [Xu+20]. These dropouts are challenging for downstream analysis, masking gene-gene relationships in complex niches or blurring detection of rare genetic variations ([Qiu20], [Kim+21] and [Lei+20]).

As RNA-Seq, the library preparation is a destructive process ([Tan22]). [Lam+18] posits that particular cell types, such as neutrophiles, were more susceptible to undergo deadly damage, resulting in a consistent underestimation of their expression. In addition, scRNA-Seq can not be applied on biological tissues that are strongly intertwined, such as brain neurons ([Sos+21]). Finally, isolation can induce "ectopic" expression, resulting from the stress induced by the lysis process ([vdBri+17]).

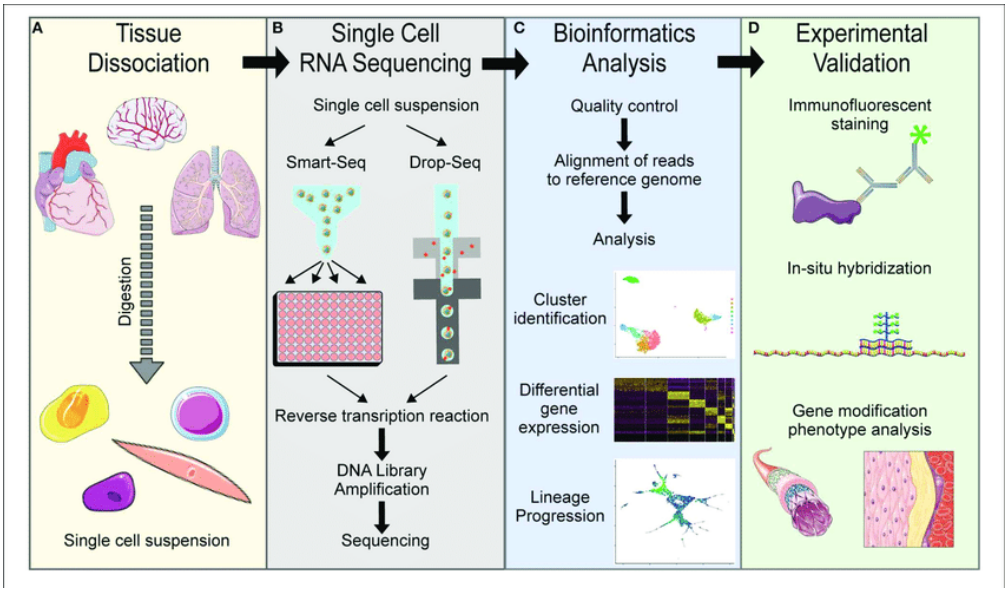
The annotation of each isolated cell, either mapped to a known cell type or to a novel one is challenging. Specifically, semi-supervised clustering techniques dedicated to that task must not confuse rare cell subtypes with aberrant transcriptomic profiles or alterations proceeding from the stochastic and plastic nature of transcriptomics regulation ([PY15], [Sha+14]).

Paired captured cells at the same spot, termed "doublets", generate a hybrid profile, that might wrongly interpreted as a intermediate developmental stage, especially when proceeding from distinct cell types. Deconvolution algorithms, such as DoubletFinder [Han+21] and DoubletDecon [XL21], can automatically remove multiplets and dead cells upstream.

Spatial transcriptomics While scRNA-Seq enables to deconvolve cell populations within a tissue, it does not capture their spatial distribution, nor reveal in real time, *in situ* cellular mechanisms. On the contrary, ST methods aim at replacing variations of expression profiles in



(a) Common applications of single-cell RNA sequencing.(a) Deconvolving heterogeneous cell populations. The single-cell resolution level can enhance the identification of rare cell species or subtypes. (b) *Trajectory analysis* of cell state transitions. (c) Dissecting the intricacy of transcription kinetics, inherently stochastic. (d) Network inference. Genes can be clustered by expression profiles to identify modules of co-regulated genes, further unravelled through studying the covariance matricial structure to infer gene regulatory networks. Reproduced from [LT16, Fig .1].



(b) Overview of a standard single cell RNA sequencing pipeline.(A) Tissue dissociation at the cellular level (B) Single cell RNA sequencing of the single cell suspension. (C) Bioinformatics analysis of the library of reads sequenced. (D) Experimental validation. Reproduced from [CH20, Fig. 2].

their spatial context. ST encompass two main techniques, each with its strengths and limitations (see also Section 1.2.3).

Image-based approach: It comprises *in situ* sequencing (ISS, [YGN20]) and *in situ* hybridisation (ISH, [Vic+19]), both methods using probes to target specific genes. ISH notably encompass seqFISH+ [Eng+19] and MERFISH [Che+15])protocols, in which target sequences are hybridised with a complementary fluorescent probe. Recently developed multiplexing methods do not rely anymore on an unique fluorescent barcode for each transcript, significantly increasing their sequencing analysis. For instance, HighPlex RNA imaging ([He+22])) is accordingly able with n distinct fluorescent colours and k sequential rounds of hybridisation cycles to unambiguously set apart k^n distinct transcripts. While these methods exhibit a lower coverage of the whole genome, requiring a prior selection of target genes, and strong noisiness, resulting from the *molecular crowding* phenomena [6], their single cell resolution and their conservation approach make them relevant to track complex temporal dynamics.

in-situ capture (Spatial Transcriptomics [Stå+16] and Slide-seq [Rod+19]) is a contemporary method that measures the transcriptome level for each of the individual spots of a finely-tuned lattice and tag it with an unique spatial "ID" barcode, with the spatial coordinates of the associated spot. Then, the sequencing protocol reconstructs the nucleotides sequence, while keeping track of its original localisation, a technique consistently termed *spatial barcoding* (Section 1.2.3). One of the current best performing methods, Visium, released by 10x Genomics, displays an increased resolution (55 µm in diameter) and sensitivity ($\sim$ 10 000 transcripts per spot) [NSH20]. Compared to the Image-based approach, spatially resolved transcriptomics is cheaper, and enables enhanced and agnostic coverage of the transcriptome.

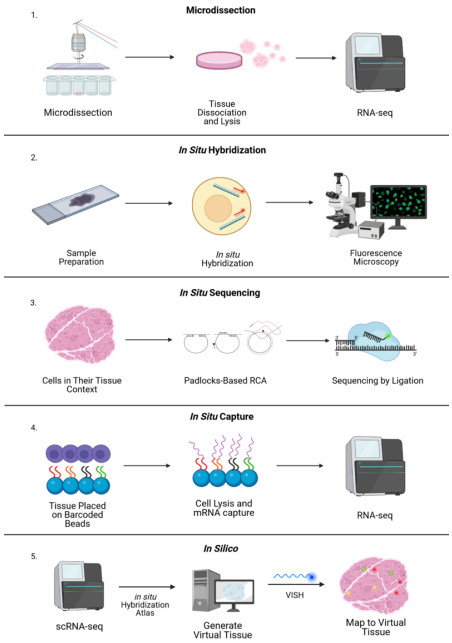
Both techniques return a three-dimensional tensor of transcriptomic expressions, each gene being represented by a matrix of intensities representing its 2D spatial expression profile, as pictured in Section 1.2.3. The resolution and conservancy of ST is likely to reveal new biological insights. Spatial transcriptomics has already been successively applied to survey dysregulated expression patterns, induced by neurodegenerative disorders (Alzheimer's disease [Che+20], AML [War+20]) to immuno-inflammatory affections (influenza, [Cur+21], sepsis, [Jan+21]). See also [Rao+21, Fig. 2] for a graphical summary of biological applications enabled by ST, at any layer of the organism.

1.2.4 Conclusion: The Significance of Transcriptomic Data in Computational Medicine

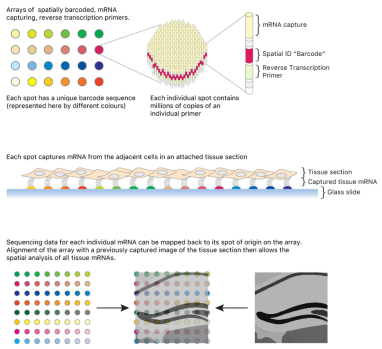
High-throughput sequencing analysis, streamlined by RNA-Seq methodologies, is double-edged: it unveils valuable biological insights, yet the inherent noisiness and high-dimensionality of the generated datasets demand finely-tuned and dedicated statistical methods (Appendix A).

To that end, a variety of downstream analyses (see Appendix A.4) have specifically been developed, encompassing *differential analysis* for identify genes impacted by biological state, patient stratification (Chapter 4) ensuring patients receive the most effective therapies, and identification of key biomarkers associated with disease progression or treatment response.

[6] Molecular crowding designs the spatial overlap of fluorescence signals, restraining the number of fluorescent tags used to a few dozens.



(a) Overview of Spatial Transcriptomics Methods.(1) Microdissection method. (2) *in-situ* Hybridisation method. (3) *in-situ* Sequencing method. (4) *in-situ* Capture method, alternatively named spatial barcoding. (5) *in-silico* method. Reproduced from [Sli23, Fig .1], using the BioRender software [].



(b) Zoom on the *spatial barcoding* technique.Reproduced from [Che23, Fig. 3].

In addition, the relative affordability of RNA-Seq technologies, compared to methods quantifying the proteome, make them relevant to approximate the proteome. Yet, post-transcriptional events (Section 1.1.1) mitigate the precision and reproducibility of such studies, as discussed in [Mei+13].

By pinpointing dysregulated pathways, downstream analyses streamline the development of targeted therapies that focus on the molecular drivers of a disease. More generally, real-time monitoring of the variations of gene expression can allow meticulous adjustment of drug doses, uphold switching to alternative treatments or design robust clinical trials, with more homogeneous cohorts.

Overall, transcriptomic data is a major input of computational medicine, enabling a deeper understanding into the gene expression patterns and regulatory mechanisms that underlie the variability of disease profiles and individual treatment responses.

Nonetheless, bulk RNA-Seq analysis is oblivious to variations in cell-type composition, averaging the contributions of individual cell subtypes. In the following chapter 2, we go down a biological strata, by specifically detailing the interactions that occur across immune cell types. We will notably see that the intertwined network between the innate and adaptive system are essential for repelling efficiently harmful invaders.

We have opted to focus on immune cell populations for two primary reasons. First, all the studies presented in this manuscript harness whole blood samples, in which white cells represent the primary contributors to transcriptomic expression. Red cells and platelets, being enucleated cells, are unable to engage in protein synthesis or mRNA transcription. Secondly, the progression of diseases under scrutiny, notably primary Sjögren's disease, are predominantly driven by dysregulated immuno-inflammatory processes.

Chapter 2

Introduction to the Immune System

Biological Perspective: Introduction to the Immune System In this chapter, we shed light on the actors of the immune system. The immune system plays a central role in defending the body against pathogens and maintaining homeostasis. Notably, we highlight throughout this chapter the significant role of transcriptomic signalling pathways, which finely tune interactions between cellular populations and play a pivotal role in the flexibility and effectiveness of the human immune system. Specifically, we emphasize the importance of cooperation between innate immune actors, which are generalist but presenting numerous harmful side effects, and adaptive immune actors, highly specialised, but requiring upon prior activation.

However, we recall in a second phase that a meaningless hitch in this well-oiled machinery can lead to immune system dysregulation, and in the worst-case scenario, leads to uncontrolled metastatic proliferation or the development of debilitating autoimmune diseases, depending on the circumstances.

2.1 Key actors of the Immune System

Overview: the immune system, the keeper of our body The immune system is an intertwined network of cells, and molecules that collaborate in protecting the body against infectious agents. Such threats are collectively named *pathogens*, ranging from viruses to bacteria and fungi.

Its primary role is to early recognise these intruders, in order to eliminate them quickly, while avoiding as much as possible side effects against innocuous molecules or self entities. Specific identification of harmful elements is enabled by *molecular recognition*, and perturbations in its underlying mechanism is at the root of most immune disorders. The immune system accordingly plays a crucial role in keeping the homoeostasis of the human body.

The immune system is traditionally split between the actors of the *innate immune system* (Section 2.1.1) and those of the *adaptive immune system* (Section 2.1.2).

2.1.1 The innate system

The innate immune system is the first line of defence and provides immediate, *non-specific* protection against invading pathogens. This includes:

- **Barrier defences:** *Physical barriers*, such as the skin and mucous membranes, are the first line of defence, preventing pathogens from accessing the inner environment. However, entire sealing is impossible, since the human organism must interact with its environment, through gas exchange and nutrition to maintain homoeostasis and perform vital biochemical functions.

  Thus, these protections are completed with chemical secretions trapping pathogens, specialised organelles such as lysosomes and reservoir of innate cells poised to fend off potential invasion.

  The mucous membranes of epithelium tissues, such as the digestive or airway tract, are the ultimate barriers against infection. Mucus, saliva and other enzymes secreted prevents colonisation by fungi and bacteria, by destroying their cell walls and increasing the acidity of the ambient medium.

- **Naive and generalist Immune cells** These cell populations recognise proteins universally shared among pathogens that bound to the Toll-like receptors (TLRs). TLRs, for which the Nobel Prize in Physiology was awarded in 2011, is remarkably widespread and preserved in the animal kingdom [AT04a]. In details, the TLR3 complex binds to double-stranded RNA, a common nucleic organisation in viruses [MS08] while the TLR4 and TLR5 respectively target the lipopolysaccharide found on the surface of many bacteria [Tak+99] and flagellin [Gew+01], the main protein of bacterial flagella.

  Upon recognition, *macrophages* ("big eaters") and *neutrophiles* wipe out pathogens that breach the natural physical barriers by phagocytosis. Generally, influx of neutrophils precedes the arrival of monocytes that rapidly differentiate into macrophages within tissues.

  Natural killer (NK) cells are a type of immune cells known for their ability to induce the death of virus-infected or abnormal cells. They do so by detecting surface receptor proteins, like stress signals or tumour antigens, and releasing toxic molecules that initiate *apoptosis*, a programmed cellular death process.

  Mast cells contribute to the inflammatory response but are also involved in aberrant reactions such as allergies, alongside *eosinophils* and *basophils*. Their primary function was the protection against multicellular parasites, such as worms.

  Finally, *dendritic cells* are more specialised, and ensure the coordination between the innate and the adaptive response (section 2.1.3).

- **Signalling and antimicrobial proteins** The *complement system* is a set of of 30 identified proteins in blood plasma that interact with each other in a highly coordinated and sequential manner. This cascade of biochemical reactions leads to the formation of protein complexes that trigger the *lysis* of invading cells (destruction through bursting of the membrane), the opsonisation (marking for destruction) and the recruitment of immune cells to the site of infection. The complement system additional plays a pivotal role in tissue repair and development.

  Interferons interfere with cells hosting virus, whose activation patterns depend on their type: α, β and γ. Interferons are released by virus-infected cells and by binding to neighbouring immune cells, they inhibit viral replication or promote the phagocytic ability of macrophages.

  Cytokines, released by various immune cells (and even non-immune ones), play a dual role in inflammation. Some, like interleukin-1 (IL-1) and tumour necrosis factor alpha (TNF-alpha), are pro-inflammatory and can recruit for instance neutrophils. Others, on the contrary, resolve inflammation and facilitate tissue repair, like interleukin-10 (IL-10) and transforming growth factor beta (TGF-beta).

- The *inflammatory response* (from Latin *inflammare*, "set on fire") is the set of events that modify the microenvironment, triggered by an intertwined cellular signalling released upon infection. The first stage often involves mast cells secreting cytokines, such as *histamine*, that promote growth, migration, and activation of endothelial cells, thus contributing to *vasodilation* [RTA20]. The dilatation of blood vessels causes the well-known, localised inflammatory response, characterised by the increase of the skin temperature, redness, and enhanced blood flow.

  Subsequently, the released cytokines promote the migration of immune cells towards the inflamed region through a process called *chemoattraction* (movement of a cell following the chemical gradient of a signalling molecule). Then, the coordinate interaction between signalling mechanisms and cellular responses keep on sustaining the inflammation process, with the deployment of the complement system and the recruitment of additional immune cells [Mol+20].

Unfortunately, this generalist first line of defence may reveal insufficient to fend off particularly harsh pathogens strengthened by millions of years of co-evolution. Indeed, some pathogens evolve specifically to overbalance the immune defences (a set of mechanisms referred to as *immune escape*). For instance, some bacteria have an outer capsule that prevents recognition, while others are resistant to breakdown by lysosomes.

2.1.2 The adaptive system

The adaptive immune system, on the other hand, provides a targeted response to pathogens that have already overwhelmed the innate system, through receptors specific to the intruders. Adaptive immunity relies on two types of lymphocytes: B cells and T cells [1].

All receptor proteins on a single B or T cell share the same antigen receptor and recognise the same *antigen*, any foreign molecule detected as non-self and able to elicit recognition. To recognise any potential antigen, millions of distinct lymphocytes coexist in the body, each with its own recognition pattern that is able to bind to a protruding antigen surface or a circulating agent, such as microbial toxins.

The adaptive immune response decomposes into four stages:

(a) **cell diversity** It is a crucial step for generating the diverse array of lymphocyte subtypes, which, in turn, gives them the capacity to recognize a broad range of antigens [2]. Random alternative splicing and gene recombination processes, specifically referred to as "V(D)J recombination", play a pivotal role in creating unique sequence arrangements.

(b) **Self-tolerance** This step aims at eliminating self-reactive lymphocytes, namely those that recognise own molecules as non self, and thus could trigger improper immune reaction against the body's own molecules and cells. These self-reactive lymphocytes are either destroyed through apoptosis, or rendered non functional.

[1] Like all blood cells, lymphocytes originate from stem cells in the bone marrow, but while T cells migrate to the thymus, B cells undergo their maturation stage in the bone marrow. Generally, Hematopoiesis is the biological process involved in renewing all the cell populations circulating in the blood, while the corresponding scientific field focuses on resolving their lineage relationships.

[2] [BSA02] and [Guo12] demonstrate that one million of different B cell antigen receptors and 10 million different T cell antigen receptors coexist in the human organism.

(c) **Clonal selection and Cell Proliferation** The activation of a unique set of B or T cells hinges on the binding between the antigen's epitope and an antigen receptor, primarily taking place within the lymph nodes ([Cam+20, Figure 6, Chapter 43]).

Upon activation by binding to an antigen, lymphocytes undergo *proliferation*, forming a clonal population of cells carrying thousands of receptors targeting the same antigen. Subsequently, these cells differentiate into "effector cells", including (1) T CD4 or Th Helper cells that coordinate the adaptive response through the clonal amplification of effective lymphocytes, (2) T CD8 cells responsible for eliminating virus-infected host cells, and (3) activated B cells, also known as "plasma cells", which produce soluble proteins called antibodies circulating within body fluids. This entire process is commonly referred to as *clonal selection*.

Helper T cells, by their multifaceted role, notably in coordinating and conducting the actors of the adaptive response, play a pivotal position. Upon recognition of antigens, presented by class II MHC molecules of antigen-presenting cells (APC) (dendritic cells (DC), but also macrophages or B cells), T CD4 cells activate metabolic pathways secreting cytokine. Excreted signalling molecules, in turn, enhances clonal amplification and triggers two immune responses, namely the humoral and cell-mediated immune response, depending on the nature of the antigen ([Cam+20, Figure 18, Chapter 43]).

(d) **Immunological Memory** The immunological memory is responsible for the long-term protection elicited by a prior infection. This protective mechanism differs from the *primary immune response* by a faster and prolonged response (peak intensity within two days against 10 days), with a greater magnitude (the concentration of antibodies or killer T cells increases by two or three folds).

This heightened secondary response to the same antigen relies on a subset of the effector lymphocytes, called *memory cells*. Upon initial activation following the exposure of an antigen, memory lymphocytes are preserved in storage tissues. Indeed, while the majority of lymphocytes are eliminated once the infection is overcome, by **regulatory** T cells, their longer life places memory cells at the forefront to initiate clonal amplification of thousands of highly-specialised effector cells in case of repeated exposure.

This mechanism finds direct application in the process of *immunization*, which involves the deliberate introduction of antigens into the body to induce the production of memory cells. Better known as *vaccination*, this concept dates back to the late 1700s, witho Jenner's observation of better protection conferred by cowpox early exposure, against the deadliest smallpox. In modern times, vaccines have evolved from first-generation formulations containing killed or weakened pathogens to third-generation advancements, such as mRNA vaccines by Pfizer and Moderna, and overall contribute to significant reductions of infectious and crippling diseases, such as polio and measles.

2.1.3 Exchange of goodwill between the two immune systems

Contrary to the standard approach which opposes the innate system, shared by all animals, to the seemingly more efficient adaptive system, only developed among jawed vertebrates, we show in this point that it is the constant interplay between lymphocytes and other innate cells that together provides this efficient, coordinated and constantly remodelled protection against a variety of pathogens.

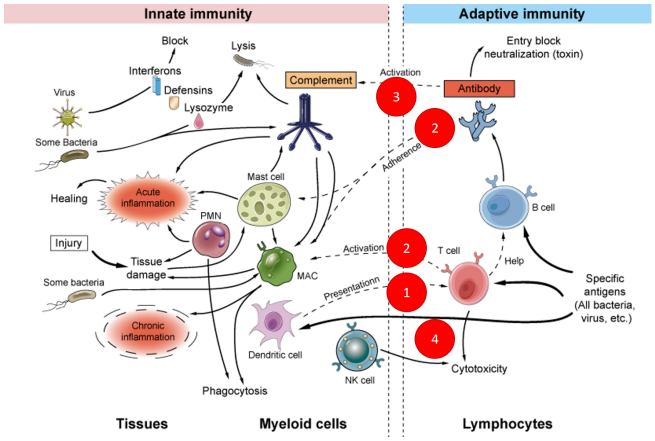


Figure 2.1: Cooperation mechanisms between the innate and the adaptive immunity. This iconography shows the inner mechanisms intervening in the innate (left) and adaptive (right) response, each able to trigger a cellular (lower half) or humoral reaction (upper half). Reproduced from [Dia20, Fig .1].

The cooperation between innate and adaptive system generates a positive feedback loop whose main stages are reported on Figure 2.1, emphasising how much these two systems are closely intertwined:

The multiple interactions occurring between the older and generalist innate system with the adaptive and highly-specific immune system include:

- A local inflammatory response generally produces *pus*, a mixture of white blood cells, and dead pathogens, that is flushed away through the lymphatic system towards the lymph nodes. There, residential macrophages, or circulating dendritic cells, can engulf antigenic fragments, then exhibit them on their surface [Cam+20, Figure 12, Chapter 43]. Precisely, the *phagocytosis* pathway cradles antigens in the MHC II complex.

  Finally, the interaction of the MHC with its antigen fragment and the TCR receptor of a TCD4 cell [Cam+20, Figures 12, 13 and 23, Chapter 43], triggers an adaptive immune response, through either a cell-mediated or humoral response.

- Antibodies directly facilitate phagocytosis, by aggregating toxins or pathogens and marking them to macrophages and neutrophils. In return, the phagocytosis enables macrophages and dendritic cells to capture antigens and ultimately stimulate helper T cells, which activate the very B cells whose antibodies contribute to phagocytosis. Indirectly, cytotoxic T cells, by bursting out cells hosting virus, exposes viral contents and increase the likelihood that APC or antibodies trap foreign peptides, which would have remained out of reach otherwise.

- Complement to neutralisation and opsonisation mechanisms, antibodies interplay with the proteins of the complement system [Cam+20, Figure 21, Chapter 43]. The associated cascade of biochemical reactions is likely to promote cell lysis.

- The virus uses the cell's biosynthetic machinery to replicate whom some of them can appear on the cell surface through the MHC class II. Recognition of these protruding epitopes by the complementary antibodies could possibly promote the recruitment of NK cells. Notably, [GR14] premises that T cells could act as antigen-specific sensors to amplify the local immune response of *innate lymphocytes* (alternative name for NK cells, as non specialised cytotoxic T cells).

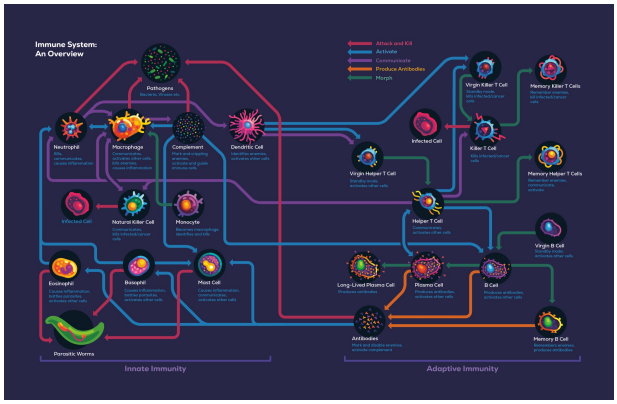


Figure 2.2: An overview of the immune system.The red and the blue edges display direct cell-cell interactions, requiring ligand-receptor bound. On the contrary, the purple and orange edges characterise long-range interactions, through the release of respectively chemical signals or antibodies. Finally, the green edge depicts differentiation from one developmental stage to another for a given cell lineage. This figure is reproduced from [Det21, Figure 1, Chapter 42].

From the previous subsections, we understand the importance of cross-talks between distinct immune cell subtypes in order to provide a balanced immune reaction against any intruder. Exploring all interactions within this intricate machinery would necessitate a whole volume of an encyclopedia, where researchers are still debating regarding the functions of certain immune cell types.

We summarised in Figure 2.2 the interactions occurring between the main actors of the immune system:

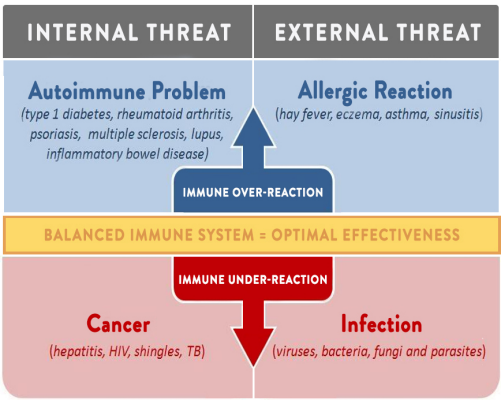
In the following section 2.1.4, we instead focus on the disorders of the immune system, when it is not anymore in its operating order.

2.1.4 Immune dysregulation

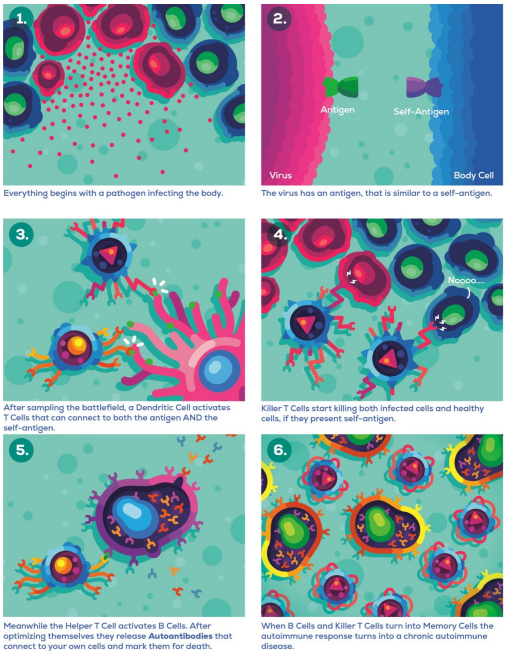
While the immune system is essential to fend off pathogens or wipe out tumoral cells, the strength of the response, especially when occurring at the scale of the organism, results in detrimental side effects. Hence, a heightened inflammatory response triggered by global tissue damage or blood contamination, is likely to lead to a life-threatening condition, called *septic shock*. It is notably characterised by an increase by several folds of the number of white blood cells, a higher temperature resulting in *fever* at the human organism scale and low blood pressure. We discuss of one of these exacerbated mechanisms in the context of patients infected by a COVID-19 strain, in Appendix F.

Additionally, not only pathogens can be recognised as non-self, but also any foreign molecule, even issued from other human bodies. The huge number of variations of the MHC molecules [3] between two individuals prompts transplants or grafts rejection. The only way to counterbalance rejection is to pair match the MHC molecules of the donor and the receptor as much as possible, and to use immuno-suppressor drugs. A similar mechanism is involved in blood transfusions, but instead of the HLA complex, glycoproteins on the surface of red blood cells are recognised as foreign, eliciting lysis of the transferred blood cells, necrosia and kidney failure ([OC08]). For

[3]The MHC complex not only displays antigen fragments, but a subset of the proteins composing it, the Human leukocyte antigens (HLA) act as an identity card that asserts the immune system that the investigated cell belongs to self



(a) Autoimmune diseases and cancers: the two sides of the same coin. Regarding cancer, the main mechanism involving immune cells is *tumour escape*, in which the immune response is incapable of eliminating self-cells that have undergone transformation. Conversely, autoimmune diseases are characterised by an over-active immune response directed towards self-particles wrongly recognised as antigens, resulting in tissue damage and chronic inflammation. However, autoimmunity and cancer both hinge on a failure of the immune system in controlling abnormal cell proliferation (respectively auto-reactive immune cells or tumour cells). Reproduced from [Sar18, Fig .4].



(b) Aetiological factors of auto-immune diseases Dysregulation at several layers of the immune system, such as the complement system, interferon or cytokine production, can contribute to a variety of autoimmune diseases (Crohn's disease) and inflammatory disorders. Cross-reactivity is thought to be one of the major aetiological factor initiating auto-immune diseases. Reproduced from [Det21, Fig. 1, Chap. 40].

long, the only way to overcome this process was to control, prior to the transfusion, that the the patient and the donor had compatible blood groups.

Nonetheless, all previous situations correspond to an expected behaviour of the immune system against foreigners, and it is either medicine progress or particularly virulent pathogens that trigger life-threatening conditions. However, the immune system can display aberrant behaviour, often elicited by a failure of the regulation mechanisms (see Section 2.1.4 and [Sch+20]).

Bad regulation of the immune system often come as the two sides of a coin. *Dysregulation* of the immune system leads to an over-activation of the immune system [RRA15], while *misregulation*, an under-activation of the immune system plays a prominent role in the evolution of the cancer, acting as the prime actor of immune escape [Cha+22].

Although auto-inflammatory diseases, triggered by an over-activation of the immune system, are typically not fatal, they are debilitating affections. The Sjögren's primary syndrome, one of the disease addressed by Servier, is notably described in Chapter 4.

Briefly, we display in Section 2.1.4, the cross-reactivity mechanism that elicits the production of self-antigens, believed to underscore the parthenogenesis of most immuno-modulated diseases.

If there a key message to keep in mind, the reader should recall that the immune system is an

intricate and interconnected network of cell populations and cytokines interplaying altogether to safeguard the body against foreign invaders. However, even a slight disruption in one of these mechanisms can result in a potentially life-threatening condition. Hence, fine-tuning the immune system is a highly challenging task, requiring a comprehensive and systemic understanding of the actors involved.

Related with auto-immune subjects, and having personally to debate on the hypothetical benefits of alternative therapies over complex and costly treatments designed by pharmaceutical companies, you should never fall for the miracle effects claimed by the homeotherapy branch. To quote the author of the `Kurzgesagt - In a Nutshell initiative`, [Det21]:

> At least for now, there are no scientifically proven ways to directly boost your immune system with any products that are easily available. And if there were, it would be very dangerous to use them without medical supervision.

2.2 Physical methods for studying changes of cellular Composition

After providing a comprehensive description of the components of the immune system and their interactions, we present a proficient array of tools for estimating cell populations. No special consideration is required for quantifying immune cell populations, particularly in whole blood samples, where the isolation of individual cells is relatively straightforward.

These methods are typically categorized into two groups: flow cytometry methods, which require the physical separation of cell populations, and imaging methods, which enable the identification of cell types "in-situ", using fluorescent markers. While the former are known for their lower cost and superior processing capacity, we also demonstrate that the latter allow for the preservation of the spatial organization of cell subtypes.

2.2.1 Cytometry analyses

Cytometry analyses quantify the relative frequencies of cell subsets (both their number and type) in blood or previously disaggregated tissues in a high-throughput manner, enabling consistent and extended comparisons across samples or conditions. To that purpose, individual cells must be first isolated, then marked to identify and classify them.

Gating specifically designs this process of selecting subpopulations of cells for quantifying them. Gating is important because it allows researchers to focus on specific cell populations and exclude unwanted cells from analysis, based on their fluorescent [4] or light scattering characteristics. Gating is a crucial step in cytometry analysis as it allows for the identification and characterization of specific cell populations [Sta+19]. Gating used to be performed manually based on expert knowledge of cell characteristics, using physical methods, such as Fluorescence-Activated Cell Sorting (FACS) or Laser Capture Microdissection.

FACS require a specific set of markers, usually cell-surface proteins, and corresponding antibodies for each set of cell population included in the cytometry analysis, which might not be available for closely related cell populations. FACS are also intrusive methods, damaging cell structure and resulting in numerous dead cells, and requires a large amount of biological material.

[4] In that case, fluorophore-conjugated antibodies are used to target markers of interest. Identification of each cell's marker is then performed by detecting the stimulation of the fluorophore by a laser emitted at a specific wavelength.

Interestingly, FACS technologies pioneered the quantification of tumour-infiltrating immune cells, with early description of dendritic cells (DC) [Thu+96] or myeloid-derived suppressor cells (MDSC) (Veglia, Perego, and Gabrilovich [VPG18]).

CyTOF (cytometry by Time-Of-Flight, also termed mass cytometry) [Nom+94] is an alternative method for the analysis of single cells, close to FACS, in which heavy metals-barcoded antibodies binding to the cell-surface-expressed proteins enable unequivocal identification of the cell type. Cell content is then ionised in a plasma state and analysed using a quadrupole time-off-light (TOF) mass spectrometer. Hence, CyTOF utilises mass spectrometry to detect cellular markers while FACS uses fluorescence-based detection, enabling to characterise simultaneously a much higher number of markers compared to classical FACS owing to a reduced spectral overlap ([RE04], [Bis81], [Mae+20] and [Fri+21]). CyTOF can serve specific biological purposes, such as surveying the evolution of stereo-selective enzymes or measuring enantiopurity (see [TH18], [hos23] and Section 2.2.2).

To overcome the first bottleneck, namely isolation of cells limitations, novel disassociation systems relying on the automatic discrimination of cell sizes by means of microfluidics [Huh+05] [PG14] or dielectrophoretic separation [Wan+00] have been developed. And nowadays, the most recent FACS-gating methods is able to quantify up to 30 markers and 10 000 cells per second [SK14].

Regarding the simultaneous annotation of thousands of cells, automated gating strategies based on machine-learning approaches have been developed to improve the accuracy and high-throughput of cytometry analyses. By training these algorithms on a gold-standard dataset of manually annotated gating definitions, machine learning pipeline reveal consistently better oracles to annotate cell types than pathologists experts, overcoming the non-standardised and non-interpreted nature of gating definitions ([Li+17] and the `DeepCyTOF` approach, [Bec+19] and its `Hypergate` protocol).

LCM is a lab technique used to isolate specific cells from a histological section. It involves the user of a laser to precisely cut and capture the target cells, while leaving the surrounding tissue intact ([Lok+90]). Facing the same limitations as flow cytometry-based techniques, it requires manual annotation of the captured cells ([MO19], [Sta+19] and [McK18]), limiting its throughput capacity. In addition, LCM is more costly, resource and time-consuming than FACS [ZW21]. Nonetheless, this technique is better suited to preserve the integrity of the tissue [AS19].

2.2.2 Imaging methods

On the one hand, *Immunohistochemistry*(IHC, see Section 2.2.2) [Ju+13] and closely related immune fluorescence (IF) techniques, along with the in-situ hybridization (ISH, see for instance the FISH-Flow protocol in [Kuh+11]), enable an `in-situ` and accordingly a spatial characterization of the cell composition, contrary to the previously described cytometry workflow (see Section 2.2.1). We detail in next section the principle underlying ISH, along with its main limitations.

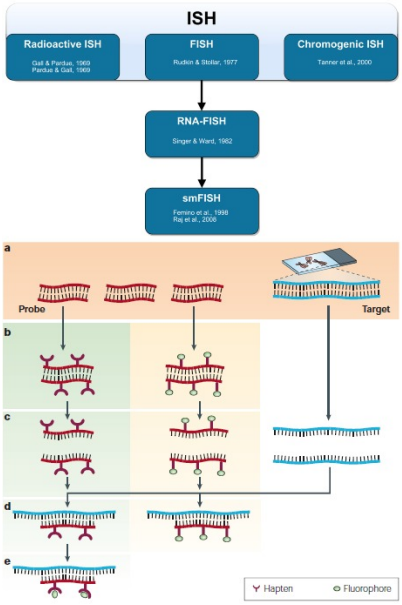
In order to fulfil an optimal microscopic visualisation at single-cell resolution, it is imperative to use tissue slides of approximately one-cell width, such as the ones returned by formalin-fixed paraffin-embedded (FFPE) samples. Individual cells are first detected by segmenting the raw images, and then classified by detecting signal emitted from the markers (can be present in nucleus, cytoplasm or on within the membrane). To spot markers, these methods use one, or more commonly, a combination of two antibodies, for an enhanced signal: the primary one targets the cell marker of interest while the secondary antibody, conjugated to either a catalytic agent (IHC) or a fluorophore (IF, see Section 2.2.2), amplifies the signal. The staining is then observed through a microscope and spatial identification of the marked cell types is performed by pathologists or

with comparable performance by a deep deconvolution neural network. For instance, [Lah+19] developed a deep learning method, UNET, that automatically segments and annotates digitised slide images into compartments of tumour, healthy tissue, and necrosis. The method utilizes a (CNN, convolutional neural network) and color deconvolution to handle staining variability. Finally, the combination of different markers conjugated with a given antibody can be used to unequivocally assign any stained cell to a specific population [Tau+18].

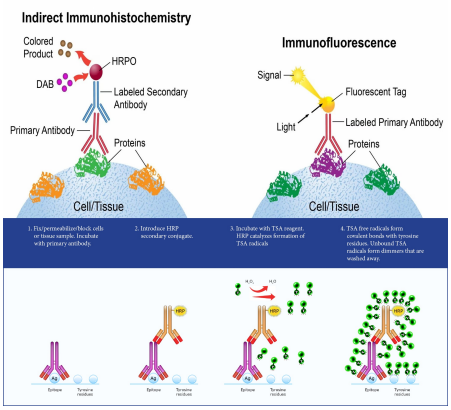
Until recently, this methodology was limited to a small number of markers, owing to *cross-reactivity* phenomena. Indeed, the secondary antibodies should be precisely tailored to target and recognise specifically epitopes of their paired first antibodies, without binding to other species' immunoglobulins that may exhibit similar ligand structures. This cross-reactivity may result in wrong signal from a non-target primary antibody: in [Ver+15a], the cross-reaction between the epitopes of HER2 and EGFR antibodies biased the markers identified in the progression of breast carcinomas.

The traditional way consists in staining consecutive tissue slides with different antibodies. However, correctly realigning and combining single slices is error-prone, losing the cell-cell distance information. To overcome it, a recent methodology, the tyramide signal amplification (TSA) system (see top subfigure, Section 2.2.2), allowed to increase the number of markers to seven colours that could be stained simultaneously. In this system, TSA free radicals catalysed by conjugation of the horseradish peroxidase to the secondary antibody allows isolation of the complex formed by the primary and secondary antibodies. It decreases then the risk of antibody cross-reactivity when adding new antibodies to stain distinct markers. In contrast to other methods, the multiplexed analysis of several markers reveals in detail the anatomical structure, including cell types' location, detection of lymphoid structures or formation of blood vessels related to angiogenesis [Lim+18].

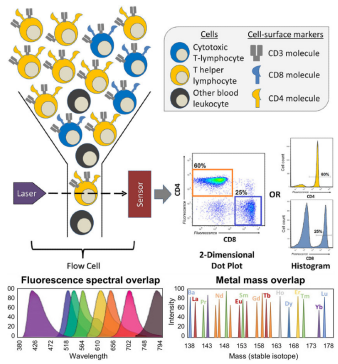
Other IF-based methods that use a combination of sequential staining with innovating staining approaches to enable highly multiplexed quantification of biomarkers include multiplexed immunofluorescence (MxIF)[Ger+13] and multiepitope ligand cartography (MELC) [Sch+06]. They use photobleaching to stop cell fluorescence from the previous marker and can stain 100 markers in a row. Finally, combination of mass cytometry with IHC enable to get a deeper resolution at the subcellular level (imaging mass cytometry (IMC) [Gie+14], Multiplexed ion beam imaging: MIBI [Ang+14]. Nonetheless, both technologies require heavy instrumentation, with a weak throughput. Finally, [Gol+18] developed a simpler technique, CODEX, that requires less material. It allows for highly multiplexed single-cell quantification of membrane protein expression in densely packed solid lymphoid tissue, which was previously considered impossible, enabling deep characterization of cellular niches and their dynamics during autoimmune disease. The CODEX technique involves the use of dye-labelled nucleotides inserted into the DNA tag of the antibodies (inspired from spatial barcoding techniques, see Section 1.2.3), combined with an image-based deconvolution algorithm that quantifies the expression of multiple membrane proteins in individual cells, and has been successfully applied to three-color fluorescence images.



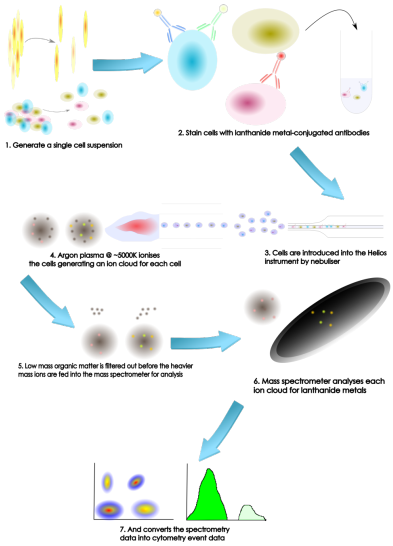
(a) ISH (in-situ hybridisation) illustration. Top figure, reproduced from [YJW20, Fig .1], describes the main technical stages of in situ hybridisation development. Bottom figure, reproduced from [SC05, Fig .1], describes the most popular ISH method, namely FISH (Fluorescence in situ hybridisation).



(b) IHC (Immuno Histo Chemistry) illustration. Top figure, comparison between IHC and immunofluorescence (IF), from [mer22, Fig .1]. Bottom figure, zoom on tyramide signal amplification (TSA) system, reproduced from autocite[Fig .1]aatbio20.



(c) FACS Top picture, reproduced from [Ver+15b, Fig. 1], illustrates FACS guidelines in order to isolate T helper from cytotoxic T-lymphocytes. Bottom picture, reproduced from [BFM19, Fig. 35.15], compares the overlap of the blurred fluorophore-related peaks used in a FACS panel with the clearly distinct peaks from a Cytof output.



(d) CyTOF illustration. [hos23] illustrates the principle of "mass cytometry", alternatively referred to as Cytometry by Time-Of-Flight (CyTOF), see also [Col20].

Figure 2.4: Overview of physical methods to infer cell composition.

To conclude this section, a systematic and comparative review, with their respective advantage and main limitations, are summarised in Section 2.2.2:

		Number of markers	Throughput	Spatial organization	Precise quantification	Many public datasets available
IHC	Brightfield	Low	Low	Yes	Yes	No
	Immunofluorescence	Low to medium	Low	Yes	In some settings	No
Cytometry	Flow Cytometry	Low to medium	Medium	No	Yes	No
	Mass Cytometry	Medium	Medium	No	Yes	No
Transcriptomics	RNA-Seq and micro-arrays	High	High	No	Yes	Yes
	Single-cell transcriptomics	High	High	In some settings	No	Yes

Green, well fitted; yellow, medium; red, poorly fitted.

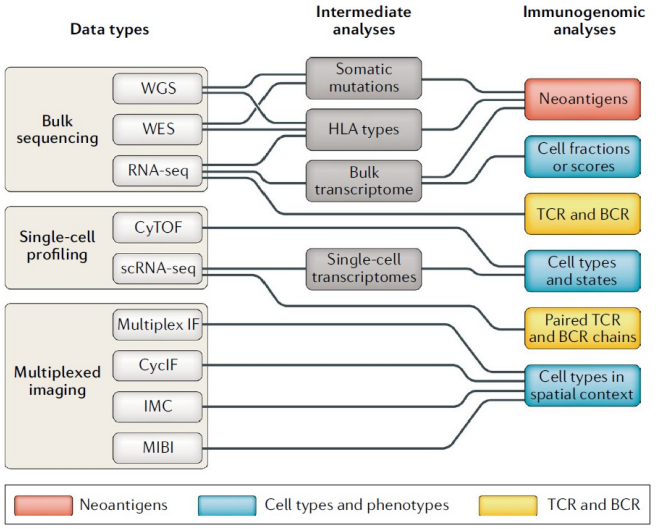


Figure 2.5: Comparison of the main experimental methodologies used to analyse complex biological environments. Top figure is reproduced from [Pet+18, Table. 1] and bottom figure, more directly connected to the study of the TME and its relation with the immune system, is drawn from [Fin+19b, Fig. 2].

After this comprehensive biological introduction, describing the main features of the inputs provided to my models, the remainder, and core part, of my manuscript, elaborates on a set of statistical endeavours for analysing in a consistent manner the primary sources of variation in patients' molecular profiles. Specifically, we focus on mixture models and deconvolution algorithms for delineating variations of global bulk expression profiles across individuals (Part II) and dissecting heterogeneous samples within tissues (Part III), respectively.

Chapters 3 and 4 focus on decomposing multimodal distributions, using parametric mixture models. In particular, GMM assumes that a set of seemingly heterogeneous observations can be clustered into subgroups, each following its own Gaussian distribution. Overall, mixture models are thus valuable when dealing with heterogeneous molecular profiles, in which the underlying phenotypical condition is unobserved, and conducted by a latent and hidden discrete variable.

In Chapters 5 and 6, we introduced deconvolution methods, which are designed to unveil the actors of the variability within an observation instead. These methods consider indeed that the global expression profile, measured at the tissue level, conceals the inherent intricacy of heterogeneous biological samples. Specifically, in order to uncover the internal constituents of this "mixture", deconvolution methods posit that the bulk expression profile can be reconstructed by

aggregating the individual contributions of the cell types believed to constitute the sample and contribute to the transcriptome library.

For both approaches, we start by reviewing existing solutions, and subsequently propose a practical biological application and a putative statistical framework avenues to bolster the robustness and accuracy of gold-standard methods.

Part II

Transcriptome and mixture models

Chapter 3

Article 1: Gaussian Mixture Models in R

Methodological Objective: Conduct a benchmark analysis to evaluate the performance of a wholesome ecosystem of R packages in retrieving parameters of Gaussian mixture models. Mixture models are widely used statistical tools within the realm of precision medicine, facilitating the clustering of patients into distinct subgroups known as "endotypes". Each class is additionally assumed to follow a unique distribution, with its own and specific characteristics. Hence, mixture models are particularly valuable when dealing with heterogeneous and multimodal molecular profiles, in which the underlying phenotypical condition is unobserved.

The fundamental objective of employing mixture models is subsequently to determine the appropriate number of levels/groups composing the latent variable, and employed the parameters inferred to assign each patient to its most probable subgroup. Among the various statistical distributions available for this purpose, Gaussian Mixture Modelss (GMM)) are highly favoured for their appealing statistical properties, ease of interpretation, and straightforward implementation.

A wholesome array of R packages has been developed to infer the parameters of GMMs. However, a comprehensive comparative review, thoroughly benchmarking their performance and choices of algorithmic implementation, is lacking. In light of this situation, we systematically compared the performance of seven R packages: `bgmm`, `EMCluster`, `GMKMcharlie`, `flexmix`, `mclust`, `mixtools`, and `Rmixmod`, in Article 3.1.

We selected them on the basis of their popular reporting in numerous biological analyses, and for the shared optimisation strategy adopted for inferring parameters. All of them indeed retrieve the parameters of the mixtures using the Expectation-Maximization (EM) algorithm, which is widely popular for guaranteeing the asymptotic consistency, convergence, and efficiency of the parameters returned.

Specifically, our comparison delves into comparing their statistical and computational performances as a function of the choice of initialization algorithm and the complexity of the mixture, in various numerical *bootstrap*/bagging simulations. Notably, while the EM algorithm for mixtures of Gaussian distributions is relatively straightforward to implement in pure R programming, it turned out that seemingly small differences in the implementation of the EM algorithm result ultimately in significant variations in statistical performance.

3.1 Article 1

Gaussian Mixture Models in R

by Bastien Chassagnol, Antoine Bichat, Cheïma Boudjeniba, Pierre-Henri Wuillemin, Mickaël Guedj, David Gohel, Gregory Nuel, and Etienne Becht

Abstract Gaussian mixture models (GMMs) are widely used for modelling stochastic problems. Indeed, a wide diversity of packages have been developed in R. However, no recent review describing the main features offered by these packages and comparing their performances has been performed. In this article, we first introduce GMMs and the EM algorithm used to retrieve the parameters of the model and analyse the main features implemented among seven of the most widely used R packages. We then empirically compare their statistical and computational performances in relation with the choice of the initialisation algorithm and the complexity of the mixture. We demonstrate that the best estimation with well-separated components or with a small number of components with distinguishable modes is obtained with REBMIX initialisation, implemented in the **rebmix** package, while the best estimation with highly overlapping components is obtained with *k*-means or random initialisation. Importantly, we show that implementation details in the EM algorithm yield differences in the parameters' estimation. Especially, packages **mixtools** (Young et al. 2020) and **Rmixmod** (Langrognet et al. 2021) estimate the parameters of the mixture with smaller bias, while the RMSE and variability of the estimates is smaller with packages **bgmm** (Ewa Szczurek 2021) , **EMCluster** (W.-C. Chen and Maitra 2022) , **GMKMcharlie** (Liu 2021), **flexmix** (Gruen and Leisch 2022) and **mclust** (Fraley, Raftery, and Scrucca 2022). The comparison of these packages provides R users with useful recommendations for improving the computational and statistical performance of their clustering and for identifying common deficiencies. Additionally, we propose several improvements in the development of a future, unified mixture model package.

1 Introduction to Mixture modelling

Formally, let's consider a pair of random variables (X, S) with $S \in \{1, \ldots, k\}$ a discrete variable and designing the component identity of each observation. When observed, S is generally denoted as the labels of the individual observations. k is the number of mixture *components*. Then, the density distribution of X is given in Equation (1):

$$\begin{aligned} f_\theta(X) &= \sum_S f_\theta(X, S) \\ &= \sum_{j=1}^{k} p_j f_{\zeta_j}(X), \quad X \in \mathbb{R} \end{aligned} \tag{1}$$

where $\theta = (p, \zeta) = (p_1, \ldots, p_k, \zeta_1, \ldots, \zeta_k)$ denotes the parameters of the model: p_j is the proportion of component j and ζ_j represents the parameters of the density distribution followed by component j. In addition, since S is a categorical variable parametrized by p, the prior weights must enforce the unit simplex constraint (Equation (2)):

$$\begin{cases} p_j \geq 0 \quad \forall j \in \{1, \ldots, k\} \\ \sum_{j=1}^{k} p_j = 1 \end{cases} \tag{2}$$

In terms of applications, mixture models can be used to achieve the following goals:

- *Clustering*: hard clustering consists in determining a complete partition of the n observations $x_{1:n}$ into k disjoint non-empty subsets. In the context of *mixture model-based clustering*, this is done by assigning each observation i to the cluster $\hat{s_i} = \arg\max_j \eta_i(j)$ that maximises the posterior distribution (MAP) (see Equation (3)):

$$\eta_i(j) := \mathbb{P}_\theta(S_i = j | X_i = x_i) \tag{3}$$

- *Prediction*: the purpose is to predict a response variable Y from an explanatory variable X. The dependent variable Y can either be discrete, taking values in classes $\{1, \ldots, G\}$ (*classification* task) or continuous (*regression* task). In this paper, we do not extensively discuss application of mixture models to regression purposes but refer the reader to Bouveyron and Girard (2009) for mixture classification and Shimizu and Kaneko (2020) for mixtures of regression models.

In section Univariate and multivariate Gaussian distributions in the context of mixture models, we describe the most commonly used family, the Gaussian Mixture Model (GMM). We then present the MLE estimation of the parameters of a GMM, introducing the classic EM algorithm in section Parameter estimation in finite mixtures models. Finally, we introduce bootstrap methods used to evaluate the quality of the estimation and metrics used for the selection of the best model in respectively appendices *Derivation of confidence intervals in GMMs* and *Model selection*.

Univariate and multivariate Gaussian distributions in the context of mixture models

We focus our study on the finite Gaussian mixture models (GMM) in which we suppose that each of the k components follows a Gaussian distribution.

We recall below the definition of the Gaussian distribution in both univariate and multivariate context. In the finite univariate Gaussian mixture model, the distribution of each component $f_{\zeta j}(X)$ is given by the following univariate Gaussian p.d.f. (probability density function) (Equation (4)):

$$f_{\zeta j}(X = x) = \varphi_{\zeta_j}(x|\mu_j, \sigma_j) := \frac{1}{\sqrt{2\pi}\sigma_j} \exp^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} \tag{4}$$

which we note: $X \sim \mathcal{N}(\mu_j, \sigma_j)$.

In the univariate case, the parameters to be inferred from each component, ζ_j, are: μ_j, the *location* parameter (equal to the mean of the distribution) and σ_j, the *scale* parameter (equal to the standard deviation of the distribution with a Gaussian distribution).

Following parsimonious parametrisations with respect to univariate GMMs are often considered:

- *homoscedascity*: variance is considered equal for all components, $\sigma_j = \sigma, \forall j \in \{1, \ldots, k\}$, as opposed to heteroscedascity where each sub-population has its unique variability.
- *equi-proportion* among all mixtures: $p_j = \frac{1}{k} \quad j \in \{1, \ldots, k\}$ [1]

In the finite multivariate Gaussian mixture model, the distribution $f_{\zeta j}(\boldsymbol{X})$ of each component j, where $\boldsymbol{X} \in \mathbb{R}^D = (X_1, \ldots, X_D)^\top$ is a multivariate random variable of dimension D, is given by the following multivariate Gaussian p.d.f. (Equation (5)):

$$f_{\zeta j}(\boldsymbol{X} = \boldsymbol{x}) = \det(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top\right) \tag{5}$$

which we note $\boldsymbol{X} \sim \mathcal{N}_D(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The parameters to be estimated for each component can be decomposed into:

- $\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{1j} \\ \vdots \\ \mu_{Dj} \end{pmatrix} \in \mathbb{R}^D$, the D-dimensional mean vector.
- $\boldsymbol{\Sigma}_j$, the $\mathcal{M}_D(\mathbb{R})$ positive-definite [2]covariance matrix, whose diagonal terms are the individual variances of each feature and the off-diagonal terms are the pairwise covariance terms.

Three families of multivariate GMMs are often considered:

- the *spherical* family, $\boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}_D$, with $\sigma_j \in \mathbb{R}_+^*$, refers to GMMs whose covariance matrix is diagonal with an unique standard deviation term. The corresponding volume representation is a $D-$*hypersphere* of radius σ_j.

[1] A rarer constraint considered implies to enforce a linear constraint over the clusters' means, of the following general form: $\sum_{j=1}^{k} a_j \mu_j = 0$, with $\{a_1, \ldots, a_k\}$. For instance, the R package **epigenomix** considers a $k = 3$ component mixture in the context of transcriptomic (differential analyses) and epigenetic (histone modification) to automatically identify undifferentiated, over and under-expressed genes between case and control samples. A common constraint then is to enforce the distribution of fold changes corresponding to the undifferentiated expressed genes to have a distribution centred on 0. Combining equality of means and equality of variances is irrelevant, as the model is then degenerate. Additionally, setting constraints on the means makes the estimation of the parameters challenging, as detailed in Appendix *Extensions of the EM algorithm to overcome its limitations*.

[2] The positive-definiteness constraint can be interpreted from a probabilistic point of view as a necessary condition such that the generalised integral of the multivariate distribution is defined and sum-to-one over $\mathbb{R}$ or from the statistical definition of the covariance. A symmetric real matrix $\boldsymbol{X}$ of rank D is said to be *positive-definite* if for any non-zero vector $\mathbf{v}, \in \mathbb{R}^D$, the following constraint $\mathbf{v}^\top \boldsymbol{X}\mathbf{v} > 0$ is enforced.

- the *diagonal* family, $\mathbf{\Sigma}_j = \text{diag}\left(\sigma_{1j}^2, \ldots, \sigma_{1D}^2\right)$, with $\sigma_j \in \mathbb{R}_+^D$, refers to GMMs whose covariance matrix is diagonal. Its associated volume representation is an ellipsoid whose main axes are aligned with the D canonical basis of $\mathbb{R}^D$. Of note, the null constraint imposed over the off-diagonal terms in the spherical and diagonal families imply that the multivariate distribution can be further decomposed and analysed as the product of univariate independent Gaussian realisations.
- the *ellipsoidal* family, also named the *general* family, refer to GMMs whose covariance matrix, $\mathbf{\Sigma}_j$, can be any arbitrary positive-definite $D \times D$ matrix. Thus, the corresponding clusters for each component J are ellipsoidal, centred at the mean vector $\boldsymbol{\mu}_j$, and volume and orientation respectively determined by the eigenvalues and the eigenvectors of the covariance matrix $\mathbf{\Sigma}_j$.

In the multivariate setting, the volume, shape, and orientation of the covariances can be constrained to be equal or variable across clusters, generating 14 possible parametrisations with different geometric characteristics (Banfield and Raftery 1993; Celeux and Govaert 1992). We review them in Appendix *Parameters estimation in a high-dimensional context* and Table 5. Of note, the correlation matrix can be easily derived from the covariance matrix with the following normalisation: $\text{cor}(\boldsymbol{X}) = \left(\frac{\text{cov}(x_l, x_m)}{\sqrt{\text{var}(x_l)} \times \sqrt{\text{var}(x_m)}}\right)_{(l,m) \in D \times D}$. Correlation if strictly included between -1 and 1, the strength of the correlation is given by its absolute value while the type of the interaction is returned by its sign. A correlation of 1 or -1 between two features indicates a strictly linear relationship.

For the sake of simplicity and tractability, we will only consider the fully unconstrained model in both the univariate (heteroscedastic and unbalanced classes) and multivariate dimension (unbalanced and complete covariance matrices for each cluster) in the remainder of our paper.

Parameter estimation in finite mixtures models

A common way for estimating the parameters of a parametric distribution is the *maximum likelihood estimation* (MLE) method. It consists in estimating the parameters by maximising the likelihood, or equivalently the log-likelihood of a sample. In what follows, $\ell(\theta|x_{1:n}) = \log(f(x_{1:n}|\theta))$ is the log-likelihood of a n-sample. When all observations are independent, it simplifies to $\ell(\theta|x_{1:n}) = \sum_{i=1}^{n} \log(f(x_i|\theta))$. The MLE consists in finding the parameter estimate $\hat{\theta}$ which maximises the log-likelihood $\hat{\theta} = \arg\max \ell(\theta|x_{1:n})$.

Recovering the maximum of a function is generally performed by finding the values at which its derivative vanishes. The MLE in GMMs has interesting properties, as opposed to the *moment estimation* method: it is a consistent, asymptotically efficient and unbiased estimator (J. Chen 2016; McLachlan and Peel 2000).

When S is completely observed, for pairs of observations $(x_{1:n}, s_{1:n})$, the log-likelihood of a finite mixture model is simply given by Equation (6):

$$\ell(\theta|X_{1:n} = x_{1:n}, S_{1:n} = s_{1:n}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \left[\log\left(f_{\zeta_j}(x_i, s_i = j)\right) + \log(p_j)\right]_{\mathbf{1}_{s_i=j}} \tag{6}$$

where an analytical solution can be computed provided that a closed-form estimate exists to retrieve the parameters ζ_j for each components' parametric distribution. The MLE maximisation, in this context, involves the estimation of the parameters for each cluster, denoted as ζ_j. The corresponding proportions, p_j, can be straightforwardly computed as the ratios of observations assigned to cluster j relative to the total number of observations, n.

However, when S is unobserved, the log-likelihood, qualified as incomplete with respect to the previous case, is given by Equation (7):

$$\ell(\theta|x_{1:n}) = \sum_{i=1}^{n} \log\left(\underbrace{\sum_{j=1}^{k} p_j f_{\zeta_j}(x_i)}_{\text{sum of of logs}}\right) \tag{7}$$

The sum of terms embed in the log function (see underbrace section in Equation (7)) makes it intractable in practice to derive the null values of its corresponding derivative. Thus, no closed form of the MLE is available, including for the basic univariate GMM model. This is why most parameter estimation methods derive instead from the *EM algorithm*, first described in Dempster, Laird, and Rubin (1977). We describe its main theoretical properties, the reasons for its popularity, and its main limitations in the next section.

The EM algorithm

In cases where both S and the parameters associated to each cluster are unknown, there is no available closed-form solution that would jointly maximise the log-likelihood, as defined in Equation (7), with respect to the set of parameters (θ, S). However, when either S or θ are known, the estimation of the other parameters is straightforward. Hence, the general principle of EM-like algorithms is splitting this complex non-closed joint MLE estimation of (S, θ) into the iterative estimation of S_q from $\hat{\theta}_{q-1}$ and X (expectation phase, or *E-step* of the algorithm) and the estimation of $\hat{\theta}_q$ from $(S_q$ and X (maximisation phase, or *M-step*), with $\hat{\theta}_{q-1}$ being the estimated parameters at the previous step $q-1$, until we reach the convergence.

The EM algorithm sets itself apart from other commonly used methods by taking into account all possible values taken by the latent variable S. To do so, it computes the expected value of the log likelihood of θ, conditioned by the posterior distribution $\mathbb{P}_{\hat{\theta}_{q-1}}(S|X)$, also named as the *auxiliary function*. Utilising the assumption of independence among observations in a mixture model, the general formula of this proxy function of the incomplete log-likelihood is given in finite mixture models by Equation (8).

$$\begin{aligned} Q(\theta|\hat{\theta}_{q-1}) &:= \mathbb{E}_{S_{1:n}|X_{1:n},\hat{\theta}_{q-1}}\left[\ell(\theta|X_{1:n}, S_{1:n})\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \eta_i(j)\left(\log(p_j) + \log(\mathbb{P}(X_i|S_i = j, \theta))\right) \end{aligned} \tag{8}$$

with $\hat{\theta}_{q-1} = \hat{\theta}$ the current estimated parameter value.

In practice, the EM algorithm consists in performing alternatively E-step and M-step until convergence, as described in the pseudocode below (Box 1):

Box 1: the EM algorithm

- *step E*: determine the posterior probability function $\eta_i(j)$ for each observation of X for each possible discrete latent class, using the initial estimates $\hat{\theta}_0$ at step $q = 0$, or the previously computed estimates $\hat{\theta}_{q-1}$. The general formula is given by Equation (9):

$$\eta_i(j) = \frac{p_j f_{\zeta_j}(x_i)}{\sum_{j=1}^{k} p_j f_{\zeta_j}(x_i)} \tag{9}$$

- *step M*: compute the mapping function $\hat{\theta}_q := M(\theta|\hat{\theta}_{q-1}) = \arg\max Q(\theta|\hat{\theta}_{q-1})$ which maximises the auxiliary function. One way of retrieving the MLE associated to the auxiliary function is to determine the roots of its derivative, namely solving Equation (10)[a]:

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \theta} = 0 \tag{10}$$

[a] To ensure that we reach a maximum, we should assert that the Hessian matrix evaluated at the MLE is indeed negative definite.

Interestingly, the decomposition of the incomplete log-likelihood associated to a mixture model $\ell(\theta|X)$ reveals an entropy term and the so-called auxiliary function (Dempster, Laird, and Rubin 1977). It can be used to prove that maximising the auxiliary function at each step induces a bounded increase of the incomplete log-likelihood. Namely, the convergence of the EM algorithm, defined by comparisons of consecutive log-likelihood, is guaranteed, provided the mapping function returns the maximum of the auxiliary function. Yet, the convergence of the series of estimated parameters $(\theta_q)_{q\geq 0} \underset{i\to+\infty}{\longrightarrow} \hat{\theta}$ is harder to prove but has been formally demonstrated for the *exponential family* (a superset of the Gaussian family), as stated in Dempster, Laird, and Rubin (1977).

Additionally, the EM algorithm is *deterministic*, meaning that for a given initial estimate θ_0 the parameters returned by the algorithm at a given step q are fixed. However, this method requires the user to provide an initial estimate, denoted as θ_0, of the model parameters and to specify the number of components in the mixture. We review some classic initialisation methods in Initialisation of the EM algorithm and some algorithms used to overcome the main limitations of the EM algorithm in the Appendix *Extensions of the EM algorithm to overcome its limitations*.

Finally, the prevalent choice of Gaussian distributions to characterize the distribution of random observations is guided by a set of interesting properties. In particular, Geary (1936) has shown that the Normal distribution is the only distribution for which the Cochran's theorem (Cochran 1934) is guaranteed, namely for which the the mean and variance of the sample are independent of each other. Additionally, similar to any distribution proceeding from the exponential family, the MLE statistic is *sufficient*[3].

Initialisation of the EM algorithm

EM-like algorithms require an initial estimate of the parameters, θ_0, to optimise the maximum likelihood. *Initialisation* is a crucial step, as a bad initialisation can possibly lead to a local sub-optimal solution or trap the algorithm in the boundary of the parameter space. The most straightforward initialisation methods, such as random initialisation, are standalone and do not require any additional initialisation algorithms, whereas *meta-methods*, such as short-EM, still need to be initialised by alternative methods. The commonly-used initialisation methods encompass:

- The *Model-based Hierarchical Agglomerative Clustering* (MBHC) is an agglomerative hierarchical clustering based on MLE criteria applied to GMMs (Scrucca and Raftery 2015). First, the MBHC is initialised by assigning each observation to its own cluster. Next, the pair of clusters that maximises the likelihood of the underlying statistical model among all possible pairs is merged. This procedure is repeated until all clusters are merged. The final resulting clusters are then simply the last k cuts of the resulting dendrogram. When the data is univariate and homoscedastic, or when the underlying distribution has a diagonal covariance matrix, the merging criterion performs similarly to *Ward's criterion*, in that merging of the two clusters also simultaneously minimizes the sum of squares. As opposed to the other initialisation methods described hereafter, MBHC is a deterministic method which does not require careful calibration of hyperparameters. However, as acknowledged by the author of the method (Fraley 1998), the resulting partitions are generally suboptimal compared to other initialisation methods.
- The conventional *random* initialization method, frequently employed for the initialization step of the k-means algorithm, involves the random selection of k distinct observations, which are referred to as *centroids*. Subsequently, each observation is assigned to the nearest centroid, a process reminiscent of the C-step in the CEM algorithm (Biernacki, Celeux, and Govaert 2003). This is the method used in this paper, unless otherwise stated. Alternative versions of this method have been developed: for instance, the package **mixtools** draws the proportions of the components from a Dirichlet distribution, whose main advantage lies in respecting the unit simplex constraint (Equation (2))[4], but uses binning methods to guess the means and standard deviations of the components. Similarly, Kwedlo (2013) proposes a method in which the means of the components are randomly chosen, but with an additional constraint of maximising the Mahalanobis distance between the selected centroids. This enables to cover a larger portion of the parameters' space.
- k-means is a CEM algorithm, in which the additional assumption of balanced classes and homoscedascity implies that each observation in the E-step is assigned to the cluster with the nearest mean (the one with the shortest Euclidean distance). K-means is initialised by randomly selecting k points, known as the *centroids*. It is often chosen for its fast convergence and memory-saving consumption.
- The *quantile* method sorts each observation x_i in an increasing order and splits them into equi-balanced quantiles of size $1/k$. Then, all observations for a given quantile are assumed to belong to the same component. [5]
- The *Rough-Enhanced-Bayes mixture* (REBMIX) algorithm is implemented in the **rebmix** (Nagode 2022) package and the complete pseudo-code is described thoroughly in (Nagode 2015; Panic, Klemenc, and Nagode 2020). The key stages implemented by the rebmix algorithm for initialising the parameters of GMMs encompass:

[3]The Pitman–Koopman–Darmois theorem (Koopman 1936) states that only the exponential family provides distributions whose statistic can summarize arbitrary amounts of iid draws using a finite number of values

[4]Without prior knowledge favouring one component over another, the Dirichlet distribution is generally parametrised by $\alpha = \frac{1}{k}$, implicitly stating that any observation has equal chance to proceed from a given cluster. In that case, the corresponding distribution is parametrised by a single scalar value α, called the *concentration parameter*.

[5]This method is only available in the univariate framework, since it is not possible to define a unique partition of the observable space into k-splits. For example, in bivariate setting, a binning with $k = 2$ components on each axis leads to a total of $2 \times 2 = 4$ binned regions, which raises the selection issue of the best k hyper-squared volumes for the initial parameters estimation. More generally, $\binom{D}{k}$ binning choices are possible in the multivariate setting.

  - First, the observations are processed using one of these three methods: *k*-nearest neighbours (KNN), Parzen kernel density estimation, or binned intervals. With the binned interval method, the observations are initially divided into $\sqrt{n}^D$ intervals of equal lengths. The mode of one of the components' distribution is subsequently determined by the midpoint of the interval with the highest frequency. The observations lying within the interval are used as preliminary estimates, referred to as "rough" parameters in Nagode (2015).
  - All other observations and intervals are then iteratively assigned to the currently estimated component or to residual components, the ones that have not yet been characterised. The decision to assign an interval to either the currently estimated component or one of the residual components depends on the magnitude of the discrepancy between the observed and the expected frequency within the interval.
  - Finally, all intervals assigned to the currently estimated component (and not only the interval including the mode of the distribution) are used to determine the parameters of the associated Gaussian distribution. Since this step relies on a more comprehensive number of observations for parameter estimation, guaranteeing in principle more robust estimates, this stage is referred to as "enhanced" estimation in Nagode (2015). The algorithm terminates when all intervals have been assigned to a cluster, and the parameters of the various distribution components have been estimated.

The rebmix algorithm can thus be seen as a natural extension of the quantile method, with more rigorous statistical support. Two drawbacks of the algorithm include the need for intensive calibration of hyperparameters and its inadequacy for the estimation of highly overlapping or high dimensional mixture distributions[6].

- The *meta-methods* consist generally in short runs of EM-like algorithms, namely CEM, SEM and EM (see Appendix B: *Extensions of the EM algorithm to overcome its limitation*), with alleviated convergence criterion. The main idea is to use several random initial estimates with shorter runs of the algorithm to explore larger regions of the parameter space and avoid being trapped in a local maximum. Yet, these methods are highly dependent on the choice of the initialisation algorithm (Biernacki, Celeux, and Govaert 2003).
- In the high-dimensional setting, if the number of dimensions D exceeds the number of observations n, all previous methods must be adjusted, usually by first projecting the dataset into a smaller, suitable subspace and then inferring prior parameters in it. In particular, **EMMIXmfa**, in the mixture of common factor analysers (MCFA) approach, initialises the shared projection matrix $\boldsymbol{Q}$ by either keeping the first d eigen vectors generated from standard principal component analysis or uses custom random initialisations (Baek, McLachlan, and Flack 2010).

Following this theoretical introduction, we empirically evaluate the performance of the aforementioned R packages, considering various initialization algorithms and the complexity of the GMMs distributions. Precisely, we outline the simulation framework used to compare the seven packages in Methods and report the results in Results. We conclude by providing a general simplified framework to select the combination of package and initialisation method best suited to its objectives and the nature of the distribution of the dataset.

2 A comprehensive benchmark comparing estimation performance of GMMs

We searched CRAN and Bioconductor mirrors for packages that can retrieve parameters of GMM models. Briefly, out of 54 packages dealing with GMMs estimation, we focused on seven packages that all estimate the MLE in GMMs using the EM algorithm, were recently updated and allow the users to specify their own initial estimates: **bgmm**, **EMCluster**, **flexmix**, **GMKMcharlie**, **mclust**, **mixtools** and **Rmixmod**. The complete inclusion process is detailed in Appendix C, *the meta-analysis workflow for the final selection of CRAN and Bioconductor platforms*. The flowchart summarising our choices is represented in Figure 1.

[6]The method we describe here to preprocess the observations in order to estimate the empirical density estimation, namely the "histogram method" is not well suited for high dimensional data, as the exponential growth of the volume with respect to dimensionality leads to data sparsity, related to the well-known issue of the "curse of dimensionality". Indeed, $\sqrt{n}^D$ distinct intervals will be parsed by the method and the probability with an increasing number of features and decreasing number of observations that no clear local maximum emerges converges to 1. In high-dimensional context, the Parzen window or the KNN method should be favoured, see (Nagode 2015), p. 16.

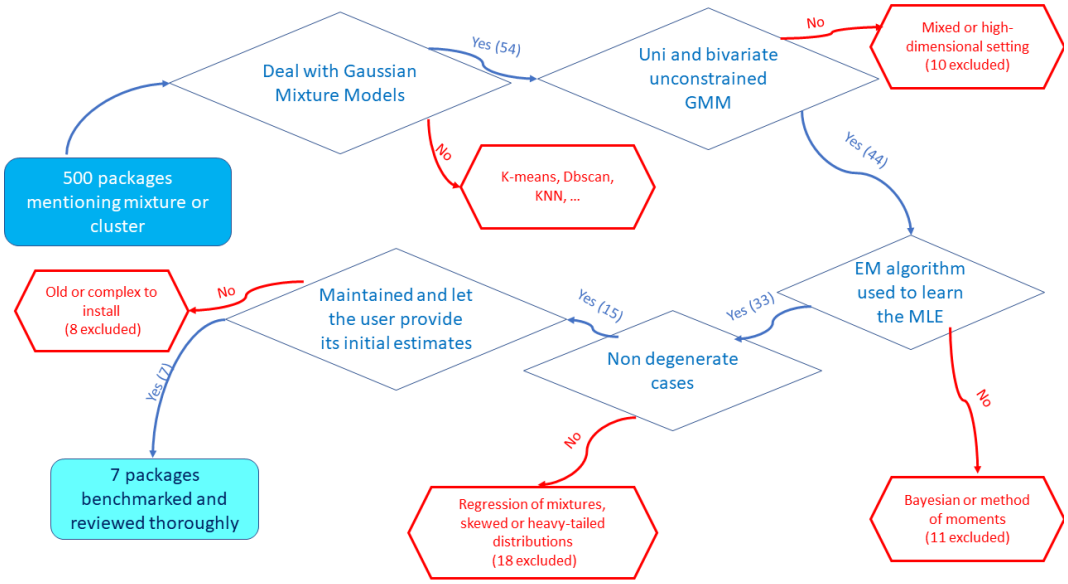


Figure 1: A minimal roadmap used for the selection of the packages reviewed in our benchmark.

We also include two additional packages dedicated specifically to high-dimensional settings, namely **EMMIXmfa** (Rathnayake et al. 2019) and **HDclassif** (Berge, Bouveyron, and Girard 2019) to compare their performance with standard multivariate approaches in complex, but non degenerate cases. We summarise the main features and use cases of the seven + two reviewed packages in Table 1. The three most commonly used packages are **mixtools**, **mclust** and **flexmix**. However, the **mclust** package is by far the most complete with many features provided to visualise and evaluate the quality of the GMM estimate. **bgmm** has the greatest number of dependencies, while **mclust** only depends of base R packages. Additionally, in parallel to clustering tasks, **flexmix** and **mixtools** packages perform regression of mixtures and implement mixture models using other parametric distributions or non-parametric methods via kernel-density estimation.

Table 1: Main features of the reviewed packages, sorted by decreasing number of daily downloads. *Downloads per day* returns the daily average number of downloads for each package on the last 2 years. *Recursive dependencies* column counts the complete set of non-base packages required, as first-order dependencies depend on other packages as well.

Package	Version	Regression	Implemented models	Downloads per day	Last update	Imports	Recursive dependencies	Language
mclust	5.4.7	☒	☒	5223	31/10/2022	R (≥ 3.0)	0	Fortran
flexmix	2.3-17	☑	Poisson, binary, non-parametric, semi-parametric	3852	07/06/2022	R (≥ 2.15.0), modeltools, nnet, stats4	3	R
mixtools	1.2.0	☑	multinomial, gamma, Weibull, non-parametric, semi-parametric	178	05/02/2022	R (≥ 3.5.0), kernlab, segmented, survival	6	C
Rmixmod	2.1.5	☒	☒	39	18/10/2022	R (≥ 2.12.0), Rcpp, RcppEigen	4	C++
EMCluster	0.2-13	☒	☒	33	12/08/2022	R (≥ 3.0.1), Matrix	3	C
bgmm	1.8.4	☒	☒	27	10/10/2021	R (≥ 2.0), mvtnorm, combinat	77	R
GMKMcharlie	1.1.1	☒	☒	12	29/05/2021	Rcpp, RcppParallel, RcppArmadillo	3	C++
EMMIXmfa	2.0.11	☒	☒	12	16/12/2019	NA	0	C
HDclassif	2.2.0	☒	☒	35	12/10/2022	rARPACK	13	R

We further detail features specifically related to GMMs in Table 2. We detail row after row its content below:

- The parametrisations used to provide parsimonious estimation of the GMMs are reviewed in Parameter estimation in finite mixtures models and summarised in rows 1 and 2 (Table 2) for the univariate and multivariate setting. We refer to the package as "canonical" when it implements both homoscedastic and heteroscedastic parametrisations in the univariate setting, and the 14 parametrisations listed in Supplementary Table 3 in the multivariate setting. Adding the additional constraint of equi-balanced clusters results in a total to $14 \times 2 = 28$ distinct models and $2 \times 2 = 4$ parametrisations, respectively in the univariate and multivariate setting. Since

EMMIXmfa and **HDclassif** are dedicated to the analysis of high-dimensional datasets, they project the observations in a smaller subspace and are not available in the univariate setting. Given an user-defined or prior computed intrinsic dimension, we can imagine using any of the standard parametrisations available for instance in the **mclust** package, and listed in Appendix *Parsimonious parametrisation of multivariate GMMs*. In addition, **HDclassif** allows each cluster j to be represented with its own subspace intrinsic dimension d_j, as we describe in further details in Appendix *Parameters estimation in a high-dimensional context*.

- The EM algorithm is the most commonly employed method for estimating the parameters of GMMs, however, alternative algorithms based on the EM framework, are reviewed in Appendix B: *Extensions of the EM algorithm to overcome its limitations* and row 3 of Table 2. Especially, GMMs estimation is particularly impacted by the presence of outliers, justifying a specific benchmark (see Appendix *A small simulation to evaluate the impact of outliers*). We briefly review the most common initialisation algorithms in section Initialisation of the EM algorithm and row 4 of Table 2, a necessary and tedious task for both the EM algorithm and its alternatives.
- To select the best parametrisations and number of components that fit the mixture, several metrics are provided by the reviewed packages (*Model selection* and row 5). Due to the complexity of computing the true distribution of the estimated parameters, bootstrap methods are commonly used used to derive confidence intervals (see Appendix *Derivation of confidence intervals in GMMs* and row 6 in Table 2).
- Six packages supply several functions for visualisation, summarised in the last row of Table 2, to display either the distributions corresponding to the estimated parameters or compare quickly the performance across packages. However, **mclust** is by far the most complete one, with density plots (in the univariate setting) and isodensity plots (bi-dimensional in the bivariate setting or in higher dimensions after appropriate dimensionality reduction), with the option to plot custom confidence intervals and critical regions, and finally boxplot bootstrap representations for displaying the distribution of the benchmarked estimated parameters.

High-dimensional packages provide specific representations adjusted to the high-dimensional settings, notably allowing the user to visualise the projected factorial representation of its dataset in a two or three-dimensional subspace. They also provide specialised performance plots, notably scree plots or BIC scatter plots to represent in a compact way numerous projections and parametrisations.

Table 2: Custom features associated to GMMs estimation for any of the benchmarked packages.

	mclust	**flexmix**	**mixtools**	**Rmixmod**	**EMCluster**	**bgmm**	**GMKMcharlie**	**EMMIXmfa**	**HDclassif**
Models Available (univariate)	canonical	unconstrained	canonical	canonical	unconstrained	canonical	unconstrained	NA	NA
Models Available (multivariate)	canonical	unconstrained diagonal or general	unconstrained	canonical	unconstrained	4 models (diagonal and general, either component specific or global)	unconstrained	4 models (either component-wise or common, on the intrinsic and diagonal residual error covariance matrices)	canonical on projected dimension
Variants of the EM algorithm	VBEM	SEM, CEM	ECM	SEM, CEM	☒	☒	CW-EM, MML	AECM	SEM, CEM
Initialisation	hierarchical clustering, quantile	short-EM, random	random	random, short-EM, CEM, SEM	random, short-EM	k-means, quantile	k-means	k-means, random, heuristic	short-EM, random, k-means
Model or Cluster Selection	BIC, ICL, LRTS	AIC, BIC, ICL	AIC, BIC, ICL, CAIC, LRTS	BIC, ICL, NEC	AIC, BIC, ICL, CLC	GIC	☒	☒	BIC, ICL, CV
Bootstrap Confidence Intervals	☑	☑	☑	☒	☒	☒	☒	☒	☒
Visualisation	performance, histograms and boxplots of bootstrapped estimates, density plots (univariate), scatter plots with uncertainity regions and boundaries (bivariate), isodensity (bivariate , 2D projected PCA or selecting coordinates)	☒	density curves	density curves, scatter plots with uncertainty boundaries	☒	performance, scatter plots with uncertainty boundaries	☒	projected factorial map	projected factorial map, performance (Cattell's scree plot, BIC performance, slope heuristic)

Methods

In addition to the the seven packages selected for our benchmark, we include a custom R implementation of the EM algorithm used as baseline, referred to as *RGMMBench*, and for the high-dimensional

setting we select packages **EMMIXmfa** and **HDclassif**, on the basis of criteria detailed in Appendix C, *General workflow*. Code for *RGMMBench* is provided in Appendix *Application of the EM algorithm to GMMs*. To compare the statistical performances of these packages, we performed *parametric bootstrap* (*Derivation of confidence intervals in GMMs*) and built an experimental design to cover distinct mixture distributions parameter configurations, using prior user-defined parameters.

For each experiment, we assign each observation to an unique cluster by drawing n labels $S_{1:n}$ from a multinomial distribution whose parameters were the prior user-defined proportions $p = (p_1, \ldots, p_k)$. Then, each observation x_i assigned to hidden component j is drawn from a Normal distribution using the `stats::rnorm()` function for the univariate distribution and `MASS::mvrnorm` for the multivariate distribution. The complete code used for simulating data is available on GitHub at RGMMBench. Finally, we obtain an empirical distribution of the estimated parameters by computing the MLE of each randomly generated sample.

For all the packages, we used the same convergence threshold, 10^{-6}, and maximum of 1,000 iterations, as a numerical criterion for convergence. We generated simulated data with $n = 200$ observations in the univariate setting and $n = 500$ observations in the bivariate setting. We set the number of observations in order to minimise the probability of generating a sample without drawing any observations from one of the components[7]. Unless stated explicitly, we kept the default hyper-parameters and custom global options provided by each package. For instance, the **flexmix** package has a default option, *minprior*, set by default to 0.05, which removes any component present in the mixture with a ratio below 0.05. Besides, the fully unconstrained model was the only one which we implemented both in the univariate and multivariate settings, as it is the only parametrisation implemented in all the seven packages.

We compared the packages' performances using five initialisation methods: random, quantile, k-means, rebmix and hierarchical clustering in the univariate setting. We benchmarked the same initialisation methods in the multivariate setting, except for the quantile method which has no multivariate equivalent (see section Initialisation of the EM algorithm):

- We used the function `EMCluster::rand.EM()` with 10 random restarts and minimal cluster size of 2 for the random initialisation. The method implemented by **EMCluster** is the most commonly used, described in details in Biernacki, Celeux, and Govaert (2003) and in section Initialisation of the EM algorithm.
- To implement the k-means initialisation, we used the `stats::kmeans()` function with a convergence criterion fo 10^{-2} and maximum of 200 iterations. The initial centroid and covariance matrix for each component were computed by restricting to the sample observations assigned to the corresponding component. The approach is close to the one adopted by the CEM algorithm (see Appendix B: *Extensions of the EM algorithm to overcome its limitations*).
- We used the `mclust::hcV()` function for the MBHC algorithm. This method has two main limitations: just like the k-means implementation, it only returns a cluster assignment to each observation instead of the posterior probabilities, and the splitting process to generate the clusters sometimes results in clusters composed of only one observation. To avoid this, we added a small epsilon to each posterior probability.
- We used in the univariate setting `bgmm::init.model.params` for the quantiles initialisation.
- To implement the rebmix method, we used the `rebmix::REBMIX` function, using the *kernel density estimation* for the estimation of the empirical density distribution coupled with *EMcontrol* set to one to prevent the algorithm from starting EM iterations.
- Any of the seven packages could be used to implement the small EM method. We decided to use the `mixtools::normalmixEM` as it is the closest one to our custom implementation. We specified 10 random restarts, a maximal number of iterations of 200 and an alleviated absolute threshold of 10^{-2}. Preliminary experiments have led us to consider the removal of the small EM initialization method from the simulation benchmark. This decision is based on the observation that the differences of performance observed between the packages were no longer significant (see supplementary Figure 9).

We sum up in Table 3 the general configuration used to run the scripts. Additionally, all simulations were run with the same R (R Core Team 2023) version 4.0.2 (2020-06-22).

Preliminary experiments suggested that the quality of the estimation of a GMM is mostly affected by the overlap between components' distribution and level of unbalance between components. We quantified the overlap between two components by the following overlap score (OVL, see Equation (11)), with a smaller score denoting well-separated components:

[7] It is especially critical in cases of highly unbalanced configurations, as detailed in Appendix *Practical details for the implementation of our benchmark*

Table 3: Global options shared by all the benchmarked packages.

Initialisation methods	Algorithms	Criterion threshold	Maximal iterations	Number of observations
midrule hc, kmeans, small EM,rebmix, quantiles, random	EM R, Rmixmod, bgmm, mclust, flexmix, EMCluster, mixtools, GMKMCharlie	10^{-6}	1000	100, 200, 500, 1000, 2000, 5000, 10000

$$\text{OVL}(i,j) = \int \min(f_{\zeta_i}(x), f_{\zeta_j}(x))dx \quad \text{with } i \neq j \tag{11}$$

We may generalise this definition to k components by averaging the individual components' overlap. We use the function `MixSim::overlap` from the **MixSim** package (Melnykov, Chen, and Maitra 2021) that approximates this quantity using a Monte-Carlo based method (see appendices *An analytic formula of the overlap for univariate Gaussian mixtures* and *Practical details for the implementation of our benchmark* for further details).

The level of imbalance may be evaluated with entropy measure (Equation (12)):

$$H(S) = -\sum_{j=1}^{k} p_j \log_k(p_j) \tag{12}$$

with k is the number of components and $p_j = \mathbb{P}(S = j)$ is the frequency of class j.

We considered 9 distinct configuration parameters, associated with distinct values of OVL and entropy in the univariate setting, 20 configurations in the bivariate setting, and 16 configurations in the high-dimensional setting. Briefly, in the univariate setting, we simulated components with the same set of four means (0, 4, 8, and 12), three sets of mixture proportions $[(0.25, 0.25, 0.25, 0.25); (0.2, 0.4, 0.2, 0.2); (0.1, 0.7, 0.1, 0.1)]$ and three variances: $(0.3, 1, 2)$. In the bivariate setting, we consider two sets of proportions: $[(0.5, 0.5); (0.9, 0.1)]$, two sets of coordinate centroids: $[(0; 20), (20, 0)]$ and $[(0; 2), (2, 0)]$, the same variance of 1 for each feature and for each component for illustrative purposes of the direct relation linking the correlation and the level of OVL and five sets of correlation:$[(-0.8, -0.8); (0.8, -0.8); (-0.8, 0.8); (0.8, 0.8); (0, 0)]$.

Finally, we tested eight configurations in the high-dimensional framework, setting to $D = 10$ the number of dimensions. We modified the level of overlap (definition is reported in Equation (11)) and the imbalance between the component proportions across our simulations. Additionally, we tested two types of constraints on the covariance matrix: fully parametrised and spherical (see Appendix *Parsimonious parametrisation of multivariate GMMs*). Each of the parameter configurations tested in the high-dimensional setting was evaluated with $n = 200$ observations and $n = 2000$ observations. Additionally, instead of manually defining the parameters for the high-dimensional simulation, we used the `MixSim` function from the **MixSim** package (Melnykov, Chen, and Maitra 2021). This function returns the user a fully parametrised GMM, with a prior defined level of maximum or average overlap[8].

The complete list of parameters used is reported respectively in Table 4 for the univariate setting, Table 5 for the bivariate setting and 6 for the high-dimensional setting. We benchmarked simulations where the components were alternatively very distinct or instead very overlapping, as well as of equal proportions or instead very unbalanced. The adjustments made to meet the specific requirements of high dimensional packages are detailed in *Practical details for the implementation of our benchmark*.

We report the most significant results and features and the associated recommendations in next section Results.

Results

All figures and performance overview tables are reported in *Supplementary Figures and Tables in the univariate simulation* for the univariate setting, *Supplementary Figures and Tables in the bivariate simulation* for the bivariate scenario and *Supplementary Figures and Tables in the HD simulation* for the high dimensional scenario.

Balanced and non-overlapping components

[8]Unfortunately, as discussed in further details in Appendix *An analytic formula of the overlap for univariate Gaussian mixtures*, the MixSim package does not compute the global distribution overlap, but instead returns the mean of pairwise overlap between any component (however, with two components, these two alternative definitions match.) Finally, it is not possible to set the proportions of the components before the generation of the parameters, except for clusters with equal proportions, and contrary to the expect behaviour of additional parameter *PiLow*, supposed to define the smallest mixing proportion.

In the univariate setting, with balanced components and low OVL (scenario U1 in Table 4), the parameter estimates are identical in most cases across initialisation methods and packages, notably the same estimates are returned with *k*-means or rebmix initialisation. However, the random initialisation method leads to a higher variance and bias on the parameter estimates than other methods (Supplementary Figure 4 and Supplementary Table 6), with some estimates fitting only local maxima, far from the optimal value.

Similarly, the scenarios in the bivariate setting (configurations B6-B10 in Table 5), with a focus on B6, B7 and B10 visualised in Supplementary Figure 16, feature well-separated and balanced components. Consistent with conclusions from the corresponding univariate configurations, all benchmarked packages return the same estimates across initialisation methods.

Unbalanced and non-overlapping components

However, with unbalanced classes and low OVL (scenario U7 in 4), the choice of the initialisation method is crucial, with quantiles and random methods yielding more biased estimates and proned to fall in local maximum. Rebmix initialisation provides the best estimates, with the smallest MSE and bias across packages (Supplementary Figure 5 and supplementary Table 7, always associated with the highest likelihood. Overall, with well-discriminated components, most of the differences on the estimation originate from the choice of initialisation method, while the choice of the package has only small impact.

In the bivariate framework, two configurations featured both strongly unbalanced and well-separated components, similarly to scenario U3 in Table 4: the configurations B12 (Supplementary Figure 12 and Table 12) and B14 (Supplementary Figure 13 and Supplementary Table 13). Similarly, configurations B16, B17 and B20 (Table 5) with similar characteristics are summarised in supplementary Figure 17. In all these configurations, neither the initialisation method nor the package have a statistical significant impact on the overall performance.

Similarly, configurations HD1a-HD4b in Table 6) in the high dimensional setting display well-separated clusters, with a representative outcome represented in Supplementary Figure 19 and Supplementary Table 16. Consistent with the results obtained in the analogous univariate and bivariate scenarios, in the unbalanced and non-overlapping framework, the majority of the benchmarked packages produce highly consistent and similar estimates when hierarchical clustering and k-means were used for parameter initialisation. However, **bgmm** and **EMCluster** clearly perform worse when the rebmix initialization method is used (however, overall, rebmix performs poorly, regardless of the package used for estimation). Notably, initialisations with the rebmix package tend to display a much larger number of poor estimations, some of which can be identified with the local maxima associated with parameter switching between the two classes. Finally, the two additional packages dedicated to high-dimensional clustering display the worst performances, with **EMMIXmfa** returning the most biased parameters and **HDclassif** the most noisy estimates. **EMMIXmfa** is the only package that returned highly biased estimates of the components' proportions in this setting.

Balanced and overlapping components

When the overlap between components increases, the bias and variability of the estimates tends to increase, and the choice of initialisation method becomes more impactful. The least biased and noisy estimations with balanced components in the univariate setting (scenario U3 in Table 4) are obtained with the *k*-means initialisation (Supplementary Figure 3 and Table 8) while the rebmix initialisation returns the most biased and noisy estimates. Similar results are found in the bivariate setting with a balanced and highly overlapping two-component GMM (configurations B1-B5 from Table 5), with the best performance reached with the *k*-means initialisation method, followed by MBHC. This is emphasised in supplementary Figure 16, in the top three most complex configurations, namely B1, B2 and B5. If the shape of the covariance matrix is well-recovered, no matter the package, the Hellinger distances are significantly higher (and thus the estimate further away from the target distribution) with the random and rebmix methods.

Similarly, in the high-dimensional scenario HD7 of Table 6), presenting balanced but highly overlapping clusters with a full covariance structure, the best performance was obtained with *k*-means initialisation, while the rebmix initialisation returned the most biased and noisy estimates. While **EMMIXmfa** performed well when it converged, it returned an error in most cases (see Column *Success* of supplementary Table 17). The least biased estimates were returned by **mixtools** and **Rmixmod** and the least noisy by **flexmix**, **mclust** and **GMKMCharlie** (smaller MSE). Interestingly, in the high-dimensional setting, the packages **EMCluster** and **bgmm** exhibited worse performance. In particular, as can be seen in panel E of supplementary Figure 20, the proportions of the components recovered the $]0-1[^k$ simplex.

Conversely, the **EMCluster** package, and to a lesser extent, the **bgmm** package, performed surprisingly well when datasets were simulated with an underlying spherical covariance structure, even though the estimation was not performed explicitly with this constraint (Supplementary Table 19). Indeed, it seems like that the off-diagonal terms tended to converge towards 0, as showcased in

Supplementary Figure 21, in Panel C, in which the fourth row from top represents the bootstrap intervals associated to the pairwise covariance between dimension 1 and 2 of each cluster.

Unbalanced and overlapping components

With unbalanced components and high OVL (scenario U9 in Table 4), all packages, no matter the initialisation method, provided biased estimates, with a higher variability of the parameter estimates compared to other configurations. The least biased estimates were obtained with *k*-means or random initialisation, but with a higher variability on the estimates with random initialisation (Supplementary Table 9). Delving further into the individual analysis of the parameter estimates associated to each component, we found out that the least biased estimates were achieved with rebmix initialisation for the most distinguishable components. For instance, in our configuration, the clusters 2 and 4 (see Supplementary Figure 7 and Table 9) were better characterised with the rebmix method. This observation aligns with the rebmix's underlying framework, using the modes of the distribution for initialising the component (Nagode 2015). With highly-overlapping distributions and unbalanced components, both the choice of the initialisation algorithm and the package have a substantial impact on the quality of the estimation of this mixture.

Two configurations in our bivariate simulation feature distributions with both strong OVL and unbalanced components. Especially, the scenario B11 (Table 5) has the strongest OVL overall, with notably a risk of wrongly assigning minor component 2 to major component 1 of 0.5 (a random method classifying each observation to cluster 1 or 2 would have the same performance).

First, we observe that the the random and rebmix initialisation methods have similar performance, significantly better than *k*-means or MBHC (Supplementary Figure 11). Specifically, the rebmix method returns the least biased estimates, while the random method is associated with the lowest MSE (respectively for configurations B11 and B15, the supplementary Tables 11 and 14). Second, the estimates differ across packages only in these two complex configurations, with packages **Rmixmod** and **mixtools** returning more accurate estimates than the others. It it is particularly emphasised in Scenario B15, where the component-specific covariance matrices are diagonal with same non-null input, and thus should present spherical density distributions. However, only the first class of packages correctly recovers the spherical bivariate 95% confidence regions while they are slightly ellipsoidal with the second class of packages (Panel B, Supplementary Figure 14).

With full covariance structures and unbalanced proportions, as depicted in the high-dimensional Scenario HD8a) and b) of Table 6, the general observations stated in the previous subsection for the high dimensional setting hold, namely that the least biased estimates are returned by packages not specifically designed for high-dimensional data, with the *k*-means initialisation (Supplementary Table 12 and supplementary Figure 22). Furthermore, the **EMCluster** and **bgmm** packages and the two packages dedicated to high-dimensional, perform similarly with $n = 200$ observations (sub-scenario a) and $n = 2000$ observations (sub-scenario b), whereas we would expect narrower and less biased confidence intervals by increasing the number of observations by a factor of 10.

Finally, with spherical covariance structures and unbalanced proportions, the best performances, both in terms of bias and variability, are obtained with **flexmix**, **mclust** and **GMKMCharlie**. Indeed, as detailed later in Conclusions, these packages are more sensitive to the choice of the initialisation method and have a greater tendency to get trapped in the neighbourhood of the initial estimates (Supplementary Table 19 and supplementary Figure 22). Accordingly, *k*-means initialisation performs best since it assumes independent and homoscedastic features for each cluster. Furthermore, **EMMIXmfa** is the package that best estimates the off-diagonal terms in this setting, as highlighted in supplementary Table 19.

Identification of two classes of packages with distinct behaviours

By summarizing the results obtained across all simulations, we identify two classes of packages with distinct behaviours (Figure 2):

- The first class of packages, represented by **Rmixmod** and **mixtools**, returns similar estimates to our baseline EM implementation. The estimates returned by these packages are less biased but at the extent of a higher variability on the estimates. Additionally, with overlapping mixtures, they tend to be slower compared to the second class, since they require additional steps to reach convergence.
- The second class of packages, composed of the other reviewed packages, is more sensitive to the initialisation method. This leads to more biased but less variable estimates, especially when assumptions done by the initialisation algorithm are not met.

Panels A, B and C display, respectively in the univariate, bivariate and high-dimensional setting, the heatmap of the Pearson correlation between the estimated parameters across the benchmarked packages for the most discriminative scenario (the one featuring the most unbalanced and overlapping components): scenario U9, Table 4 in the univariate setting, scenario B11, Table 5, for the bivari-

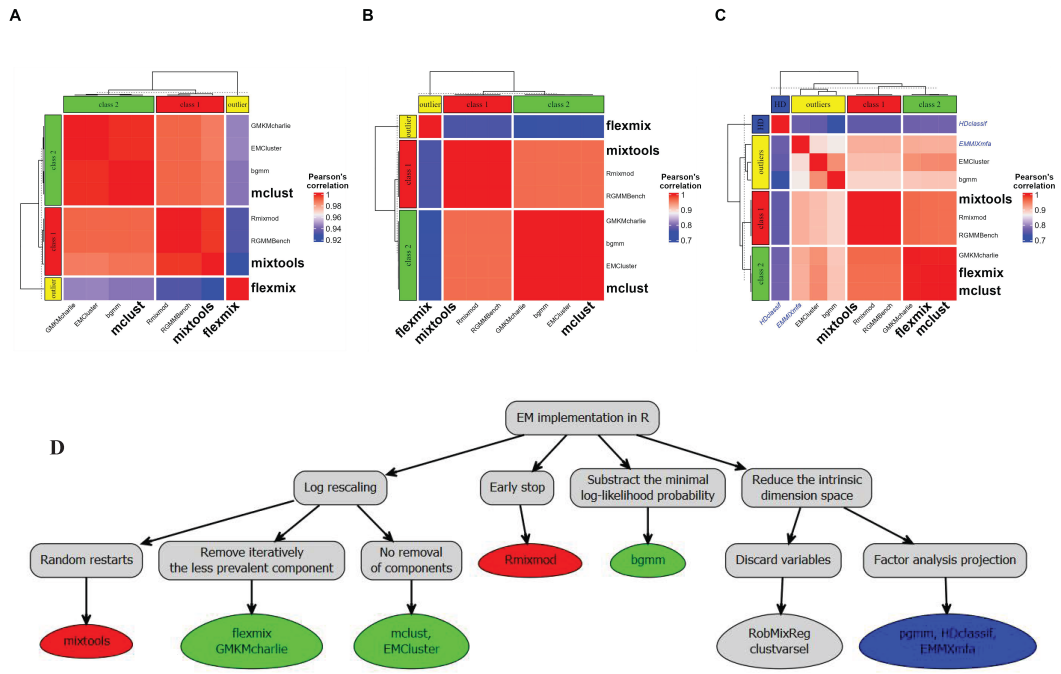


Figure 2: Panels A, B and C show respectively the heatmap of the Pearson correlation in the univariate, bivariate and high-dimensional framework between the parameters estimated by the packages, evaluated for the most discriminating and complex scenario. The correlation matrix was computed using the function `stats::cor` with option *complete* to remove any missing value related to a failed simulation, and the heatmap generated with the Bioconductor package **ComplexHeatmap**. Panel D represents a tree summarising the main differences between the benchmarked packages, in terms of the EM implementation. They are discussed in more detail in Appendix *EM-implementation differences across reviewed packages*.

ate simulation and scenario HD8, Table 6 for the high-dimensional simulation, with the k-means initialisation.

We further identified with this representation minor differences for the estimation of the parameters between **Rmixmod** and **mixtools**, while three subgroups can be identified in the second class of packages: the first subset with **bgmm** and **mclust**, the second subset with **EMCluster** and **GMKMcharlie** packages and the **flexmix** package, which clearly stands out from the others, as being the one most likely to be trapped at the boundaries of the parameter space. After examining the source codes of the packages, we attribute this differences to custom implementation choices of the EM algorithm, such as the way numerical underflow is managed or the choice of a relative or absolute scale to compare consecutive computed log-likelihoods (see Appendix *EM-implementation differences across reviewed packages* and Panel D, Figure 2). In the high-dimensional setting, the second class of packages showed additional heterogeneity, with **EMCluster** and **bgmm** setting themselves apart from the other three packages.

Failed estimations

Finally, in some cases, neither the specific EM algorithm implemented by each package nor the initialisation method were able to return an estimate with the expected number of components, or converged to a degenerate distribution (e.g., with infinite or zero variances). In that case, we considered the estimation as failed and accordingly we did not include it into the visualisations and the summary metric tables.

Most of the failed estimations occurred with the rebmix initialisation. Indeed, an updated version of the package forced the user to provide a set of possible positive integer values for the number of components, with at least a difference of two between the model with the most components and the model with the least components (we therefore set the parameter *cmax* to $k+1$ and *cmin* to $k-1$).In scenarios where the distributions associated with each cluster exhibit significant overlap, there is an increased risk of incorrectly estimating the number of components. This arises from the inherent difficulty of discerning the modes within the overall distribution. For instance, in the most complex scenario B11, characterized by strong overlap and imbalanced clusters (refer to Table 5), up to 20% of initialisations were unsuccessful. Similarly, in the second most challenging scenario, B15, approximately 10% of initializations failed against an averaged number of 4% of the simulations

exhibiting an inaccurate estimation of the number of components.

Removing errors proceeding from the initialisation method, only the **flexmix** package failed in returning an estimate matching the user criteria in some configurations of the univariate and bivariate settings. In both cases, the strong assumption that any cluster with less than 5% of the observations is irrelevant, results in trimming one or more components[9]. This strong agnostic constraint on component proportions led to failures in configurations featuring strongly overlapping clusters, with up to 20% failed estimations with the random initialisation method in scenario B11 (Table 5) and 80% failed estimations in the univariate setting[10] with the rebmix initialisation with scenario U9, Table 4.

In a relatively high dimensional framework, as tested on our third benchmark (Table 6), none of the algorithms that were initialised with the random method (`EMCluster::rand.EM()`) converged successfully. Indeed, of the 16 configurations tested, the covariance returned during the initialisation was systematically non-positive definite for at least one of the components, violating the properties of covariance matrices. Furthermore, an analysis of summary metrics in scenarios HD1 and HD8, reported in supplementary Tables 20 and 21, revealed a notably higher rate of failures when employing rebmix initialisation in conjunction with packages tailored for high dimensionality, such as **HDclassif** **EMMIXmfa**. This discrepancy was in stark contrast to the more reliable and consistent initial estimates returned by *k*-means or hierarchical clustering.

Furthermore, as shown by the comparison of summary metrics with $n = 200$ and $n = 2000$ observations in supplementary Tables 20 and 21, respectively for the simplest scenario HD1 and the most complex one HD8, the rebmix initialisation on the one hand, and the packages dedicated to high dimensionality or those of the second class of packages that show a particular behaviour, present much more failures than the *k*-means or hierarchical clustering initialisation.

3 Conclusions

There are many packages that implement the EM algorithm for estimating the parameters of GMMs. But only few are regularly updated, implement both the unconstrained univariate and multivariate GMM, and enable the user to provide its own initial estimates. Hence, among the 54 packages dealing with GMMs available on CRAN or Bioconductor repositories, we focused our review on 7 packages which implement all of these features. We believe that our in-depth review of the packages can help users to quickly find the best package for their clustering pipeline and highlight limitations in the implementation of some packages. Our benchmark covers a much broader range of configurations than the previously-published studies (Nityasuddhi and Böhning 2003; Lourens et al. 2013; Leytham 1984; Xu and Knight 2010), as we studied the impact of the level of overlap and the imbalance of the mixture proportions on the quality of the estimation.

Interestingly, the EM algorithm occasionally yields biased and inefficient estimates when the components overlap a lot, which agrees with the past literature (Lourens et al. 2013; Leytham 1984; Xu and Knight 2010). This appears to go counter to the theoretical results presented by Leytham (1984), which demonstrated the asymptotic consistency, unbiasedness, and efficiency of maximum likelihood estimates of GMMs. However, it's important to note that this theoretical demonstration relies on the definition of a "local" environment, necessitating the prior setting of boundaries within which the theorem's conditions are met (in other words, the definition of the *support*, which delineates the region where the initial values can be sampled from). It's not then surprising that the EM algorithm struggles in reaching the global maximum of the distribution in the presence of multiple local extremes.

When all components are well-separated or have a relatively small number of components (three or fewer), we found that the best estimation (lowest MSE and bias) is reached with the latest initialisation method developed, namely the rebmix one. Notably, the global maximum is always properly found in our simulations with distinguishable components. Yet, with overlapping components, the least biased and variable estimates overall are obtained with *k*-means initialisation, enforcing the robustness of the method while the assumptions for using it are not met.

On the contrary, with unbalanced and numerous components (above three), the quantiles initialisation leads to the most biased estimates while the rebmix initialisation induces the highest variability. Indeed, rebmix initialisation is not fit for highly overlapping mixtures, since one of its most restrictive assumption is that each generated interval of the empirical mixture distribution can be associated unambiguously to a component (see Initialisation of the EM algorithm and Nagode (2015)).

Furthermore, rebmix is not particularly adjusted to deal with high-dimensional mixtures, displaying systematically poorer performance compared to other initialisation strategies, such as *k*-means or hierarchical clustering, as illustrated by the summary metrics listed in Appendix *Supplementary Figures*

[9] With a two-components mixture like our bivariate scenario, this even implies to consider an unimodal distribution of the dataset

[10] the gap proceeds from the stronger level of imbalance and the greater number of components

and Tables in the HD simulation. Higher risk of returning a sub-optimal extremum likely arises from the increased data sparsity in high dimensional datasets, which grows as the square root of the number of dimensions $\sqrt{D}$ (Convergence of distance definitions). Thus, we expect that most of the equally-large intervals binning the sampling space and that are used to initiate the rebmix algorithm contain either no or only observation, preventing from retrieving the numerically defined mode of the distribution and increasing the risk of initiating the algorithm in a spurious neighbourhood.

About the remaining initialisation strategies, we observed that, even in the well-separated case, random initializations can sometimes yield highly biased estimates, far from the true parameter values. Consistent with our observations, it was shown in Jin et al. (2016) that the probability for the EM algorithm to converge from randomly initialised estimates to a local suboptimal maximum is non null above two components, increasing with the number of components. Additionally, the local maximum of the likelihood function obtained can be arbitrarily worse than the global maximum. Finally, hierarchical clustering does not take into account any uncertainty on the assignment for an observation to a given class, which explains its rather bad performances with overlapping components. Overall, there is always an initialisation algorithm performing better than the hierarchical clustering, and further it is also by far the slowest and most computationally intensive initialisation method (see supplementary Figure 10).

Our study reveals that while the EM algorithm is supposed to be deterministic, the estimates obtained from its implementations can differ across packages. We were able to link these differences with custom choices of EM implementations across the benchmarked packages. Two distinct classes of packages emerge, each with specific approaches to address certain limitations of the EM algorithm. The first class, exemplified by **mixtools** and **Rmixmod** typically yields smaller but less biased estimates that exhibit lower sensitivity to the choice of initialization method. However, these estimates tend to have higher variability and require longer running times to achieve convergence. The second class, composed of the remaining packages, provide estimates with reduced MSE, but at the extent of a higher bias on the estimates. One plausible explanation is that the first class of packages, comparing absolute iterations of the function to be maximised, tends on average to perform more iterations. The estimated results are accordingly more consistent and closer to the true MLE estimation but at the expense of an increased risk of getting trapped in a local extrema or a plateau, explaining the greater number of outliers observed. Among them, **flexmix** stands out by choosing an unbiased but non MLE-estimate of the covariance matrix, without any clear improvement of the overall performance in our simulations.

Based on these results, we design a decision tree indicating the best choice of package and initialisation method in relation with the shape of the distribution, displayed in Figure 3. Interestingly, our conclusions are consistent between the univariate and bivariate settings. Furthermore, most of the general recommendations on the best choices of packages with respect to the characteristics of the mixture model generally hold in a relatively higher dimensional setting[11]. From this, we assume that projection into a lower-dimensional space is only beneficial in a very high-dimensional setting, for example when the number of dimensions exceeds the number of observations, or when unrestricted parameter estimation (with the full covariance structure) is practically infeasible for computational reasons.

Comparing all these packages suggest several improvements.

1. The use of C++ code speeds up the convergence of the EM algorithm compared to a native R implementation.
2. All packages dealing with GMMs should use *k*-means for overlapping, complex mixtures and rebmix initialisation for well-separated components, provided that the dimension of the dataset is relatively small. The final choice between these two could be set after a first run or visual inspection aiming at determining roughly the level of entropy across mixture proportions and the degree of overlap between components.
3. The packages should allow the user to set their own termination criteria (either relative or absolute log-likelihood or over the estimates after normalisation). Interestingly, **EMMIXmfa** is the only package among those examined that allows the user to consider an absolute or relative convergence endpoint of the EM algorithm, through the `conv_measure` attribute, with `diff` and `ratio` options respectively.
4. With a great number of components or complex overlapping distributions, the optimal package should integrate prior information when available, e.g. via Bayesian estimation.

While **mclust** appeared as the most complete package to model GMMs in R, none of the packages reviewed in this report features all the characteristics mentioned above. We thus strongly believe that

[11]We should note, however, that a larger sample space revealed that the packages **bgmm** and **EMCluster** display more biased and noisy parameters compared to the other packages benchmarked and that their performance was surprisingly unaffected by the number of simulated realisations

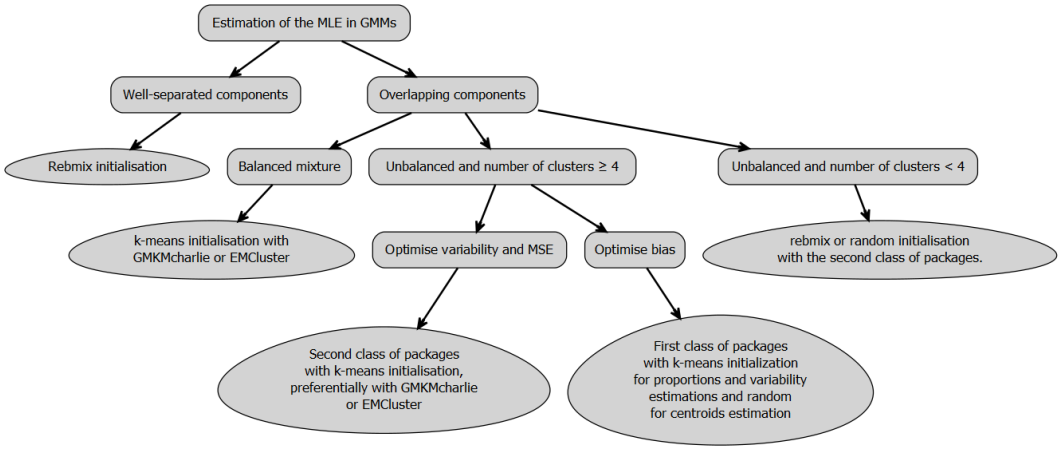


Figure 3: A decision tree to select the best combination of package and initialisation method with respect to the main characteristics of the mixture. It's worth pointing that in both univariate and low dimension multivariate settings, the recommandations are similar.

our observations will help users identify the most suitable packages and parameters for their analyses and guide the development or updates of future packages.

4 Bibliography

Baek, Jangsun, Geoffrey J. McLachlan, and Lloyd K. Flack. 2010. "Mixtures of Factor Analyzers with Common Factor Loadings: Applications to the Clustering and Visualization of High-Dimensional Data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2009.149.

Banfield, Jeffrey D., and Adrian E. Raftery. 1993. "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics*. https://doi.org/10.2307/2532201.

Berge, Laurent, Charles Bouveyron, and Stephane Girard. 2019. *HDclassif: High Dimensional Supervised Classification and Clustering*.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert. 2003. "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models." *Computational Statistics & Data Analysis*. https://doi.org/10.1016/S0167-9473(02)00163-9.

Bouveyron, Charles, and Stéphane Girard. 2009. "Robust Supervised Classification with Mixture Models: Learning from Data with Uncertain Labels." *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2009.03.027.

Celeux, Gilles, and Gérard Govaert. 1992. "A Classification EM Algorithm for Clustering and Two Stochastic Versions." *Computational Statistics & Data Analysis*. https://doi.org/10.1016/0167-9473(92)90042-E.

Chen, Jiahua. 2016. "Consistency of the MLE Under Mixture Models." https://doi.org/10.1214/16-STS578.

Chen, Wei-Chen, and Ranjan Maitra. 2022. *EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution*.

Cochran, W. G. 1934. "The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance." *Mathematical Proceedings of the Cambridge Philosophical Society*. https://doi.org/10.1017/S0305004100016595.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm." *Journal of the Royal Statistical Society*. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Ewa Szczurek, Przemyslaw Biecek &. 2021. *Bgmm: Gaussian Mixture Modeling Algorithms and the Belief-Based Mixture Modeling*.

Fraley, Chris. 1998. "Algorithms for Model-Based Gaussian Hierarchical Clustering." *SIAM Journal on Scientific Computing*. https://doi.org/10.1137/S1064827596311451.

Fraley, Chris, Adrian E. Raftery, and Luca Scrucca. 2022. *Mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*.

Geary, R. C. 1936. "The Distribution of "Student's" Ratio for Non-Normal Samples." *Supplement to the Journal of the Royal Statistical Society*. https://doi.org/10.2307/2983669.

Gruen, Bettina, and Friedrich Leisch. 2022. *Flexmix: Flexible Mixture Modeling*.

Jin, Chi, Yuchen Zhang, Sivaraman Balakrishnan, et al. 2016. "Local Maxima in the Likelihood of

Gaussian Mixture Models: Structural Results and Algorithmic Consequences." Curran Associates, Inc. https://doi.org/https://doi.org/10.48550/arXiv.1609.00978.

Koopman, B. O. 1936. "On Distributions Admitting a Sufficient Statistic." *Transactions of the American Mathematical Society*. https://doi.org/10.2307/1989758.

Kwedlo, Wojciech. 2013. "A New Method for Random Initialization of the EM Algorithm for Multivariate Gaussian Mixture Learning." In *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*, edited by Robert Burduk, Konrad Jackowski, Marek Kurzynski, Michał Wozniak, and Andrzej Zolnierek. Springer International Publishing. https://doi.org/10.1007/978-3-319-00969-8/_8.

Langrognet, Florent, Remi Lebret, Christian Poli, et al. 2021. *Rmixmod: Classification with Mixture Modelling*.

Leytham, K. M. 1984. "Maximum Likelihood Estimates for the Parameters of Mixture Distributions." *Water Resources Research*. https://doi.org/10.1029/WR020i007p00896.

Liu, Charlie Wusuo. 2021. *GMKMcharlie: Unsupervised Gaussian Mixture and Minkowski and Spherical k-Means with Constraints*.

Lourens, Spencer, Ying Zhang, Jeffrey D Long, et al. 2013. "Bias in Estimation of a Mixture of Normal Distributions." *Journal of Biometrics & Biostatistics*. https://doi.org/10.4172/2155-6180.1000179.

McLachlan, Geoffrey, and David Peel. 2000. *Finite Mixture Models: McLachlan/Finite Mixture Models*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471721182.

Melnykov, Volodymyr, Wei-Chen Chen, and Ranjan Maitra. 2021. *MixSim: Simulating Data to Study Performance of Clustering Algorithms*.

Nagode, Marko. 2015. "Finite Mixture Modeling via REBMIX." *Journal of Algorithms and Optimization*.

———. 2022. *Rebmix: Finite Mixture Modeling, Clustering & Classification*.

Nityasuddhi, Dechavudh, and Dankmar Böhning. 2003. "Asymptotic Properties of the EM Algorithm Estimate for Normal Mixture Models with Component Specific Variances." *Computational Statistics & Data Analysis*. https://doi.org/10.1016/S0167-9473(02)00176-7.

Panic, Branislav, Jernej Klemenc, and Marko Nagode. 2020. "Improved Initialization of the EM Algorithm for Mixture Model Parameter Estimation." *Mathematics*. https://doi.org/10.3390/math8030373.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rathnayake, Suren, Geoff McLachlan, David Peel, et al. 2019. *EMMIXmfa: Mixture Models with Component-Wise Factor Analyzers*.

Scrucca, Luca, and Adrian E. Raftery. 2015. "Improved Initialisation of Model-Based Clustering Using Gaussian Hierarchical Partitions." *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-015-0220-z.

Shimizu, Naoto, and Hiromasa Kaneko. 2020. "Direct Inverse Analysis Based on Gaussian Mixture Regression for Multiple Objective Variables in Material Design." *Materials & Design*. https://doi.org/10.1016/j.matdes.2020.109168.

Xu, Dinghai, and John Knight. 2010. "Continuous Empirical Characteristic Function Estimation of Mixtures of Normal Parameters." *Econometric Reviews*. https://doi.org/10.1080/07474938.2011.520565.

Young, Derek, Tatiana Benaglia, Didier Chauveau, et al. 2020. *Mixtools: Tools for Analyzing Finite Mixture Models*.

5 Simulation settings

For ease of reading, we reproduce below the parameter configurations used to run the three benchmarks, respectively for the univariate (Table 4), bivariate (5) and high dimensional setting (Table 6).

Table 4: The 9 parameter configurations tested to generate the samples of the univariate experiment, with $k = 4$ components.

ID	Entropy	OVL	Proportions	Means	Correlations
U1	1.00	3.3e-05	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U2	1.00	5.7e-03	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U3	1.00	2.0e-02	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	2 / 2 / 2 / 2
U4	0.96	3.3e-05	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U5	0.96	5.8e-03	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U6	0.96	2.0e-02	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	2 / 2 / 2 / 2
U7	0.68	2.7e-05	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U8	0.68	4.4e-03	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U9	0.68	1.5e-02	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	2 / 2 / 2 / 2

Table 5: The 20 parameter configurations tested to generate the samples of the bivariate experiment.

ID	Entropy	OVL	Proportions	Means	Correlations
B1	1.00	0.15000	0.5 / 0.5	(0,2);(2,0)	-0.8 / -0.8
B2	1.00	0.07300	0.5 / 0.5	(0,2);(2,0)	-0.8 / 0.8
B3	1.00	0.07300	0.5 / 0.5	(0,2);(2,0)	0.8 / -0.8
B4	1.00	0.00078	0.5 / 0.5	(0,2);(2,0)	0.8 / 0.8
B5	1.00	0.07900	0.5 / 0.5	(0,2);(2,0)	0 / 0
B6	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	-0.8 / -0.8
B7	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	-0.8 / 0.8
B8	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0.8 / -0.8
B9	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0.8 / 0.8
B10	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0 / 0
B11	0.47	0.06600	0.9 / 0.1	(0,2);(2,0)	-0.8 / -0.8
B12	0.47	0.01600	0.9 / 0.1	(0,2);(2,0)	-0.8 / 0.8
B13	0.47	0.05000	0.9 / 0.1	(0,2);(2,0)	0.8 / -0.8
B14	0.47	0.00045	0.9 / 0.1	(0,2);(2,0)	0.8 / 0.8
B15	0.47	0.03900	0.9 / 0.1	(0,2);(2,0)	0 / 0
B16	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	-0.8 / -0.8
B17	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	-0.8 / 0.8
B18	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0.8 / -0.8
B19	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0.8 / 0.8
B20	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0 / 0

6 Additional files

- Additional files related to the univariate setting
  - **S1.** Bootstrap distributions of the estimated parameters for each scenario described in 4.
  - **S2.** Mean, standard deviation, bias and MSE for each individually estimated parameter in configurations listed in 4.
  - **S3.** Distribution of the running times taken for the EM estimation of the parameters of the GMM, across all nine configurations described in 4, for each benchmarked package. We selected the *k*-means algorithm to initialise the EM algorithm, as being the least variable for a given package and scenario.
  - **S4.** Distribution of the time computations taken by the six initialisation methods listed in Table 3.
- Additional files related to the outliers setting:
  - **S5.** Bootstrap distributions of the estimated parameters used to generate Supplementary Figure 2. We additionally include the **otrimle** package, dedicated to these extreme distributions. Two configurations were tested, introducing 2% and 4% of outliers drawn from

Table 6: The 16 parameter configurations tested to generate the samples in a high dimensional context. The first digit of each ID index refers to an unique parameter configuration (identified by its level of overlap, entropy and topological structure, either circular or ellipsoidal, of the covariance matrix, while the lowercase letter depicts the number of observations, a) with $n = 200$ and b) with $n = 2000$.

ID	OVL	Number of observations	Proportions	Spherical
HD1a	1e-04	200	0.5 / 0.5	☑
HD1b	1e-04	2000	0.5 / 0.5	☑
HD2a	1e-04	200	0.19 / 0.81	☑
HD2b	1e-04	2000	0.19 / 0.81	☑
HD3a	1e-04	200	0.5 / 0.5	☒
HD3b	1e-04	2000	0.5 / 0.5	☒
HD4a	1e-04	200	0.21 / 0.79	☒
HD4b	1e-04	2000	0.21 / 0.79	☒
HD5a	2e-01	200	0.5 / 0.5	☑
HD5b	2e-01	2000	0.5 / 0.5	☑
HD6a	2e-01	200	0.15 / 0.85	☑
HD6b	2e-01	2000	0.15 / 0.85	☑
HD7a	2e-01	200	0.5 / 0.5	☒
HD7b	2e-01	2000	0.5 / 0.5	☒
HD8a	2e-01	200	0.69 / 0.31	☒
HD8b	2e-01	2000	0.69 / 0.31	☒

an improper uniform distribution.

  - **S6.** Mean, standard deviation, bias and MSE for each individually estimated parameter in both configurations visualised on Supplementary Figure 2, for each combination of package and initialisation method.

- Additional files related to the bivariate benchmark:

  - **S7.** Bootstrap distributions of the estimated parameters for each scenario described in 5.
  - **S8.** Mean, standard deviation, bias and MSE for each individually estimated parameter in configurations listed in 5.
  - **S9.** Distribution of the running times taken for the EM estimation of the parameters of the GMM, across all twenty configurations described in 5, for each benchmarked package. We selected the *k*-means algorithm to initialise the EM algorithm, as being the least variable for a given package and scenario.

- Additional files related to the high-dimensional benchmark:
  - **S10.** Bootstrap distributions of the estimated parameters for each scenario described in 6.
  - **S11.** Mean, standard deviation, bias and MSE for each individually estimated parameter in configurations listed in 6.
  - **S12.** Distribution of the running times taken for the EM estimation of the parameters of the GMM, across all twenty configurations described in 6, for each benchmarked package. We selected the *k*-means algorithm to initialise the EM algorithm, as being the least variable for a given package and scenario.

7 References

Bastien Chassagnol
Laboratoire de Probabilités, Statistiques et Modélisation (LPSM), UMR CNRS 8001
4 Place Jussieu Sorbonne Université
75005, Paris, France
ORCiD: 0000-0002-8955-2391
bastien_chassagnol@laposte.net

Antoine Bichat
Les Laboratoires Servier
50 Rue Carnot
92150, Suresnes, France
https://rdrr.io/github/abichat/abutils/
ORCiD: 0000-0001-6599-7081
antoine.bichat@servier.com

Cheïma Boudjeniba
Systems Biology Group, Dept. of Computational Biology, Institut Pasteur
25 Rue du Dr Roux
75015 Paris
cheima.boudjeniba@servier.com

Pierre-Henri Wuillemin
Laboratoire d'Informatique de Paris 6 (LIP6), UMR 7606
4 Place Jussieu Sorbonne Université
75005, Paris, France
http://www-desir.lip6.fr/~phw/
ORCiD: 0000-0003-3691-4886
pierre-henri.wuillemin@lip6.fr

Mickaël Guedj
Les Laboratoires Servier
50 Rue Carnot
92150, Suresnes, France
https://michaelguedj.github.io/
ORCiD: 0000-0001-6694-0554
mickael.guedj@gmail.com

David Gohel
ArData
59 rue Voltaire
92800, PUTEAUX, France
https://www.ardata.fr/expertise-r/
ORCiD: 0000-0003-2837-8884
david.gohel@ardata.fr

Gregory Nuel
Laboratoire de Probabilités, Statistiques et Modélisation (LPSM), UMR CNRS 8001
4 Place Jussieu Sorbonne Université

75005, Paris, France
http://nuel.perso.math.cnrs.fr/
ORCiD: 0000-0001-9910-2354
Gregory.Nuel@math.cnrs.fr

Etienne Becht
Les Laboratoires Servier
50 Rue Carnot
92150, Suresnes, France
ORCiD: 0000-0003-1859-9202
etienne.becht@servier.com

3.2 Main results

Overall, we observed a significant decrease in the precision and robustness of Gaussian mixture estimation as the overlap, entropy (level of unbalancedness between cluster proportions), and dimensionality of the problem increased. In details, we have identified two distinct behavioural categories. `mixtools` and `Rmixmod` exhibit smaller bias in estimating mixture parameters, while the remaining packages display the smallest Root Mean Square Error (RMSE) and variability on average.

However, these differences become even more pronounced in high dimensions, with the statistical performances of `bgmm` and `EMCluster` significantly lagging behind their counterparts. Interestingly, we demonstrated that software packages specifically designed to address high dimensionality, such as `EMMIXmfa` or `HDclassif`, certainly achieve significantly improved computational efficiency and enhanced result interpretability, but at the expense of increased bias and reduced accuracy compared to conventional clustering approaches, when applied to moderately-large datasets.

This work is currently in press, and should be published in the next release of the R Journal, in November 2023. The R Journal is an open-access and peer-reviewed journal dedicated to statistical development with real-world applications, programmed in the R language.

Our paper aligns well with the R community's commitment to promoting best statistical practices and sharing comprehensive documentation and guidelines. Indeed, this comprehensive review provides new guidelines to R users for choosing the best package features based on the characteristics of the dataset analysed. Ultimately, we provide recommendations for the development of a comprehensive, and multifaceted mixture model package, concatenating the most promising cutting-edge clustering innovations introduced by the packages benchmarked in this analysis.

3.3 Perspectives

While Gaussian distributions exhibit a large array of interesting theoretical properties and are relevant to describe most biological phenomena, they perform badly to approximate distributions with outliers, high-dimensional datasets (especially when the number of features exceeds the number of observations) and flawed distributions (quasi-Gaussian distributions displaying a significant skewness or kurtosis, bounded datasets, ...).

To expand the scope of our paper, we conducted additional benchmark simulations, closer to biological scenarios:

- Description of variants of the EM-algorithm tailored to account for the presence of an unknown (or user-provided) number of outliers. These tools grapple with two significant challenges: characterising the distribution of outliers, usually modelled by an improper probability distribution (typically, the uniform distribution with infinite support), and striking the correct balance between identifying aberrant data points and unveiling insightful biological phenomena occurring within a small subpopulation (please refer to Appendix B for comprehensive details).
- Robust inference of parameters within a high-dimensional framework. We employed two complementary strategies for this purpose: parsimonious parametrisations of the covariance matrix structure (with the most stringent approach considering homoscedastic noise and

independent features, corresponding to the identity covariance matrix modulo a constant multiplicative factor), and projection into a lower-dimensional space using Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) decomposition approaches (see Appendix B).

- Inference of quasi-Gaussian distributions exhibiting strong skewness or kurtosis.
- Utilization of mixtures of left-truncated Gaussian distributions within the $\log_2$ space of RNA-Seq counts. Alternatively, we could directly employ zero-inflated log-Normal distributions within the raw or TPM transcriptomic space. We could leverage these methods to automate the removal of background noise inherent to any transcriptomic output (see Appendix A.3.2 for more details). In that benchmark scenario, it would be interesting to include additional discrete probabilistic models for inferring the parameters the bimodal distribution of raw counts. For instance, Negative Binomial or Poisson distributions naturally captures the discrete nature of RNA-Seq counts, while the former methods only adhere to the positive constraint of transcriptomic expression.

We investigate in next Chapter 4 an immediate application of mixture models, assuming Gaussian-distributed data, to recover subcategories of individuals sharing similar transcriptomic profiles within a versatile and heterogeneous auto-immune pathology. We will particularly observe that in the medical lingo, this classification of individuals, based on molecular fingerprints, are referred to as “endotypes”.

In part III, in Chapter 6, we explore another statistical extension of multivariate Gaussian distributions. Precisely, we employed a convolution-based approach rather than a mixture-based approach, in order to describe other biological drivers of the variability observed across bulk RNASeq samples, namely cell populations. Nonetheless, the benchmark analysis conducted in this chapter revealed insightful for the creation of this pioneering deconvolution algorithm, by benchmarking the best methods to retrieve the parameters controlling the individual multivariate Gaussian distribution underlying each purified cell population.

Chapter 4

Article 2: A new molecular classification in primary Sjögren's syndrome

Clinical Objective: Establishing a Mapping Between the Diversity of Therapeutic Responses in Primary Sjögren's Syndrome and Molecular Fingerprints

Personalised medicine: a brief overview In recent years, a paradigm shift has moved clinical research from a traditional "one-size-fits-all" approach to the realm of personalized medicine, often referred to as precision medicine ([Twy+23] and [Guc17]).

Personalised medicine tailors healthcare decisions based on individual patient characteristics, encompassing genetic, molecular, and clinical features. Individuals sharing similar molecular profiles are often clustered together, and referred to as endotypes. Precision medicine is hereby expect to reduce adverse drug reactions and improving health patient outcomes, by providing targeted treatments.

To retrieve latent variables underlying the classification of patients, mixture models, already mentioned in previous chapter 3, are widely advocated by clinicians and biostatisticians, these tools streamlining the exploration of disease heterogeneity across individual patient observations.

Primary Sjögren's syndrome: A Debilitating Chronic Disease of Daily Life My former pharmaceutical company, Servier, addressed three immune-mediated disorders: Systemic Lupus Erythematosus (SLE), Multiple Sclerosis (MS), and Primary Sjögren's Disease (pSD), for which the immunopathological mechanisms remain largely unclear. In addition, these ailments typically exhibit strong heterogeneity, characterised by a wide spectrum of clinical symptoms and varying degrees of severity.

In addition, currently developed treatments are primarily tailored to enhance patients' quality of life and alleviate disabling symptoms, but most of them have not been formally validated in clinical trials, nor impact disease progression on the long run.

In this paper, we focused on pSD, which mostly differs from the other diseases by lymphoid infiltration of exocrine glands and Sicca's syndrome. To better unravel the protean clinical symptoms characterising this ailment, we conducted a comprehensive stratification study on a

collection of blood samples, derived from a cross-sectional cohort of 304 patients. Of note, this project is part of, and funded by the PRECISESADS IMI overarching consortium [1].

By integrating several orthogonal modalities, encompassing genetics, epigenomics, and transcriptomics, alongside immuno-phenotypic analyses via flow cytometry, we expect enhanced performance in retrieving meaning endotypes, compared to more traditional classification approaches, based solely on Interferon (IFN) activation scores or clinical features. The latter indeed tend to nelgect the intrinsic molecular heterogeneity of pSD ([Li+13], [GT16]).

This analysis was complemented by a collection, from [Cha+08], of immune-related transcriptomic modules.

Statistical Framework for Unsupervised Patient Stratification Using Gaussian Mixture Models The clustering algorithm was performed on pre-processed RNAseq data, for $N = 304$ individuals. It encompasses the following steps:

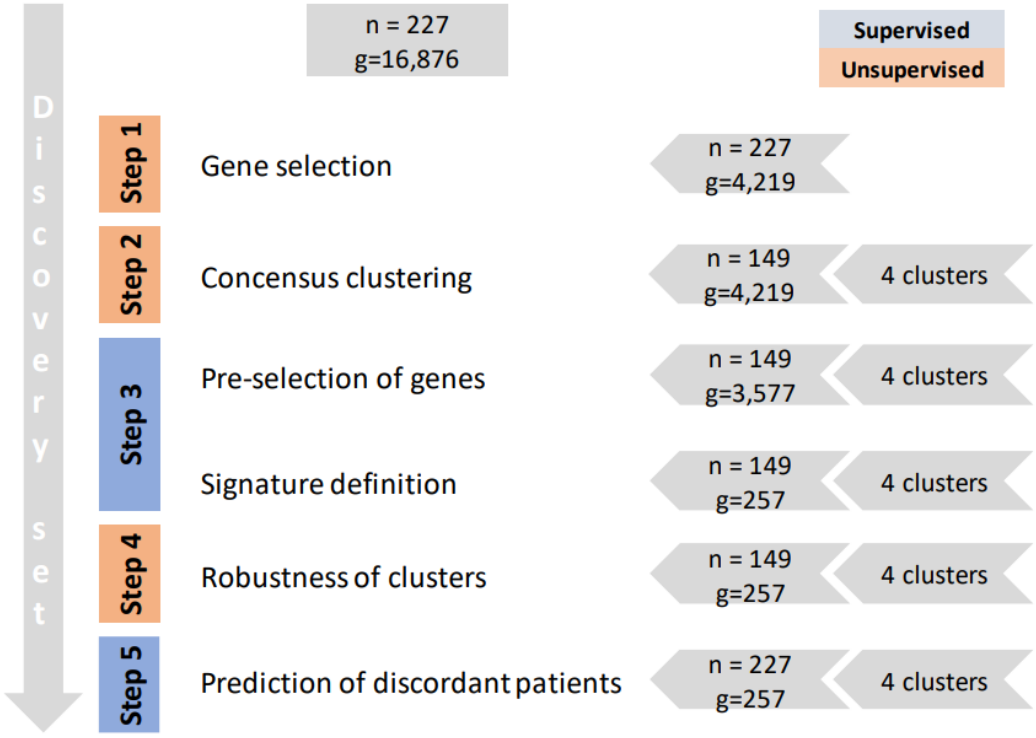
1. The top 25% most variant genes, defined by their degree of variation coefficient(CV), were selected to perform the clustering analysis.

2. The robust clustering, originally outlined in [Gue+12], was leveraged to cluster patients based on their transcriptomic profiles. This method capitalises on two other clustering methods, namely *Agglomerative Hierarchical Clustering* ([Scr+16]) and k-means clustering [Mac67], in addition to standard Gaussian mixture clustering. All these methods are implemented in the versatile toolkit of the `mclust` package [Scr+16]. We selected the hyperparameters that achieved the best overarching consensual cluster assignment for the three clustering algorithms overall (Figure 4.1(a)).

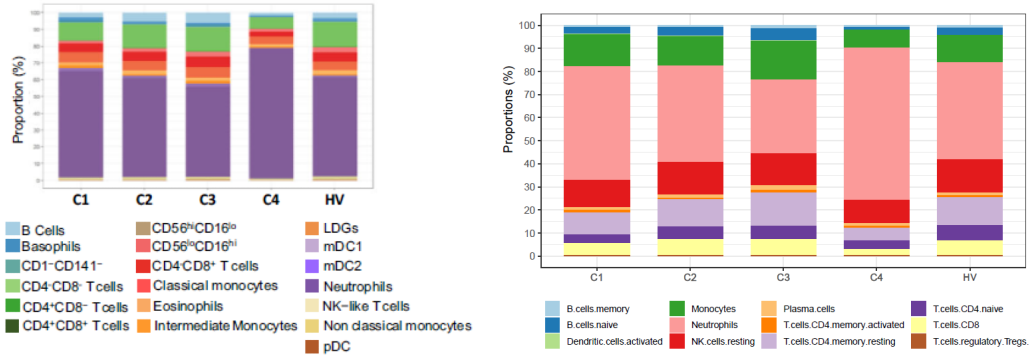
3. From the initial cohort of 304 patients, a subset of 149 "core" patients exhibited consistent cluster assignments for all three clustering methods employed. They have subsequently been selected for deriving a minimal transcriptomic signature of 257 top discriminating genes, after initial gene feature selection using ANOVA and further reduction refinement by a Random Forest approach [CW22].

4. Finally, patients inconsistently assigned for the 3 clustering methods were mapped to their respective closest centroid. Of note, since the overall classification process consists of unsupervised and supervised steps, it is commonly referred as "semi-supervised".

[1] The consortium notably focuses on the discovery of discriminative biomarkers, coupling cutting-edge statistical frameworks with numerous modalities, for accurately predicting disease progression. Once a promising set of putative targets have been identified, the Necessity group endeavours to bring about transformative clinical trial designs, referred to as "multi-arm multi-stage platform", and relying intensively on prior patient stratification [Bar+18]



(a) Flow chart describing the "semi-supervised" clustering approach used for patient stratification. A bagging ensemble approach, inspired from [Gue+12], has notably been used for increasing the robustness of the classification.



(b) Cell population composition in blood samples from the 4 identified clusters.. The bar charts denotes the relative cellular composition for each cluster. To the right, the cell ratios inferred using Fluorescence-Activated Cell Sorting (FACS) analyses (Section 2.2.1), and to the left, numerically estimated cell ratios using the CIBERSORT deconvolution algorithm ([New+15]).

Figure 4.1: Visual summary of Chapter 4, Article 2.

4.1 Article 2

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23472-7 OPEN

A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome

Perrine Soret[1,29], Christelle Le Dantec[2,29], Emiko Desvaux[1,2], Nathan Foulquier[2], Bastien Chassagnol[1], Sandra Hubert[1], Christophe Jamin[2,3], Guillermo Barturen[4], Guillaume Desachy[1], Valérie Devauchelle-Pensec[2,3], Cheïma Boudjeniba[1], Divi Cornec[2,3], Alain Saraux[2,3], Sandrine Jousse-Joulin[2,3], Nuria Barbarroja[5], Ignasi Rodríguez-Pintó[6], Ellen De Langhe[7], Lorenzo Beretta[8], Carlo Chizzolini[9], László Kovács[10], Torsten Witte[11], PRECISESADS Clinical Consortium*, PRECISESADS Flow Cytometry Consortium*, Eléonore Bettacchioli[3], Anne Buttgereit[12], Zuzanna Makowska[12], Ralf Lesche[12], Maria Orietta Borghi[13], Javier Martin[14], Sophie Courtade-Gaiani[1], Laura Xuereb[1], Mickaël Guedj[1], Philippe Moingeon[1], Marta E. Alarcón-Riquelme[4], Laurence Laigle[1] & Jacques-Olivier Pers[2,3]✉

There is currently no approved treatment for primary Sjögren's syndrome, a disease that primarily affects adult women. The difficulty in developing effective therapies is -in part- because of the heterogeneity in the clinical manifestation and pathophysiology of the disease. Finding common molecular signatures among patient subgroups could improve our understanding of disease etiology, and facilitate the development of targeted therapeutics. Here, we report, in a cross-sectional cohort, a molecular classification scheme for Sjögren's syndrome patients based on the multi-omic profiling of whole blood samples from a European cohort of over 300 patients, and a similar number of age and gender-matched healthy volunteers. Using transcriptomic, genomic, epigenetic, cytokine expression and flow cytometry data, combined with clinical parameters, we identify four groups of patients with distinct patterns of immune dysregulation. The biomarkers we identify can be used by machine learning classifiers to sort future patients into subgroups, allowing the re-evaluation of response to treatments in clinical trials.

[1] Institut de Recherches Internationales Servier, Departments of Translational Medicine and Immuno-Inflammatory Diseases Research and Development, Suresnes, France. [2] LBAI, UMR1227, Univ Brest, Inserm, Brest, France. [3] CHU de Brest, Brest, France. [4] Department of Medical Genomics, Center for Genomics and Oncological Research (GENYO), Granada, Spain. [5] Reina Sofia Hospital, Maimonides Institute for Research in Biomedicine of Cordoba (IMIBIC), University of Cordoba, Cordoba, Spain. [6] Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Catalonia, Spain. [7] Skeletal Biology and Engineering Research Center, KU Leuven and Division of Rheumatology, UZ Leuven, Belgium. [8] Scleroderma Unit, Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca'Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy. [9] Immunology & Allergy, University Hospital and School of Medicine, Geneva, Switzerland. [10] University of Szeged, Szeged, Hungary. [11] Klinik für Immunologie und Rheumatologie, Medical University Hannover, Hannover, Germany. [12] Pharmaceuticals Division, Bayer Pharma Aktiengesellschaft, Berlin, Germany. [13] Università degli studi di Milano, Milan, Italy. [14] Institute of Parasitology and Biomedicine López-Neyra, Consejo Superior de Investigaciones Científicas (IPBLN-CSIC), Granada, Spain. [29] These authors contributed equally: Perrine Soret and Christelle Le Dantec. *Lists of authors and their affiliations appear at the end of the paper. ✉email: pers@univ-brest.fr

Primary Sjögren's syndrome (pSS) is a chronic, disabling, complex systemic autoimmune disease that mostly affects adult women and still lacks a specific therapy. Although the involvement of salivary and lachrymal glands is the hallmark of the disease, during pSS progression, various organs and systems can be involved including joints, lungs, kidneys, liver, nervous and musculoskeletal system[1]. Thus, the clinical spectrum of the disease ranges from a benign slowly progressive autoimmune exocrinopathy to a severe systemic disorder with significant symptom heterogeneity and scattered complications. The diagnosis of pSS is currently based upon a combination of clinical, serological, histological, and functional parameters which are most often only satisfied at a late stage of the disease, i.e., when glandular dysfunction and symptoms already severely affect a patient's overall quality of life. Moreover, one fifth of pSS patients may present major organ involvement with potentially severe end-organ damage[2] and five percent of patients may also develop non-Hodgkin's lymphoma[3]. Primary SS is one of the few prototypic diseases to link autoimmunity, cancer development and infections, offering unique insights in many areas of basic science and clinical medicine. However, the pathogenesis of the disease remains elusive. Specifically, limited knowledge of existing pSS disease variants arguably represents the greatest obstacle to improve patients' diagnosis and identify patients' subsets in view of early stratification and personalized treatment[4]. It was recently shown in the PRECISESADS IMI JU project that systemic autoimmune diseases exhibit a diverse spectrum and a complex nuanced or overlapping molecular phenotype with four clusters identified, representing 'inflammatory', 'lymphoid', 'interferon' and 'healthy-like' patterns each including all diagnoses and defined by genetic, clinical, serological and cellular features[5]. Many of them share susceptibility genes[6] and an overexpression of interferon (IFN) inducible genes known as the IFN signature is observed in many of these patients[7]. Such autoimmune diseases are driven by numerous environmental factors, therefore displaying a marked variability in their natural course as it relates to their initiation, propagation and flares.

The present study was undertaken to establish a precise molecular classification of patients affected by pSS into more homogeneous clusters whatever their disease phenotypes, activity or treatment. We report herein on the integrated molecular profiling of 304 pSS patients compared to 330 matched healthy volunteers (HV) performed using high-throughput multi-omics data collected within the PRECISESADS IMI JU project (genetic, epigenomic, transcriptomic, combined with flow cytometric data, multiplexed cytokines, as well as classical serology and clinical data). We identify 4 groups of patients with distinct patterns of immune dysregulation. The Cluster 1 (C1), C3 and C4 display a high IFN signature reflecting the pathological involvement of the IFN pathway, but with various Type I and II IFN gene enrichment. C1 has the strongest IFN signature with both Type I and Type II gene enrichment when compared to C3 (intermediate) and C4 (lower). C4 has a Type II gene enrichment stronger than Type I and equivalent to C3 while C3 has the opposite composition. C2 exhibits a weak Type I and Type II IFN signature with no other obvious distinguishable profile relative to HV. We further characterized C1, C3 and C4 using multi-omics and clinical data. C1 patients present a high prevalence of SNPs, C3 patients an involvement of B cell component more prominent than in the other clusters and especially an increased frequency of B cells in the blood while C4 patients have an inflammatory signature driven by monocytes and neutrophils, together with an aberrant methylation status. Algorithms derived from machine learning discriminate the 4 clusters based on distinct biomarkers that can be easily used in a composite model to stratify patients in clinical trials. This composite model is validated by using an independent inception cohort of 37 pSS patients. In conclusion, this work provides a clear understanding of pSS heterogeneity providing clinically and immunopathologically relevant signatures to guide precision medicine strategies. Decision trees coming from this patient classification have an immediate application to re-evaluate response to treatments in clinical trials.

Results

Four functional molecular clusters of pSS patients were identified. Our initial study population comprised 382 pSS patients enrolled in the PRECISESADS cross-sectional study. Following complete quality control and diagnosis validation (each patient had to present either anti-SSA/Ro antibody positivity or focal lymphocytic sialadenitis with a focus score of ≥1 foci/mm^2), 78 patients were removed (Supplementary Fig. 1a–c). Patient characteristics are presented in Table 1. To perform the clustering of the remaining 304 samples, transcriptomics data were analyzed with a semi-supervised robust approach previously applied to breast cancer[8] that iterates unsupervised and supervised steps and relies on the concordance between 3 methods of clustering (see Methods). Samples were divided into a discovery set and an independent validation set, representing 75 and 25% of samples, respectively. The discovery set allowed to cluster patients in four groups, as confirmed in the validation set (Fig. 1a). When the two sets were merged, Cluster 1 (C1) contained 101 patients (33.2%), Cluster 2 (C2) 77 patients (25.3%), Cluster 3 (C3) 88 patients (28.9%) and Cluster 4 (C4) 38 patients (12.5%). The supervised step allowed to select a subset of 257 top genes discriminating the 4 clusters of patients (Supplementary Fig. 2) and divided into 3 modules: M.a (105 genes), M.b (20 genes) and M.c (132 genes). An enrichment analysis was used to annotate each gene module, showing that M.a was enriched in IFN signaling, M.b in lymphoid lineage pathways and M.c in inflammatory and myeloid lineage transcripts (Supplementary Fig. 3). C1, and to a lesser extent C3, presented overexpression of gene module M.a, whereas C3 showed overexpression of M.b as well and C4 strong overexpression of M.c (Fig. 1a). Because C2 had no obvious discernible pattern, healthy volunteers (HV) were assigned to the 4 molecular clusters distance to centroids (Fig. 1b). When projected into the patient population, HV did not constitute a separate cluster but mainly matched with C2 (0.5%, 93%, 4% and 2.5% of HV merged with C1, C2, C3, and C4, respectively). This means that the C2 transcriptional signature is not different from HV, at least at the blood level. Interestingly, our data are consistent with the previous observation of a healthy-like patient group detected in a pooled population of 7 different autoimmune diseases[5].

We then assessed whether covariates like systemic treatments could drive the transcriptome-based clustering. Indeed, half of the pSS patients were treated with either anti-malarials, immunosuppressants, or steroids at the time of the visit with a statistically significant difference in the distribution among the four clusters (p-values were respectively 0.002 for anti-malarials, <0.001 for immunosuppressants and steroids) (Table 2). When compared to the 3 other clusters, a higher proportion of patients treated with anti-malarials in C2 and a higher proportion of patients receiving immunosuppressants or steroids in C4 were observed. Importantly, sensitivity analyses of treated versus untreated patients in each cluster showed no impact of treatments on cluster distribution (Supplementary Fig. 4).

In depth functional pathway analysis of individual pSS clusters. To investigate molecular processes and their biological function underlying each of the pSS patients' clusters, specific differentially expressed genes (DEG) signatures compared to HV were assessed using Limma in the 4 clusters. Ingenuity Pathway Analysis (IPA)

Table 1 Healthy volunteers (HV) and Primary Sjögren's syndrome (pSS) patient characteristics.

			HV (*N* = 330)	pSS Discovery (*N* = 227)	pSS Validation (*N* = 77)	pSS All (*N* = 304)
Demography						
Age		*n*	330	227	77	304
		Mean ± SD	53.294 ± 10.998	58.524 ± 13.440	58.039 ± 13.554	58.401 ± 13.448
Gender		*n*	330	227	77	304
	Female	*n* (%)	302 (91.52)	211 (92.95)	71 (92.21)	282 (92.76)
Obesity (BMI > = 30)		*n*	328	218	74	292
	Yes	*n* (%)	24 (7.27)	30 (13.76)	3 (4.05)	33 (11.30)
Race		*n*	330	227	77	304
	Asian	*n* (%)	2 (0.61)	1 (0.44)	1 (1.30)	2 (0.66)
	Black/African American	*n* (%)	—	—	1 (1.30)	1 (0.33)
	Caucasian/White	*n* (%)	328 (99.39)	224 (98.68)	74 (96.10)	298 (98.03)
	Other	*n* (%)	—	2 (0.88)	1 (1.30)	3 (0.99)
Diagnostic criteria						
Focus score > 1		*n*	—	82	27	109
	Yes	*n* (%)	—	73 (89.02)	24 (88.89)	97 (88.99)
Anti-SSA positivity		*n*	—	227	77	304
	Yes	*n* (%)	—	205 (90.30)	69 (89.61)	274 (90.13)
Disease activity						
Disease duration, years		*n*	—	225	77	302
		Mean ± SD	—	10.788 ± 7.535	11.094 ± 9.620	10.866 ± 8.101
Disease activity (PGA*)		*n*	—	211	75	286
		Mean ± SD	—	25.687 ± 18.976	24.840 ± 20.984	25.465 ± 19.488
ESSDAI (**)		n	—	133	60	193
		Mean ± SD	—	4.609 ± 5.358	4.850 ± 5.495	4.684 ± 5.388
ESSPRI (**)		*n*	—	106	44	150
		Mean ± SD	—	5.176 ± 2.286	4.568 ± 2.648	4.998 ± 2.405

n: Number of patients with available information.
(*) PGA: Physician Global Assessment.
(**) collected in a substudy.

was subsequently applied to determine the most significantly dysregulated canonical pathways with Benjamini–Hochberg false discovery rate (FDR) adjusted p-value ≤ 0.05 and absolute fold change (FC) ≥ 1.5. As a result, 284 DEG were found significant in C1, 301 DEG in C3 and 1686 DEG in C4 (Supplementary Data 1).

Since no DEG were noticed in C2 when compared to HV, only C1, C3, and C4 were functionally annotated. Top 20 significant canonical pathways within each DEG signature are presented in Supplementary Data 2 and pathways related to the most significantly enriched immunological responses are reported as radar plots in Fig. 1c. While all 3 clusters were enriched in genes involved in antiviral and anti-bacterial responses indicative of an innate-mediated activation profile, C1 was mainly enriched with IFN-related pathways including IFN signaling, role of pattern recognition receptors for bacteria and viruses and Interferon Regulatory Factor (IRF) activation. Notably, C3 and C4 were further characterized by alterations in biological networks linked to adaptive immunity. Specifically, significant activation of canonical pathways related to B cell activation such as B cell receptor signaling, and B cell development were observed in C3. In addition, comparative analyses provided evidence for IL7-signaling up-regulation and LXR/RXR activation in C3 compared to C1.

Interestingly, C4 was the endotype with the highest number of DEG compared to HV with highly heterogeneous dysregulated canonical pathways. Ingenuity pathway analysis confirmed the activation of T and B lymphocyte related pathways reflecting Th1 and Th2 activation, B cell receptor signaling, together with prominent inflammatory signatures most particularly linked to cytokine signaling (IL-6 and IL-10 signaling, IL-15 production, STAT-3 pathway).

Further upstream regulator analysis predicted significant activation of IFN-α in all three clusters, as well as CpG ODN in C3 and LPS, IFNγ, TNF-α, and IL-4 in C4, further highlighting B cell activity and inflammatory responses in C3 and C4, respectively.

Noteworthy, while C2 displayed no DEG compared to HV, 14 genes were differentially expressed in C2 patients positive for SSA antibodies compared to HV whereas only 2 DEG were found in SSA-negative C2 patients. These SSA-positive C2 patients were characterized by significant enrichment in IFN-related genes compared to HV including *IFI44, IFI44L, IFI6, IFIFT1, IFIT3, ISG15, MX1, OAS3, SERPING1,* and *SIGLEC1* (Supplementary data 1).

To further characterize patient cluster variability at a molecular level, we then used the blood transcriptome modular repertoire recently established on an expended range of disease and pathological states. The latter includes 382 transcriptome modules based on genes co-expression patterns across 16 diseases and 985 unique transcriptome profiles[9]. Again, no aggregate was found differentially expressed in C2 confirming the healthy-like profile of these patients, whereas an up-regulated IFN signature dominated in C1, C3, and C4 (Fig. 2). In C4, the most induced modules include genes associated with inflammation and neutrophils. As the highest inflammatory phenotype, C4 is associated with a hypercytokinemia/hyperchemokinemia

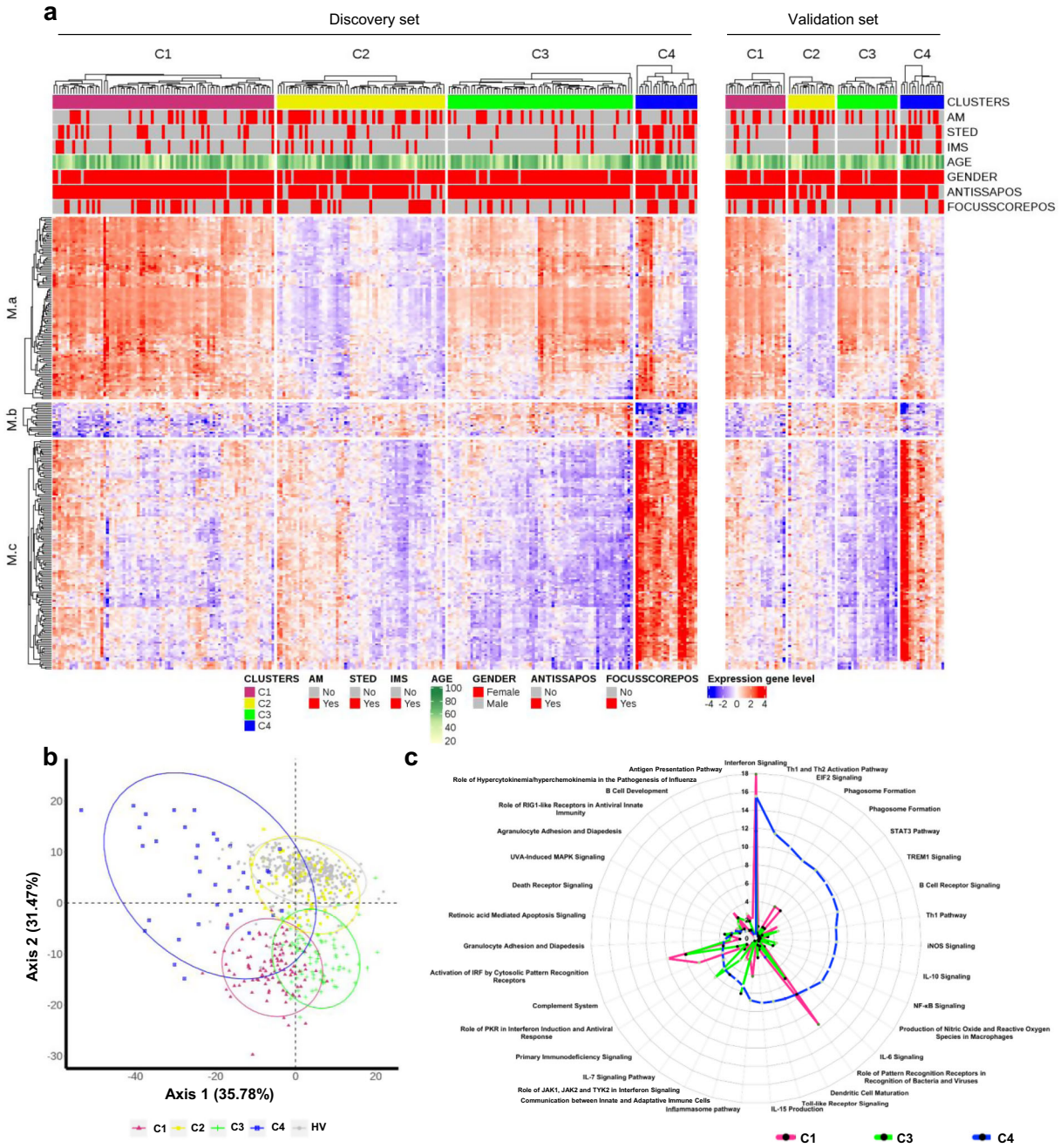


Fig. 1 Molecular pattern distribution is represented by 4 clusters of pSS patients with different canonical pathways. a Heatmap performed for 304 pSS patients (Discovery set: 227, Validation set: 77) showing the distribution of gene transcripts across the 4 clusters. In columns patients are grouped by cluster assignment and in rows genes are grouped by functional modules. Each subset of patients (discovery set on the left and validation set on the right) is presented separately. Red represents overexpression and blue represents under-expression. At the top of the figure annotations show: each of the treatment groups for each individual (AM: antimalarials, STED: steroids and IMS: immunosuppressors, red represents patients with treatment and gray represents patients without treatment), age (levels of yellow to green with yellow for younger patients and dark green for older patients), gender (red represents woman and gray represents man), ANTISSAPOS: anti-SSA/Ro antibody positivity, FOCUSSCOREPOS: focus score of ≥1 foci/mm^2 (red represents focus score of ≥1 foci/mm^2 and gray represents focus score of <1 foci/mm^2). **b** Scatterplot of the first two components PCA (performed for 304 pSS patient and 330 HV) model showing clearly defined clusters in signature gene. HV (gray dot) are confused with C2 cluster (yellow dot). **c** Top 20 most significant canonical pathways for each cluster. Radar plots are represented according to –log (*p*-value) (Fisher's exact test) associated to the most significant pathways of each cluster; C1 (pink), C3 (green), C4 (blue).

observed in modules (M13.16, M15.84, M16.80) consistent with an upregulation of the TNF-associated module (M16.47) and a downregulation of the TGFβ-associated module (M16.65) (Fig. 2). Some modules were under-expressed, such as those associated with both protein synthesis (M12.7, M11.1, M13.28, M14.80), B cells (M13.27, M12.8) and T cells (M15.38, M14.42, M12.6). Genes mainly overexpressed in C1 were also implicated in inflammatory responses and neutrophils (A33, A35), in parallel with down-regulated B and T cell signatures (Supplementary Fig. 5). Moreover, distinct sub-modules expressed in opposite

Table 2 Descriptive analysis of the clinical parameters by primary Sjögren's syndrome cluster.

			C1 ($n = 101$)	C2 ($n = 77$)	C3 ($n = 88$)	C4 ($n = 38$)	*p*-value
Age, years		*n*	101	77	88	38	
		Mean ± SD	57.327 ± 13.705	58.805 ± 13.688	57.250 ± 12.032	63.105 ± 14.790	0.10
Gender		*n*	101	77	88	38	
	Female	*n* (%)	96 (95.05)	71 (92.21)	81 (92.05)	34 (89.47)	0.70
Age at onset, years		*n*	101	76	88	37	
		Mean ± SD	45.663 ± 14.475	50.428 ± 14.532	47.606 ± 12.687	51.739 ± 16.053	0.071
Disease duration, years		*n*	101	76	88	37	
		Mean ± SD	12.247 ± 8.921	8.965 ± 7.336	10.183 ± 7.210	12.625 ± 8.524	0.029
Disease activity (PGA*)		n	94	71	85	36	
		Mean ± SD	27.245 ± 20.535	22.718 ± 17.698	23.212 ± 18.766	31.556 ± 20.646	0.092
ESSDAI		n	70	52	44	27	
		Mean ± SD	5.029 ± 5.959	3.731 ± 4.594	4.227 ± 4.017	6.370 ± 6.828	0.10
ESSPRI		n	56	43	30	21	
		Mean ± SD	4.833 ± 2.460	5.031 ± 2.429	5.300 ± 2.703	4.937 ± 1.803	0.87
Arthritis		*n*	98	77	86	38	
	Past	*n* (%)	39 (39.80)	18 (23.38)	20 (23.26)	12 (31.58)	0.016
	Present	*n* (%)	2 (2.04)	3 (3.90)	4 (4.65)	5 (13.16)	
Focus score > 1		*n*	96	29	21	14	
	Yes	*n* (%)	39 (40.63)	28 (96.55)	17 (80.95)	12 (85.71)	0.4
Anti-SSA positivity		*n*	101	77	88	38	
	Yes	*n* (%)	99 (99.00)	56 (72.72)	87 (98.86)	31 (81.57)	<0.001
Anti-SSB positivity		*n*	100	77	86	38	
	Yes	*n* (%)	61 (61.00)	12 (15.58)	39 (45.35)	11 (28.95)	<0.001
Hypergammabulinemia		*n*	97	73	86	38	
	Past	*n* (%)	23 (23.71)	8 (10.96)	9 (10.47)	3 (7.89)	<0.001
	Present	*n* (%)	44 (45.36)	10 (13.70)	41 (47.67)	7 (18.42)	
Abnormal inflammatory indexes		*n*	100	77	87	38	
	Past	*n* (%)	28 (28.00)	13 (16.88)	20 (22.99)	12 (31.58)	0.003
	Present	*n* (%)	35 (35.00)	11 (14.29)	22 (25.29)	10 (26.32)	
Reduced C3 levels		*n*	93	74	82	35	
	Past	*n* (%)	13 (13.98)	5 (6.76)	11 (13.41)	4 (11.43)	0.8
	Present	*n* (%)	7 (7.53)	4 (5.41)	5 (6.10)	3 (8.57)	
Reduced C4 levels		*n*	93	74	82	35	
	Past	*n* (%)	13 (13.98)	3 (4.05)	9 (10.98)	4 (11.43)	0.10
	Present	*n* (%)	10 (10.75)	3 (4.05)	3 (3.66)	4 (11.43)	
Abnormal Creatinine		*n*	98	77	88	38	
	Past	*n* (%)	10 (10.20)	4 (5.19)	-	2 (5.26)	0.009
	Present	*n* (%)	5 (5.10)	2 (2.60)	7 (7.95)	6 (15.79)	
Proteinuria		*n*	65	58	56	25	
	Moderate	*n* (%)	5 (7.69)	2 (3.45)	1 (1.79)	3 (12.00)	0.093
	Past	*n* (%)	5 (7.69)	—	3 (5.36)	—	
Current use of antimalarials		*n*	101	77	88	38	
	Yes	*n* (%)	33 (32.67)	42 (54.55)	24 (27.27)	15 (39.47)	0.002
Current use of Immunosuppressants		*n*	101	77	88	38	
	Yes	*n* (%)	17 (16.83)	14 (18.18)	7 (7.95)	15 (39.47)	<0.001
Current use of steroids		*n*	101	77	88	38	
	Yes	*n* (%)	23 (22.77)	14 (18.18)	10 (11.36)	23 (60.53)	<0.001

n: Number of patients with available information, (*) PGA: Physician Global Assessment.
Statistical tests performed: chi-square test of independence for categorial variable and Kruskal–Wallis test for continue variable.

directions allows to functionally discriminate C1 and C3. Patients from C3 demonstrated a significant under-expression of modules related to erythrocytes (A37; M9.2, M11.3) and cytokines/chemokines (A35; M15.84, M13.16) and an increased expression in some of the B cell modules (A1; M12.8) (Supplementary Fig. 5 and Fig. 2).

IFN signatures. Consistent with the literature, the most significantly enriched pathway confirmed to be up-regulated in all three clusters was the IFN signaling pathway (Fig. 2, Supplementary Fig. 5). In SLE, Chiche et al. have previously identified three strongly up-regulated IFN-annotated modules (M1.2, M3.4, and M5.12) from peripheral blood transcriptomic data, with for each module a distinct activation threshold[10]. Genes within the M1.2 module are induced by IFNα, while other genes from both M1.2 and M3.4 are up-regulated by IFNβ, corresponding to a type I IFN signature. The M5.12 genes are poorly induced by IFNα and IFNβ alone but are rather up-regulated by IFNγ characterizing a type II IFN signature[11]. Moreover, transcripts belonging to M3.4 and M5.12 were only fully induced by a combination of Type I and Type II IFNs. Kirou et al. made similar observations and identified genes preferentially induced by IFNα or IFNγ[12]. The different z-scores were then calculated accordingly to characterize further the IFN signature observed in the various clusters (Fig. 3). All IFN z-scores were increased to some extent in C2 when compared to HV. In line with the strong signal observed, C1 patients had the highest Type I and type II scores. Interestingly, C3 had higher Type I IFN score than C4 but these 2 clusters were not different for Type II IFN score.

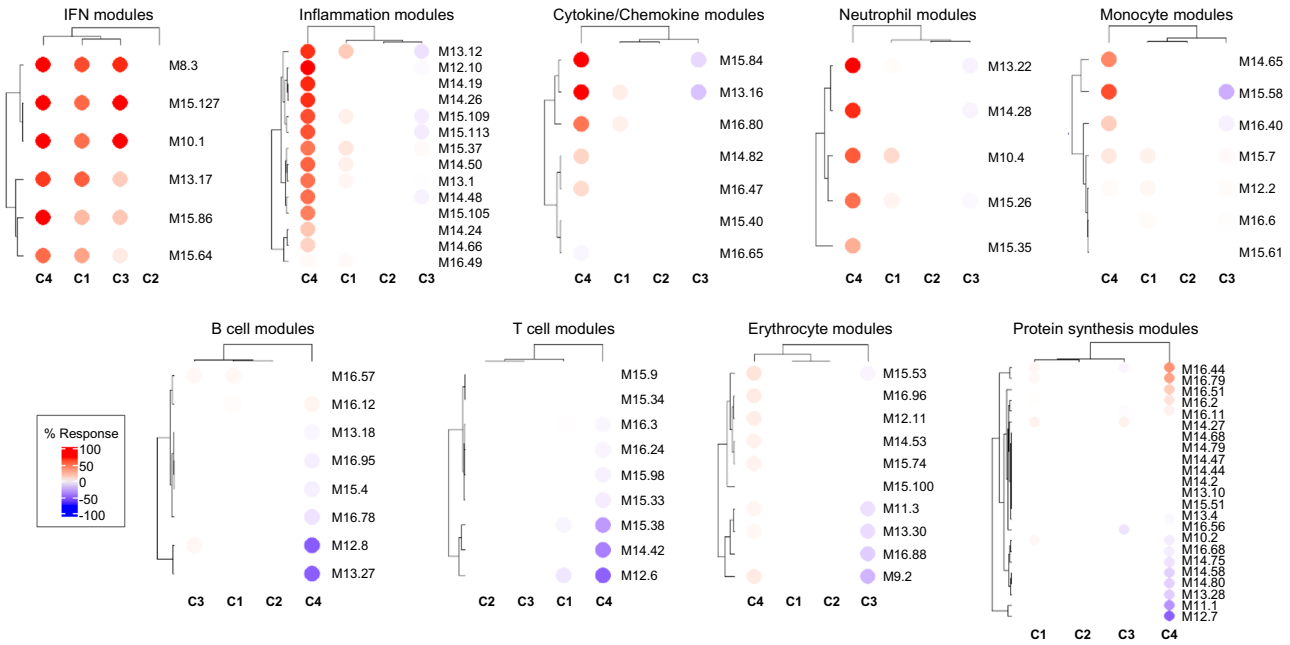


Fig. 2 Patterns of abundance of the different modules distinguish the four pSS clusters. Each heatmap, achieved with BloodGen3Module R package[9], represents one of the most significant patterns differentiating the four clusters of 304 pSS patients (C1: 101, C2: 77, C3: 88, and C4: 38) compared to 330 healthy volunteers (HV). These patterns correspond to modules associated with IFN, neutrophils, inflammation, cytokines/chemokines, protein synthesis, erythrocytes, monocytes, B cells and T cells. Columns on this heatmap corresponds to clusters. Each row corresponds to one of the modules associated with the pattern. For each module, the percentage of increased genes (from 0 to 100) and decreased genes (from 0 to 100) were calculated. A red spot on the heatmap indicates an increase in abundance of transcripts comprising a given module for a given cluster. A blue spot indicates a decrease in abundance of transcripts. The absence of color indicates no changes.

Upstream analysis of C4 DEG predicted IFNγ as an important regulator suggesting that Type II IFN activation was prominent in C4.

Genome-wide association study analysis. We investigated whether clusters showed any differences in the genetic contribution of risk alleles known to be associated with pSS[13–15]. Even in the mid-size cohort of patients analyzed (304 pSS and 330 HV), we unambiguously detected (with signals genome wide significance level $<5 \times 10^{-8}$) 35 single nucleotide polymorphisms (SNPs) in C1 compared to only six in C3 and one in C4 (Fig. 4a, Supplementary Data 3). Interestingly, no significant enrichment was found in C2. The 35 SNPs assessed in C1 are found within genes associated with either the immune system (*HLA-DQB1, HLA-DQA1, HLA-DRA, HLA-C, HLA-G*), signal transduction (*NOTCH4*), developmental biology (*POU5F1*), gene expression (*DDX39B*) or cell cycle (*TUBB*). The presence of such significant genetic associations was already found in clusters of systemic autoimmune disease patients whose molecular disease pathway is the Type I IFN pathway[5]. Moreover, a strong association of SNPs with HLA class II genes was reported in SLE patients with a high level of autoantibodies[16]. One SNP (rs2734583) was common to C1 and C3 and is associated to the *DDX39* gene. Of note, DDX39B, the protein encoded by this gene, is required for the prevention of dsRNA formation during influenza A virus infection, thereby preventing the activation of the Type I IFN system[17]. The five others SNPs in C3 are nearby *HLA-DQA*, *HLA-DRA* (2 SNPs), *BTNL2* and *HCG23*. The only SNP (rs2247056) found in C4, also common with C1, is located in intron 1 of the *LINC02571* gene and was previously associated with a risk for developing SLE.

Linkage disequilibrium is a non-random association of alleles at different loci in a given population. When analyzing linkage disequilibrium (Fig. 4b) in the loci of the 35 SNPs detected in C1 and located on chromosome 6 (from base 29809362 to 32681631), three SNPs were strongly associated in *HLA-DQA1* locus (rs9272219, rs9271588, rs642093), five SNPs in *HLA-DRA* | *HLA-DQA1* locus (rs7195, rs1041885, rs3129890, rs9269043, rs7749057) and three SNPs in *HCG27* | *HLA-C* locus (rs3130473, rs2394895 and rs3130467). Two other regions contain strongly associated SNPs. The *NOTCH4* | *C6orf10* locus presented 5 associated SNPs (rs3130347, rs204991, rs3132935, rs7751896, rs9268220) as well as the *IER3* | *DDR1* locus (rs3094122, rs6911628, rs3094112, rs2517576, rs3095151).

Methylation analysis. The methylation analysis was performed with a Benjamini Hochberg FDR < 0.1 and absolute ΔBeta > 0.075. Only two differentially methylated positions (DMPs) corresponding to two genes were found in C2. Those DMPs were common with the 3 other clusters (Fig. 5a) and were located in the TSS1500 shore of the *NLRC5* gene and in the 5'UTR of the gene encoding *MX1*, two genes involved in the IFN signature. NLRC5 plays a role in cytokine response and antiviral immunity through inhibition of NF-kappa-B activation and negative regulation of Type I IFN signaling pathways[18]. *MX1* encodes an IFN induced dynamic-like GTPase with antiviral activity which was proposed as a clinically applicable biomarker for identifying systemic Type I IFN in pSS[19].

145 DMPs corresponding to 87 genes and 96 DMPs corresponding to 56 genes were found in C1 and C3 respectively, whereas an aberrant methylation status with 8,445 DMPs corresponding to 3,636 genes characterized C4 (Fig. 5a). In order to test whether the methylation defect in C4 was associated with steroids treatment, we compared the 9 untreated to the 17 treated patients. No CpG with a Benjamini-Hochberg FDR adjusted *p*-value < 0.1 was found to be differentially methylated in treated versus untreated patients. A global hypomethylation of CpG was observed for all clusters (89.6% in C1, 100% in C2, 67.7% in C3

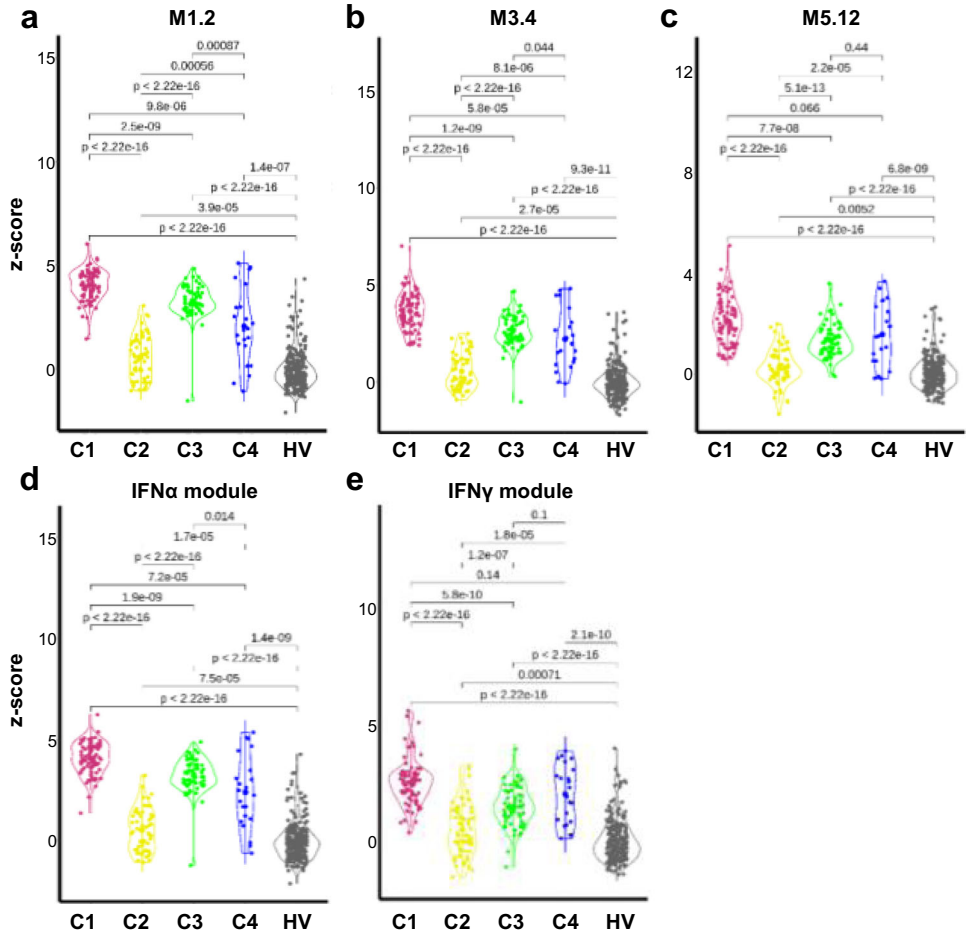


Fig. 3 The 4 pSS clusters show typical IFN signature according to modular IFN z-scores. IFN score analyses were performed for 304 pSS patients and 330 healthy volunteers (HV). Repartition of samples from the 4 pSS clusters are shown according to the most characterized IFN module z-scores. The genes (*IFI44*, *IFI44L*, *IFIT1* and *MX1*) of the M1.2 module (**a**) are induced by IFNα, while genes from both M1.2 and M3.4 (**b**) (*ZBP1*, *IFIH1*, *EIF2AK2*, *PARP9* and *GBP4*) are up-regulated by IFNβ. **c** The genes (*PSMB9*, *NCOA7*, *TAP1*, *ISG20* and *SP140*) from the M5.12 module are poorly induced by IFNα and IFNβ alone while they are up-regulated by IFNγ. Moreover, transcripts belonging to M3.4 and M5.12 were only fully induced by a combination of Type I and Type II IFNs[10]. Other modules identified genes preferentially induced by IFNα (*IFIT1*, *IFI44* and *EIF2AK2*) (**d**) or IFNγ (*IRF1*, *GBP1* and *SERPING1*) (**e**)[12]. Two-tailed pairwise Wilcoxon-rank sum test results are shown. Plots show median with error bars indicating ± interquartile range.

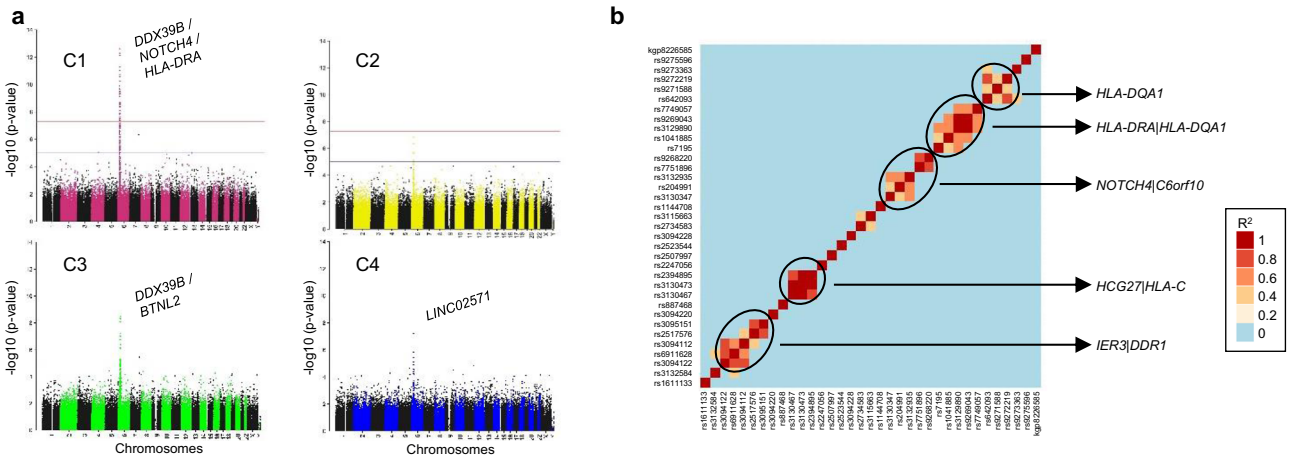


Fig. 4 Cluster genome-wide association analyses (GWAS). GWAS analysis was performed using Plink, an open-source whole genome association analysis toolset, using a logistical regression for 304 pSS (C1: 101, C2: 77, C3: 88 and C4: 38) patients and 330 healthy volunteers (HV) and each cluster was compared to HV. **a** Manhattan plots for each cluster are shown. **b** Linkage disequilibrium analysis in the loci of the 35 SNPs detected in C1 and located on chromosome 6 from base 29809362 to base 32681631. The R^2 correlation coefficient and linkage disequilibrium heatmap were obtained with Plink, and oncofunco R package, respectively. Strongest associations between SNPs are annotated.

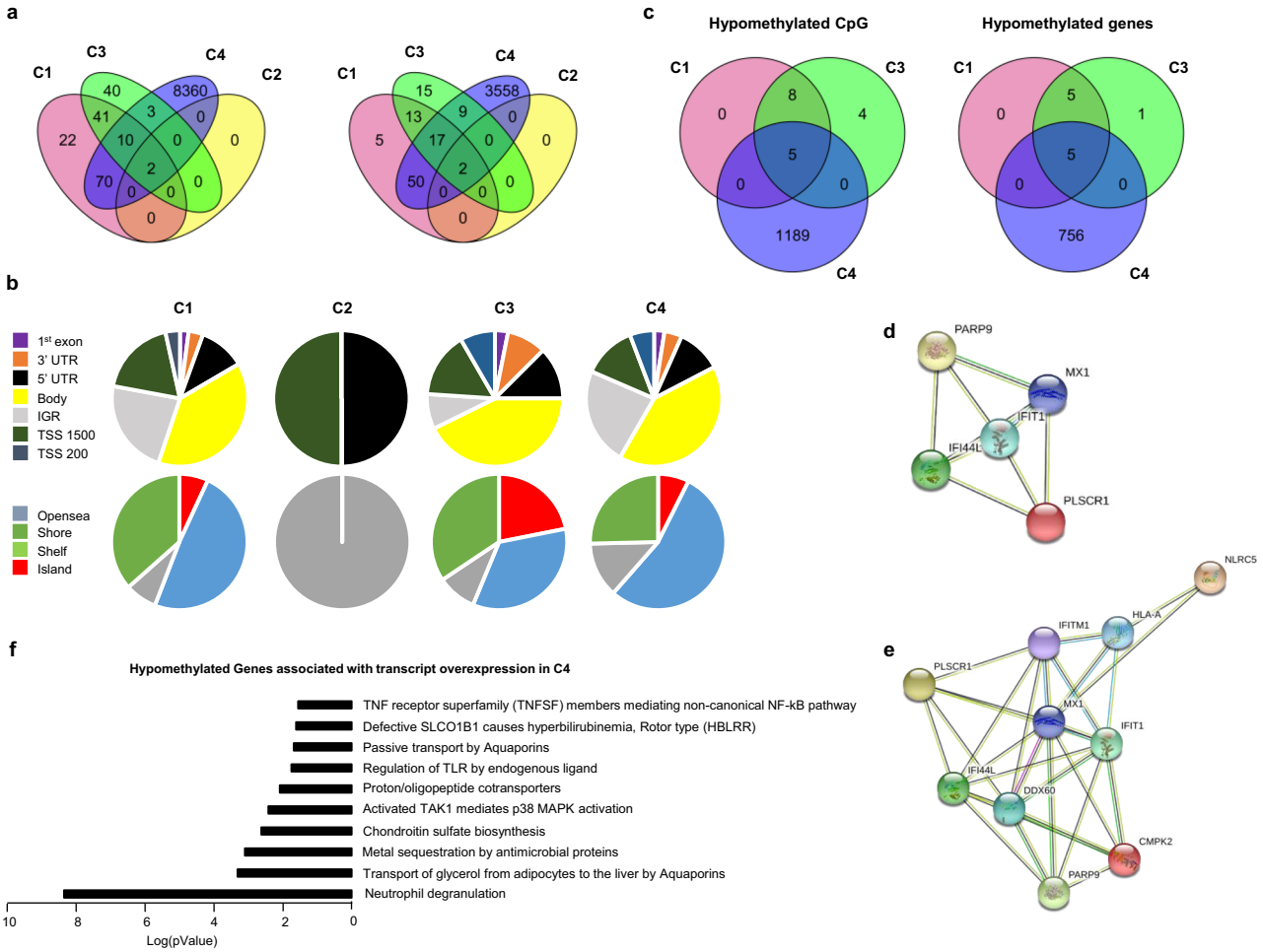


Fig. 5 Methylation analysis confirms the strong IFN signature in C1 and C3 and reveals an aberrant methylation status in C4. Whole blood methylation analysis was performed for 226 pSS patients (C1: 81, C2: 57, C3: 62, and C4: 26) and 175 healthy volunteers (HV) doing pairwise comparisons between each cluster and HV. **a** Venn diagram showing the overlap of differentially methylated CpG sites and genes between the 4 clusters with absolute ΔBeta > 0.075. **b** DMP distribution across the different genomic regions (gene body, 3'UTR, intergenic (IGR), 5'UTR, Exon 1, TSS 1500 and TSS 200; and according to the CpG density to CpG island, shelf, shore, and open sea. **c** Venn diagram showing the overlap of hypomethylated CpG and genes with absolute ΔBeta > 0.15 between the three IFN clusters. **d** Interaction network of these 5 genes common to the three clusters by STRING analysis with a confidence cut-off of 0.4 reveals a common IFN signature. **e** Interaction network of the 10 genes hypomethylated common to C1 and C3 by STRING analysis with a confidence cut-off of 0.4. **f** Reactome analysis[22] of the functional pathways enriched for the 126 genes hypomethylated and over expressed in C4 (absolute ΔBeta > 0.15, FC ≥ 1.5).

and 80.4% in C4). Because functionally important DNA methylation occurs in promoter regions and in CpG islands[20], DMP distribution across the different genomic regions was investigated (Fig. 5b). A higher representation of DMPs in the promoter region was found in C3 (36.4%) and C1 (33.1%) when compared to C4 (29.1%). The consequence was a lower representation of DMPs in intergenic regions for C3 (8.8%) compared to C1 (22.8%) and C4 (23.1%). To gain insight on this pattern, we divided the probes according to CpG islands; shores (regions up to 2 kb from CpG island), shelves (regions from 2 to 4 kb from CpG island) and open sea (the rest of the genome). Interestingly, 21.8% of the DMPs for C3 were located in CpG islands versus 6.9 and 7.4% for C1 and C4, respectively.

To identify the most robust and significant signature of hypo- and hyper-methylated genes, we fixed the ΔBeta cut-off at 0.15. Regarding hypomethylated CpGs, 13 DMPs were found in C1, 17 in C3 and 1,194 in C4, corresponding to 10, 11 and 761 hypomethylated genes, respectively. Five genes with hypomethylated DMPs were common to these 3 clusters (*IFI44L*, *IFIT1*, *MX1*, *PARP9* and *PLSCR1*) (Fig. 5c), corresponding to genes reported to present strong interactions (Fig. 5d). Interestingly, these genes were also significantly hypomethylated in C2 when compared to HV (Supplementary Fig. 6). Of note, 5 additional genes (*HLA-A*, *DDX60*, *CMPK2*, *IFITM1* and *NLRC5*) were common to C1 and C3 and were also strongly associated with the previous ones, reinforcing the IFN signature in these two clusters (Fig. 5e). These common 10 hypomethylated genes are implicated in defense responses to virus and are induced by IFN[21].

The remaining 756 hypomethylated genes in C4 were mainly associated with the neutrophil degranulation pathway. Regarding hypermethylated CpGs, 41 DMPs corresponding to 25 genes were only found in C4. Those genes are mainly implicated in translocation of ZAP-70 to the immunological synapse, phosphorylation of CD3 chains including zeta, platelet activation, signaling and aggregation, homeostasis and PD-1 signaling.

Combining transcriptomic (FC ≥ 1.5) and methylomic (absolute ΔBeta > 0.15) analyses, the transcripts of 8, 8 and 126 genes were found to be increased in association with a decreased methylation status in C1, C3 and C4, respectively. Interestingly, the previously isolated 5 common hypomethylated genes

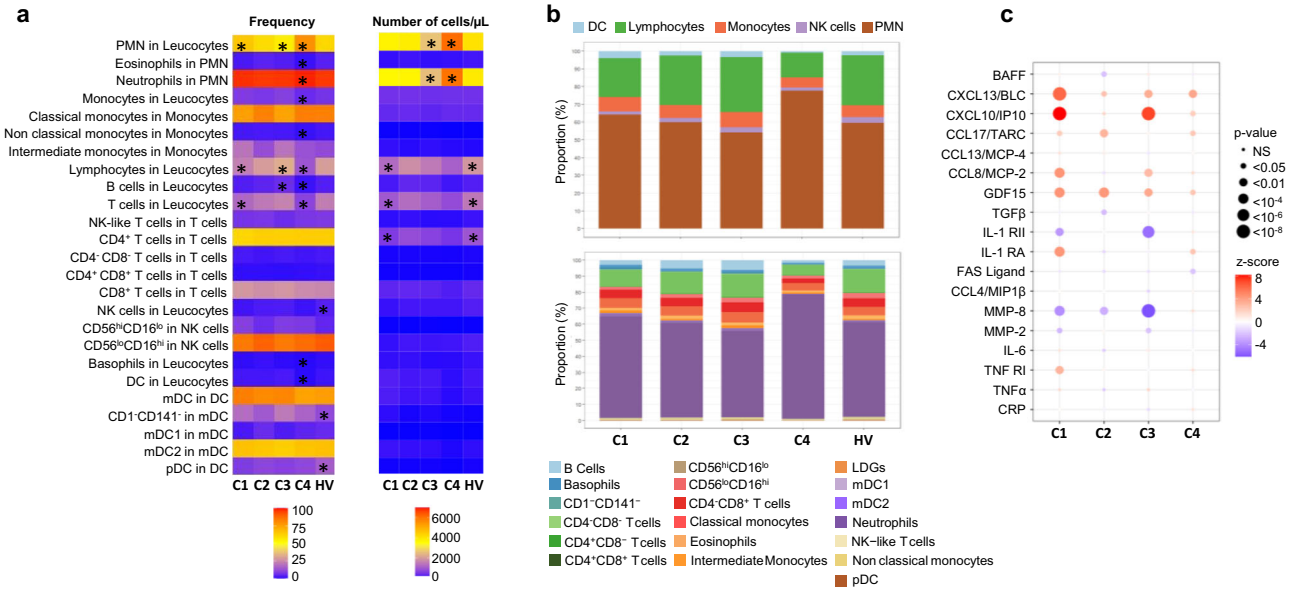


Fig. 6 Cell subset distribution in blood and cytokines, chemokines and inflammatory mediators in serum in the 4 clusters and healthy volunteers (HV). **a** Flow cytometry analysis was performed for 283 patients (C1: 96, C2: 71, C3: 80, and C4: 36) and 309 HV. The 2 heatmaps show the mean distribution of blood cell subsets in frequency (0–100%) and in absolute numbers (per µL of blood) across the 4 clusters and HV assessed by flow cytometry. Columns represent clusters and HV and rows the different cell subsets. The asterisk means that the cluster (or HV) is statistically different from all the others. **b** Flow cytometry data represented by bar charts cell types proportion per cluster. **c** Serum mediators were analyzed for 192 pSS patients (C1: 67, C2:48, C3: 61, C4:16) and 171 HV. Patient and HV distribution according to each analyzed variable is described in Methods. CXCL13/BLC, FAS Ligand, GDF-15, CXCL10/IP-10, CCL8/MCP-2, CCL13/MCP-4, CCL4/MIP-1β, MMP-8, CCL17/TARC, IL-1 RII, TNF-RI, and IL1-RA were measured using the Luminex system and expressed as pg/ml. Soluble MMP-2, CRP, TNFα, IL-6, BAFF, and TGFβ were measured by the quantitative sandwich enzyme immunoassay technique and expressed as pg/ml. Cytokine or chemokine concentration levels for each cluster were compared to HV. Statistical significance is determined using a one-way ANOVA followed by post-hoc Tukey's test. The significance between the cluster and HV is represented as bullet ranging from small (non-significant) to big (significant). The direction of the association is shown as the z-score where red bullet is up-regulated, and blue bullet is down-regulated.

implicated in IFN signaling were also overexpressed at the transcriptional levels in the 3 clusters. Transcript overexpression was strongly associated with hypomethylation in C1 (8/10) and C3 (8/11) and to a less extend in C4 (126/761). Among the 126 genes from C4, 21 were implicated in neutrophil degranulation which constitutes the most relevant pathways according to Reactome Pathway Database[22] (Fig. 5f). Only 6/25 transcripts were repressed in association with an increased methylation status of their genes in this cluster (*CD247*, *CD3G*, *CDC25B*, *CXCR6*, *TBC1D4*, *UBASH3A*).

Flow cytometry analysis. As significant alterations in patterns of peripheral blood leukocytes have been previously described[23,24], we then investigated the composition of leukocyte subsets in the various clusters. (Fig. 6a, b, Supplementary Fig. 7). In C2, the frequency and absolute numbers were similar to HV in all the different subsets analyzed. An increase in the frequency of monocytes and lymphocytes characterized C3, in association with a marked increase in the frequency of B cells. At the same time, a lymphopenia affecting mainly T cells was found in C1. Finally, the most distinguishable cluster in terms of distribution and absolute number of cells is C4. Specifically, C4 was characterized by higher percentages and absolute numbers of PMN (especially neutrophils) in peripheral blood in comparison with those in other clusters and HV. Conversely, the percentages of lymphocytes (B and T cells) and monocytes were markedly decreased in C4 compared to either the controls or the other clusters. Finally, lower frequencies and absolute numbers of basophils and DCs were also found in this cluster.

An in-depth analysis of the different cell subpopulations was then conducted. First, monocytes represent a heterogeneous cell population in terms of both phenotype and function. Based on the expression of CD14 and CD16, 3 monocyte subsets can be defined, including classical ($CD14^{++}CD16^{-}$), intermediate ($CD14^{++}CD16^{+}$) and non-classical ($CD14^{+}CD16^{++}$). Classical monocytes are critical for the initial inflammatory response, can differentiate into macrophages in tissue and contribute to chronic disease. Intermediate monocytes are highly phagocytic cells that produce high levels of ROS and inflammatory mediators. Non classical monocytes have been widely viewed as anti-inflammatory, as they maintain vascular homeostasis and constitute a first line of defense in recognition and clearance of pathogens[25]. Interestingly, the frequency and absolute number of intermediate monocytes were increased in C1 and C3 whereas the frequency of classical monocytes was decreased when compared to the 2 others and the nonclassical subset was markedly decreased in C4, in line with the inflammatory response observed in these different clusters.

Second, NK cells are defined by the expression of CD56 and the lack of CD3-TCR complex. Moreover, based on CD16 and CD56 expression levels, they are classified in two subsets: $CD56^{hi}CD16^{lo}$ and $CD56^{lo}CD16^{hi}$. The latter NK cell subset mediates natural and antibody-dependent cellular cytotoxicity, exhibiting high levels of perforin and enhanced killing. In contrast, $CD56^{hi}CD16^{lo}$ NK cells are characterized by low levels of perforin, and are primarily specialized for cytokine production including IFN[26,27]. Accordingly, the frequency of $CD56^{hi}CD16^{lo}$ NK cells subset over $CD56^{lo}CD16^{hi}$ was increased in C4, C1, C3 and to a lower extent in C2. This may partly explain the up-regulation of cytokines and interferon pathways in disease clusters. Although plasmacytoid dendritic cells (pDCs) are thought to represent the main IFNα producing cells, no differences were observed between clusters and their reduction was confirmed in peripheral blood of pSS patients when compared to HV[28].

Cytokine analysis. We subsequently assessed whether pSS clusters also showed differences in systemic parameters of inflammation, such as cytokines, chemokines and other soluble factors (Fig. 6c and Supplementary Fig. 8). The IFNγ-induced protein (CXCL10/IP-10) as well as CCL8/MCP-2 and TNFα were increased in C1 and C3, i.e. the two main clusters associated with a strong IFN signature. At the same time, IL-1 RII, the decoy receptor for cytokine belonging to the IL-1 family, was down regulated in C1 and C3. Overall, C1 was largely enriched in CXCL13/BLC, IL-6, and IL-1RA. Levels of MMP-8, a protease mainly expressed by neutrophils, were not different from HV in C4 but lower in the other clusters. Of note, many cytokines such as CXCL10/IP-10, CXCL13/BLC, BAFF, and GDF15 were increased in all clusters including C2 when compared to HV. However, no differences between clusters were found for CRP, Fas Ligand, CCL13/MCP-4, CCL4/MIP-1β, CCL17/TARC and TGFβ.

To confirm that patients with an active IFN signature have elevated circulating Type I IFN, we measured levels of IFNα in plasma using Simoa Single Molecule Array Technology in pSS patients and HV. Median levels of IFNα in plasma were 807 (177–1744) fg/ml and 530 (106–1033) fg/ml in C1 and C3, respectively, while circulating levels in the other clusters and HV were close to the lower limit of quantification (Supplementary Fig. 9a). Interestingly, IFNα in serum was positively correlated with the two IFN transcriptomic modules (M1.2 and IFNα module) described in Fig. 3, especially in C1 and to a lesser extent in C3, confirming the Type I IFN signature observed in these patients (Supplementary Fig. 9b). Of note, half of the patients in C2 received antimalarials and previous studies have also shown that hydroxychloroquine use can reduce the levels of circulating Type I[29,30] and Type II[31,32]; IFN z-scores. IFNα in serum was not associated with ESSDAI (Supplementary Fig. 9b) but higher levels of serum IFNα were associated with hematological and biological domains of ESSDAI (Supplementary Data 4).

Clinical symptoms and serological characteristics. Patient medical history and disease characteristics including clinical and serological parameters were collected for the 304 pSS patients. Details are displayed in Table 2 and Supplementary Data 5. Patients from C2 had a lower disease duration when compared to patients from other clusters.

Although the Physician Global Assessment (PGA) was collected for the whole population, ESSDAI and ESSPRI were only assessed in expert centers (Barcelona, Brest, Cordoba, Geneva, Hannover, Leuven, Milano, Porto and Szeged) in a subset of 193 and 150 respectively of the 304 pSS studied patients (70/101 and 56/101 from C1, 52/77 and 43/77 from C2, 44/88 and 30/88 from C3 and 27/38 and 21/38 from C4, Supplementary Data 5).

The lowest mean ESSDAI score was observed in C2 and the highest ESSDAI and PGA mean scores in C4 (Table 2, Fig. 7a) but there were no statistically significant differences between the 4 clusters. No clear difference in the ESSDAI components nor in the objective measures of ocular and salivary dryness was observed between the 4 clusters. Moreover, there was no significant difference for the global ESSPRI score and its 3 components (i.e. dryness, pain and fatigue) except between SSA-positive C2 patients who reported lower ESSPRI scores (p-value < 0.001) compared to the SSA-negative patients (Supplementary Data 6).

Statistically significant differences in the distribution of reported arthritis (p-value = 0.016), rate of cancer history (p-value = 0.028), coronary artery disease (p-value = 0.002) and chronic obstructive pulmonary disease (p-value = 0.016) were observed between the four clusters. (Supplementary Data 7). Interestingly, patients from C4 reported more severe clinical symptoms compared to the 3 other clusters.

Some serological characteristics were significantly different across the 4 clusters, hypergammaglobulinemia (p-value < 0.001) (Table 2), extractable nuclear antigen (ENA) antibodies (p-value <0.001), the presence of serum anti-SSA52/anti-SSA60 autoantibodies (p-value < 0.001) and higher circulating kappa and lambda free light chains (cFLC) (p-value < 0.001) (Fig. 7b, and Supplementary Data 8). C1 and C3 were associated with higher levels of these parameters when compared to C2 and C4. Moreover, C2 and C4 were enriched in patients with glandular manifestations of the disease assessed by a positive focus score in the absence of anti-SSA antibodies (Table 2).

In addition, the levels of rheumatoid factor (p-value < 0.001) and complement C4 fraction levels (p-value = 0.003) were statistically different between the four clusters. C1 was characterized by a higher rheumatoid factor and by a reduced complement C4 fraction levels compared to the other clusters. While some patients presented anti-dsDNA antibodies in C1 and C3 and anti-CCP antibodies in C4, almost none of these autoantibodies were present in the other clusters (Supplementary Data 8).

Prediction of patient membership to each of the four clusters. We then developed through machine learning approaches a composite model able to predict, according to a small number of variables, to which of the 4 clusters each patient belongs (see Methods). The proposed composite model was built with a 2-step approach to allocate patient to the right cluster (Supplementary Fig. 10). The final sets of selected features were composed of 10 genes for the C4 prediction model (first step) and 31 genes for the C1, C2, and C3 classification model (second step). The distribution among clusters of the variance stabilizing transformation (vst) normalized expression for all these transcripts is shown in Supplementary Fig. 11. The validation set (Fig. 1 and Table 1) was used for training, due to the heterogeneity of C4 pSS patients in this set, and the composite model was then run on the discovery set. The accuracy of the model was 95.15%, with 99.12% and 95.57%, for the first and the second steps respectively. The confusion matrix, the corresponding discriminant function analysis, and the probabilities to belong to one of the 4 clusters are shown in Fig. 8a, b, and Supplementary Data 9, respectively.

To generalize the composite model, we used an independent inception cohort of 37 pSS patients. After prediction, C1 contained 16 patients (43.2%), C2 6 patients (16.2%), C3 7 patients (18.9%) and C4 8 patients (21.6%). The corresponding discriminant function analysis and the probabilities for a patient to belong to one of the 4 clusters are shown in Fig. 8c and Supplementary Data 10, respectively. We then used the minimal list of 257 discriminative genes signature previously selected in Fig. 1a to generate a heat map with the prediction established by the composite model (Supplementary Fig. 12a). The clusters observed had the same profile than those identified in the discovery set and observed again in the validation set (Fig. 1a), confirming once more the clustering model. Furthermore, the predicted patients showed a distribution of the IFN signatures (Supplementary Fig. 12b) consistent with the one characterizing the identified clusters (Fig. 3). Altogether, these observations strengthen the validation of our composite model.

Finally, in order to allow our model to process other cohorts of patients, we implement an interpolation function based on 6 genes presenting a constant expression across all 4 clusters and HV (Supplementary Fig. 13). The composite model is integrated into an analysis tool available on the laboratory's github repository[33].

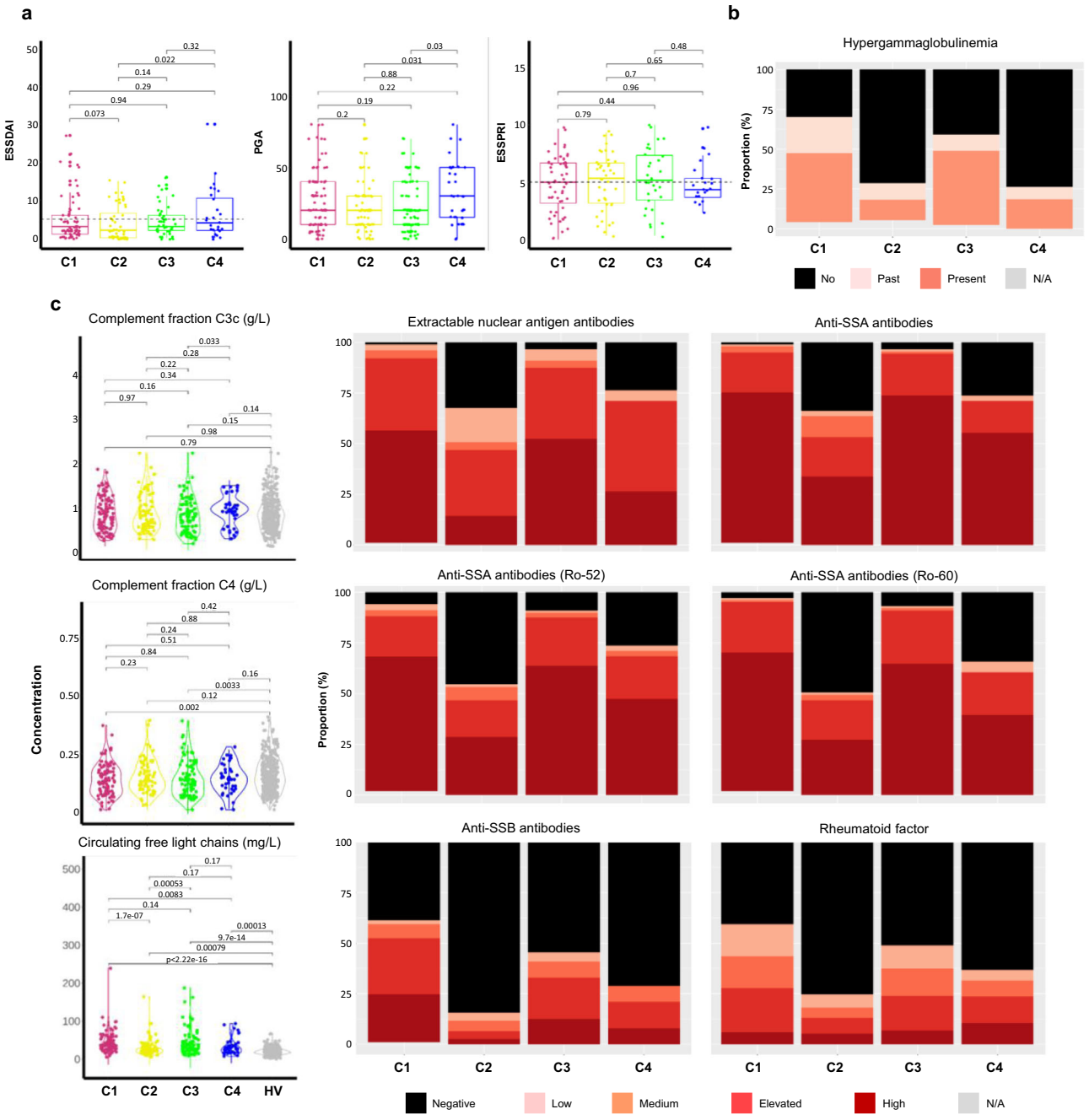


Fig. 7 Disease activity and serological distributions in the 4 clusters. a ESSDAI collected for 193 pSS patients (C1: 70, C2: 52, C3: 44, C4: 27), PGA collected for 286 pSS patients (C1: 94, C2: 71, C3: 85, C4: 36,) and ESSPRI collected for 150 pSS patients (C1: 56, C2: 43, C3: 30, C4: 21) distributions are shown in the 4 clusters. Two-tailed pairwise Wilcoxon-rank sum test results are shown. **b** The barplot shows the proportion of past (light orange) or present (orange) hypergammaglobulinemia (C1: 97, C2: 73, C3: 86, C4: 38) in each cluster. **c** Extractable nuclear antigen antibodies, anti-SSA antibodies, anti-SSA antibodies (Ro-52), anti-SSA antibodies (Ro-60), anti-SSB antibodies, rheumatoid factor were performed for 304 pSS patients (C1: 101, C2:77, C3:88, C4:38) and 330 HV and measured in serum, at the same center, using an automated chemiluminescent immunoanalyzer (IDS-iSYS). Barplots show the proportion of concentration level in each cluster (black: negative, light pink: low, orange: medium, red: elevated and dark red: high). Turbidimetry was used for rheumatoid factor (RF), complement fractions C3c and C4 determination and circulating free light chains. Statistical significance is determined by two-tailed pairwise Wilcoxon-rank sum test. Plots show median with error bars indicating ± interquartile range. Patient and HV distribution according to PGA and biological parameters analyzed variable is described in Methods.

Discussion

Over the last decade, numerous targeted immunomodulatory therapies for pSS have failed to show a benefit in clinical trials, hence no disease-modifying therapy has yet been approved for this disease[34–39]. The heterogeneous nature of pSS and its non-linear development, with flares of activity and subsequent remission associated to a very heterogeneous clinical presentation may explain clinical trial failures[40]. In this context, there is growing interest in the identification of well-characterized subgroups of patients, a prerequisite to the identification of molecular biomarkers predictive of treatment response[41].

We report herein on a large molecular profiling study carried out in pSS patients, a comprehensive molecular profiling of these patients irrespective of their clinical phenotypes. Previous studies

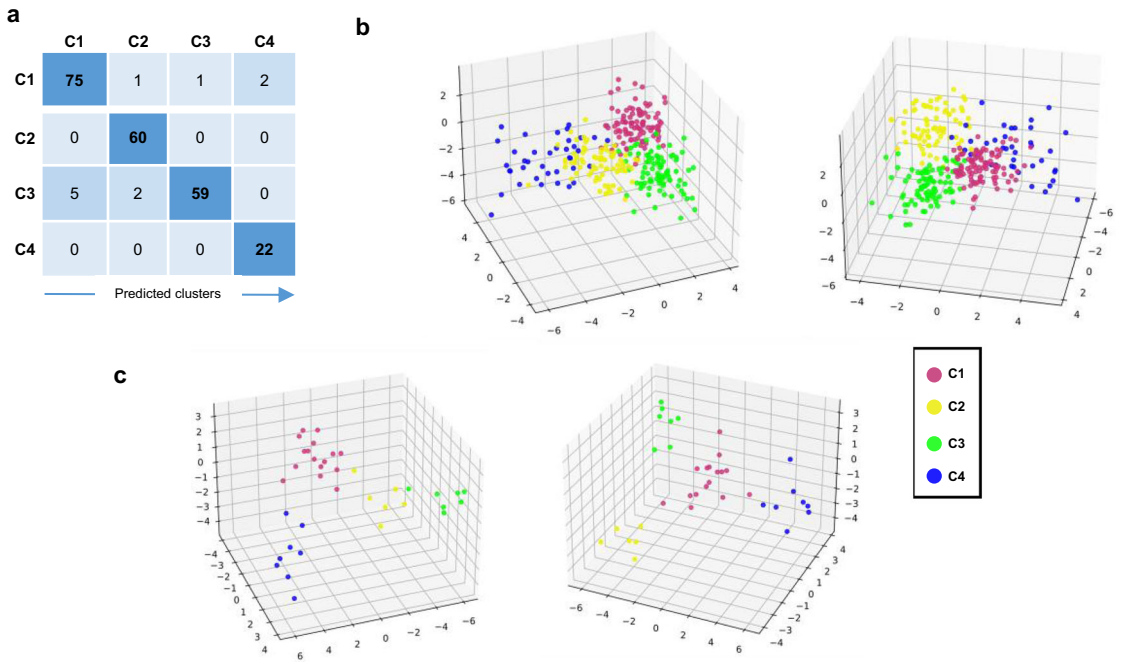


Fig. 8 Development of a composite model to predict the belonging of a patient to one of the 4 clusters. a Confusion matrix of the composite model in the discovery cohort performed for 227 pSS patients (C1: 79, C2: 60, C3: 66, and C4: 22) is shown. **b** Discriminant function analysis (DFA) of the predicted patients from the discovery cohort shows clearly separated clusters. Two different views of the same DFA are shown. **c** DFA of the predicted patients from the inception cohort shows clearly separated clusters. Two different views of the same DFA are shown. Thirty-seven pSS patients from the inception cohort were analyzed and predicted as C1: 16, C2: 6, C3: 7, and C4: 8.

in pSS focus particularly on the IFN signaling involvement[11]. Thereby, pSS patients could be stratified in interferon negative, Type I or Type I + II positive subgroups with higher prevalence of anti-SSA and anti-SSB among those with IFN activation without relation with systemic activity. Another group[42] performed a clustering analysis of blood gene expression microarray which classified the 47 pSS patients in three clusters characterized by IFN and inflammation with no discriminant clinical features Moreover, four subgroups of patients with similar patients' clinical characteristics were identified based on absolute cell counts per µL of blood[23]. Lastly, a stratification based on patient clinical phenotypes characterized a posteriori at the molecular level was proposed[43]. These works provide good basis for building a molecular taxonomy of pSS. Our integrative approach using multi-omics and patient clinical characteristics allows going further in understanding pSS heterogeneity.

We identified transcriptional modules allowing to separate pSS patients into four distinct clusters, irrespective of their treatment, reflecting specific patterns of immune dysregulation, with disease activity and patient reported symptom mean scores similar to naturalistic cohorts like ASSESS[44] and UKPSSR [45].

Patients from C2 displayed a healthy-like profile which nonetheless encompasses bona fide pSS patients reporting a similar level of objective symptoms of dryness, pain and fatigue, albeit a lower ESSDAI compared to the 3 other clusters. C2 was also enriched in patients with glandular manifestations of the disease assessed by a positive focus score and no anti-SSA antibodies. A similar cluster was recently described[42] with no increase in the IFN modules and minimal activity of inflammation-related gene modules. Noteworthy, all molecular profiling data reported here were obtained from blood samples which could affect interpretation of some of the results. For example, the reduction of peripheral blood pDCs of pSS patients when compared to HV already reported[28] does not consider that pDC are enriched in the salivary glands and the possibility that tissue sites may be the major source of IFNα in these individuals[46]. Extending in the future those analyses to the salivary gland will provide a more complete picture of the pathophysiology of the disease, especially in C2.

The three other clusters exhibited significant differences with HV and in particular a prominent IFN gene signature. These findings add to the growing evidence towards a significant role of the IFN pathways in the pathogenesis of systemic and organ-specific disorders including pSS. Whereas Type I IFN were proposed as predominant contributors in the pathogenesis of pSS, a role of Type II IFN in disease pathogenesis has also been highlighted[6,47]. Interestingly, our results show that the IFN signature in the 3 IFN-driven clusters is different. C1 patients had the highest Type I and Type II IFN scores, C3 a higher Type I IFN score than C4, these 2 clusters having similar Type II IFN score. Thus, C4 IFN score was mainly driven by IFN Type II activation. Consequently, C1 and C3 were similar to the IFN cluster recently described by James et al.[42] also associated with high blood protein levels of CXCL10/IP-10.

In line with observed IFN scores, circulating serum levels of IFNα were positively correlated with Type I IFN signature (Supplementary Fig. 9 and Fig. 3) especially in C1 and to a lesser extent in C3. However, levels of IFNα in serum were not correlated with ESSDAI global score, but higher levels of serum IFNα were associated with hematological and biological domains of ESSDAI.

While C1 was mainly driven by IFN, an increase in frequency of B lymphocytes in the blood associated with a significant activation of canonical pathways related to B cell activation such as B cell receptor signaling, and B cell development were observed in C3. Main biological features associated with C3 but also C1 were hypergammaglobulinemia, anti-nuclear antibodies, the presence of serum anti-SSA52/anti-SSA60 autoantibodies and higher cFLC confirming what was already reported in autoantibody-positive pSS patients[21]. Finally, SNPs associated with HLA class II genes

were mainly reported in patients from C1 and C3 presenting a positive IFN signature and high levels of autoantibodies as already shown in SLE[16].

Patients from C4 exhibited a more severe clinical phenotype compared to the others with an inflammatory transcriptomic signature particularly linked to cytokine signaling from the acute phase response. C4 was also characterized by a massive lymphopenia and high levels of neutrophils. The neutrophil-to-lymphocyte ratio (NLR) has been previously shown to correlate with disease activity in systemic autoimmunity[48,49] and elevated NLR are thought to represent a pro-inflammatory state. Indeed, in a study of 483 adult patients with multiple sclerosis, NLR could differentiate between relapsing-remitting and primary progressive multiple sclerosis and predict worsening disability[50]. Further studies are required in pSS to evaluate the importance of this ratio.

Because the main current challenge in clinical trials of new therapies for pSS is the selection of the appropriate patients, we propose here a combination of molecular parameters allowing patient classification by endotypes (Supplementary Fig. 14). We then developed a composite model derived from machine learning, based on the use of a limited number of transcripts from whole blood RNASeq and validated in an independent data set from a pSS inception study, to allow a reanalysis of the previous and ongoing clinical trials to depict predictors of treatment response.

These findings have major implications for the treatment of pSS patients, providing a rationale for both optimal drug positioning and combinations of drugs with complementary mechanisms of action. Specifically, our findings provide a strong rationale for treating patients with either a C1, C3, or C4 profile with inhibitors of type I IFN responses alone or in combination as they support the relevance of B cells as potential therapeutic targets in C3 patients. Trials with B cell depleting antibodies (rituximab) have shown promising results primarily in reducing systemic activity in pSS[51].

Areas requiring further investigation have been identified. First, although our identified cluster gene signatures are strong enough to overcome the disequilibrium in blood cell counts and are not associated with disease duration, except for C2, RNA-Seq analysis is oblivious to sample cell-type composition[52]. Further analyses are on-going, using deconvolution approaches. Second, as hypotheses were derived from a cross-sectional study and a small inception cohort, findings need to be confirmed in longitudinal cohorts to clarify whether patients will stay longitudinally in their initial cluster whatever the disease activity level and the treatments received, or whether treatments decrease disease activity by modifying the extent and scope of gene signaling dysregulations.

Altogether, our results can improve pSS treatment strategies allowing a patient centric approach. This paradigm already implemented in the oncology field will increase the probability of trial successes and boost the development of new efficient drugs against pSS.

Methods

Computational tools. Except when indicated, data analyses were carried out using either an assortment of R system software (http://www.R-project.org, V2.10.1) packages including those of Bioconductor or original R code. R packages are indicated when appropriate. For GWAS analysis, we used Plink, an open-source whole genome association analysis toolset. Machine learning approaches were carried out using python programs (v3.8.5) based on the following modules: scikit-learn, numpy and xgboost.

Patient population. The present study was conducted in patients with pSS and HV included in the European multi-center cross-sectional study of the PRECISESADS IMI consortium which involved patients from seven systemic autoimmune diseases. This study was a pre-planned substudy to be specifically conducted in the pSS population and fulfill the STROBE statements (Supplementary note). Diagnosis of pSS was made according to the 2002 American-European Consensus Group classification criteria, with at least the presence of anti-SSA and/or a positive focus on a minor salivary gland biopsy. Choice of the patient analysis set is detailed in Supplementary Fig. 1a. Recruitment was performed between December 2014 and October 2017 involving 19 institutions in 9 countries (Austria, Belgium, France, Germany, Hungary, Italy, Portugal, Spain and Switzerland). The composite model was validated using transcriptome of 37 pSS newly diagnosed patients recruited in the inception study also obtained from the PRECISESADS consortium. Inception patients were recruited by 10 institutions in Spain, Belgium, France, Italy, Germany and Switzerland. Eligible patients were diagnosed within less than a year since pSS diagnosis.

The two studies (cross-sectional and inception) adhered to the standards set by International Conference on Harmonization and Good Clinical Practice (ICH-GCP), and to the ethical principles that have their origin in the Declaration of Helsinki (2013). Each patient signed an informed consent prior to study inclusion. The Ethical Review Boards of the 19 participating institutions approved the protocol of the cross-sectional study. Moreover, the protocol of the inception study was approved by the ethical committees of the 10 participating institutions. These 10 sites were also participating to the cross-sectional study, therefore these ethical committees reviewed both protocols. The ethical committees involved were: Comitato Etico Milano, Italy; Comité de Protection des Personnes Ouest VI Brest, France; Louvain, Comité d'Ethique Hospitalo-Facultaire, Belgium; Comissao de ética para a Saude—CES do CHP Porto, Portugal; Comité Ética de Investigación Clínica del Hospital Clínic de Barcelona, Spain; Commissie Medische Ethiek UZ KU Leuven/Onderzoek, Belgium; Geschaftsstelle Ethikkommission, Cologne, Germany; Ethikkommission Hannover, Germany; Ethik Kommission. Borschkegasse, Vienna, Austria; Comité de Ética e la Investigación de Centro de Granada, Spain; Commission Cantonale d'éthique de la recherche Hopitaux universitaires de Genève, Switzerland; Csongrad Megyei Kormanyhivatal, Szeged, Hungary; Ethikkommission, Berlin, Germany; Andalusian Public Health System Biobank, Granada, Spain.

The protection of the confidentiality of records that could identify the included subjects is ensured as defined by the EU Directive 2001/20/EC and the applicable national and international requirements relating to data protection in each participating country. The cross-sectional and inception studies are registered in ClinicalTrials.com with respectively number NCT02890121 and number NCT02890134.

For each individual, blood samples as well as biological and clinical information were collected as described in the next Methods sections. For more technical details on sample and data collection, please refer to the main PRECISESADS paper [5].

After quality control on transcriptomics RNAseq data (described below), verification of the ARC/EULAR classification criteria (focus score ≥ 1 foci/mm² and anti-SSA/Ro antibody positivity), and match of the HV to the patients based on age and gender, our final study cohort comprises 304 patients with pSS and 330 HV. This selection is detailed in Supplementary Fig. 1. Among the 304 pSS, 227 (75%) were used for the discovery step and 77 (25%) were kept for validation (Table 1).

Available data. High-dimensional omics genotype, transcriptome, DNA methylome and proportions of relevant cell types using flow cytometry custom marker panels were analyzed from whole blood samples. Low dimensional information was obtained from serum samples, including selected serology information such as autoantibodies, cytokines, chemokines and inflammatory mediators. Of note, except for samples collected for flow cytometry analysis, all samples were shipped by the clinical sites to a Central Biobank (Granada) for processing, storage, and onward shipment to the analysis sites, where the various determinations were performed. Flow cytometry was managed at each center on fresh blood after a multi-center harmonization of flow cytometers to ensure mirroring of all instruments[53,54], thereby allowing subsequent integration of all the data obtained across the different sites and instruments. Consequently, all the different omics samples were processed with the same protocols at the same site (RNA-Seq at Bayer, cytokines at UNIMI, autoantibodies and integrated analyses of flow cytometry at UBO, methylome at IDIBELL, GWAS at CSIC which guarantees the high quality of the data generated.

Methods used for RNA sequencing, quality control, data processing, and expression profiling are detailed below and in Supplementary Fig. 1c.

RNA-Seq. Methods used for RNA sequencing, quality control, data processing, and expression profiling are detailed below and in Supplementary Fig. 1c. Total RNA was extracted from whole blood samples collected in Tempus tubes using Tempus Spin technology (Applied Biosystems). 1857 samples were processed in batches of 384, randomized to four 96-well plates with respect to patient diagnosis, recruitment center and RNA extraction date. The samples were depleted in alpha- and beta-globin mRNAs using globinCLEAR protocol (Ambion) and 1 µg of total RNA was used as input. Subsequently, 400 ng of globin-depleted total RNA was used for library synthesis with TruSeq Stranded mRNA HT kit (Illumina). The libraries were quantified using qPCR with PerfeCTa NGS kit (Quanta Biosciences), and equimolar amounts of samples from the same 96-well plate were pooled. Four

pools were clustered on a high output flow cell (two lanes per pool) using HiSeq SR Cluster kit v4 and the cBot instrument (Illumina). Subsequently, 50 cycles of single-read sequencing were performed on a HiSeq2500 instrument using and HiSeq SBS kit v4 (Illumina). The clustering and sequencing steps were repeated for a total of three runs in order to generate sufficient number of reads per sample. The raw sequencing data for each run were preprocessed using bcl2fastq software and the quality was assessed using FastQC tools. Cutadapt[55] was used to remove 3′ end nucleotides below 20 Phred quality score and extraneous adapters, additionally reads below 25 nucleotides after trimming were discarded. Reads were then processed and aligned to the UCSC Homo sapiens reference genome (Build hg19) using STAR v2.5.2b[56]. 2-pass mapping with default alignment parameters were used. To produce the quantification data, we used RSEM v1.2.31[57] resulting in gene level expression estimates (Transcripts Per Million, TPM and read counts).

For sample filtering, samples were filtered in at least one of the following situations: (i) the total sum of count is too low (<5000,000), (ii) they were extracted with another method than Tempus Spin, and (iii) the RIN (RNA Integrated Number) value of the sample is below 6.5, (iv) samples with RNAseq inferred gender inconsistent with clinical data, and (v) there was a disagreement between genotypes inferred from RNA-Seq and those obtained from GWAS genotyping.

For normalizations and batch correction, read counts were normalized by the variance stabilizing transformation vst function from DESeq2 (v1.30.0) R package[58]. To reduce the effect of the RIN, a correction was applied using the ComBat function from sva (v3.38.0) R package[59], after categorization of RIN values into 7 classes: (7.5, 8], (8.5, 9], (9.5, 10], (8, 8.5], (7, 7.5], (9, 9.5], (6.5, 7].

For Gene filtering, among the 55,771 genes detected in the data, those with 0 count over all the samples or having an expression level below 1 in more than 95% were filtered. At the end, our final RNA-Seq data comprises 16,876 genes. This selection is detailed in Supplementary Fig. 1.

Molecular subgroups discovery. Our rational was to produce a robust classification scheme and to ensure the greatest possible homogeneity within identified subgroups. To this aim, subgroup discovery was based on the pre-processed RNA-seq data of the discovery set (after vst transformation). We implemented a strategy already applied in breast cancer that iterates unsupervised and supervised steps, which was, therefore, designated as "semi-supervised" approach[8]. It is described hereafter and summarized in Supplementary Fig. 2.

Step 1: Unsupervised gene selection

The coefficient of variation ($CV_g = \frac{\sigma_g}{\mu_g}$, with σ_g is the standard deviation of the gene g and μ_g the mean of the gene g estimated on discovery population) and its robust version ($rCV_g = \frac{\gamma_g}{\mu_g}$, with γ_g is the median absolute deviation) were calculated for each gene. Both were highly concordant. The top 25% most variants were selected to perform the subsequent clustering analysis.

Step 2: Robust consensus clustering

To determine the number of clusters, a consensus clustering between three methods was performed: (i) Agglomerative Hierarchical Clustering (hclust function from stats v4.0.2 R package) with Pearson correlation as a similarity measure and the Ward's linkage method, (ii) K-means clustering (kmeans function from stats R package) with 4 groups and (iii) Gaussian mixture clustering (mclust function from mclust v5.4.6 R package).

Step 3: Identification of molecular signature

A supervised analysis was performed on the 149 patients with consistent cluster assignments between the three clustering methods (considered as "core" molecular profils), in order to identify the most discriminating signature of the 4 clusters. The first signature of set of 3577 genes was selected from a classical one-way ANOVA (FDR < 1e-10), and then reduced by Random Forest to 257 top discriminating genes (randomForest function from randomForest v4.6-14 R package[60]).

Step 4: Robustness classification

To validate the robustness of our clustering, we re-applied Step 2 on our discovery set and on the final signature.

Step 5: Classification of discordant patients

Patients assigned to different groups with the 3 clustering methods were assigned to one of the 4 clusters by applying a distance-to-centroid method based on Pearson correlation.

Molecular subgroup validation. Validation datasets were independently classified in the pSS molecular subgroups by applying a classical distance-to-centroid approach based on correlation. Following the same approach, HV did not constitute a separate cluster but mainly matched with C2 (0.5% in C1, 93% in C2, 4% in C3, and 2.5% in C4) pSS molecular subgroups by applying a classical distance-to-centroid approach based on correlation. The final clustering (without HV) is represented with heatmap using the Heatmap function from ComplexHeatmap (v2.6.2) R package. Clusters are separately constrained for better visualization. This method allows to spotlight heterogeneous intra-clusters. The principal component analysis (PCA) representation will explore the clearly defined clusters and the matching between C2 and HV.

Half of the pSS patients was treated with either anti-malarial, immunosuppressant, or steroids at the time of the visit. When compared to the 3 other clusters, we observed higher proportion of treated patients in C4. To investigate the impact of the treatment on the clustering, we compared treated patients and untreated patients. For this, we apply a hierarchical clustering on treated patients and untreated patients and compare the cluster distribution. The heatmap (Supplementary Fig. 4) of treated vs untreated patients were highly similar which shows that the final clustering is not driven by treatments.

Enrichment analysis. Enrichment analysis was performed by applying a two-tailed Fisher-exact test[61] against different sources of gene modules or pathways: (i) 3 strongly upregulated IFN-annotated modules from[10] (M1.2, M3.4, and M5.12) determined from peripheral blood transcriptomic data with for each a distinct activation threshold, (ii) genes preferentially induced by IFNα or IFNγ identified by[10], (iii) canonical pathway from Ingenuity Pathway Analysis (IPA, Release Date: 2020-06-01), (iv) repertoire recently established on an expended range of disease and pathological states (382 transcriptome modules based on genes co-expression patterns across 16 diseases and 985 unique transcriptome profiles) by [9].

Differential gene expression analysis. To identify genes differentially expressed between pSS subgroups and HV, we performed a linear model (lmFit function from limma v3.46.0 R package[62]) on vst transformation gene expression dataset. Resulting p-values were adjusted for multiple hypothesis testing and filtered to retain DE genes with FDR adjusted p-value ≤ 0.05 and a |Fold-Change (FC)| ≥ 1.5.

Genome-wide association study. Genome-wide association studies (GWAS) were performed for each pSS subgroups (C1: 101, C2: 77, C3: 88, and C4: 38) versus 330 HV. After DNA extraction, the samples were genotyped using HumanCore-24 v1.0 and Infinium CoreExome-24 v1.2 genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA). Individuals were excluded on the basis of incorrect gender assignment, high missingness (>10%), non- European ancestry (<55% using Frappe15 and REAP), and high relatedness (PLINK v1.9[45], pi_hat >0.5)[63]. Genotypes were filtered before imputation due to high missingness (>2%), Hardy–Weinberg equilibrium (HWE) < 0.001, minor allele frequency (MAF) <1%, and AT/CG changes with MAF >40%. PLINK v1.9[45] was used to carry out quality control (QC) measures, genotype data filtering. The basic association for a cluster trait locus, based on comparing allele frequency between patients from each cluster vs HV, was also obtained with this toolset thanks to computational resources from the Roscoff Bioinformatics platform ABiMS. Genotypes were phased using Eagle v2.3 and imputed using Minimac3 against the HRC v1.1 Genomes reference panel from the Michigan Imputation Server platform. Genotypes were filtered after imputation to have HWE p-value > 0.001, MAF > 1% and imputation info score > 0.7 and resulted in 6,664,685 imputed genotypes. Statistical analysis of association for each cluster versus HV was performed by logistic regression under the additive allelic model. The GWAS significant level was fixed at p-value < 5×10^{-8}. SNP annotations and Manhattan plot were obtained using the web-based tool SNP snap from the Broad Institute[64] and qqman (v0.1.8)[65] R packages respectively.

Methylation. Whole blood methylation analysis was performed for 226 pSS patients (C1: 81, C2: 57, C3: 62, and C4: 26) and 175 healthy volunteers (HV). DNA was extracted using a magnetic-bead nucleic acid isolation protocol (Chemagic DNA Blood Kit special, CHEMAGEN) automated with chemagic Magnetic Separation Module I (PerkinElmer) from K2EDTA blood tube (lavender cap, BD Vacutainer) of 10 ml (extractions were performed on 3 ml). 2 µg of DNA were sent for DNA methylation assay. The samples were analyzed using Infinium Human Methylation 450 K BeadChip (Illumina, Inc., San Diego, CA, USA) which covers more than 400,000 CpG sites. DNA samples were bisulfite-converted using the EZ DNA methylation kit (Zymo Research, Orange, CA, USA). After bisulfite conversion, the remaining assay steps were performed following the specifications recommended by the manufacturer. The array was hybridized using a temperature gradient program, and arrays were imaged using a BeadArray Reader (Illumina Inc., San Diego, CA, USA). Sample QC and functional normalization were completed using minfi (v3.3) R package[66]. Briefly, during QC steps, subjects were removed based on outliers for methylated vs unmethylated signals, deviation from mean values at control probes, and high proportion of undetected probes (using minfi default parameters). DNA methylation probes that overlapped with SNPs (dbSNPs v147), located in sexual chromosomes or considered cross-reactive were removed. Additionally, only probes quality controlled and shared between both arrays were used in the subsequent analysis (368,607 probes). Measure of methylation level (B values) were produced for each CpG probe and ranged from 0 (0% molecules methylated at a particular sites) to 1 (100% molecules methylated).

To identify differentially methylated positions (DMPs) between HV and each pSS subgroups (C1 to C4), the champ.DMP function of ChAMP (v2.18.3) R package[67] was implemented doing pairwise comparison between each cluster and HV. Many Δ-beta thresholds were described in the literature and the most frequently used for whole blood studies in autoimmune diseases were 0.05 (5% difference) and 0.1 (10% difference). In order to fix the best threshold for our study, we tested the values of 0.05, 0.075, 0.1, and 0.15 for the absolute ΔBeta. Supplementary Data 11 presents the numbers of DMPs and genes obtained with these different thresholds.

Then, we decided to analyze the data in two steps: the first step with a significant adjusted p-value (Benjamini Hochberg) at 0.1 and an absolute ΔBeta > 0.075. We assumed that a threshold of 0.05 was too low and it would have been

very difficult to interpret the signification of these defects in methylation for C4. If we had applied a ΔBeta threshold of 0.1 in the first intention, we could have missed DMPs. In the second step in order to identify the most robust and significant signature of hypo and hyper methylated genes, a significant adjusted *p*-value (Benjamini Hochberg) at 0.1 and an absolute ΔBeta > 0.15 were applied.

For network viewing, we tested gene lists onto the STRING 9.1 Network of Known and Predicted Protein–Protein Interactions (http://string-db.org/)[68].

Flow cytometry. Multi-parameter flow cytometry analyses have been performed in eleven different centers from the PRECISESADS consortium. Therefore, the integration of all data in common bioinformatical and biostatistical investigations has required a fine mirroring of all instruments[54]. The calibration procedure elaborated to achieve this prerequisite and the antibody panels used have been previously described[53].

The antibody panels, specificities, and clones used are shown in Supplementary Fig. 15a.

The strategy developed to avoid any redundancy in the different cell subsets and to increase the accuracy of the phenotypes has been automated by AltraBio (Lyon, France). The generated automatons have been validated in a preliminary study on 300 patients comparing data from automated gating to data manually gated by the same operator (coefficient of correlation 0.9996). The gating strategy was as follows: after exclusion of debris, dead cells and doublets, frequencies and absolute numbers of $CD15^{hi}CD16^{hi}$ neutrophils, $CD15^{hi}CD16^{+}$ eosinophils, $CD14^{+}CD15^{hi}$ LDGs, $CD14^{++}CD16^{-}$ classical monocytes, $CD14^{++}CD16^{+}$ intermediate monocytes, $CD14^{+}CD16^{++}$ non classical monocytes, $CD3^{+}$ T cells (with $CD4^{+}CD8^{-}$, $CD4^{+}CD8^{+}$, $CD4^{-}CD8^{+}$, $CD4^{-}CD8^{-}$ T cell subsets), $CD19^{+}$B cells, $CD3^{-}CD56^{+}$ NK cells (with $CD16^{lo}CD56^{hi}$ and $CD16^{hi}CD56^{lo}$ NK cell subsets), $CD3^{+}CD56^{+}$ NK-like cells, Lin-$HLA\text{-}DR^{+}$ DCs (with $CD11c^{-}CD123^{+}$ pDCs, $CD11c^{+}CD123^{-}$ mDCs (with $CD141^{-}CD1c^{+}$ mDC1, $CD141^{+}CD11c^{-}$ mDC2 and $CD141^{-}CD1c^{-}$ mDC subsets)) and $CD123^{+}HLA\text{-}DR^{-}$ basophils were automatically extracted from FCS and LMD files of 283 patients and 309 HV and sent in an Excel flow cytometry workflow. The mean distribution of blood cell subsets in frequency (0–100%) and absolute numbers by clusters are compared using a Kruskal–Wallis test.

Gating strategies of the automatons are shown in Supplementary Fig. 15b. For all instruments, the data from the flow cytometry files are analyzed with a similar strategy by one automaton for panel 1 and another automaton for panel 2, and then specifically for each instrument from the gate [S4] to account for the variability of FSC and SSC signals. The desired cell populations are identified by gating strategies identical for all instruments for panel 1 and panel 2 stainings. The mean distribution of blood cell subsets in frequency and absolute numbers are shown in Supplementary Data 12 and 13, respectively.

Cytokines. Cytokines were measured on serum samples. CXCL13/BLC, FAS Ligand, GDF15, CXCL10/IP-10, CCL8/MCP-2, CCL13/MCP-4, CCL4/MIP-1β, MMP-8, CCL17/TARC, IL-1 RII, TNF RI, and IL1-Ra were measured using the Luminex system. The 12-analyte customized panel was built using human premixed multi-analyte Luminex assay (R&D Systems). Samples were thawed on the day of analysis and tested in batches. Soluble MMP-2, CRP, TNFα, IL-6, BAFF, and TGFβ were measured using ELISA assay. Descriptive statistics are shown in Supplementary Data 14. We measured levels of IFNα in plasma using Simoa Single Molecule Array Technology. Results were calculated referring to a standard curve created using a four parameters logistic curve fit and were expressed as pg/ml. For more technical details on sample and data collection, please refer to the main PRECISESADS study[5]. The differential cytokine concentration between subgroups vs HV was performed using a one-way ANOVA followed by post-hoc Tukey's test (function ghlt from multicomp multcomp v1.4-13 R package[69]). The z-score indicate the direction of the concentration between the cluster and the HV. A z-score > 0 means that the cluster has an overexpression compare to HV. A z-score < 0 means that the cluster has a lower expression compare to HV (Fig. 6). Concentration distribution by subgroup is represented in Supplementary Fig. 8. Two-tailed pairwise Wilcoxon-rank sum tests have been computed.

Autoantibodies. Autoantibodies (Extractable nuclear antigen antibodies, anti-SSA antibodies, anti-SSA antibodies (Ro-52), anti-SSA antibodies (Ro-60), Anti-SSB antibodies), were measured in serum using an automated chemiluminescent immunoanalyzer (IDS-iSYS). After processing, the final result is indicative of the concentration of the specific autoantibody present in the sample. Rheumatoid factor (RF), complement C3c, C4, and individualized (kappa, lambda) free light chains (Combilite and freelight, respectively) were measured in serum using a turbidimetric immunoassay method according to manufacturer's recommendations (SPAPLUS analyser). For more technical details on sample and data collection, please refer to the main PRECISESADS study[5]. Autoantibodies and RF distribution have been described by concentration level (Negative/Low/Medium/Elevated/High) and a Fisher's exact test was applied to compare the proportion and the concentration across the 4 clusters. Complements C3 and C4 and circulating free light chains have been described in continued concentration expressed in g/L and mg/L respectively and a Krukal–Wallis test was applied to compare the concentration level across the 4 clusters. Descriptive statistics are described in Supplementary Data 8.

Clinical data. Clinical data on 304 patients with pSS and 330 HV describing the disease phenotype was collected using an electronic case report form (eCRF). A working group of experts on systemic autoimmune diseases was established and the desired items were selected via a Delphi technique. A final set of items was created, digitalized and pilot tested divided into 8 domains (constitutional symptoms, gastrointestinal, vascular, heart and lung, nervous system, skin and glands, musculoskeletal, therapy). After the confirmation of patient inclusion, clinical data were collected including patient's age, sex, ethnicity, dates of first disease manifestation (disease onset), clinical and biological characteristics at baseline, the physician global assessment of disease activity, comorbidity, and current use of treatments.

Another working group of pSS pathology experts was established to select pSS disease-specific items, mainly pSS disease activity scales like ESSDAI and its components, and ESSPRI and its components. These items were collected on a pSS sub-population ($n = 193$).

To characterize pSS subgroups, association test was performed with clinical data. A two-tailed Fisher's exact test (fisher.test function from stats R package) or chi-square test (chisq.test function from stats R package) as appropriate was applied to evaluate the association between the pSS supbgroups and a qualitative clinical factor. A Kruskal–Wallis test (Kruskal.Wallis function from stats R package) was used to evaluate the association between pSS subgroups and quantitative clinical variables.

Development of the composite model for cluster prediction. This feature selection process is composed of two distinct parts: (i) identify a subset of genes potentially interesting to predict the 4 clusters, (ii) use these previously identified subsets to actually craft a prediction model and extract the features used by the model to increase its precision. In the first part, with FC ≥1.5 and FDR ≤0.05, we selected the DEGs according to the following 7 combinations: C2 vs C1, C3 vs C1, C4 vs C1, C4 vs C2, C3 vs C2, C4 vs C2, C4 vs C3. We identified 14,240 and selected those common to all combinations representing 1154 DEGs.

We used the Boruta algorithm[70] on all dataset (discovery and validation sets) to extract features that significantly contributed to predict the patient's cluster.

The algorithm started to extend the dataset by adding copies of each feature in the original dataset. These features were called "shadow features" and consisted in random permutation of the modality of the original feature, in order to remove any correlation with the target variable, in our case, the cluster assignment. Once shadow features were crafted, a random forest classifier was run on the whole dataset and z-scores were computed for all features (real and "shadow"). Shadow features were then sorted according to their z-score and the maximum score was kept in memory as a threshold. The algorithm assigned a hit to each real feature that had a z-score above this threshold. Finally, Boruta marked the features which had a z-score significantly lower than the shadow with maximum z-score as "unimportant" and removed them from the dataset, before removing all shadow features and returning a clean dataset.

This process allowed us to identify variables in the dataset that were significantly more contributing to the classification problems than noisy variables and random artefacts emulated by the original variable modality permutation, ensuring the use of robust features for the second step of our feature selection strategy.

The relatively small size and heterogeneity of C4 in comparison to the other clusters can impact the feature selection process, therefore we chose to solve two classification problems: (i) identify C4 versus all clusters, (ii) discriminate between C1, C2, and C3.

The operation was performed twice: one to predict C4 cluster versus all other clusters and one to discriminate between C1, C2, and C3. In both cases, the algorithm ran over 100 iterations with a max depth of 5 and balanced classes for initializations of random forests.

The two sets of selected features were respectively composed of 255 genes for the C4 prediction dataset and 597 genes for the C1, C2, and C3 prediction dataset.

We then used xgboost-tree[71] approach, to train a model on the dataset with a binary logistic objective function to predict C4 vs all (using the 255 genes previously identified by the Boruta algorithm) and to extract features that have been used by the algorithm to craft the decision tree of the model.

The model can be summarized by $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$ where $\hat{y}_i$ is the cluster prediction for the patient i, x_i the vector describing the patient i (composed of the selected features), F the set of estimators for the model (4 in our case, one for each cluster) and K the number of trees by estimator which is 3 for C4 and 4 for C1, C2, and C3. In this context, f_k refers to the tree number k of the estimator f where $f \in F$. K has been manually refined in order to find a compromise between good predictive performance and a low complexity model.

We performed the same approach with a softmax objective function in a multi-classification context to predict the C1, C2, and C3 cluster based on the 597 features previously highlighted by Boruta for this specific classification problem.

The final sets of selected features were composed of 10 genes for the C4 prediction model and 31 genes for the multi-classification (C1, C2, or C3) model (Supplementary Fig. 10). The accuracies of the models, during the training phase perform on the validation set (Table 1) were 94.81% for the C4 prediction model and 96.72% for the multi-classification model.

We then created a composite model, using the combinatorial results of the C4 predictor model and the multi-classification model to predict all 4 clusters on the patients of the discovery set.

Patients were first evaluated by the C4 predictor model. If C4 was not assigned, the patients were evaluated by the multi-classification model.

In order to allow our model to process other cohorts of patients we implemented an interpolation function described by (2). We selected 6 genes with FC ≤ 1.1 and FDR ≥ 0.05 based on their constant expression across all 4 clusters and HV. Their expressions were between 4 and 14 vst normalized counts [*SPIRE* (4), *NUP210L* (6), *GATAD1* (8), *HVCN1* (10), *ENO* (12), and *FLNA* (14)] (Supplementary Fig. 13). This set of genes was denoted G. The interpolated value of a gene x, $I(x)$ was computed as $I(x) = I(a) + (I(b) - I(a)) \times \frac{x-a}{b-a}$ with a and b representing the vst normalized expression value of two genes such as genes $a, b \in G$, $a < x < b$ and $b \neq a$.

The composite model is integrated into an analysis tool available[33] and the pseudocode description is reported in Supplementary Fig. 16.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data included in our study is available upon request at ELIXIR Luxemburg, except the GWAS data that cannot be anonymized, with the permanent link: https://doi.org/10.17881/th9v-xt85 and access procedure is described on the ELIXIR data landing page. The PRECISESADS Consortium committed to secure patient data access through the ELIXIR platform. This commitment was formerly given by written to all patients at the end of the project and to the involved Ethical Committees. The future use of the Project database was framed according to the scope of the patient information and consent forms, where the use of patient data is limited to scientific research in autoimmune diseases. ELIXIR reviews applicants requests and prepares Data Access Committee's decisions on access to Data, communicates such decisions to the Data Providers, who have 10 days to exercise their right to veto; otherwise access is granted to the User.

Code availability

Except when indicated, data analyses were carried out using either an assortment of R system software (http://www.R-project.org, V4.0.1) packages including those of Bioconductor or original R code. R packages are indicated when appropriate. For GWAS analysis, we used Plink, an open-source whole genome association analysis toolset. Machine learning approaches were carried out using python programs (v3.8.5) based on the following modules: scikit-learn, numpy, and xgboost. The composite model designed to predict the patient's cluster is integrated into an analysis tool available on the laboratory's github repository at the following address: https://lbai-infolab.github.io/SjTree/(33).

Received: 2 November 2020; Accepted: 30 April 2021;
Published online: 10 June 2021

References

1. Brito-Zerón, P. et al. Sjögren syndrome. *Nat. Rev. Dis. Primers* **2**, 16047 (2016).
2. Baldini, C. et al. Primary Sjogren's syndrome as a multi-organ disease: impact of the serological profile on the clinical presentation of the disease in a large cohort of Italian patients. *Rheumatology (Oxford)* **53**, 839–844 (2014).
3. Qin, B. et al. Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Ann Rheum. Dis.* **74**, 1983–1989 (2015).
4. Goules, A. V. & Tzioufas, A. G. Primary Sjögren's syndrome: clinical phenotypes, outcome and the development of biomarkers. *Autoimmun. Rev.* **15**, 695–703 (2016).
5. Barturen, G. et al. Integrative analysis reveals a molecular stratification of systemic autoimmune diseases. *Arthritis Rheumatol.* (2020) https://doi.org/10.1002/art.41610 https://doi:10.17881/th9v-xt85.
6. Li, H., Ice, J. A., Lessard, C. J. & Sivils, K. L. Interferons in Sjogren's syndrome: genes, mechanisms and effects. *Front Immunol.* **4**, 290 (2013).
7. Bennett, L. et al. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* **197**, 711–723 (2003).
8. Guedj, M. et al. A refined molecular taxonomy of breast cancer. *Oncogene* **31**, 1196–1206 (2012).
9. Rinchai, D. et al. BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. *Bioinformatics* **btab121**, 1–8, https://doi.org/10.1093/bioinformatics/btab121 (2021).
10. Chiche, L. et al. Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures. *Arthritis Rheumatol.* **66**, 1583–1595 (2014).
11. Bodewes, I. L. A. et al. Systemic interferon type I and type II signatures in primary Sjögren's syndrome reveal differences in biological disease activity. *Rheumatology (Oxford)* **57**, 921–930 (2018).
12. Kirou, K. A. et al. Coordinate overexpression of interferon-alpha-induced genes in systemic lupus erythematosus. *Arthritis Rheum.* **50**, 3958–3967 (2004).
13. Lessard, C. J. et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nat. Genet.* **45**, 1284–1292 (2013).
14. Li, Y. et al. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11.23. *Nat. Genet.* **45**, 1361–1365 (2013).
15. Le Pottier, L., Amrouche, K., Charras, A., Bordron, A., Pers, J-O. Sjögren's syndrome. In: Martín, J., Carmona, F. (eds) Genetics of Rare Autoimmune Diseases. Rare Diseases of the Immune System. (Springer, Cham, 2019), https://doi.org/10.1007/978-3-030-03934-9_4.
16. Morris, D. L. et al. MHC associations with clinical and autoantibody manifestations in European SLE. *Genes Immun.* **15**, 210–217 (2014).
17. Wisskirchen, C., Ludersdorfer, T. H., Müller, D. A., Moritz, E. & Pavlovic, J. The cellular RNA helicase UAP56 is required for prevention of double-stranded RNA formation during influenza A virus infection. *J. Virol.* **85**, 8646–8655 (2011).
18. Tong, Y. et al. Enhanced TLR-induced NF-κB signaling and type I interferon responses in NLRC5 deficient mice. *Cell Res.* **22**, 822–835 (2012).
19. Naomi, M. et al. MXA as a clinically applicable biomarker for identifying Type 1 interferon in primary Sjogren's syndrome. *Ann. Rheum. Dis.* **73**, 1052–1059 (2014).
20. Irizarry, R. A. et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
21. Imgenberg-Kreuz, J. et al. Genome-wide DNA methylation analysis in multiple tissues in primary Sjögren's syndrome reveals regulatory effects at interferon-induced genes. *Ann. Rheum. Dis.* **75**, 2029–2036 (2016).
22. Fabregat, A. et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinf.* **18**, 142 (2017).
23. Davies, R. et al. Patients with primary Sjögren's syndrome have alterations in absolute quantities of specific peripheral leucocyte populations. *Scand. J. Immunol.* **86**, 491–502 (2017).
24. d'Arbonneau, F. et al. BAFF-induced changes in B cell antigen receptor-containing lipid rafts in Sjögren's syndrome. *Arthritis Rheum.* **54**, 115–126 (2006).
25. Mukherjee, R. et al. Non-Classical monocytes display inflammatory features: validation in sepsis and systemic lupus erythematous. *Sci. Rep.* **5**, 13886 (2015).
26. Schleinitz, N., Vély, F., Harlé, J. R. & Vivier, E. Natural killer cells in human autoimmune diseases. *Immunology* **131**, 451–458 (2010).
27. Aramaki, T. et al. A significantly impaired natural killer cell activity due to a low activity on a per-cell basis in rheumatoid arthritis. *Mod. Rheumatol.* **19**, 245–252 (2009).
28. Wildenberg, M. E., van Helden-Meeuwsen, C. G., van de Merwe, J. P., Drexhage, H. A. & Versnel, M. A. Systemic increase in type I interferon activity in Sjögren's syndrome: a putative role for plasmacytoid dendritic cells. *Eur. J. Immunol.* **38**, 2024–2033 (2008).
29. van den Hoogen, L. L. et al. Monocyte type I interferon signature in antiphospholipid syndrome is related to proinflammatory monocyte subsets, hydroxychloroquine and statin use. *Ann. Rheum. Dis.* **75**, e81 (2016).
30. Xourgia, E. & Tektonidou, M. G. Type I interferon gene expression in antiphospholipid syndrome: pathogenetic, clinical and therapeutic implications. *J. Autoimmun.* **104**, 102311 (2019).
31. Wallace, D. J., Gudsoorkar, V. S., Weisman, M. H. & Venuturupalli, S. R. New insights into mechanisms of therapeutic effects of antimalarial agents in SLE. *Nat. Rev. Rheumatol.* **8**, 522–533 (2012).
32. van den Borne, B. E., Dijkmans, B. A., de Rooij, H. H., le Cessie, S. & Verweij, C. L. Chloroquine and hydroxychloroquine equally affect tumor necrosis factor-alpha, interleukin 6, and interferon-gamma production by peripheral blood mononuclear cells. *J. Rheumatol.* **24**, 55–60 (1997).
33. Foulquier N. A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome, SjTree, https://doi.org/10.5281/zenodo.4643639 (2020).
34. Gottenberg, J. E. et al. Effects of hydroxychloroquine on symptomatic improvement in primary Sjögren syndrome: the JOQUER randomized clinical trial. *JAMA* **312**, 249–258 (2014).
35. Mariette, X. et al. Efficacy and safety of belimumab in primary Sjögren's syndrome: results of the BELISS open-label phase II study. *Ann. Rheum. Dis.* **74**, 526–531 (2015).
36. Devauchelle-Pensec, V. et al. Treatment of primary Sjögren syndrome with rituximab: a randomized trial. *Ann. Intern. Med.* **160**, 233–242 (2014).
37. Bowman, S. J. et al. Randomized controlled trial of rituximab and cost-effectiveness analysis in treating fatigue and oral dryness in primary Sjögren's syndrome. *Arthritis Rheumatol.* **69**, 1440–1450 (2017).
38. Meiners, P. M. et al. Abatacept treatment reduces disease activity in early primary Sjögren's syndrome (open-label proof of concept ASAP study). *Ann. Rheum. Dis.***73**, 1393–1396 (2014).
39. St Clair, E. W. et al. Clinical efficacy and safety of baminercept, a lymphotoxin β receptor fusion protein, in primary Sjögren's syndrome: results from a phase

II randomized, double-blind, placebo-controlled trial. *Arthritis Rheumatol.* **70**, 1470–1480 (2018).
40. Gandolfo, S. & De Vita, S. Emerging drugs for primary Sjögren's syndrome. *Expert Opin. Emerg. Drugs* **24**, 121–132 (2019).
41. Barturen, G., Beretta, L., Cervera, R., Van Vollenhoven, R. & Alarcón-Riquelme, M. E. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nat. Rev. Rheumatol.* **14**, 180 (2018).
42. James, J. A. et al. Unique Sjögren's syndrome patient subsets defined by molecular features. *Rheumatology (Oxford)* **59**, 860–868 (2020).
43. Tarn, J. R. et al. Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials. *Lancet Rheumatol.* **1**, e85–e94 (2019).
44. Carvajal Alegria, G. et al. Epidemiology of neurological manifestations in Sjögren's syndrome: data from the French ASSESS Cohort. *RMD Open* **2**, e000179 (2016).
45. Lewis, I., Hackett, K. L., Ng, W. F., Ellis, J. & Newton, J. L. A two-phase cohort study of the sleep phenotype within primary Sjögren's syndrome and its clinical correlates. *Clin. Exp. Rheumatol.* **37**, S78–S82 (2019).
46. Hillen, M. R. et al. Plasmacytoid DCs from Patients with Sjögren's syndrome are transcriptionally primed for enhanced pro-inflammatory cytokine production. *Front. Immunol.* **10**, 2096 (2019).
47. Nezos, A. et al. Type I and II interferon signatures in Sjogren's syndrome pathogenesis: contributions in distinct clinical phenotypes and Sjogren's related lymphomagenesis. *J. Autoimmun.* **63**, 47–58 (2015).
48. Toro-Domínguez, D. et al. Differential treatments based on drug-induced gene expression signatures and longitudinal systemic lupus erythematosus stratification. *Sci. Rep.* **9**, 15502 (2019).
49. Han, B. K. et al. Neutrophil and lymphocyte counts are associated with different immunopathological mechanisms in systemic lupus erythematosus. *Lupus Sci. Med.* **7**, e000382 (2020).
50. Hemond, C. C., Glanz, B. I., Bakshi, R., Chitnis, T. & Healy, B. C. The neutrophil-to-lymphocyte and monocyte-to-lymphocyte ratios are independently associated with neurological disability and brain atrophy in multiple sclerosis. *BMC Neurol.* **19**, 23 (2019).
51. Devauchelle-Pensec, V. et al. Gene expression profile in the salivary glands of primary Sjögren's syndrome patients before and after treatment with rituximab. *Arthritis Rheum.* **62**, 2262–2271 (2010).
52. Shen-Orr, S. S. et al. Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
53. Jamin, C. et al. Multi-center harmonization of flow cytometers in the context of the European "PRECISESADS" project. *Autoimmun. Rev.* **15**, 1038–1045 (2016).
54. Le Lann, L. et al. Standardization procedure for flow cytometry data harmonization in prospective multicenter studies. *Sci. Rep* **10**, 11567 (2020).
55. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
56. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
57. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
58. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
59. Leek, J. T. et al. sva: Surrogate variable analysis. *R package version* **3**, 882–883 (2017).
60. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
61. Gold, D. L., Coombes, K. R., Wang, J. & Mallick, B. Enrichment analysis in high-throughput genomics—accounting for dependency in the NULL. *Brief Bioinformatics* **8**, 71–77 (2007).
62. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
63. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Johnson, A. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **15**, 2938–2939 (2008).
65. Turner, S. D. qqman: a R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
66. Aryee, M. J. et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
67. Morris, T. J. et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).
68. Franceschini, A. et al. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
69. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. *Biom J.* **50**, 346–363 (2008).
70. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 11 (2010).
71. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785 (2016).

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under the Grant Agreement Number 115565 (PRECISESADS project), resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007–2013) and EFPIA companies' in-kind contribution. LBAI was supported by the Agence Nationale de la Recherche under the "Investissement d'Avenir" program with the Reference ANR-11-LABX-0016-001 (Labex IGO) and the Région Bretagne. The authors would like to particularly express their gratitude to the patients, nurses, technicians and many others who helped directly or indirectly in the consecution of this study. They are grateful to the Institut Français de Bioinformatique (ANR-11-INBS-0013), the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr) for providing computing and storage resources and the Hypérion platform at LBAI (Brest, France) for flow cytometry facilities. Finally, this work is now supported by ELIXIR Luxembourg via its data hosting service.

Author contributions

P.S., C.L.D., E.D., B.C., S.H., and C.B. performed the computational studies and carried out the analysis, N.F. performed the computational studies and developed the composite model, C.J., G.B., G.D., PRECISESADS Flow Cytometry Consortium, E.B., J.M., A.B., Z. M. R.L., M.O.B. performed the experimental studies. V.D.P., D.C., A.S., S.J.J., N.B.P., I.R. P., E.D.L., L.B., C.C., L.K., T.W., and PRECISESADS Clinical Consortium contributed to the recruitment of patients. S.C.G., L.X., M.G., P.M. contributed to the edition of the manuscript, MEAR supervised the PRECISESADS consortium, L.L. and J.O.P. supervised the work and wrote the manuscript. All the authors have approved the content of this paper and its related supplementary files and have agreed to the Nature Communications submission policies.

Competing interests

While engaged in the research project, R.L., F.M., and Z.M. were regular employees of Bayer A.G. At present, R.L. and Z.M. are regular employees of Nuvisan ICB GmbH, a company providing contract research services. P.S., S.H., S.C.G., L.X., M.G., P.M., and L. L. were regular employees of Institut de Recherches Internationales Servier at the time of the research project. B.C., C.B., and E.D. were phD students financed by Institut de Recherches Internationales Servier when they contributed to the research project. All other authors confirmed signing the ICMJE form for Disclosure of Potential Conflicts of Interest and none of them have any conflict of interest related to this work.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23472-7.

Correspondence and requests for materials should be addressed to J.-O.P.

Peer review information *Nature Communications* thanks A. Darise Farris, Zhan-Guo Li, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PRECISESADS Clinical Consortium

Lorenzo Beretta[8], Barbara Vigone[8], Jacques-Olivier Pers[2,3], Alain Saraux[2,3], Valérie Devauchelle-Pensec[2,3], Divi Cornec[2,3], Sandrine Jousse-Joulin[2,3], Bernard Lauwerys[15], Julie Ducreux[15], Anne-Lise Maudoux[15], Carlos Vasconcelos[16], Ana Tavares[16], Esmeralda Neves[16], Raquel Faria[16], Mariana Brandão[16], Ana Campar[16], António Marinho[16], Fátima Farinha[16], Isabel Almeida[16], Miguel Angel Gonzalez-Gay Mantecón[17], Ricardo Blanco Alonso[17], Alfonso Corrales Martínez[17], Ricard Cervera[6], Ignasi Rodríguez-Pintó[6], Gerard Espinosa[6], Rik Lories[7], Ellen De Langhe[7], Nicolas Hunzelmann[18], Doreen Belz[18], Torsten Witte[11], Niklas Baerlecken[11], Georg Stummvoll[19], Michael Zauner[19], Michaela Lehner[19], Eduardo Collantes[5], Rafaela Ortega-Castro[5], Ma Angeles Aguirre-Zamorano[5], Alejandro Escudero-Contreras[5], Ma Carmen Castro-Villegas[5], Yolanda Jiménez Gómez[5], Norberto Ortego[20], María Concepción Fernández Roldán[20], Enrique Raya[21], Inmaculada Jiménez Moleón[21], Enrique de Ramon[22], Isabel Díaz Quintero[22], Pier Luigi Meroni[13], Maria Gerosa[13], Tommaso Schioppo[13], Carolina Artusi[13], Carlo Chizzolini[9], Aleksandra Zuber[9], Donatienne Wynar[9], Laszló Kovács[10], Attila Balog[10], Magdolna Deák[10], Márta Bocskai[10], Sonja Dulic[10], Gabriella Kádár[10], Falk Hiepe[23], Velia Gerl[23], Silvia Thiel[23], Manuel Rodriguez Maresca[24], Antonio López-Berrio[24], Rocío Aguilar-Quesada[24], Héctor Navarro-Linares[24], Yiannis Ioannou[25], Chris Chamberlain[26], Jacqueline Marovac[26], Marta Alarcón Riquelme[4] & Tania Gomes Anjos[4]

[15]Pôle de pathologies rhumatismales systémiques et inflammatoires, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium. [16]Centro Hospitalar do Porto, Port, Portugal. [17]Servicio Cantabro de Salud, Hospital Universitario Marqués de Valdecilla, Santander, Spain. [18]Klinikum der Universitaet zu Koeln, Cologne, Germany. [19]Medical University Vienna, Vienna, Austria. [20]Complejo hospitalario Universitario de Granada (Hospital Universitario San Cecilio), Granada, Spain. [21]Complejo hospitalario Universitario de Granada (Hospital Virgen de las Nieves), Granada, Spain. [22]Hospital Regional Universitario de Málaga, Málaga, Spain. [23]Charite, Berlin, Germany. [24]Andalusian Public Health System Biobank, Granada, Spain. [25]UCB Pharma (PRECISESADS Project office), Slough, UK. [26]Chromatin and Disease Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain.

PRECISESADS Flow Cytometry Consortium

Christophe Jamin[2,3], Concepción Marañón[4], Lucas Le Lann[2], Quentin Simon[2], Bénédicte Rouvière[2,3], Nieves Varela[4], Brian Muchmore[4], Aleksandra Dufour[9], Montserrat Alvarez[9], Carlo Chizzolini[9], Jonathan Cremer[7], Ellen De Langhe[7], Nuria Barbarroja[5], Chary Lopez-Pedrera[5], Velia Gerl[23], Laleh Khodadadi[23], Qingyu Cheng[23], Anne Buttgereit[12], Zuzanna Makowska[12], Aurélie De Groof[14], Julie Ducreux[14], Elena Trombetta[8], Tianlu Li[26], Damiana Alvarez-Errico[26], Torsten Witte[11], Katja Kniesch[11], Nancy Azevedo[15], Esmeralda Neves[15], Sambasiva Rao[27], Pierre-Emmanuel Jouve[28] & Jacques-Olivier Pers[2,3]

[27]Sanofi Genzyme, Framingham, MA, USA. [28]AltraBio SAS, Lyon, France.

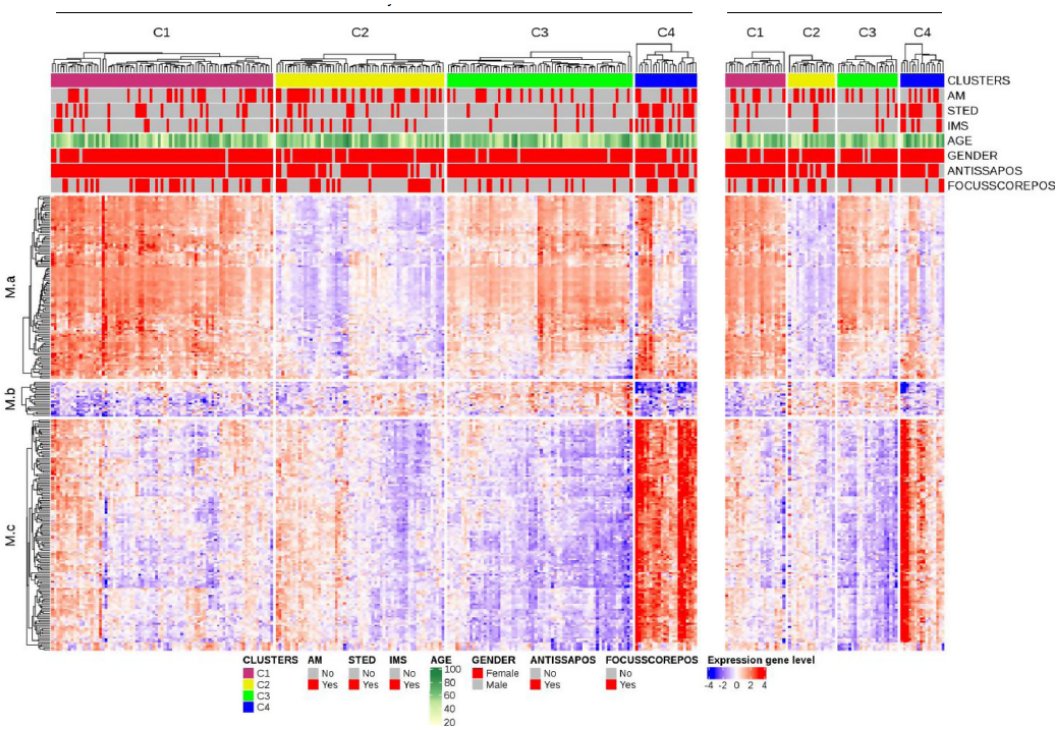


Figure 4.2: Heatmap resulting from stratification of patients with primary Sjögren's syndrome. Patients, in columns, are grouped according to their cluster assignments, while genes are classified into functional modules (three identified). The transcriptomic expression level is color-coded, with red indicating higher expression values. At the top, additional phenotypic annotations are provided, such as treatment (AM: antimalarials, STED: steroids, IMS: immunosuppressants), sex, and antinuclear autoantibodies. This figure is reproduced from [Sor+21, Fig. 1].

4.2 Main results

Four endotypes identified using data-driven approaches The unsupervised patient stratification identified four different clusters of pSS patients, each defined by distinct molecular signatures. The four clusters are represented with the corresponding Heatmap expression profiles in Creffig:heatmap-sjogren,

and were further annotated through integrative analysis as:

- The C2 cluster displays a healthy-like profile, gathering patients with on average, lower disease activity and no Anti-SSA and anti-SSB autoantibodies.
- In contrast, cluster C4 displays the most distinct and severe clinical phenotype, marked by a pronounced Type II inferferon signature activation signature, significant lymphopenia, and elevated neutrophil counts, collectively indicative of an inflammatory profile. This inflammatory pattern was further confirmed by methylation analysis whicih revealed significant hypomethylation of genes associated with IFN signalling [2]. Additionally, it's worth noting that C4 is the smallest cluster, making up only for 12.5% of the total cohort.

[2]High levels of gene expression are often associated with low promoter methylation ([Wag+14])

- The cluster C1 showcases the highest Type I and Type II IFN scores. Genome-wide association study (GWAS) analysis only pinpoints significant differences in genomic sequences in this endotype, particularly for genes associated with the immune system, and signal transduction.
- Cluster C3 exhibits a salient Type I IFN score, alongside significant activation of B cell-related pathways, highlighting the potential therapeutic significance of addressing B cells proliferation for patients assigned to this cluster.

Composite model for refining patient molecular fingerprints A composite model has further been developed to map each patient to one of the four clusters identified, achieving an impressive overall accuracy of 95%, assigning membership to a given cluster employing only ten genes. The robustness of the model was confirmed using an external validation dataset of pSS patients, and thus, revealed helpful in selecting patients, prior to clinical trials, based on their molecular profile and related predictive therapeutic outcome.

In conclusion, the biological insights gained from this data-driven clustering approach shed light on the intricate role of pathways, particularly IFN signalling, in driving the complex heterogeneity of pSS patients. Besides, dissecting the heterogeneity of pSS patients should promote the development of personalised therapies, for instance, targeting B cells using depleting antibodies, like rituximab, has already proved efficient in reducing immuno-inflammatory activity in C3-like patients ([Gri+19]).

Personal Contributions I personally contribute to this paper by standardising a pipeline for processing RNASeq data (see Details in Section 3.3), by providing guidelines for designing a robust clustering method and finally, I performed the statistical analyses to evaluate whether the cellular composition differs significantly between the identified clusters.

To analyse the pool of cell populations, I combine mass cytometry datasets with numerical estimation of cellular ratios. Flow cytometry analyses were conducted in eleven medical centres by the PreciseSADS consortium, using the same calibrations to avoid batch effect. In particular, the automated gating protocol outperformed against the manual gating. We used CIBERSORT [New+15], as this deconvolution method proved the most robust and accurate in a variety of benchmark studies performed recently ([Stu+19] and [Fa+20]).

We specifically utilized the LM22 signature matrix, derived from whole blood samples of Systemic Lupus Erythematosus (SLE) patients. Indeed, the performance of supervised deconvolution algorithms rely partly on the closeness of the signature used with the patient phenotype profile, and SLE patients exhibit similar molecular profiles to those affected by Sjögren's disease. In addition, LM22 offered better characterization of the B cell population, a key contributor to inflammation in C3-like patients, with further classification into naive, plasma, and memory subtypes.

However, we excluded certain ectopic cell lines, such as macrophages and mast cells, from the original purified signature, as they are not typically found in circulating blood. [3].

We ultimately checked the concordance between the numerically inferred ratios with flow cytometry outputs using RMSE scores. Since cell populations were not always available at the

[3]We recommend in addition the fusion of subsets of cell types, for instance, combining "resting" and "activated" dendritic cells, or merging subsets of T cell helpers (naive, central and effector memory, follicular helper) could be considered. This suggestion is based on the observation that the transcriptomic differences between these subcategories are nearly insignificant, and integrating highly correlated cell lines might introduce potential "spillover" effects.

same scale, we summed children cell lines to reconstitute consistently cell lineages. Kruskal-Wallis tests, the unparametric equivalent of the ANOVA test, wered used to evaluate whether global differences of cell composition across cluster assignment were statistically significant.

Biological interpretation of numerical deconvolution outputs, focusing on B cells Flow cytometry analyses have already exhibited a stronger proportion of B cells in cluster C3 and a down-regulation of pathways involved in B cell development in cluster C4, resulting in lymphopenia.

However, the increased granularity provided by the LM17 signature matrix allows for a deeper exploration of potential mechanistic factors underlying the alteration of B cell composition (Figure 4.1(b), right panel). Notably, the C3 cluster sets apart by a higher proportion of naive B cells, which aligns with elevated activity of the IL-7 pathway ([Cla+14], known for its involvement in B cell lymphopoiesis and signalling). The presence of self-reactive memory B cells, combined with chronic inflammation, indicates the possibility of a cross-reactivity phenomenon (Section 2.1.4). Lastly, cluster C3 displays an elevated proportion of plasma cells, known for producing substantial quantities of antinuclear autoantibodies, as emphasized in [Mor+14].

I personally hypothesized that different mechanisms underlie the significant alterations of B cell composition observed for both clusters C3 and C4.

The cluster C4 aggregates the patients with the most severe clinical features, which are thus more likely to take immunosuppressant medications, such as corticosteroids. These treatments tend to down-regulated the pathway involved in the maturation of naive B cells, [Cri+21] notably observed that relapses in B cell-mediated autoimmune diseases, treated by rituximab treatment, were associated with the reactivation of ancestral memory B cells.

In addition, [Mau+23] and [DGvM23] demonstrated that autoimmune diseases, in their acute phases, induce specific alterations of the immune cell composition. Notably, it has been shown that the relative proportion of plasma cells and autoreactive memory B cells, even without antigen, increases compared to naive B cells. Possibly, similar mechanisms altering the B cell homoeostasis explain the contradictory overarching lymphopenia observed in cluster C4, with a general decrease of B cell populations, but a comparatively increase of plasma and memory B cells.

4.3 Limitations and perspectives

The scope of these preliminary findings is limited by the cross-sectional design of the study itself. Longitudinal studies, on the contrary, would enable to characterise the stability of cluster assignment, and the impact of treatments on gene signalling dysregulation over time.

In addition, only blood samples have been used to cluster patients into endotypes, while the Sjögren's disease is a systematic auto-immune disease, affecting several organs and tissues, notably the salivary glands [BW17]. This may explain the lack of significant correlation between the observed clinical features and the patients clustered on the basis of transcriptomic data.

Secondly, k-means and hierarchical clustering are rather used for initialising the parameters of Gaussian mixture models, as they make stronger assumptions on the distribution of the data and thus are not able to capture as meaningful patterns. For instance, k-means assumes equi-balanced and homoscedastic clusters (see `Initialisation of the EM algorithm`, in Chapter 3).

Simultaneously, the model's robustness could have been performed through hyper-parameter tuning using a bootstrap approach, for selecting the overarching parametrisation and final number

of clusters selected. In addition, the resulting bootstrap estimates could have been used to derive confidence intervals, thereby assessing the statistical uncertainty associated with the parameters specific to each cluster (see sections "Model selection" and "Derivation of confidence intervals in GMMs" in Chapter 3).

Ultimately, standard parametric mixture models are not tailored for high-dimensional datasets, especially when the number of variables significantly exceeds the number of observations, as it is the case in this patient stratification. By contrast, lower-dimensional projection (e.g., Principal Component Analysis) techniques or parsimonious model parametrisations could enhance the discriminative ability of Gaussian mixture models in identifying patient subgroups and streamline biological interpretation of the inferred endotypes(see Appendix B).

Thirdly, we only leverage transcriptomic data to classify Sjögren's patients into molecular endotypes, possibly explaining our failure to relate the clusters with a specific set of clinical symptoms. Expanding the repertoire and diversity of biological tissue origins, including salivary glands—a primary site of involvement in Sjögren's syndrome, and using multiple orthogonal modalities, hence hold significant promise in retrieving the aetiological factors contributing to Sjögren's syndrome pathophysiology, a challenge still remnant.

In particular, as detailed in Section 4.2, we observed that the inference of clusters and the variations of transcriptomic expression across patients were mostly driven by changes of the cell composition in whole blood samples, rather than intrinsic changes of the RNA-Seq expression within native cell groups.

In next chapter 5, we precisely review several numerical methods, aimed at retrieving the cellular composition of heterogeneous samples. We notably exhibit in next part how such approaches, by unravelling the intrinsic heterogeneity of tissues, improve the accuracy of downstream analyses, and contribute to streamline the exploration of complex niches, such as the tumoral micro-environment.

Part III

Cell populations and deconvolution algorithms

Chapter 5

Article 3: review of cellular deconvolution methods

Methodological Objective: Evaluating the Performance of Deconvolution Algorithms in the Endeavour of Biological Features of the sample As we have seen in previous Chapter 4, the biological relevance of unsupervised stratification, based solely on transcriptomic datasets, for patients afflicted by Sjögren's primary syndrome, is limited by the nature of bulk RNA-Seq analysis itself. By averaging transcriptomic measures over a heterogeneous biological mixture, such measures are indeed oblivious to variations in cell-type composition, hindering partially the identification of the key cell subtypes involved in disease progression and diverse response to treatments (see also index Heteroskedasticity).

In section 5.1, we present an up-to-date review of deconvolution algorithms designed to automatically infer the cellular composition of a biological sample. We notably focus on so-called "partial" numerical approaches, which aim to deduce the relative abundances of cellular components within a heterogeneous mixture, harnessing purified cellular expression profiles.

While the majority of the analysed deconvolution models assume that the total transcriptomic expression of the mixture can be reconstructed by summing the individual contributions of each cell subtype, weighted by their abundance, we emphasize that these models differ in their objectives and biological constraints. For instance, some algorithms have been developed to estimate the characteristics of an unknown cell population, often a tumoral clone, while others concentrate on the robustness of the outputs, by incorporating additional feature selection steps in the refinement of signature profiles.

We subsequently broaden our focus to encompass the entire deconvolution process, from data retrieval to statistical evaluation and biological interpretation of the results. In this section, we underscore the critical importance of data collection, preprocessing, and cleaning in shaping the final result's quality. To paraphrase [Fin+19a], obtaining high-quality purified transcriptomic expression profiles, in sufficient quantities and closely aligned with the biological context under study, is arguably more crucial than the choice of the algorithm itself.

We conclude this review with a more specific discussion regarding the future of cellular deconvolution methods. Notably, we emphasize their complementarity with single-cell RNA sequencing (scRNA) technologies to enhance their own accuracy, alongside the resolution of spatial transcriptomics methods.

5.1 Article 3

An updated State-of-the-Art Overview of transcriptomic Deconvolution Methods

Bastien Chassagnol[1,2,*], Grégory Nuel[2], Etienne Becht[1]

1 Institut De Recherches Internationales Servier (IRIS), FRANCE
2 LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), Sorbonne Université, 4, place Jussieu, 75252 PARIS, FRANCE

*** bastien_chassagnol@laposte.net**

Abstract

Although bulk transcriptomic analyses have significantly contributed to an enhanced comprehension of multifaceted diseases, their exploration capacity is impeded by the heterogeneous compositions of biological samples. Indeed, by averaging expression of multiple cell types, RNA-Seq analysis is oblivious to variations in cellular changes, hindering the identification of the internal constituents of tissues, involved in disease progression. On the other hand, single-cell techniques are still time, manpower and resource-consuming analyses.

To address the intrinsic limitations of both bulk and single-cell methodologies, computational deconvolution techniques have been developed to estimate the frequencies of cell subtypes within complex tissues. These methods are especially valuable for dissecting intricate tissue niches, with a particular focus on tumour microenvironments (TME).

In this paper, we offer a comprehensive overview of deconvolution techniques, classifying them based on their methodological characteristics, the type of prior knowledge required for the algorithm, and the statistical constraints they address. Within each category identified, we delve into the theoretical aspects for implementing the underlying method, while providing an in-depth discussion of their main advantages and disadvantages in supplementary materials.

Notably, we emphasise the advantages of cutting-edge deconvolution tools based on probabilistic models, as they offer robust statistical frameworks that closely align with biological realities. We anticipate that this review will provide valuable guidelines for computational bioinformaticians in order to select the appropriate method in alignment with their statistical and biological objectives.

We ultimately end this review by discussing open challenges that must be addressed to accurately quantify closely related cell types from RNA sequencing data, and the complementary role of single-cell RNA-Seq to that purpose.

1 Introduction

The transcriptome refers to the complete set of RNA transcripts, expressed within a biological sample. By providing a snapshot of gene expression patterns, studying its variations across phenotypical conditions provide valuable insights into the regulatory mechanisms of gene expression that underlie disease progression and individual responses to treatments.

The main biological sources of transcriptomic expression, between individuals and within tissues, proceed from three main biological factors, summarised in Figure 1: the global environmental and topic condition of the sample, encompassing disease state and tissue location; the genotype condition, involving single-nuclear polymorphisms, haplotypes, and comparable genetic aspects; and the cellular composition. Changes of cell composition are notably driven by intertwined physiological processes activating *cell motility* and *cell differentiation* mechanisms ([SG13]). In addition, the pertinent biological signal is often entangled with extraneous technical noise, requiring specific corrections in subsequent downstream analyses.

In addition, intrinsic heterogeneity is also present at the cell population level itself, arising from the presence of unspecified and infrequent population subtypes, coexistence of different developmental *cell states* or asynchronous biological processes (such as the cell cycle or circadian rhythm). Lastly, the kinetics of transcriptome regulation is inherently stochastic [Bue+15] (see Figure 1).

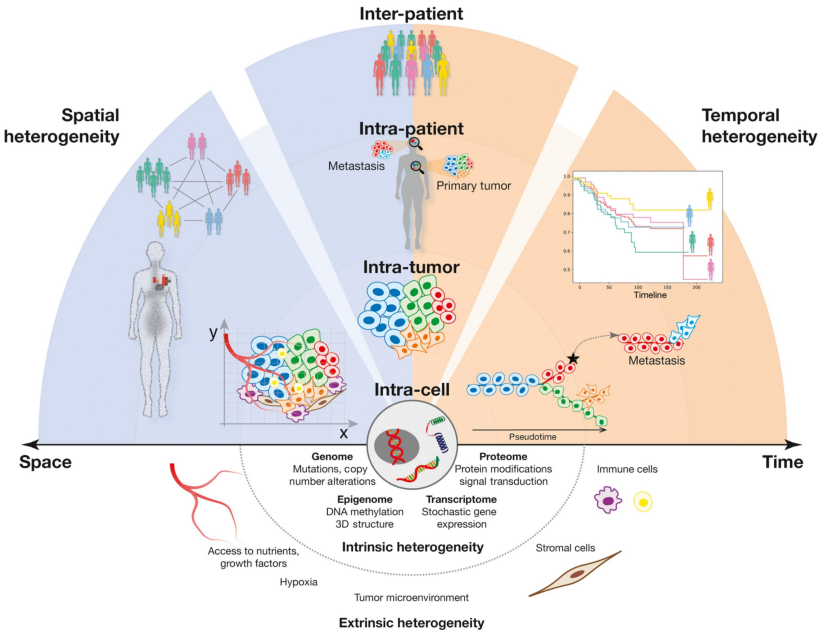


Figure 1. Main sources of transcriptomic variability, illustrated by the the intricacy of tumoral environments. The diversity of molecular profiles proceeds from a combination of intrinsic and extrinsic factors. *Intrinsic factors* encompass stochastic genetic, transcriptional, and proteomic mechanisms, while *extrinsic factors* include interactions between the resident cell populations and the surrounding microenvironment. The interconnection between these factors requires a systematic and multi-layered approach to comprehensively understand the intricacy of such biological environments. Figure reproduced from [Kas+22, Fig. 1]

While the analysis of the transcriptome through bulk RNA-Seq reveals meaningful co-expression patterns, by averaging measurements over several cell populations, it tends to ignore the intrinsic heterogeneity and complexity inherent to biological samples. Accordingly, bulk RNA-based methods are usually not able to determine whether significant changes in gene expression stem from a change of cell composition, from phenotype-induced variations or a combination of these factors ([Kuh+12]).

Hence, failure to account for changes of the cell composition is likely to result in a loss of *specificity* (genes mistakenly identified as differentially expressed, while they only reflect an increase in the cell

population naturally producing them) and *sensibility* (genes expressed by minor cell populations are amenable being masked by highly variable expression from dominant cell populations), as simply illustrated in Section 1. Overall, the intrinsic heterogeneity of complex tissues, above all tumoral ones, reduces the robustness and reproducibility of downstream analyses, notably differential gene expression analysis or clustering of co-expression networks [1].

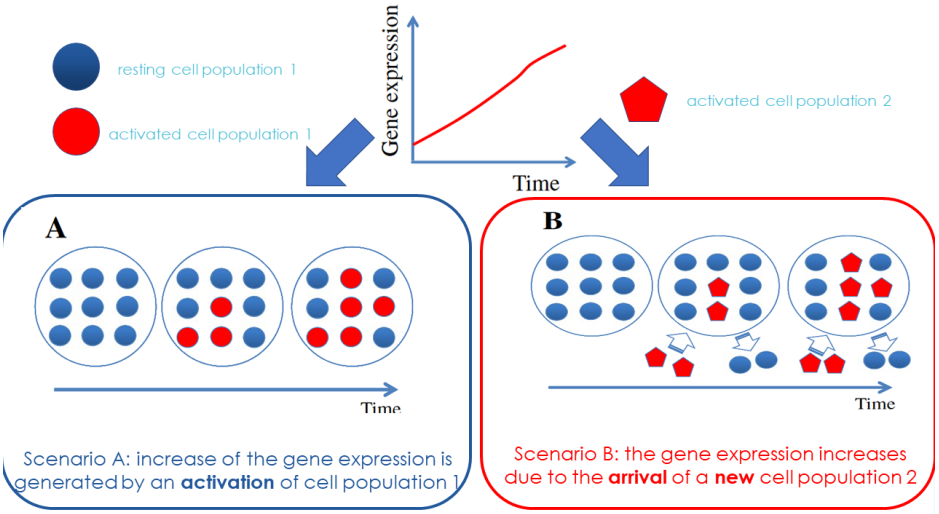


Figure 2. Changes in cell composition impact the transcriptomic expression. Here, at least two distinct biological mechanisms can likely explain the increased expression of transcriptomic activity observed for a given marker gene. In the scenario (A), the cell composition is unchanged, but previously inactivated cells are stimulated and released the TF in the biological medium. In scenario (B), there is a change of cell composition, with the infiltration of a second cell type in the sample. Reproduced from [Sho+12, Fig. 1].

Various computational methodologies have emerged in recent years to estimate automatically cell type proportions in biological samples from bulk transcriptomic profiles, alleviating the high costs of single-cell RNA-Seq technologies or enabling the exploitation of archived patient datasets whose original material is not anymore available [Avi+18]. Furthermore, by requiring prior isolation of cell populations single-cell technologies hinder the analysis of interactions occurring between them.

In contrast to bulk RNA-Seq and single cell methodologies, computational techniques (see Section 2) can simultaneously capture systemic and cell-specific information, respectively ([SG13]). Accordingly, by dissecting the intricacy of tissues, they reveal a strong potential to identify causal drivers and provide insights on regulation mechanisms.

In section 2, we present an updated review of deconvolution algorithms, with a specific focus on *partial methods* designed to automatically infer the cellular composition of heterogeneous biological samples utilising purified cellular expression profiles. These methods are usually categorised into those using a common purified expression profile, as discussed in section 2.1 (Reference-Based Approaches), and algorithms relying on gene markers specific to unique cell populations, as outlined in section 2.2 (Marker-Based Approaches). We also briefly touch upon reference-free methods in section 2.3 (Unsupervised and Reference-Free Deconvolution Methods), which are applicable when no prior information is available about the composition or characteristics of the mixture.

Throughout this section, we emphasise that adaptations and variations from the original deconvolution framework address specific biological questions and challenges, such as identifying rare tumoral clones or reconciling discrepancies between cell ratios estimated from transcriptomic expression data and those measured by physical cytometry technologies.

[1] [Whi+03] notably exhibits that most of the variability of gene expression in whole blood samples proceeds from relative changes of the composition in neutrophils, the most abundant immune cell type.

We subsequently broaden our perspective to cover the entire deconvolution process, spanning from data collection to the statistical evaluation and biological interpretation of the results, as detailed in section 3 (Deconvolution Pipeline). In this section, we emphasise the pivotal significance of meticulous data collection, preprocessing, and quality control in shaping the quality of the deconvolution output.

We conclude this review with a general discussion in the fate of cellular deconvolution methods in the context of expanding use of single-cell RNA sequencing and spatially resolved transcriptomics (SRT) technologies in section 4. We specifically underscore the relevance of integrating single-cell-based profiles into the spatial deconvolution framework to enhance transcriptomics resolution.

2 Overview of Numerical Deconvolution Methods

Deconvolution generally speaking names the process that consists in retrieving from a mixture its individual sub-components, popularised as the "cocktail party problem" [Che53]. In a biological sample (whole blood, tissue, ...), this consists generally in retrieving the distinct cell populations (immune, stromal...) composing it, but it can be directly extended to identify the different sources of the RNA production (for instance, many studies investigate on estimating a tumour purity score returning the proportion of malignant cells in [Yos+13]) or, at higher resolution, identify the cycle stages within a cell population (see Figure 3).

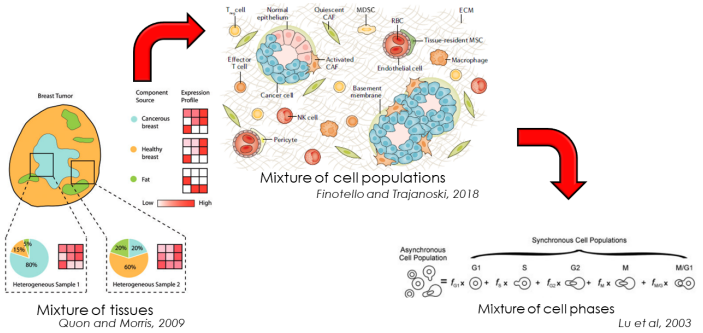


Figure 3. We detail some common applications of deconvolution methods, ordered by tier of resolution, from the least detailed resolution: *tissue* level ([QM09, Fig .1]), to the most detailed one, *cell cycles* ([LNM03, Fig .1]), through the *cell population* strata ([Fin+19a, Fig .1]).

Traditionally, deconvolution models assume that the total bulk expression is linearly related to the individual cell profiles. Precisely, they posit that the global expression can reconstructed by summing the distinct contributions of every cellular population weighed by their respective abundance within the sample (see Equation (1) and graphical illustration in Section 2):

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X} \times \boldsymbol{p}_i \quad \text{matricial form} \\ y_{gi} &= \sum_{j=1}^{J} x_{gj} \times p_j \quad \text{algebraic form} \end{aligned} \tag{1}$$

, with the following notations:

- $(\boldsymbol{y} = (y_{gi}) \in \mathbb{R}_+^{G \times N}$ is the global bulk transcriptomic expression, measured in N individuals.
- $\boldsymbol{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G \times J}}$ the signature matrix of the mean expression of G genes in J purified cell populations.
- $\boldsymbol{p} = (p_{ji}) \in]0, 1[^{J \times N}$ the unknown relative proportions of cell populations in N samples

Overall, the system includes G linear equations with J unknowns (the cellular proportions). In addition, most deconvolution problems explicitly integrate the *compositional* nature of cell ratios, enforcing in the estimation process the *unit-simplex constraint* (Equation (2)):

$$\begin{cases} \sum_{j=1}^{J} p_{ji} = 1 \\ \forall j \in \widetilde{J} \quad p_{ji} \geq 0 \end{cases} \tag{2}$$

Implicitly, Equation (2) implies that no other, unknown cell population could contribute to the measured bulk mixture. The main classes of deconvolution methods, defined on the basis of their biological objectives, are summarised in Figure 4(b), ranging from the approaches requiring the most information to the most unsupervised approaches:

In the following Section 2.1, we focus on *partial deconvolution* methods, that require individual cellular expression profiles to infer cell composition [Stu+04]. Besides, in the remainder of this paper, we posit, as most deconvolution algorithms, that the samples are uncorrelated with each other (independence assumption), allowing simultaneous and parallel cell ratio estimations. While this assumption reduces computational complexity, [Efr09] demonstrates cross-correlation across samples in real-world transcriptomic profiles.

2.1 Reference-based Approaches: Deciphering Cell Mixture through Expression Signatures

2.1.1 Regression-based approaches

OLS Regression The system of linear equations, given in Equation (1) rarely holds in practice, due to technical noise or unaccounted environmental variations. Most deconvolution algorithms model explicitly the error with a residual unobserved term, added to each individual transcriptomic measure, ϵ_g.

Subsequently, the usual approach is to retrieve the ordinary least squares (OLS) estimate which minimise the sum of squares (SSE) between predicted values fitted by the linear model: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{p}$ and the actually observed and measured values: $\boldsymbol{y}$:

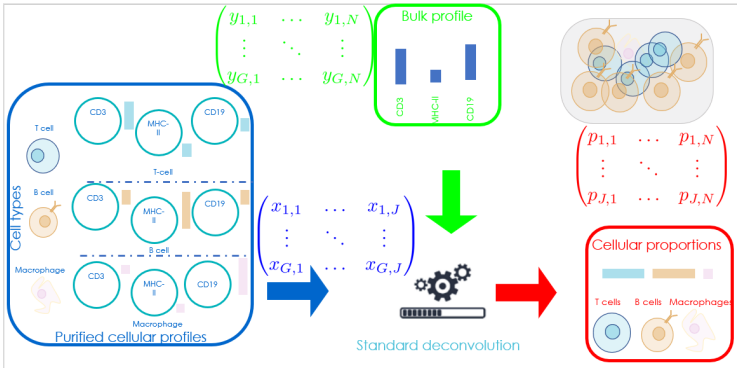
$$\hat{\boldsymbol{p}}^{\text{OLS}} \equiv \arg\min_{\boldsymbol{p}} ||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2 = \arg\min_{\boldsymbol{p}} ||\boldsymbol{X}\boldsymbol{p} - \boldsymbol{y}||^2 = \sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} x_{gj} p_j \right)^2 \tag{3}$$

with $\hat{\boldsymbol{p}}$ the unknown *coefficients* to estimate, $\boldsymbol{y}$ known as the *predicted, response variable* in a linear regression context and $\boldsymbol{X}$ the *design matrix*, storing the J purified profiles. Note that the "Rouché-Capelli" theorem states that the uniqueness of a solution to Equation (3) requires that the number of genes is at least equal to the number of cell ratios to estimate. The OLS estimator,$\hat{p}_{\text{OLS}}$ is explicitly given by the *Normal equations* (see Theorem A.1):

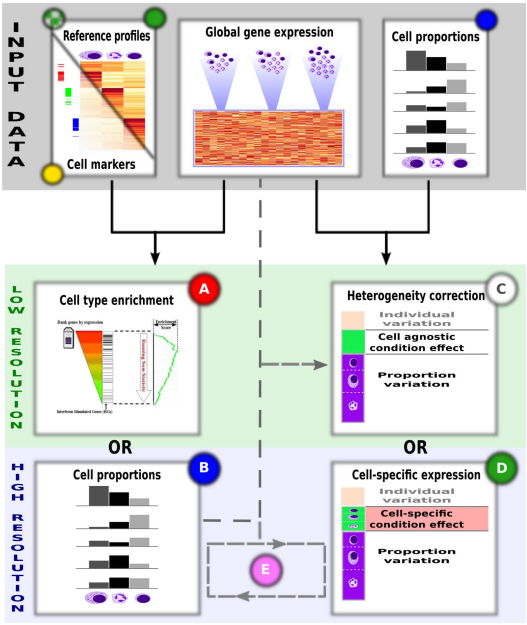
Interestingly, if we consider a generative approach, in which the error term is described by a white-*Gaussian* process (homoscedastic, null-centred), the *Gaussian-Markov* theorem (see Theorem A.2) states that the OLS estimate is unique and equal to the Maximum Likelihood Estimate (see Proof A.3).

Linear modelling, whose cellular ratios are the ones given by the Normal Equations (Theorem A.1), has first been used as such in [Abb+09] paper, using the lsfit function. The same method is used in [Li+16],to identify subgroups of melanomas characterised by varying levels of TCD8 subsets and correlate them with prognostic factors. To avoid accounting for tumoral cells when asserting ratios of infiltrated cells, only genes both highly correlated to the cell types of the sample and negatively correlated to the *tumour purity*, defined as the ratio of *aneuploid cells* exhibiting a non canonical number of chromosomes.

However, assumption of homoscedasticity of the residuals makes standard linear approaches sensitive to outliers, while they do not endorse explicitly the unit-simplex constraint (Equation (2)), requiring posterior normalisation of the coefficients.



(a) Graphical abstract, illustrating the fundamental linear assumption of bulk mixture construction underlying the cellular deconvolution framework.



(b) The deconvolution methods are classified according to their input data requirements as well as the output type and resolution they provide. Supervised, alternatively named partial methods, methods utilise markers, signatures, or cytometry proportions, to achieve cell detection (A), estimating cell proportions (B), correcting heterogeneity (C), or estimating cell type-specific expression profiles (D), ranked from the simplest to the most challenging task.On the other hand, complete deconvolution methods (E) simultaneously estimate cellular proportions and purified expression profiles. Reproduced from [SG13, Fig. 3].

Weighted linear approaches The presence of an unknown cell population might be relaxed by including a constant intersection term p_0, adding in practice a column of ones in the design matrix. To account for potential heteroscedascity (variance of the errors depends on the gene value), weighted linear approaches allow users to add prior weights to modify the *leverage* (contribution) of each gene to the computation of the OLS estimate. Considering $\boldsymbol{W}$ the diagonal matrix of weights, the Weighted version of the Least Square estimate is given by Equation (4):

$$\hat{p}_{\text{wOLS}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{y} \tag{4}$$

EPIC [Rac+17] combines this weighted approach with the addition of a column characterising the tumour profile in the signature matrix. [Rac+17] notably provides two signatures of circulating and tumour-infiltrating immune cells, CAFs (cancer-associated fibroblasts) and epithelial cells, respectively designed for whole-blood and solid tumoral tissues, aggregating bulk and scRNA-Seq data.

Instead, the quanTIseq [Fin+19a] algorithm integrates an additional constant intersection term to quantify the contribution of the unknown tumoral content. In addition, to address the issue of cell "drop-outs" (cell populations, generally infrequent and/or exhibiting a strong correlation with other cell types, that are wrongly estimated as absent), a heuristic approach is employed whereby the final Tregs estimate is computed as the average of two Tregs measures, in the presence and absence of the TCD4+ subset in the design matrix. Tregs are indeed highly correlated with TCD4+ cell populations.

In weighted linear approaches, individual gene contributions are usually provided by the user. Without prior knowledge, the usual approach is then to give less importance to genes exhibiting strong variability within a cell population. However, assigning appropriate weights to each gene typically necessitates either prior knowledge or strong assumptions about the dataset's distribution. We subsequently review in next Section 2.1.1 robust linear regression methods that compute the weights or trim outlying gene expression in a automated manner.

Robust Linear Regression and SVR Approaches for Automated Selection of Transcriptomic Markers In the previously described approaches, the inclusion of all genes in the regression framework may yield biased estimates when the expression of some genes significantly differ, due to significant changes of sequencing protocol or phenotype condition between the bulk mixture and purified expression profiles. Unfortunately, outlying genes in least-square approaches have the strongest influence on the parameters estimation, in reason of the Euclidean metric used to evaluate the prediction error.

Several robust methods, making a compromise between *efficiency* and *robustness* of the estimate, have been proposed. They are usually classified into *M-estimates* (see Definition A.4), whereby an adaptive function is enforced on the residuals, giving less weights to those with strong leverage, and *LTS estimates*, where a user-provided ratio of aberrant genes is automatically identified and trimmed (see Definition A.5).

With both methods, the weights assigned to each observation depend on the estimator which in turn depend on the weights. As a result, the robust estimator must be computed sequentially, these methods are accordingly referred to as Iteratively Reweighted Least Squares (IRLS) approaches. Uniform weights are usually assigned to each observation, subsequently, a standard least regression estimate is computed. Once the OLS obtained, each observation is reweighted, using the transformation induced by the *influence function*, and which usually depends on its leverage on the regression framework. The subsequent IRLS estimates are then computed with those new weights, and the process continues until convergence [Yoh87]).

The RCR (Robust Computational Reconstitution) deconvolution algorithm, by [Hof+06], notably couples the LAD (see Definition A.5) regression framework while adhering to the unit-simplex constraint (eq. (2)).

A variant of the LTS (least trimmed squares) approach has been implemented by the FARDEEP algorithm [Hao+19]. It has notably been modified to ensure convergence towards a final set of trimmed observations, in a linearly growing number of iterations. However, the algorithm is highly

sensitive to the tuning parameter that controls the final number of observations trimmed during the regression. And while convergence and consistency of the algorithm is guaranteed, there's no theoretical guarantee that the final estimate returned is indeed optimal.

Overall, all the variants proposed in this section are proned to overfitting. Indeed, since these weights are derived from the model's performance, they are highly sensible to dataset-specific patterns, leading to potential inconsistent and poor results on newly observed datasets. In addition, they are less efficient than the standard OLS estimate in case the Gauss-Markov assumptions hold.

Support-vector-regression are supervised machine learning algorithm featuring an alternative strategy to select genes. It turned out that in real-world experiences, they tend to exhibit increased robustness to noisy observations. The first historical mention to SVR approach, termed ϵ-SVR [CV95], uses a insensitive loss function, whose parameter ϵ is provided by the user to control the error rate tolerated on the outputs (see Definition A.6).

CIBERSORT (Cell Type Identification By Estimating Relative Samples Of RNA Transcripts), developed by [New+15], utilises the the $\nu-$SVR ([CC02]) variant. Instead of optimising the precision (error rate tolerance), the ν parameter controls the proportion of Support Vectors integrated in the regression framework ([Sch+00]) [2]. Compared to standard robust linear regression approaches, [New+15] exhibits the better performance of SVR methods with "spillover effects" (see Section 3.2), enabling them to integrate more closely related cell types in their analysis while providing a more robust and explainable model.

In practice, CIBERSORT implements the $nu-$ SVR approach with the svm function from R package `e1071` ([Mey+21]). CIBERSORT additionally provides a standalone web application, and relevant purified signatures. The most popular is the LM22 profile, a meta transcriptomic collection of 6 studies of 22 distinct immune cell types (see Section 3). The ImmuCC algorithm ([Che+17]) harnesses the implementation from CIBERSORT algorithm, with a new reference signature aggregating 25 cell types and tailored for murine deconvolution.

Correcting the Uncoupling Between RNA and Cytometry Fractions It appears that most of the existing deconvolution algorithms estimate the fraction of mRNA coming attributable to each cell type, rather than the underlying cell proportion itself. In other words, they assume *homogeneous* cell populations, e.g. they consider that each cell subtype exhibits the same RNA library depth ([Sos+21]). However, in real-world settings, this premise usually does not hold, for both technical and biological reasons. For instance, the RNA extraction efficiency may depend on the cell type, and its survival capacity to the lysis and extraction phase. Once the average production of total transcriptomic expression has been estimated (or phsyically measured), it becomes feasible to subsequently re-normalise the inferred cellular transcriptomic ratios, such that they align with the anticipated, biologically interpretable cellular ratios (see Equation (5)):

$$\hat{p_j^*} = K\frac{\hat{p_j}}{r_j}, \quad K = \frac{1}{\sum_{j=1}^{J}\frac{\hat{p_j}}{r_j}} \tag{5}$$

with r_j the average number of transcripts extracted per cell type, and K the normalisation constant.

Post-correction of this uncoupling is accounted in [Rac+17] and [Fin+19a] studies, with direct measures of the total expression of cell subtypes, as quantified with RNAeasy mini kit (Qiagen) and he Proteasome Subunit Beta 2, respectively [3].

When direct measures are not available, the MMAD (microarray microdissection with analysis of differences, [LHP14]) proposes an iterated approach for estimating the coefficient extraction efficiency, r_j. Yet, the regression framework is not anymore linear, and the new cellular estimate is computed using a non-linear conjugate gradient search algorithm.

[2] [CC02] demonstrates the equivalence between the two approaches: increasing the ν hyper-parameter results in a smaller ϵ-tube and a higher precision on the results. Asymptotically, determining the ν-proportion of support vectors reaching a given precision $\hat{\epsilon}$, is even equal to the output of the ϵ-SVR with that degree of precision.

[3] In the back-end, they utilise the expression of the *housekeeping genes* as a surrogate variable of the absolute number of transcripts produced by the cell population

Linear Regression Approaches with Explicit Unit-Simplex Constraint All the previously described algorithms do not explicitly integrate the unit-simplex constraint Equation (2) during the estimation process, and re-normalise instead, posterior to the estimation, the inferred ratios.

The NNLS (Non Negative Least Squares) estimate relies on the Lawson Hanson algorithm [HH81], and its output is often provided as a reference in most review papers benchmarking deconvolution algorithms ([Stu+19], [JL21]). The nnls function from the R `limSolve` package can be used to solve this optimisation problem.

The Least Squares with Equality and Inequality Constraints (LSEI) generalises this approach by enforcing both non-negativity and sum-to-one constraints. The *lsei* function in R, from limSolve package, can be used to solve the corresponding optimisation problem. The Matlab *lsqlin* function, returning the same output as *lsei*, is used by the Bioconductor package `DeconRNASeq` ([Gon+11], [GS13]) .

Both algorithms belong to the class of *QP* (quadratic programming), which aims at optimising a system of linear, convex functions, with a guaranteed unique solution.

Regularised linear regression When the number of cell types J exceeds the number of transcripts G, the deconvolution problem stated in Equation (1) is *undetermined*, with potential infinite set of solutions verifying the set of G equations. Several regularised linear approaches have been implemented to deal specifically with problems where the number of unknowns exceeds the number of variables (see Definition A.7).

The DCQ algorithm [Alt+14] uses in particular the **Elastic Net** regularisation, a compromise between the L1 and L2 penalties proposed by the Lasso and Ridge methods. In R, the *glmnet* [Fri+11] offers a straightforward and versatile implementation of the method. The benchmark study led by [JL21] exhibits the reduced performance of deconvolution methods applying these regularised approaches. However, a comprehensive analysis of the settings used to conduct the benchmark study show that they somehow miss the point: penalised linear regression approaches are not intended to retrieve the cell ratios of a given biological sample, but rather retrieve the optimal *support* of cell populations that induce transcriptomic variations from a biological state to another. Implicitly, these methods assume that the proportions of most cell populations do not vary over time.

To illustrate the point, DCQ has been used to identify the dynamical evolution of immune cell ratios during influenza infection. Indeed, dozens of immune cell types coordinate their efforts to maintain tissue homeostasis. Precisely, DCQ studied the evolution dynamics of up to to 213 immune cell subpopulations in mice lungs for ten time points and retrieve significant changes in 70 immune cell type ratios.

Two years after, the ImmQuant package [Fri+16] offers a user-friendly tool for inferring immune cells in both human and mice organisms. The pipeline includes automatic data import and cleansing, selection of the marker genes, deconvolution of the biological samples provided and visualisation of the output.

2.1.2 Probabilistic-based approaches

The second family of methods for inferring cellular ratios from purified reference profiles utilises probabilistic models to capture the generative process underlying the bulk expression production. Interestingly, these approaches naturally address the unit-simplex constraint (Equation (2)), provide a more accurate representation of the discrete nature of transcript counts and can even account for an unknown cell population or individual variations of the gene expression. In particular, these approaches accurately reproduce the commonly observed correlation between the mean and the variance of the gene expression ([Lob+08]).

Since a large number of parameters might be introduced in these models, it is common practice to represent the conditional independence relating them using a directed acyclic graph and the homogenised notation illustrated in Section 2.1.2.

Discrete probabilistic approaches Latent Dirichlet Allocation (LDA) is a straightforward approach to model abundances (see also [BNJ03] and Definition A.8). The NNML (Non-negative maximum likelihood model) algorithm, by [Qia+12], extends the frequentist LDA model adopting a Bayesian approach. Precisely, the prior distribution of the cell ratios is modelled by a symmetric Dirichlet distribution. This kind of distributions exhibits several advantages: it naturally endorses the unit-simplex constraint Equation (2) and streamlines the integration of prior knowledge, such as equibalanced hypothesis or inclusion of cytometry measures [4]

Extensions of the NNML algorithm introduce generative models that relax controversial assumption, such as the completeness (no unknown cell population) or the validity (no sample-specific variations of the purified signatures) of the reference profile. However, these probabilistic frameworks often require **regularisation** strategies, classified as "hard" and "soft" constraints, to ensure problem *identifiability*. Practical regularisation strategies often rely on strong constraints and assumptions about the distribution of purified expression profiles. They must balance the trade-off between introducing too much bias and risk overfitting, or insufficiently define the problem and suffer from *ill-conditioned* modelling.

To that end, the `ISOLATE` algorithm ([QM09]) assumes that the expression profile of any gene of the unknown cell type can be rewritten as the expression of one of the cell types already described, up to an additional multiplicative perturbation described by an uninformative Gamma prior. In a tumoral context, this constraint can be interpreted as a change of gene expression induced by heterotypic tumoral conditions, on an unique cell population subset, termed CSO in the paper (cancer site of origin). The basic framework described above has been extended in the `ISOpure` algorithm ([Quo+13]). Unlike the naive approach, `ISOpure` not only computes a shared cancer profile common across all samples but also refines it to incorporate sample-specific variations in tumoral expression. However, the CSO assumption only holds if the mutations concern only one cell line, an assumption that usually does not hold in intricate TMEs, wehre both tumoral and normal cell lines expression are impacted by the clonal growth.

Accordingly, the NNML_{np} algorithm ([Qia+12] and Section 2.1.2) assumes instead that the transcriptomic profile of the unknown cell type can be rewritten as a potential convex combination of all (possibly a subset) the included cell populations. Biologically, this approach hypothesises that the tumoral part of the sample is not a new cell line, but rather a mixture itself of the original cell populations, whose expression has been altered upon tumoral mutations, or changes induced by the new conditions of the medium. Their approach is nonetheless hindered by the stringent regularisation assumption that the perturbation factor for a given gene is the same across cell populations.

The PERT algorithm ([Qia+12] and Section 2.1.2) relaxes the strong assumption that the purified cell expression profiles are representative of the expression profiles of the mixture. Specifically, the vector representing the expression profile of a cell population is altered through a multiplicative perturbation factor $\boldsymbol{\rho}_G$, which is gene-specific and sampled from a non-informative Gamma distribution with an average value of 1.

TEMT (Transcript Estimation from Mixed Tissue samples, Section 2.1.2), by [LX13], harnesses directly the reads (sequence of nucleotides) themselves, instead of raw RNA-Seq counts. This approach enables to account for multiple transcripts resulting from *alternative splicing* ([Cam+20, Chap 14]) and technical biases issued from read sequencing itself [5]. The methodology is thus particularly relevant for decomposing, and correcting technical artefacts from relevant biological signal, and can be used as an alternative normalisation method for making samples comparable, regardless of the sequencing platform used.

This approach uniquely incorporates technical artefacts into the deconvolution process, addressing the assumption made by other methods that input data has been corrected for such noise. Additionally, it estimates an unknown cell profile, in a process similar to the NNML_{np} approach.

[4] To note, the Beta distribution is a variant of Dirichlet distribution with two-component mixtures, used as prior for binomial distributions.

[5] Technical artefacts in RNA-Seq encompass length, positional and amino bias. For instance, longer transcripts may yield more counts ("effective length"), while sequence-related biases include over-transcription around transcript ends.

The complexity of the likelihood or the posterior function requires specific optimisation methods to retrieve the relevant parameters: PERT and NNML uses a conjugate gradient descent algorithm, while TEMT and the ISOLATE algorithm utilise a variational online EM [DLR77]. Since diverse regularisation strategies do not address the same biological constraints, and often require different optimisation strategies, [QM09] suggests to systematically benchmark the method against manually annotated tumours, as evaluated by pathologists.

Continuous probabilistic approaches The `Demix` generative model, by [Ahn+13], and its direct `DemixT` extension,by [Wan+18], infer the proportion and expression profile of the tumoral content, in a two and three-component mixture, respectively. Briefly, Demix(T) models the distribution of the bulk expression for each gene as a convolution (sum of independent variables) univariate log Normal distributions (see Section 2.1.2), each purified profile parametrised by its own parameters, inferred prior to the study. For the sake of comparison, a generative model based on a convolution of Normal distributions is also compared to the log Normal approach. This model streamlines the estimation process as a closed-form can be derived for the log-likelihood. However, the $\log_2$-transformation required to endorse the assumptions of the model is likely to disrupt the fundamental linearity deconvolution assumption (Equation (1)).

Modelling the mixture problem as a convolution offers several advantages, including the elimination of a residual error term to account for the stochasticity of the resulting bulk profile, and the utilisation of distributions that accurately depict the inherent compositional characteristics of RNA-Seq datasets.

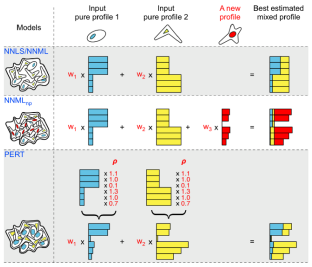
However, no explicit form for the convolution of $\log_2$-normalised variables is known, and an iterated conditional modes-like ([Bes86]) approach [6] is used to maximise the log-likelihood of the resulting generative model:

- The unknown general parameters of interest (cellular proportions and mean and variance of the tumoral profile), are determined by maximising the log-likelihood of the generative model depicting the convolution, conditioned on the previously known mean and variance for healthy cell populations. Since the closed form of the log-likelihood is not known for a convolution of log-Normal, it is approximated through numerical integration (not needed with a convolution of Normal distributions), and the MLE is obtained using a *Nelder-Mead* procedure.
- In a second time, tumoral profiles are estimated by plugging-in the parameters estimated in the previous step. With a two-component model, the unit-simplex constraint (Equation (2)) and the fundamental linear deconvolution assumption (Equation (1)), only one degree of freedom, or unknown, namely the tumoral content, must be inferred (see [Ahn+13, Eq.1]).

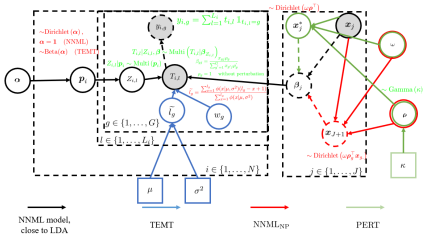
[Erk+10] implements instead a Bayesian framework, `Dsection` (see Section 2.1.2, in which the bulk expression of each gene in each sample, y_{gi}, follows a Normal distribution whose parameters are stochastic variables rather than point values. For instance, the distribution of the inverse of the variance, referred to *precision* in the paper, is modelled by a Gamma distribution.

The posterior distribution of individual cell-specific expressions and bulk gene variances is identifiable to known density distributions (*conjugate* priors). However, the posterior distribution of cellular ratios lacks a known density distribution due to the intractable integration of the normalising constant. The Metropolis-Hasting algorithm is employed to sample this posterior distribution, which is only known up to a normalising constant, while Gibbs sampling is used to retrieve simultaneously the joined posterior distributions of the whole set of parameters composing the generative model. Note that in opposition to the Demix(T) approach ([Ahn+13]), the variance of the bulk expression is uncoupled to the individual variance of the purified cellular profiles.

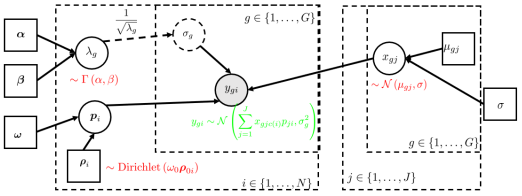
[6] The parameters are iteratively maximised, conditioned on the current updated value of the remaining subset of parameters, rather than simultaneously



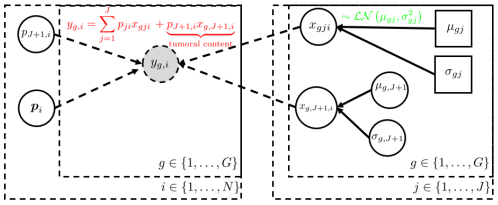
(a) To each biological requirement, its suited probabilistic model. The non-negative least squares model (NNLS) and the non-negative maximum likelihood model (NNML) can only predict proportions of pre-specified reference populations. In scenario ii), the non-negative maximum likelihood new population model (NNMLnp) can additionally account for an unknown cell population, while in scenario iii) the perturbation model (PERT) can integrate sample-specific variations. Reproduced from [Qia+12, Fig. 1].



(b) DAGs of the generative model described in section 2.1.2. All the discrete probabilistic models derived from the LDA generative framework. This DAG notably encompasses, using different colour notations, the NNML, NNML_{NP} and PERT algorithms ([Qia+12]), along wihth the TEMT model [LX13].



(c) Graphical representation of the Demix(T) ([Ahn+13] and [Wan+18]) probabilistic model.



(d) Graphical representation of the Dsection [Erk+10] probabilistic model.

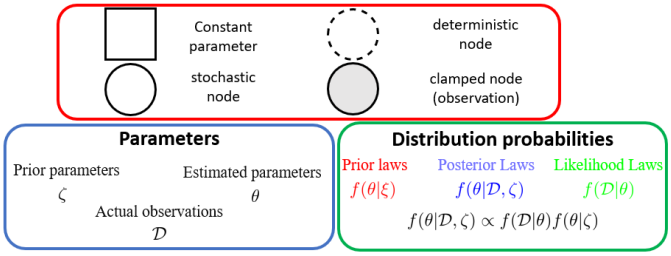


Figure 5. Partial probabilistic models to infer cellular ratios.We follow the RevBayes convention to homogenise indexes and parameters across a set of generative models. Notably, the *likelihood* density functions describing the distribution of the observations, are in green colour while the prior distributions of the parameters to estimate are in red colour.

2.2 Marker-Based Approaches: Pathway Enrichment Analysis and Hyper-Geometric Scores

Some deconvolution algorithms simplify the estimation process by adopting a marker-based paradigm. The definition of "markers" genes has gradually broadened, from designing genes uniquely expressed in a cell a population to include genes comprehensively expressed in one cell type relatively to other cell groups. Marker-based relied historically on strong definitions of *marker* genes ([GPT07], [CSC10]), however, nowadays, *weak* markers approaches are favoured (markers are only required to be consistently over-expressed in a given cell population), since they also enable to delineate closely related cell types.

These markers can be derived through either knowledge-driven approaches ([Ang+15], [Roo+15]) or data-driven methods [CZS15], [Bec+16], [Zha+17]. The initial data-driven strategy for identifying marker genes involved identifying genes whose mean expression value in a give cell population consistently exceeded the expression value measured across other cell types ([Sho+12], [CZS15]). More robust statistical approaches, evaluating the relevance of selected markers through the computing of empirically estimated$p$-values, have been developed since then, ranging from SNR (signal-to-noise) ratios [Bec+16], to the F-statistic ([Wan+10]) through the Gini index ([Zha+17]).

Integrating the definition of a gene marker into the fundamental presumption of linear deconvolution simplifies framework Equation (1)) into Equation (6):

$$
\begin{aligned}
y_{\forall g \in \widetilde{G_j}} &= \sum_{j^{'}=1}^{J} x_{gj^{'}} \times p_{j^{'}} = x_{gj} p_j, \\
&\text{since by definition } x_{gj^{'}} = 0, \forall j^{'} \neq j \\
\begin{pmatrix} \boldsymbol{y}_{\widetilde{G_1}} \\ \boldsymbol{y}_{\widetilde{G_2}} \\ \vdots \\ \boldsymbol{y}_{\widetilde{G_J}} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{x}_{\widetilde{G_1},1} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_{\widetilde{G_2},2} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{x}_{\widetilde{G_J},J} \end{pmatrix} \times \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{pmatrix}
\end{aligned}
\tag{6}
$$

with the following notations:

- $\widetilde{G} = \{1, \dots, G\}$ is the set indexing the total number of genes selected in the signature matrix (we introduce the tilde as a shorthand indicator for a set).
- $\widetilde{G_j} \subset \widetilde{G}$ is the subset of genes expressed uniquely in cell population $j \in \widetilde{J}$
- We additionally assume the unique existence of a *partition* $\widetilde{G}$, shared across samples, such that $\widetilde{G_j} \cap \widetilde{G_l} = \varnothing$, $\forall (l, j) \in \widetilde{J}$, $l \neq j$ and $\bigcup_{j=1}^{J} \widetilde{G_j} = \widetilde{G}$.
- We introduce the shorthand $\boldsymbol{y}_{\widetilde{G_j}}$ and $\boldsymbol{X}_{\widetilde{G_j},j}$ to respectively denote the measured expression of the market set $\widetilde{G_j}$ in the bulk mixture, and its respective expression in the purified cell population j.

If eq. (6) holds, the bulk expression associated to a gene marker set is proportional to the expression of the cell population associated to this marker, the multiplicative constant being the ratio associated to this cell type, p_j.

However, as already specified in Section 2.1, the presence of technical noise or intrinsic biological stochasticiy usually renders the system of equations inconsistent. Assuming the same framework detailed in Section 2.1.1, the Normal Equations give the following OLS solution (Equation (7)):

$$
\hat{p_j} = \frac{1}{|G_j|} \sum_{g \in \widetilde{G_j}} \frac{y_g}{x_{gj}} \tag{7}
$$

with $|G_j|$ the module, namely the number of genes composing the marker set of a cell population.

Once specific markers for each population have been identified, the estimation of cellular ratios relies either on *abundance score* (see Section 2.2.1) or *enrichment score* (see Section 2.2.2 and Section 2.2.3).

2.2.1 Abundance scores

Historical endeavours, by [GPT07] and [CSC10], assume the strong definition of a marker (section 2.2) holds, and the cellular ratios that were returned correspond to the estimates given in eq. (7). [CSC10] only differed by the addition of a *link function*, precisely a *log2* transformation to reduce the noise bias associated to small ratio values, applied to the bulk and purified profiles.

Later, the MCP (Micro-environment Cell Populations)-counter, by [Bec+16], adopts a weak marker paradigm, and replaces the abundance score given in Equation (7), by the geometric mean of the genes characterising a given cell population (eq. (8)):

$$\mathrm{ES}(\widetilde{G_j} \in \widetilde{G}) = \left(\prod_{g \in \widetilde{G_j}} y_j \right)^{1/|\widetilde{G_j}|} \propto p_j \tag{8}$$

2.2.2 Enrichment scores, based on KS metric

Most of the methods computing an enrichment score rely on a variant of the weighted enrichment-based method named ssGSEA, for single-sample gene set enrichment analysis ([Sub+05] and [Bar+09]). The computation of enrichment scores, based on the Kolmogorov–Smirnov metric, is reported in Definition B.1, while its main limitations.

[Yos+13] implements the ESTIMATE metric to compute immune and stromal enrichment scores in tumoral samples. Te best link function coupling the purity score (proportion of tumoral cells) with the ESTIMATE measure was computed with the https://en.wikipedia.org/wiki/Eureqa software. [ASB15] implements an extension of this method integrating orthogonal modalities. Precisely, the tumour purity score is computed from four distinct sources: the ESTIMATE score itself , ABSOLUTE (quantify the proportion of cancer cells based on the number and location of somatic copy-number mutations), LUMP (correlation between the degree of methylation and the tumour proportion) and immunehistochemistry image analysis.

[Roo+15] and [Ang+15] uses GSEA-based metrics to compute the tumoral activity and relate it to mechanisms involved in immune tumour resistance. [Ang+15] notably demonstrates the co-existence of two kinds of tumoural environments, distinguishing hypermutated tumours showing upregulation of immunoinhibitory molecules from non-hypermutated and stagnant tumours, enriched with immunosuppressive cells.

[Şen+16] infers gene markers for 24 distinct cell populations in 19 cancer types. With these enrichment scores, they demonstrate that the over-expression of Th17, CD8+ and Tregs increases chances of survival, while strong activity of Th2 cells is correlated with a negative prognostic.

Ultimately, the `xCell` algorithm, by [AHB17], claims to identify up to 64 distinct cell types, including immune and stromal ones, derived from a compendium of 1822 purified transcriptomic cell lines. *Calibration*, using a power link function to couple abundance scores with true cell ratios, and reduction of the multi-collinearity of the signature matrix to avoid "spillover" effects, underlie the originality, and robustness of the method.

Finally, TIminer, by [Tap+17], is a free Docker pipeline, aggregating the marker sets of [AHB17], [Ang+15] and [Cha+17]. It was initially designed for estimating the proportion of infiltrated immune cell types, along with neoantigen prediction and tumour immunogenicity.

2.2.3 Enrichment scores, based on alternative metrics

Alternative strategies can be employed to compute enrichment scores, such the hypergeometric test (see Definition B.2).

[BUK11] implements SPEC (Subset Prediction from Enrichment Correlation) to predict which cell population is more likely to contribute to an observed change in the gene expression, based on Pearson correlation. SPEC notably demonstrates that the main resistance mechanism of the gold-standard treatment against Hepatis C was the cross-interaction between the myeloid cells and the anti-interferon therapy.

[Sho+12] uses the z-score (negative log10 of p-value), resulting from a Fisher's exact test.

The Bioconductor package BioQC, by [Zha+17], computes abundance scores by evaluating the relevance of median differential expressions with a non-parametric *Wilcoxon-Mann-Whitney* test.

In conclusion, marker-based methodologies provide abundance scores that are only proxy of relative cellular ratios. [AHB17] and [Yos+13] attempt to mitigate this issue, by learning a link function coupling these two features. Overall, these restrictions render marker-based methods impractical for intra-sample comparisons, in contrast to the signature-based methods, discussed in previous Section 2.1 (see also Appendix B.3).

We outline the major categories of deconvolution algorithms used to estimate cell ratios in a heterogeneous biological sample in Figure 6:

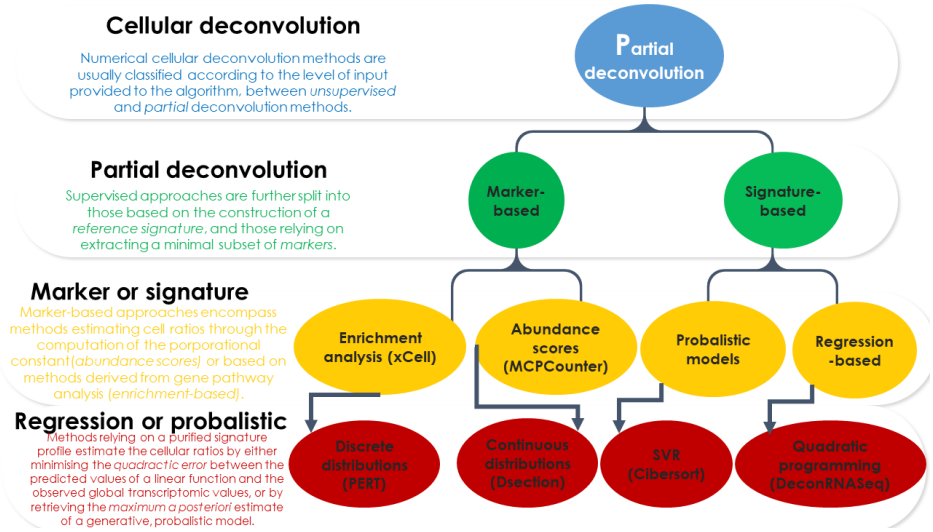


Figure 6. General classification of partial-based deconvolution algorithms.

2.3 Reference-Free Approaches: Simultaneous Deconvolution of Cell Fractions and Purified Expression Profiles

Complete deconvolution algorithms attempt to simultaneously estimate both the proportions and the pure expression profile of cell types [SG13] from the bulk profile alone, namely minimising the following quantity (Equation (9)):

$$\left(\hat{\boldsymbol{P}}, \hat{\boldsymbol{X}}\right) = \arg\min_{\boldsymbol{P},\boldsymbol{X}} \{|\boldsymbol{Y} - \boldsymbol{X} \times \boldsymbol{P}|\} \quad \boldsymbol{Y} \in \mathbb{R}_+^{G\times N}, \boldsymbol{X} \in \mathbb{R}_+^{G\times J}, \boldsymbol{P} \in \mathbb{R}_+^{J\times N} \tag{9}$$

Without further information, the system of equations described in Equation (9) is *undetermined*, having either an infinite set of solutions or no one at all. Hence, the identifiability of the unsupervised deconvolution problem require strong assumptions on the distribution.

2.3.1 Unsupervised approaches

[Ven+01] proposes the first version of a reference-free approach, inspired from Gaussian mixtures, to deconvolve colon cancer samples, from which two clusters, on a total of four identified, could be labelled with strong evidence as hematopoietic and fibroblast cells. [Ven+01] also demonstrates that the marker-based assumption (see Section 2.2) is a necessary condition for the existence and uniqueness of the system of equations (Equation (9)).

Repsilber and colleagues then extended the method proposed by [Ven+01], by solving Equation (9) using a Non-Negative Matrix Factorisation algorithm. NMF notably guarantees that both $\boldsymbol{X}$ and $\boldsymbol{P}$ are strictly non-negative (see Definition C.1 and [Rep+10]), as reported in Equation (10):

$$\begin{aligned} &\min_{\boldsymbol{P},\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{X}\|_F^2 \\ &\text{subject to the non-negativity constraints:} \\ &\boldsymbol{P} \geq 0,\, \boldsymbol{X} \geq 0 \end{aligned} \tag{10}$$

Variants of the NMF approach were used in UNDO, by [Wan+15] and CAM, by [Wan+16], methodologies. The Convex Analysis of Mixtures (CAM) enforces both the non-negativity of the outputs returned, and the unit-simplex constraint Equation (2) for the ratios. Precisely, these convex geometry-based methods project the resulting bulk expression matrix $\boldsymbol{Y}$ into a J-dimensional *polytope*, whereby each cell population profile forms a convex hull whose vertices are the marker genes of the so-called cell population. The final set of convex solutions are the ones covering the most precisely the facets of the convex hulls derived from the bulk profile. CAMTHC, by [Che19], for Convex Analysis of Mixtures for Tissue Heterogeneity Characterisation, and *CAMfree*, by [JL21], are both R package implementing the CAM methodology.

2.3.2 Semi-supervised approaches integrating prior information

Since then, semi-supervised approaches, coupling partial prior knowledge of markers associated with a cell type with numerically inferred `de-novo` molecular markers, enable to increase the identifiability of the problem by reducing the set of possible solutions. Semi-approaches directly extending [Wan+16] have been implemented in R, as packages *CAMmarker* and *CellMix* [7] The usual approach to integrate prior information is to constrain all input values of the purified expression profile to zero, except whether the gene has formally been associated with a cell population.

Closely related is the semi-CAM approach, by [Don+20]. In details, the semi-CAM approach is a two-step estimation procedure; first, it identifies the final gene partition for the deconvolution process, assigning each unlabelled gene to its most probable cell type, given the already identified marker genes. To achieve this, it enhances the k-means clustering employed by the CAMfree approach, whereby the initial centroids are the vertices covering the most the convex hulls, by incorporating known marker information into the cluster centre construction. Whenever known marker genes for partially described cell types are available, [Don+20] demonstrates that the semi-CAM method outperforms the unsupervised historical CAMfree method.

The Digital Sorting Algorithm (DSA, [Zho+13]), is another semi-supervised approach, adopting a EM-like approach. Precisely, the cellular ratios and the purified expression profiles are iteratively estimated, conditioned on the current update of the remaining parameters, until convergence. Prior information can easily be integrated as initial values for either cellular ratios or purified expression profiles. However, the identifiability of the problem still requires the marker assumption.

Overall, all the methods described in this section are much more sensitive to the quality of data provided, especially when no prior information is provided.

[7] *CellMix*, by [Gau13], benchmarks a whole set of deconvolution methods, in particular, *ssKL* and *ssFrobenius* that solve optimisation problem Equation (10) by minimising the KullBack Leibler divergence and the Frobenius norm, respectively.

3 Outline of the Cellular Deconvolution Procedure

The estimation of the composition of a biological sample is only one of the steps composing the deconvolution framework. In the remainder of the text, we define as *pipeline* this whole process, ranging from the pre-processing and collection of purified profiles to the downstream analyses, while the term "algorithm" only refers to the estimation stage itself.

A standard cellular deconvolution pipeline typically involves the following main steps:

1. **Data Collection:** This step, illustrated in the stage 1, in Section 3, involves the selection of tissues and purified cell populations, subsequently followed by their preprocessing. The choice of tissues and cell populations should be guided by the biological objective the experimenter pursues and the characteristics of the bulk profiles to deconvolve.

2. **Gene Marker Selection and the Refined Construction of Cellular Signatures:** Partial methods inferring cell ratios requires an additional step consisting of identifying and characterising a subset of genes, able to delineate all the cell populations ought to compose the mixture. Following the identification of gene markers, the purified expression profiles should undergo the same preprocessing operations and transformations as the bulk mixture samples, including the removal of unwanted batch effects induced by technical artefacts. This step is illustrated in Section 3, part 2.

3. **Parameter Estimation:** This step refers to the deconvolution algorithm itself (stage 3, Section 3). The type of tissue or/and organism to deconvolve along with the objective biological goal guide the final choice of the algorithm used.

4. **Evaluate the output:** This step involves the formulation of statistical tests to assess the presence of a cell population within the sample (intra-sample comparison) or to compare two cell fractions across different biological conditions (inter-sample comparison). Surprisingly, there is a notable absence of robust and widely accepted methods proving theoretically the consistency and precision of the outputs returned by most deconvolution methods. Alternatively, it is possible to benchmark the performance of a new deconvolution algorithm against gold-standard deconvolution methods and against cytometry data.

5. **Visualisation and biological interpretation:** Ultimately, various visualisations and expert validations play a pivotal role in verifying the precision and biological relevance of the algorithm in deciphering disease mechanisms, or providing new biomarkers (see stage 4, in Section 3).

A practical use case, with the construction and the application of the LM22 signature in conjunction with the CIBERSORT algorithm is reported in [Che+18]. In the following sections, we provide a summary of guidelines for enhancing the performance of deconvolution algorithms, based on insights drawn from recent benchmark studies.

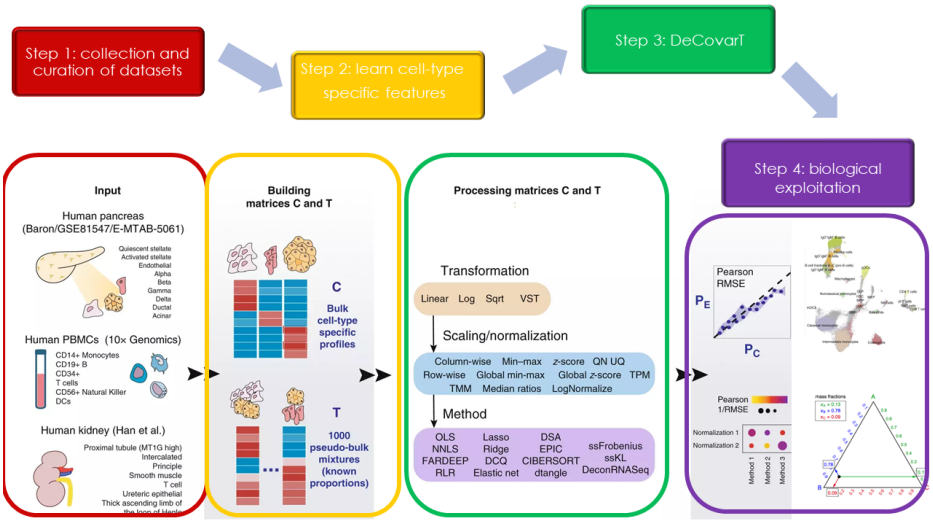


Figure 7. Workflow for bulk deconvolution methods. Inspired from [Avi+18, Fig. 1].

Since we have already elaborated in the previous section 2 about deconvolution algorithms, we now focus on preliminary steps that are likely to improve the performance and reproducibility of any deconvolution algorithms.

In section 3.1, we provide general guidelines for selecting datasets to create purified cellular signatures, underscoring the significance of obtaining samples that closely match the phenotype condition and technology platform of the bulk mixtures to be deconvoluted.

In section 3.2, we review strategies for identifying the minimal subset of genes that best discriminates the cell populations included in the signature matrix.

3.1 Guidelines for Data Collection

3.1.1 Guidelines for Cell Population Selection in Data Collection

Many deconvolution methods are highly sensitive to the absence of cell subtypes in the reference signature, yielding the best estimates when the reference profile faithfully represents the actual composition of the biological sample [Stu+19].

These discrepancies, most pronounced in the absence of closely correlated or orthogonal cellular profiles, lead to the "spillover" phenomena ([SG13], [Fa+20]). For instance, [Hao+19] demonstrates substantial reduction in estimating the cellular ratios of moncotyes, when myeloid dendritic cells are not included in the reference profile, despite being truly present in the mixture.

On the other hand, *background prediction* refers to erroneous identification of a cell population as being present in a mixture. This issue is even more pronounced with marker-based methods (section 2.2), assuming transcriptomic markers are associated with an unique cell population.

Overall, Cibersort [New+15], CibersortX [New+19] and MuSiC [Wan+19] are the least sensitive to the presence of undescribed highly-correlated or rare cell types in the mixture ([JL21]).

3.1.2 Guidelines for Phenotype and Tissue Selection in Data Collection

To mitigate the recommendations of constructing the most representative cell signature, we should highlight that comprehensive and simultaneous estimation of the whole array of cell populations composing the mixture is usually infeasible.

Firstly, some rare cell types may remain unprofiled, in particular, tumoral profiles are complex to dissect. Tumoral microenvironments display significant variability and plasticity, characterised by distinct mutation patterns, and intra-tumour heterogeneity resulting from the joint presence of diverse

tumoral subclones ([Bok+22]). In addition, somatic mutations in native cell lines may lead to the loss of certain markers, posing challenges in defining pro-metastatic immune cell subsets ([Boe+22]), especially for marker-based approaches.

TIMERtumour, by [Li+16] and *EPICabsolute*, by [Rac+17], are computational methodologies specifically tailored to quantify the level of infiltration and contamination of tumoral tissues by immune cells. Yet, none of the existing deconvolution methodologies address the intra tumoral heterogeneity, stemming from the potential presence of distinct tumoral subclones ([Yu22]).

[Rac+17] additionally pinpoints that the actual deconvolution solutions for unravelling tumoral heterogeneity are targeted towards decomposition of *solid tumours*, rather than "liquid" tumours, such as haematological malignancies (leukaemia).

Secondly, there is no unique and consistent nomenclature for identifying immune cell subsets, as translating functional insights into reliable phenotypic definitions based on protein markers is challenging ([ALH21]). It is noteworthy to mention that a suite of R packages, `ontologyX` [GRT17], specially tailored to store biological annotations in a structured and tree-like format, have been developed in order to homogenise cell nomenclature with updated ontologies, integrating updated cell atlases ([Lew20]) and dictionary of immunological terms ([Uni23]).

Thirdly, it is strongly deterred to incorporate cell populations from different hierarchical levels in the analysis, as this may lead to increased multicollinearity or even violate the independence assumption between purified expression profiles. The best results are typically achieved by constructing signature matrices at the finest level of granularity, as they mitigate "dropouts" effects by better delineating closely related cell types.

In order to compute back the contributions of the parental and higher-ranked cell lines, [Stu+19] provides the R function map_result_to_celltypes in the `immunedeconv` package, which automatically aggregates estimated descendant ratios to compute the parental fraction (or even cell lines separated by further layers of lineage).

Ultimately, bad characterisation of cell populations may stem from existing intra-variability within a cell population, which results from asynchronous dynamics, such as the coexistence of different phases of the cell cycle.

While in controlled conditions, such as cell cultures, chemical arrest or nutrient starvation can achieve synchronisation of the cell cycles [Bar+08], it becomes a challenging task when profiling living tissue [8].

Sample-specific events, such as *heterotypic* contamination (for instance, infiltrates of blood circulating immune cells, [Cha+19]), disease-induced ([Gau13]) or microenvironment dysregulations ([TPZ20]) may additionally alter the transcriptomic profiles of purified cell lines.

Accordingly, to mitigate the significant loss of performance commonly observed between artificial benchmarks and real-world conditions, it is recommended to collect purified profiles in a variety of tissues, or at least representative of the phenotype condition of the bulk profiles to deconvolve [9]. The performance of deconvolution algorithms in real conditions depends more on the representativeness of cell types profiled in the signature and environmental conditions than the choice of the regression or probabilistic framework, as discrepancies between the phenotype and tropic conditions of purified samples, compared to bulk profiles, can introduce significant bias and reduce model accuracy ([SFL20], [Cai+22]).

As a final side note, we quote [Stu+19], who believed that the "improvements made to signature matrices largely outweigh potential algorithmic improvement". We refer the reader to Section 3.1.2

[8] For instance, the CD3 marker, commonly used to define T cell subsets, may exhibit variable expression levels or even be entirely absent, depending on the cell cycle phase.

[9] Unfortunately, this recommendation is rarely observed, for instance, the expression profile of eosinophils, in the LM22 signature of Cibersort ([New+15]) was solely estimated from three distinct samples, from the same cohort.

providing general guidelines on the best signature to harness, with respect to the cell populations profiled.

Cell type	Recommended methods	Overall performance	Absolute score	No background predictions
B cell	EPIC	++	++	+
	MCP-counter	++	–	–
T cell CD4+	EPIC	++	++	–
	xCell	++	–	++
T cell CD4+ non-regulatory	quanTIseq	+	++	+
	xCell	+	–	++
T cell regulatory	quanTIseq	++	++	–
	xCell	++	–	++
T cell CD8+	quanTIseq	++	++	–
	EPIC	++	++	–
	MCP-counter	++	–	–
	xCell	+	–	++
Natural Killer Cell	EPIC	++	++	+
	MCP-counter	++	–	–
Macrophage / Monocyte	xCell	–	++	
	EPIC	+	++	+
	MCP-counter	++	–	–
Cancer-associated fibroblast	EPIC	++	++	+
	MCP-counter	++	–	–
Endothelial Cell	EPIC	++	++	+
	xCell	++	–	++
Dentricic cell	None of the methods can be recommended to estimate overall DC content. MCP-counter and quanTIseq can be used to profile mDCs.			

Figure 8. Guidelines for the selection of a deconvolution algorithm. The *overall* metric quantifies the correlation between the inferred fractions with the initial parameters used in the benchmark. The *background prediction* is a proxy of of the inclined of a deconvolution method to forecast the presence of a cellular classification, even when absent in the mixture. Reproduced from [Stu+19, Table. 2].

Figure 9. Outline and general guidelines for practical application of deconvolution algorithms.

3.2 Guidelines for the Refined Construction of Cellular Signatures

3.2.1 Best Strategy for the Selection of Marker Genes in Cellular Signatures

Regarding the construction of a signature matrix, [Avi+18] emphasises that pre-filtering genes exhibiting the strongest differences between cell types improves the robustness and reproducibility of the algorithm. These techniques for refining the final subset of genes included in the signature matrix fall under the general *feature-engineering* machine-learning framework. Feature selection usually refers to the preprocessing stage that filters irrelevant variables before applying the model itself ([GE03]).

Precisely, partial deconvolution methods based on signature profiles (Section 2.1) typically employ the "one-vs-all strategy" to identify the minimal set of transcripts consistently expressed in a given cell population, compared to all others. This strategy notably aims to reduce gene expression variance within a given cell type while simultaneously maximising the variance between different cell populations. However, once concatenated, the number of identified markers is still usually intractable to perform deconvolution tasks, and the resulting signature matrix often exhibits strong multicollinearity.

To select the genes in a global approach, the most common approach, for models based on regression optimisation, relies on optimising the *condition number* of the final reference matrix. In short, the idea is to identify the subset of quantified genes whose combined expression in the transcriptomic expression profile has the smallest condition number. This approach is particular favoured within linear regression-based methods (see Section 2.1.1), [New+15] notably demonstrates that minimising the *condition number* of the signature matrix effectively reduces its multicollinearity and improves the performance and robustness of the deconvolution algorithm (see also Appendix D).

3.2.2 Best Strategy for Normalisation and Transformation in Cellular Signatures

Several benchmarks have recently been developed to compare the performances of numerical deconvolution methods in relation with the preprocessing protocol chosen to normalise datasets ([Fa+20]) or the noise structure and magnitude ([JL21]).

[Fa+20] defines *data normalisation* as the set of techniques to make samples' distribution comparable, including universal scaling methods (min-max, z-score, row or column-wise). It also encompasses more specific methods, such as *TPM* or *FPKM*, to account for variations of the library size and depth. On the other hand, *data transformation* refers to the *link function* applied on raw datasets, such that the assumptions underlying the generative model hold.

[Fa+20] exhibits that *scaling* methods, such as *row scaling*, or *z-score*, which are used to smooth extreme values, decrease overall the performance of the deconvolution algorithms. In addition, [Fa+20] demonstrates that applying log-normalisation leads to suboptimal performances while the best results are reached without transforming the data, conclusions consistent to the findings from [Zho+13]. Indeed, [Hof+06] shows that the *log2* transformation, while better guaranteeing the normality requirements on the distribution of the residuals, breaks the fundamental linear assumption (Equation (1)).

[JL21] suggests to apply the same transformations on both the purified signature matrix and the bulk matrix expression, with the best performances obtained with TPM (Transcripts Per Million) normalisation. [Rac+17] indeed suggests that the TPM normalisation, as a *linear mapping*, naturally enforces the unit-simplex constraint Equation (2).

To counterbalance technical biases induced by the transcriptomic quantification technology, either RNA-Seq or microarray, some deconvolution methodologies, such as `CibersortX` ([New+19]) propose automated batch correction effect with the ComBat function, prior to the deconvolution process. Interestingly, [JL21] demonstrates that Cibersort [New+15], CibersortX [New+19] and MuSiC [Wan+19] were less sensitive to the choice of normalisation and sequencing platform, compared to other methods benchmarked.

In conclusion of this section, [JL21] shows that penalised regression approaches, including Lasso, Ridge and Elastic Net ([Alt+14]) approaches were usually outperformed by robust linear regression approaches (RLR, FARDEEP, SVR, see Section 2.1.1).

For readers interested in further exploration of this topic, we recommend the following review papers, which offer comprehensive insights into recently developed deconvolution approaches and outline future perspectives and unmet needs for enhancing their exploratory capabilities: [Fin+19b], [Pet+18], [Avi+18] and [Bla+21].

4 General Discussion: the Fate of Deconvolution Algorithms in the Fields of Spatial Transcriptomics and single cell RNA-Seq

4.1 Overview of Spatial Transcriptomics and Single-Cell RNA Sequencing

Spatial transcriptomics enables the simultaneous profiling of gene expression at a high spatial resolution *in-situ*, while preserving the global cellular layout. ST reveals notably useful to determine the general layout of cell populations within a tissue and to identify hotspots, also known as "niches" (localised microenvironments in which stem cells prevail over fully differentiated cell subtypes) [10].

However, the design of the lattice of spots in ST technologies, such as HDST [Vic+19] or Slide-Seq [Rod+19]), is constrained by physical limitations that directly alleviate the final *resolution* (namely the distance between capture spots). Hence, it is not uncommon that the mRNA collected at a given sport constitutes a mixture of cell types, rather than representing a single cell.

Thus, SRT techniques have to meet a middle ground between cellular resolution and the depth and coverage of the RNA library. For instance, approaches like SeqFISH+ ([Eng+19]) and MERFISH

[10] It is common to use the abbreviation "SRT", for Spatially Resolved Transcriptomics, when referring to the general spatial sequencing framework, in order to mitigate nomenclature confusion with the specific and corporate technology "Spatial Transcriptomics" ([Stå+16])/

([Che+15]) provide subcellular resolution but are limited in throughput. Conversely, Spatial Transcriptomics ([Stå+16]) and FISSEQ ([Lee+14]) exhibit larger coverage of the genome, yet they cannot achieve single-cell resolution sequencing and are further constrained by high detection thresholds [11].

Single-cell RNA sequencing (scRNA-Seq) provides a high-resolution view of the transcriptome, by quantifying RNA content at the single-cell level. scRNA-Seq enabled to uncover cellular heterogeneity, identify rare cell populations, and capture complex dynamic changes in gene expression, that were typically obscured in bulk RNA-Seq analysis.

However, scRNA-Seq is costly and time-consuming, making it challenging to scale up for large sample sizes. In addition, the sparse nature of scRNA-Seq outputs, resulting from "drop-outs" and the complexity of the technology, renders the analysis challenging and prone to higher technical biases and variability. Hence, going down to the single cell level, scRNA-Seq typically exhibits lower coverage and depth compared to bulk RNASeq (but still higher compared to SRT).

Coupling scRNA-Seq with spatial transcriptomic data streamlines the understanding of the mechanisms relating gene expression patterns with changes of cell populations within tissues, by bridging the advantages of both methodologies while mitigating their major limitations. However, *mismatch*, designing the discordance between the cell types inferred from expression profiles derived from single-cell RNA sequencing and SRT, is commonly observed. Mismatch usually results from pre-sequencing and post-sequencing artefacts. Pre-sequencing mismatch can stem from *sampling bias* of the tissue section (lower depth with spatial barcoding or lower access to intertwined tissue structures with HPRI) or from an artificial and ectopic stimuli perturbing the cellular expression profile (stress response, or less likely, alteration of cell phenotype due to the disruption of *in situ* spatial dynamics resulting from tissue dissociation).

4.2 Construction of reference signatures, based on single Cell RNA-Seq profiles

On the other hand, single-cell RNA sequencing technologies empower cellular deconvolution algorithms, by enabling the derivation of signature matrices more representative of the phenotype condition.

Indeed, by capturing gene expression profiles at the single-cell level, scRNA-Seq allows better discrimination of closely related cell types, and identification of rare cell type variants, which are likely to be confused with noise using bulk RNA-Seq.

Even better, the stronger granularity of scRNA-Seq outputs enables to capture the heterogeneity within cell populations, including unravelling asynchronous states of a cell population.

4.3 Integrating Spatial Transcriptomics with Single-Cell RNA-Seq Data Through Deconvolution Approaches

Recent alternative to mitigate the low detection threshold of scRNA in SRT and better handle mismatch issues, involve two primary approaches: *deconvolution* algorithms and *mapping* (report to Appendix E).

Spatial deconvolution tools, a close synonym to *stochastic profiling* techniques, estimate the cell composition for each capture spot. While sequencing the transcriptome at the single cell level is usually infeasible in a spatial context, aggregating the expression of a random pool of cells (usually rather small, aggregating no more than a dozen of them) automatically increases the depth and coverage of the RNA library, which in turn counterbalances the intrinsic noisiness and low resolution of scRNA-Seq methods.

[11] A minimal number of 200 mRNA molecules per cell is required to detect the expression of a transcript, excluding practically a large amount of genes involved only in specific phases of the cell cycle

Spatial deconvolution algorithms usually capitalise on reference signatures obtained from single-cell RNA sequencing profiles (see section 4.2), instead of bulk expression. The final signature is finally computed by summing the individual cellular contributions in order to reconstitute a "pseudo-bulk" mixture.

Nonetheless, spatial deconvolution algorithms necessitate specific adjustments compared to traditional approaches, as conventional deconvolution algorithms, designed for bulk transcriptome, often yield suboptimal results when dealing with sparse expression matrices, inherent to the SRT framework ([Kle+20]). In addition, spatial deconvolution methods face similar challenges to traditional deconvolution algorithms, as they too, cannot obtain absolute estimation of cell ratios, thus limiting their applicability for meaningful intra-sample comparisons.

The most population spatial deconvolution methods encompass, ranked by analytical complexity:

- The most basic methods calculate "enrichment scores" that indicate the degree of association between an individual spatial location and a specific cell type. These scores are computed using the same techniques outlined in Section 2.2. For example, in Seurat, by [Kis+17], each spatial location is assigned to the cell type whose expression profile, composed of the markers within its gene set, exhibits the highest similarity.

  Taking a more advanced approach, the Multimodal Intersection Analysis (MIA, [Mon+20]) combines gene pathway information inferred from scRNA-Seq data with gene modules that are identified as enriched through spatial barcoding techniques.

- SPOTlight [Elo+21] and SpatialDWLS [DY21] are both regression-based models that used linear solvers to estimate cellular ratios while enforcing the unit-simplex constraint, through the non-negative least squares (NNLS) algorithm.

- *Probabilistic models*, represent the mixture as a convolution of parametric distributions whose estimated cell ratios are the MLE (alternatively the MAP whereby a prior distribution is assigned to the cell ratios) of the distribution. Stereoscope ([Kho+21], also illustrated in section 4.3) and Cell2location ([Kle+20]) fit the distribution with a mixture of negative binomial(NB) distributions, while Robust cell-type decomposition (RCTD, [Cab+22]) utilises Poisson distributions.

- NMF regression (NMFref) is an unsupervised algorithm used both by SlideSeq [XHB16] and SPOTLight [Gul+13] to infer simultaneously cellular ratios and individual expression profiles.

- More exotic and recent methods explore alternative ways, such as DSTG [He+20] algorithm using *mutual nearest neighbour clustering* or deep-learning methods, with Tangram [Ber+20].

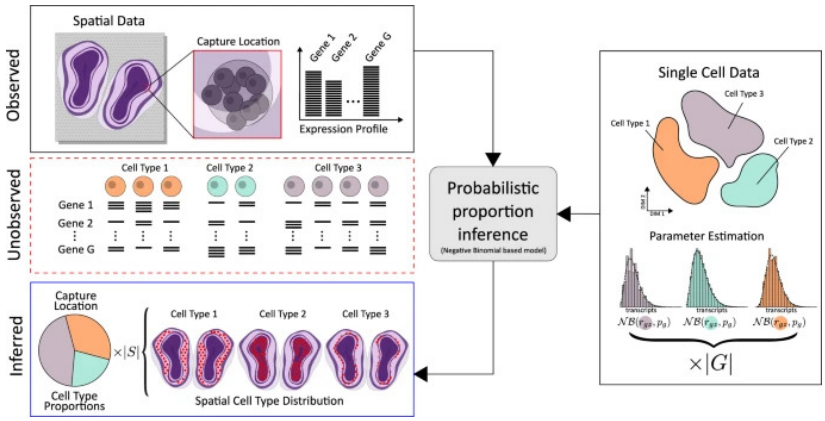


Figure 10. Illustration of a spatial deconvolution algorithm principle, with `stereoscope`. A deconvolution algorithm is used to model and infer the mixture composition of cell populations at a specific capture site using signatures derived from single-cell datasets. `stereoscope` precisely employs a convolution of Negative Binomials to model the mixture of cell types within a captured side. Reproduced from [Kho+21, Fig. 1].

Promising studies extend the investigational capacity of spatial transcriptomics, by coupling high-resolution *tissue images* with *histological annotations* (cell sizes and shapes, for instance) and *SRT* data ([Lar+22]). It is hence believed that the integration of distinct biological modalities in a spatially resolved context is poised to elucidate the as-yet-unsolved biological processes driving the spatial organisation of tissue niches([Roz+17]).

`SpaDecon`, by [Col+23], is one of the most promising spatial integrated approach, coupling histological annotations with metabolic and transcriptomic activity. `34P`, by [Occ+23], even claims to be able to dissect intra-tumour heterogeneity in luminal breast cancer by integrating morphological annotations, SRT data and whole slide images to a neural network architecture. As a complementary resource, we refer the interested readers to [Rao+21], [Lon+21], [Kre21] and [Wil+22] for a comprehensive and updated review of a whole array of pioneering methods integrating spatial transcriptomic, scRNA-Seq technologies and imagery annotations.

To close this discussion, [Tes+17] pinpointed that the lack of reproducibility and robustness observed for most deconvolution methods may be mitigated by coupling cellular estimates obtained from distinct biological sources. Yet, a comprehensive benchmark comparing the performance of deconvolution approaches with regard to the biological input still lacks.

References

[Che53] E. Colin Cherry. "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *The Journal of the Acoustical Society of America* (Sept. 1, 1953). ISSN: 0001-4966. DOI: 10.1121/1.1907229. URL: https://asa.scitation.org/doi/10.1121/1.1907229.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* (Sept. 1977). ISSN: 00359246. DOI: 10.1111/j.2517-6161.1977.tb01600.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1977.tb01600.x.

[HH81] Karen H. Haskell and Richard J. Hanson. "An Algorithm for Linear Least Squares Problems with Equality and Nonnegativity Constraints". In: *Mathematical Programming* (Dec. 1, 1981). ISSN: 1436-4646. DOI: 10.1007/BF01584232. URL: https://doi.org/10.1007/BF01584232.

[Bes86] Julian Besag. "On the Statistical Analysis of Dirty Pictures". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1986). ISSN: 0035-9246. URL: https://www.jstor.org/stable/2345426.

[Yoh87] Victor J. Yohai. "High Breakdown-Point and High Efficiency Robust Estimates for Regression". In: *The Annals of Statistics* (June 1987). ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176350366. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-15/issue-2/High-Breakdown-Point-and-High-Efficiency-Robust-Estimates-for-Regression/10.1214/aos/1176350366.full.

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Machine Learning* (Sept. 1, 1995). ISSN: 1573-0565. DOI: 10.1007/BF00994018. URL: https://doi.org/10.1007/BF00994018.

[Sch+00] Bernhard Schölkopf et al. "New Support Vector Algorithms". In: *Neural Computation* (May 1, 2000). ISSN: 0899-7667. DOI: 10.1162/089976600300015565. URL: https://doi.org/10.1162/089976600300015565.

[Ven+01] D. Venet et al. "Separation of Samples into Their Constituents Using Gene Expression Data". In: *Bioinformatics* (June 1, 2001). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.suppl_1.S279. URL: https://doi.org/10.1093/bioinformatics/17.suppl_1.S279.

[CC02] Chang Cc and Lin Cj. "Training Nu-Support Vector Regression: Theory and Algorithms". In: *Neural computation* (Aug. 2002). ISSN: 0899-7667. DOI: 10.1162/089976602760128081. URL: https://pubmed.ncbi.nlm.nih.gov/12180409/.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* (Mar. 1, 2003). ISSN: 1532-4435.

[GE03] Isabelle Guyon and André Elisseeff. "An Introduction of Variable and Feature Selection". In: *J. Machine Learning Research Special Issue on Variable and Feature Selection* (Jan. 1, 2003). DOI: 10.1162/153244303322753616.

[LNM03] Peng Lu, Aleksey Nakorchevskiy, and Edward M. Marcotte. "Expression Deconvolution: A Reinterpretation of DNA Microarray Data Reveals Dynamic Changes in Cell Populations". In: *Proceedings of the National Academy of Sciences* (Sept. 2, 2003). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1832361100. URL: https://www.pnas.org/content/100/18/10370.

[Whi+03] Adeline R. Whitney et al. "Individuality and Variation in Gene Expression Patterns in Human Blood". In: *Proceedings of the National Academy of Sciences of the United States of America* (Feb. 18, 2003). ISSN: 0027-8424. DOI: 10.1073/pnas.252784499.

[Stu+04] Robert O. Stuart et al. "In Silico Dissection of Cell-Type-Associated Patterns of Gene Expression in Prostate Cancer". In: *Proceedings of the National Academy of Sciences of the United States of America* (Jan. 13, 2004). ISSN: 0027-8424. DOI: 10.1073/pnas.2536479100.

[Sub+05] Aravind Subramanian et al. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles". In: *Proceedings of the National Academy of Sciences of the United States of America* (Oct. 25, 2005). ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

[Hof+06] Martin Hoffmann et al. "Robust Computational Reconstitution – a New Method for the Comparative Analysis of Gene Expression in Tissues and Isolated Cell Fractions". In: *BMC Bioinformatics* (Aug. 4, 2006). ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-369. URL: https://doi.org/10.1186/1471-2105-7-369.

[GPT07] Mark Gosink, Howard Petrie, and Nicholas Tsinoremas. "Electronically Subtracting Expression Patterns from a Mixed Cell Population". In: *Bioinformatics. 23* (2007).

[Bar+08] Ziv Bar-Joseph et al. "Genome-Wide Transcriptional Analysis of the Human Cell Cycle Identifies Genes Differentially Regulated in Normal and Cancer Cells". In: *Proceedings of the National Academy of Sciences* (Jan. 22, 2008). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0704723105. URL: https://www.pnas.org/content/105/3/955.

[Lob+08] Edward K. Lobenhofer et al. "Gene Expression Response in Target Organ and Whole Blood Varies as a Function of Target Organ Injury Phenotype". In: *Genome Biology* (June 20, 2008). ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-6-r100. URL: https://doi.org/10.1186/gb-2008-9-6-r100.

[Abb+09] Alexander R. Abbas et al. "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus". In: *PloS One* (July 1, 2009). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006098.

[Bar+09] David A. Barbie et al. "Systematic RNA Interference Reveals That Oncogenic KRAS-driven Cancers Require TBK1". In: *Nature* (Nov. 5, 2009). ISSN: 1476-4687. DOI: 10.1038/nature08460.

[Efr09] Bradley Efron. "Are a Set of Microarrays Independent of Each Other?" In: *The Annals of Applied Statistics* (Sept. 2009). ISSN: 1932-6157, 1941-7330. DOI: 10.1214/09-AOAS236. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-3/issue-3/Are-a-set-of-microarrays-independent-of-each-other/10.1214/09-AOAS236.full.

[QM09] Gerald Quon and Quaid Morris. "ISOLATE: A Computational Strategy for Identifying the Primary Origin of Cancers Using High-Throughput Sequencing". In: *Bioinformatics (Oxford, England)* (Nov. 1, 2009). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp378.

[CSC10] Jennifer Clarke, Pearl Seo, and Bertrand Clarke. "Statistical Expression Deconvolution from Mixed Tissue Samples". In: *Bioinformatics (Oxford, England)* (Apr. 15, 2010). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq097.

[Erk+10] Timo Erkkilä et al. "Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues". In: *Bioinformatics* (Oct. 15, 2010). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq406. URL: https://doi.org/10.1093/bioinformatics/btq406.

[Rep+10] Dirk Repsilber et al. "Biomarker Discovery in Heterogeneous Tissue Samples -Taking the in-Silico Deconfounding Approach". In: *BMC Bioinformatics* (Jan. 14, 2010). ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-27. URL: https://doi.org/10.1186/1471-2105-11-27.

[Wan+10] Yipeng Wang et al. "In Silico Estimates of Tissue Components in Surgical Samples Based on Expression Profiling Data". In: *Cancer research* (Aug. 15, 2010). ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-10-0021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411177/.

[BUK11] Christopher R. Bolen, Mohamed Uduman, and Steven H. Kleinstein. "Cell Subset Prediction for Blood Genomic Studies". In: *BMC Bioinformatics* (June 24, 2011). ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-258. URL: https://doi.org/10.1186/1471-2105-12-258.

[Fri+11] Jerome Friedman et al. *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Version 4.1-3. 2011. URL: https://CRAN.R-project.org/package=glmnet.

[Gon+11] Ting Gong et al. "Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples". In: *PloS One* (2011). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027156.

[Kuh+12] Alexandre Kuhn et al. "Cell Population-Specific Expression Analysis of Human Cerebellum". In: *BMC Genomics* (Nov. 12, 2012). ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-610. URL: https://doi.org/10.1186/1471-2164-13-610.

[Qia+12] Wenlian Qiao et al. "PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions". In: *PLOS Computational Biology* (Dec. 20, 2012). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002838. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002838.

[Sho+12] Jason E. Shoemaker et al. "CTen: A Web-Based Platform for Identifying Enriched Cell Types from Heterogeneous Microarray Data". In: *BMC Genomics* (Sept. 6, 2012). ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-460. URL: https://doi.org/10.1186/1471-2164-13-460.

[Ahn+13] Jaeil Ahn et al. "DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data". In: *Bioinformatics (Oxford, England)* (Aug. 1, 2013). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt301.

[Gau13] R. Gaujoux. "An Introduction to Gene Expression Deconvolution and the CellMix Package A Comprehensive Framework for Gene Expression Deconvolution". In: *undefined* (2013). URL: https://www.semanticscholar.org/paper/An-introduction-to-gene-expression-deconvolution-A-Gaujoux/980b8ac01435d2faa76eb1e0bc94e0a83b27b7a3.

[GS13] Ting Gong and Joseph D. Szustakowski. "DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data". In: *Bioinformatics (Oxford, England)* (Apr. 15, 2013). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt090.

[Gul+13] Ankur Gulati et al. "Association of Fibrosis with Mortality and Sudden Cardiac Death in Patients with Nonischemic Dilated Cardiomyopathy". In: *JAMA* (Mar. 6, 2013). ISSN: 1538-3598. DOI: 10.1001/jama.2013.1363.

[LX13] Yi Li and Xiaohui Xie. "A Mixture Model for Expression Deconvolution from RNA-seq in Heterogeneous Tissues". In: *BMC bioinformatics* (2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-S5-S11.

[Quo+13] Gerald Quon et al. "Computational Purification of Individual Tumor Gene Expression Profiles Leads to Significant Improvements in Prognostic Prediction". In: *Genome Medicine* (Mar. 28, 2013). ISSN: 1756-994X. DOI: 10.1186/gm433. URL: https://doi.org/10.1186/gm433.

[SG13] Shai S. Shen-Orr and Renaud Gaujoux. "Computational Deconvolution: Extracting Cell Type-Specific Information from Heterogeneous Samples". In: *Current Opinion in Immunology* (Oct. 2013). ISSN: 1879-0372. DOI: 10.1016/j.coi.2013.09.015.

[Yos+13] Kosuke Yoshihara et al. "Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data". In: *Nature Communications* (Oct. 11, 2013). ISSN: 2041-1723. DOI: 10.1038/ncomms3612. URL: https://www.nature.com/articles/ncomms3612.

[Zho+13] Yi Zhong et al. "Digital Sorting of Complex Tissues for Cell Type-Specific Gene Expression Profiles". In: *BMC Bioinformatics* (Mar. 7, 2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-89. URL: https://doi.org/10.1186/1471-2105-14-89.

[Alt+14] Zeev Altboum et al. "Digital Cell Quantification Identifies Global Immune Cell Dynamics during Influenza Infection". In: *Molecular Systems Biology* (Feb. 28, 2014). ISSN: 1744-4292. DOI: 10.1002/msb.134947. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4023392/.

[Lee+14] Je Hyuk Lee et al. "Highly Multiplexed Subcellular RNA Sequencing in Situ". In: *Science* (Mar. 21, 2014). DOI: 10.1126/science.1250212. URL: https://www.science.org/doi/full/10.1126/science.1250212.

[LHP14] David A. Liebner, Kun Huang, and Jeffrey D. Parvin. "MMAD: Microarray Microdissection with Analysis of Differences Is a Computational Tool for Deconvoluting Cell Type-Specific Contributions from Tissue Samples". In: *Bioinformatics (Oxford, England)* (Mar. 1, 2014). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt566.

[Ang+15] Mihaela Angelova et al. "Characterization of the Immunophenotypes and Antigenomes of Colorectal Cancers Reveals Distinct Tumor Escape Mechanisms and Novel Targets for Immunotherapy". In: *Genome Biology* (Mar. 31, 2015). ISSN: 1474-760X. DOI: 10.1186/s13059-015-0620-6.

[ASB15] Dvir Aran, Marina Sirota, and Atul J. Butte. "Systematic Pan-Cancer Analysis of Tumour Purity". In: *Nature Communications* (Dec. 4, 2015). ISSN: 2041-1723. DOI: 10.1038/ncomms9971.

[Bue+15] Florian Buettner et al. "Computational Analysis of Cell-to-Cell Heterogeneity in Single-Cell RNA-sequencing Data Reveals Hidden Subpopulations of Cells". In: *Nature Biotechnology* (Feb. 2015). ISSN: 1546-1696. DOI: 10.1038/nbt.3102. URL: https://www.nature.com/articles/nbt.3102.

[Che+15] Kok Hao Chen et al. "RNA Imaging. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells". In: *Science (New York, N.Y.)* (Apr. 24, 2015). ISSN: 1095-9203. DOI: 10.1126/science.aaa6090.

[CZS15] Maria Chikina, Elena Zaslavsky, and Stuart C. Sealfon. "CellCODE: A Robust Latent Variable Approach to Differential Expression Analysis for Heterogeneous Cell Populations". In: *Bioinformatics (Oxford, England)* (May 15, 2015). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv015.

[New+15] Aaron Newman et al. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles". In: *Nature methods* (Mar. 30, 2015). DOI: 10.1038/nmeth.3337.

[Roo+15] Michael S. Rooney et al. "Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity". In: *Cell* (Jan. 15, 2015). ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.12.033.

[Wan+15] Niya Wang et al. "UNDO: A Bioconductor R Package for Unsupervised Deconvolution of Mixed Gene Expressions in Tumor Samples". In: *Bioinformatics (Oxford, England)* (Jan. 1, 2015). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu607.

[Bec+16] Etienne Becht et al. "Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression". In: *Genome Biology* (Oct. 20, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1070-5. URL: https://doi.org/10.1186/s13059-016-1070-5.

[Fri+16] Amit Frishberg et al. "ImmQuant: A User-Friendly Tool for Inferring Immune Cell-Type Composition from Gene-Expression Data". In: *Bioinformatics* (Dec. 15, 2016). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw535. URL: https://doi.org/10.1093/bioinformatics/btw535.

[Li+16] Bo Li et al. "Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy". In: *Genome Biology* (Aug. 22, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1028-7.

[Şen+16] Yasin Şenbabaoğlu et al. "Tumor Immune Microenvironment Characterization in Clear Cell Renal Cell Carcinoma Identifies Prognostic and Immunotherapeutically Relevant Messenger RNA Signatures". In: *Genome Biology* (Nov. 17, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1092-z. URL: https://doi.org/10.1186/s13059-016-1092-z.

[Stå+16] Patrik L. Ståhl et al. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics". In: *Science (New York, N.Y.)* (July 1, 2016). ISSN: 1095-9203. DOI: 10.1126/science.aaf2403.

[Wan+16] Niya Wang et al. "Mathematical Modelling of Transcriptional Heterogeneity Identifies Novel Markers and Subpopulations in Complex Tissues". In: *Scientific Reports* (Jan. 7, 2016). ISSN: 2045-2322. DOI: 10.1038/srep18909. URL: https://www.nature.com/articles/srep18909.

[XHB16] Ling Xu, Dan He, and Ying Bai. "Microglia-Mediated Inflammation and Neurodegenerative Disease". In: *Molecular Neurobiology* (Dec. 2016). ISSN: 1559-1182. DOI: 10.1007/s12035-015-9593-4.

[AHB17] Dvir Aran, Zicheng Hu, and Atul Butte. "xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape". In: *Genome Biology* (Dec. 1, 2017). DOI: 10.1186/s13059-017-1349-1.

[Cha+17] Pornpimol Charoentong et al. "Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade". In: *Cell Reports* (Jan. 3, 2017). ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.12.019.

[Che+17] Ziyi Chen et al. "Inference of Immune Cell Composition on the Expression Profiles of Mouse Tissue". In: *Scientific Reports* (Jan. 13, 2017). ISSN: 2045-2322. DOI: 10.1038/srep40508. URL: https://www.nature.com/articles/srep40508.

[GRT17] Daniel Greene, Sylvia Richardson, and Ernest Turro. "ontologyX: A Suite of R Packages for Working with Ontological Data". In: *Bioinformatics* (Apr. 1, 2017). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw763. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386138/.

[Kis+17] Vladimir Yu Kiselev et al. "SC3: Consensus Clustering of Single-Cell RNA-seq Data". In: *Nature Methods* (May 2017). ISSN: 1548-7105. DOI: 10.1038/nmeth.4236.

[Rac+17] Julien Racle et al. "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data". In: *eLife* (Nov. 13, 2017). Ed. by Alfonso Valencia. ISSN: 2050-084X. DOI: 10.7554/eLife.26476. URL: https://doi.org/10.7554/eLife.26476.

[Roz+17] Orit Rozenblatt-Rosen et al. "The Human Cell Atlas: From Vision to Reality". In: *Nature* (Oct. 2017). ISSN: 1476-4687. DOI: 10.1038/550451a. URL: https://www.nature.com/articles/550451a.

[Tap+17] Elias Tappeiner et al. "TIminer: NGS Data Mining Pipeline for Cancer Immunology and Immunotherapy". In: *Bioinformatics (Oxford, England)* (June 15, 2017). DOI: 10.1093/bioinformatics/btx377.

[Tes+17] Andrew E. Teschendorff et al. "A Comparison of Reference-Based Algorithms for Correcting Cell-Type Heterogeneity in Epigenome-Wide Association Studies". In: *BMC Bioinformatics* (Feb. 13, 2017). ISSN: 1471-2105. DOI: 10.1186/s12859-017-1511-5. URL: https://doi.org/10.1186/s12859-017-1511-5.

[Zha+17] Jitao David Zhang et al. "Detect Tissue Heterogeneity in Gene Expression Data with BioQC". In: *BMC Genomics* (Apr. 4, 2017). ISSN: 1471-2164. DOI: 10.1186/s12864-017-3661-2. URL: https://doi.org/10.1186/s12864-017-3661-2.

[Avi+18] Francisco Avila Cobos et al. "Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations". In: *Bioinformatics (Oxford, England)* (June 1, 2018). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty019.

[Che+18] Binbin Chen et al. "Profiling Tumor Infiltrating Immune Cells with CIBERSORT". In: *Methods in molecular biology (Clifton, N.J.)* (2018). ISSN: 1064-3745. DOI: 10.1007/978-1-4939-7493-1_12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5895181/.

[Pet+18] Florent Petitprez et al. "Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine". In: *Frontiers in Oncology* (2018). ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00390.

[Wan+18] Zeya Wang et al. "Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration". In: *iScience* (Nov. 2, 2018). ISSN: 2589-0042. DOI: 10.1016/j.isci.2018.10.028. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249353/.

[Cha+19] Wennan Chang et al. *ICTD: A Semi-Supervised Cell Type Identification and Deconvolution Method for Multi-Omics Data*. Dec. 5, 2019. DOI: 10.1101/426593. URL: https://www.biorxiv.org/content/10.1101/426593v3. preprint.

[Che19] L. Chen. *CAMTHC: Convex Analysis of Mixtures for Tissue Heterogeneity Characterization*. 2019. URL: https://rdrr.io/bioc/CAMTHC/.

[Eng+19] Chee-Huat Linus Eng et al. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+". In: *Nature* (Apr. 2019). ISSN: 1476-4687. DOI: 10.1038/s41586-019-1049-y. URL: https://www.nature.com/articles/s41586-019-1049-y.

[Fin+19a] Francesca Finotello et al. "Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-seq Data". In: *Genome Medicine* (May 24, 2019). ISSN: 1756-994X. DOI: 10.1186/s13073-019-0638-6. URL: https://doi.org/10.1186/s13073-019-0638-6.

[Fin+19b] Francesca Finotello et al. "Next-Generation Computational Tools for Interrogating Cancer Immunity". In: *Nature Reviews Genetics* (Dec. 2019). ISSN: 1471-0064. DOI: 10.1038/s41576-019-0166-7. URL: https://www.nature.com/articles/s41576-019-0166-7.

[Hao+19] Yuning Hao et al. "Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles Using Least Trimmed Squares". In: *PLoS computational biology* (May 2019). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006976.

[New+19] Aaron M. Newman et al. "Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry". In: *Nature Biotechnology* (July 2019). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0114-2. URL: https://www.nature.com/articles/s41587-019-0114-2.

[Rod+19] Samuel G. Rodriques et al. "Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution". In: *Science* (Mar. 29, 2019). DOI: 10.1126/science.aaw1219. URL: https://www.science.org/doi/10.1126/science.aaw1219.

[Stu+19] Gregor Sturm et al. "Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology". In: *Bioinformatics (Oxford, England)* (July 15, 2019). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz363.

[Vic+19] Sanja Vickovic et al. "High-Definition Spatial Transcriptomics for in Situ Tissue Profiling". In: *Nature Methods* (Oct. 2019). ISSN: 1548-7105. DOI: 10.1038/s41592-019-0548-y. URL: https://www.nature.com/articles/s41592-019-0548-y.

[Wan+19] Xuran Wang et al. "Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference". In: *Nature Communications* (Jan. 22, 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-018-08023-x. URL: https://www.nature.com/articles/s41467-018-08023-x.

[Ber+20] Ludvig Bergenstråhle et al. *Super-Resolved Spatial Transcriptomics by Deep Data Fusion*. Mar. 13, 2020. DOI: 10.1101/2020.02.28.963413. URL: https://www.biorxiv.org/content/10.1101/2020.02.28.963413v2. preprint.

[Cam+20] Neil A. Campbell et al. *Biology: A Global Approach*. Pearson Education Limited, May 14, 2020. 1512 pp. ISBN: 978-1-292-34163-7.

[Don+20] Li Dong et al. "Semi-CAM: A Semi-Supervised Deconvolution Method for Bulk Transcriptomic Data with Partial Marker Gene Information". In: *Scientific Reports* (Mar. 25, 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-62330-2. URL: https://www.nature.com/articles/s41598-020-62330-2.

[Fa+20] Cobos Fa et al. "Comprehensive Benchmarking of Computational Deconvolution of Transcriptomics Data". In: (Jan. 10, 2020). DOI: 10.1101/2020.01.10.897116. URL: https://europepmc.org/article/ppr/ppr108248.

[He+20] Bryan He et al. "Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning". In: *Nature Biomedical Engineering* (Aug. 2020). ISSN: 2157-846X. DOI: 10.1038/s41551-020-0578-x.

[Kle+20] Vitalii Kleshchevnikov et al. *Comprehensive Mapping of Tissue Cell Architecture via Integrated Single Cell and Spatial Transcriptomics*. Nov. 17, 2020. DOI: 10.1101/2020.11.15.378125. URL: https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1. preprint.

[Lew20] Julius M. Cruse Lewis Robert E. *Illustrated Dictionary of Immunology*. 3rd ed. Boca Raton: CRC Press, June 30, 2020. 816 pp. ISBN: 978-0-429-12407-5. DOI: 10.1201/9780849379888.

[Mon+20] Reuben Moncada et al. "Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas". In: *Nature Biotechnology* (Mar. 2020). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0392-8. URL: https://www.nature.com/articles/s41587-019-0392-8.

[SFL20] Gregor Sturm, Francesca Finotello, and Markus List. "In Silico Cell-Type Deconvolution Methods in Cancer Immunotherapy". In: *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*. Ed. by Sebastian Boegel. Methods in Molecular Biology. New York, NY: Springer US, 2020. ISBN: 978-1-07-160327-7. DOI: 10.1007/978-1-0716-0327-7_15. URL: https://doi.org/10.1007/978-1-0716-0327-7_15.

[TPZ20] Daiwei Tang, Seyoung Park, and Hongyu Zhao. "NITUMID: Nonnegative Matrix Factorization-Based Immune-TUmor MIcroenvironment Deconvolution". In: *Bioinformatics* (Mar. 1, 2020). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz748. URL: https://doi.org/10.1093/bioinformatics/btz748.

[ALH21] Michiel C. van Aalderen, Rene A. W. van Lier, and Pleun Hombrink. "How to Reliably Define Human CD8+ T-Cell Subsets: Markers Playing Tricks". In: *Cold Spring Harbor Perspectives in Biology* (Jan. 11, 2021). ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a037747. URL: http://cshperspectives.cshlp.org/content/13/11/a037747.

[Bla+21] Andrea Blasco et al. "Improving Deconvolution Methods in Biology through Open Innovation Competitions: An Application to the Connectivity Map". In: *Bioinformatics* (Mar. 22, 2021). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab192. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8479655/.

[DY21] Rui Dong and Guo-Cheng Yuan. "SpatialDWLS: Accurate Deconvolution of Spatial Transcriptomic Data". In: *Genome Biology* (May 10, 2021). ISSN: 1474-760X. DOI: 10.1186/s13059-021-02362-7. URL: https://doi.org/10.1186/s13059-021-02362-7.

[Elo+21] Marc Elosua-Bayes et al. "SPOTlight: Seeded NMF Regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes". In: *Nucleic Acids Research* (May 21, 2021). ISSN: 1362-4962. DOI: 10.1093/nar/gkab043.

[JL21] Haijing Jin and Zhandong Liu. "A Benchmark for RNA-seq Deconvolution Analysis under Dynamic Testing Environments". In: *Genome Biology* (Apr. 12, 2021). ISSN: 1474-760X. DOI: 10.1186/s13059-021-02290-6. URL: https://doi.org/10.1186/s13059-021-02290-6.

[Kho+21] Combiz Khozoie et al. *scFlow: A Scalable and Reproducible Analysis Pipeline for Single-Cell RNA Sequencing Data*. preprint. Preprints, Aug. 16, 2021. DOI: 10.22541/au.162912533.38489960/v1. URL: https://www.authorea.com/users/226952/articles/480342-scflow-a-scalable-and-reproducible-analysis-pipeline-for-single-cell-rna-sequencing-data?commit=921426e3a377f7897ba262b5fc2bf0ef3680570a.

[Kre21] Ivan Krešimir Lukić. "Bioinformatics Approach to Spatially Resolved Transcriptomics". In: *Emerging Topics in Life Sciences* (Aug. 9, 2021). ISSN: 2397-8554. DOI: 10.1042/ETLS20210131. URL: https://doi.org/10.1042/ETLS20210131.

[Lon+21] Sophia K. Longo et al. "Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics". In: *Nature Reviews Genetics* (Oct. 2021). ISSN: 1471-0064. DOI: 10.1038/s41576-021-00370-8. URL: https://www.nature.com/articles/s41576-021-00370-8.

[Mey+21] David Meyer et al. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. Version 1.7-9. Sept. 16, 2021. URL: https://CRAN.R-project.org/package=e1071.

[Rao+21] Anjali Rao et al. "Exploring Tissue Architecture Using Spatial Transcriptomics". In: *Nature* (Aug. 2021). ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9. URL: https://www.nature.com/articles/s41586-021-03634-9.

[Sos+21] Olukayode A. Sosina et al. "Strategies for Cellular Deconvolution in Human Brain RNA Sequencing Data". In: (Aug. 4, 2021). DOI: 10.12688/f1000research.50858.1. URL: https://f1000research.com/articles/10-750.

[Boe+22] Maximilian Boesch et al. "OMIP 077: Definition of All Principal Human Leukocyte Populations Using a Broadly Applicable 14-Color Panel". In: *Cytometry Part A* (2022). ISSN: 1552-4930. DOI: 10.1002/cyto.a.24481. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.24481.

[Bok+22] A. A. Bokil et al. "32P Discovering Markers That Identify Pro-Metastatic Immune Cell Subsets". In: *Immuno-Oncology and Technology*. Abstract Book of the ESMO Immuno-Oncology Congress 2022 7-9 December 2022 (Dec. 1, 2022). ISSN: 2590-0188. DOI: 10.1016/j.iotech.2022.100137. URL: https://www.sciencedirect.com/science/article/pii/S2590018822000685.

[Cab+22] Dylan M. Cable et al. "Robust Decomposition of Cell Type Mixtures in Spatial Transcriptomics". In: *Nature Biotechnology* (Apr. 2022). ISSN: 1546-1696. DOI: 10.1038/s41587-021-00830-w. URL: https://www.nature.com/articles/s41587-021-00830-w.

[Cai+22] Manqi Cai et al. "Robust and Accurate Estimation of Cellular Fraction from Tissue Omics Data via Ensemble Deconvolution". In: *Bioinformatics* (May 26, 2022). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac279. URL: https://doi.org/10.1093/bioinformatics/btac279.

[Kas+22] Aditya Kashyap et al. "Quantification of Tumor Heterogeneity: From Data Acquisition to Metric Generation". In: *Trends in Biotechnology* (June 1, 2022). ISSN: 0167-7799, 1879-3096. DOI: 10.1016/j.tibtech.2021.11.006. URL: https://www.cell.com/trends/biotechnology/abstract/S0167-7799(21)00267-5.

[Lar+22] Ludvig Larsson et al. "SnapShot: Spatial Transcriptomics". In: *Cell* (July 21, 2022). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2022.06.002. URL: https://www.cell.com/cell/abstract/S0092-8674(22)00707-3.

[Wil+22] Cameron G. Williams et al. "An Introduction to Spatial Transcriptomics for Biomedical Research". In: *Genome Medicine* (June 27, 2022). ISSN: 1756-994X. DOI: 10.1186/s13073-022-01075-1. URL: https://doi.org/10.1186/s13073-022-01075-1.

[Yu22] Xiaoqing Yu. "Estimation of Tumor Immune Signatures from Transcriptomics Data". In: *Handbook of Statistical Bioinformatics*. Ed. by Henry Horng-Shing Lu et al. Springer Handbooks of Computational Statistics. Berlin, Heidelberg: Springer, 2022. ISBN: 978-3-662-65902-1. DOI: 10.1007/978-3-662-65902-1_16. URL: https://doi.org/10.1007/978-3-662-65902-1_16.

[Col+23] Kyle Coleman et al. "SpaDecon: Cell-Type Deconvolution in Spatial Transcriptomics with Semi-Supervised Learning". In: *Communications Biology* (Apr. 7, 2023). ISSN: 2399-3642. DOI: 10.1038/s42003-023-04761-x. URL: https://www.nature.com/articles/s42003-023-04761-x.

[Occ+23] N. Occelli et al. "34P Investigating Morphological Heterogeneity in Luminal Breast Cancer Integrating Artificial Intelligence and Spatial Transcriptomics". In: *ESMO Open* (May 1, 2023). ISSN: 2059-7029. DOI: 10.1016/j.esmoop.2023.101258. URL: https://www.esmoopen.com/article/S2059-7029(23)00484-2/fulltext.

[Uni23] Japan University. *Encyclopedia of Immunology*. 2023. URL: https://rnavi.ndl.go.jp/mokuji_html/000003269982.html.

5.2 Conclusion: major Limitations of existing Deconvolution Algorithms Solutions

All the the deconvolution methods we reviewed in Section 5.1 exhibit limited performance in estimating rare cell populations, or delineating cell populations with closely related molecular profiles, limitations that were also reported in the comparative review by [FT18]. We posit that the biologically erroneous assumption of the absence of gene-gene interactions is one of the major sources of the limited robustness and reproducibility of deconvolution algorithms.

To address these shortcomings, in the final chapter of this thesis (Section 6.1), we propose an innovative deconvolution algorithm, `DeCovarT`, assuming a paradigm shift by considering a multivariate and integrated approach to unravel the constituents of the "soup of transcripts".

Chapter 6

Article 4: DeCovarT, a deconvolution algorithm leveraging co-expression networks

Statistical Objective: Enhancing the Robustness of Deconvolution Algorithms through Integration of Co-Expression Networks The truly innovative statistical contribution of this thesis is underscored by the implementation of a new deconvolution algorithm, "DeCovarT", to address common limitations of recently developed deconvolution solutions. Notably, gold-standard approaches (see Chapter 5), such as CIBERSORT, tend to underperform when estimating closely related, or rare, cell populations. In order to alleviate these limitations, we relax the assumption of independence between individual gene expressions, integrating explicitly transcriptomic co-expression networks in the generative model. Finally, by constraining that the parameters controlling the individual distribution of cell profiles are known prior to the simulation, and by modelling explicitly the stochastic nature of the regulation of transcriptomic expression, we indeed expected that the method proposed would be less sensible to sample-specific variations of the transcriptome.

Specifically, we modelled the purified expression profile, characterizing the transcriptomic expression of each cell population, as a random vector following a multivariate Gaussian distribution. The associated covariance matrix explicitly encodes direct gene interactions. In addition, we assume that the bulk transcriptomic mixture can be reconstructed as the convolution of these multivariate variables describing the cell expression profiles, each weighted by unknown cellular ratio.

6.1 Article 4

DeCovarT: A Probabilistic and Multidimensional Framework for Cellular Deconvolution in Heterogeneous Biological Samples

Bastien Chassagnol[1,2,*], Grégory Nuel[2], Etienne Becht[1]

1 Institut De Recherches Internationales Servier (IRIS), FRANCE
2 LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), Sorbonne Université, 4, place Jussieu, 75252 PARIS, FRANCE

*** bastien_chassagnol@laposte.net**

Abstract

Although bulk transcriptomic analyses have greatly contributed to a better understanding of complex diseases, their sensibility is hampered by the highly heterogeneous cellular compositions of biological samples. To address this limitation, computational deconvolution methods have been designed to automatically estimate the frequencies of the cellular components that make up tissues, typically using reference samples of physically purified populations. However, they perform badly at differentiating closely related cell populations.

We hypothesised that the integration of the covariance matrices of the reference samples could improve the performance of deconvolution algorithms. We therefore developed a new tool, DeCovarT, that integrates the structure of individual cellular transcriptomic network to reconstruct the bulk profile. Specifically, we inferred the ratios of the mixture components by a standard maximum likelihood estimation (MLE) method, using the Levenberg-Marquardt algorithm to recover the maximum from the parametric convolutional distribution of our model. We then consider a reparametrisation of the log-likelihood to explicitly incorporate the simplex constraint on the ratios. Preliminary numerical simulations suggest that this new algorithm outperforms previously published methods, particularly when individual cellular transcriptomic profiles strongly overlap.

1 Introduction

The analysis of the bulk transcriptome provided new insights on the mechanisms underlying disease development. However, such methods ignore the intrinsic cellular heterogeneity of complex biological samples, by averaging measurements over several distinct cell populations. Failure to account for changes of the cell composition is likely to result in a loss of *specificity* (genes mistakenly identified as differentially expressed, while they only reflect an increase in the cell population naturally producing them) and *sensibility* (genes expressed by minor cell populations are amenable being masked by highly variable expression from major cell populations).

Accordingly, a range of computational methods have been developed to estimate cellular fractions, but they perform poorly in discriminating cell types displaying high phenotypic proximity. Indeed, most of them assume that purified cell expression profiles are fixed observations, omitting the variability and intrinsically interconnected structure of the transcriptome. For instance, the gold-standard deconvolution algorithm *CIBERSORT* [New15] applies nu-support vector regression (ν-SVR) to recover the minimal subset of the most informative genes in the purified signature matrix. However, this machine learning approach assumes that the transcriptomic expressions are independent.

In contrast to these approaches, we hypothesised that integrating the pairwise covariance of the genes into the reference transcriptome profiles could enhance the performance of transcriptomic deconvolution methods. The generative probabilistic model of our algorithm, *DeCovarT* (Deconvolution using the Transcriptomic Covariance), implements this integrated approach.

2 Model

First, we introduce the following notations:

- $(\boldsymbol{y} = (y_{gi}) \in \mathbb{R}_+^{G \times N}$ is the global bulk transcriptomic expression, measured in N individuals.
- $\boldsymbol{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G \times J}}$ the signature matrix of the mean expression of G genes in J purified cell populations.
- $\boldsymbol{p} = (p_{ji}) \in]0,1[^{J \times N}$ the unknown relative proportions of cell populations in N samples

As in most traditional deconvolution models, we assume that the total bulk expression can be reconstructed by summing the individual contributions of each cell population weighted by its frequency, as stated explicitly in the following linear matricial relationship (Equation (1)):

$$\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{p} \tag{1}$$

In addition, we consider unit simplex constraint on the cellular ratios, $\boldsymbol{p}$ (Equation (2)):

$$\begin{cases} \sum_{j=1}^{J} p_j = 1 \\ \forall j \in \widetilde{J} \quad p_j \geq 0 \end{cases} \tag{2}$$

2.1 Standard linear deconvolution model

However, in real conditions with technical and environmental variability, strict linearity of the deconvolution does not usually hold. Thus, an additional error term is usually considered, and without further assumption on the distribution of this error term, the usual approach to retrieve the best of parameters is by minimising the squared error term between the mixture expressions predicted by the linear model and the actual observed response. This optimisation task is achieved through the ordinary least squares (OLS) approach (Equation (3)),

$$\hat{\boldsymbol{p}}_i^{\mathrm{OLS}} \equiv \underset{\boldsymbol{p}_i}{\arg\min} \, ||\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i||^2 = \underset{\boldsymbol{p}_i}{\arg\min} \, ||\boldsymbol{X}\boldsymbol{p}_i - \boldsymbol{y}_i||^2 = \sum_{g=1}^{G} \left(y_{gi} - \sum_{j=1}^{J} x_{gj} p_{ji} \right) \tag{3}$$

If we additionally assume that the stochastic error term follows a *homoscedastic* zero-centred Gaussian distribution and that the value of the observed covariates (here, the purified expression profiles) is determined (see the corresponding graphical representation in Figure 1a and the set of equations describing it Equation (4)),

$$y_{gi} = \sum_{j=1}^{J} x_{gj} p_{ji} + \epsilon_i, \quad y_{gi} \sim \mathcal{N}\left(\sum_{j=1}^{J} x_{gj} p_{ji}, \sigma_i^2\right), \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{4}$$

then, the MLE is equal to the OLS, which, in this framework, is given explicitly by Equation (5):

$$\hat{\boldsymbol{p}}_i^{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}_i \tag{5}$$

and is known under the the Gauss-Markov theorem.

2.2 Motivation of using a probabilistic convolution framework

In contrast to standard linear regression models, we relax in the DeCovarT modelling framework the *exogeneity* assumption, by considering the set of covariates $\boldsymbol{X}$ as random variables rather than fixed measures, in a process close to the approach of DSection algorithm and DeMixt algorithms. However, to our knowledge, we are the first to weaken the independence assumption between observations by explicitly considering a multivariate distribution and integrating the intrinsic covariance structure of the transcriptome of each purified cell population.

To do so, we conjecture that the G-dimensional vector $\boldsymbol{x}_j$ characterising the transcriptomic expression of each cell population follows a multivariate Gaussian distribution, given by Equation (6):

$$\text{Det}(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_{.j})\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_{.j})^\top\right) \tag{6}$$

and parametrised by:

- $\boldsymbol{\mu}_{.j}$, the mean purified transcriptomic expression of cell population j
- $\boldsymbol{\Sigma}_j$, the covariance matrix of each cell population, constrained to be *positive-definite* (see Appendix A.1). Precisely, we retrieve it from inferring its inverse, known as the precision matrix, through the gLasso [Maz11] algorithm. We define $\boldsymbol{\Theta}_j \equiv \boldsymbol{\Sigma}_j^{-1}$ the corresponding *precision matrix*, whose inputs, after normalisation, store the partial correlation between two genes, conditioned on all the others. Notably, pairwise gene interactions whose corresponding off-diagonal terms in the precision matrix are null are considered statistically spurious, and discarded.

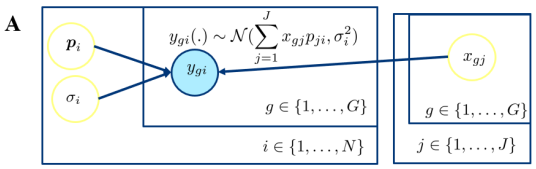
To derive the log-likelihood of our model, first we *plugged-in* the mean and covariance parameters $\zeta_j = \left(\boldsymbol{\mu}_{.j}, \boldsymbol{\Sigma}_j\right)$ estimated for each cell population in the previous step. Then, setting $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j \in \widetilde{J}} \in \mathcal{M}_{G \times J}, \quad \boldsymbol{\Sigma} \in \mathcal{M}_{G \times G}$ the known parameters and $\boldsymbol{p}$ the unknown cellular ratios, we show that the conditional distribution of the observed bulk mixture, conditioned on the individual purified expression profiles and their ratios in the sample, $\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p})$, is the convolution of pairwise independent multivariate Gaussian distributions. Using the *affine invariance* property of Gaussian distributions, we can show that this convolution is also a multivariate Gaussian distribution, given by Equation (7).

$$\boldsymbol{y}|(\boldsymbol{\zeta}, \boldsymbol{p}) \sim \mathcal{N}_G(\boldsymbol{\mu}\boldsymbol{p}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j \in \widetilde{J}}, \quad \boldsymbol{p} = (p_1, \ldots, p_J) \text{ and } \boldsymbol{\Sigma} = \sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j \tag{7}$$

The DAG associated to this modelling framework is shown in Figure 1b).

In the next section, we provide an explicit formula of the log-likelihood of our probabilistic framework, its gradient and hessian, which in turn can be used to retrieve the MLE of our distribution.

(a) Standard linear model representation.



(b) The generative model used for the DeCovarT framework.

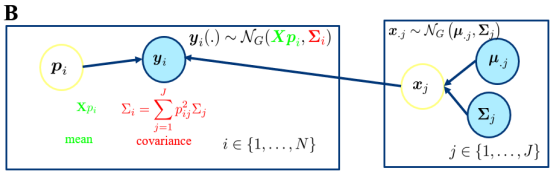


Figure 1. We use the standard graphical convention of graphical models, as depicted in RevBayes webpage. For identifiability reasons, we conjecture that all variability proceeds from the stochastic nature of the covariates.

2.3 Derivation of the log-likelihood

From Equation (7), the conditional log-likelihood is readily computed and given by Equation (8):

$$\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p}) = C + \log\left(\text{Det}\left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1}\right) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu})^\top \left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1} (\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu}) \tag{8}$$

2.4 First and second-order derivation of the unconstrained DeCovarT log-likelihood function

The stationary points of a function and notably maxima, are given by the roots (the values at which the function crosses the x-axis) of its gradient, in our context, the vector: $\nabla\ell : \mathbb{R}^J \to \mathbb{R}^J$ evaluated at point $\nabla\ell(\boldsymbol{p}) :]0,1[^J \to \mathbb{R}^J$. Since the computation is the same for any cell ratio p_j, we give an explicit formula for only one of them (Equation (9)):

$$\begin{aligned}
\frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p})}{\partial p_j} &= \tfrac{\partial \log(\text{Det}(\boldsymbol{\Theta}))}{\partial p_j} - \tfrac{1}{2}\left[\tfrac{\partial(\boldsymbol{y}-\boldsymbol{\mu p})^\top}{\partial p_j}\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p}) + (\boldsymbol{y}-\boldsymbol{\mu p})^\top \tfrac{\partial \boldsymbol{\Theta}}{\partial p_j}(\boldsymbol{y}-\boldsymbol{\mu p}) + (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta} \tfrac{\partial(\boldsymbol{y}-\boldsymbol{\mu p})}{\partial p_j}\right] \\
&= -\text{Tr}\left(\boldsymbol{\Theta}\tfrac{\partial \boldsymbol{\Sigma}}{\partial p_j}\right) - \tfrac{1}{2}\left[-\boldsymbol{\mu}_{.j}^\top \boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p}) - (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta} \tfrac{\partial \boldsymbol{\Sigma}}{\partial p_j} \boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p}) - (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j}\right] \\
&= -2p_j \,\text{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\right) + (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j} + p_j (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_j \boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p})
\end{aligned} \tag{9}$$

Since the solution to $\nabla\left(\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p})\right) = 0$ is not closed, we had to approximate the MLE using iterated numerical optimisation methods. Some of them, such as the Levenberg–Marquardt algorithm, require a second-order approximation of the function, which needs the computation of the Hessian matrix. Deriving once more Equation (9) yields the Hessian matrix, $\mathbf{H} \in \mathcal{M}_{J\times J}$ is given by:

$$\begin{aligned}
\mathbf{H}_{i,i} = \frac{\partial^2 \ell}{\partial^2 p_i} &= -2\,\text{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) + 4p_i^2\,\text{Tr}\left(\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right)^2\right) - 2p_i(\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - \boldsymbol{\mu}_{.i}^\top\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - \\
&\quad 2p_i(\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - (\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\left(4p_i^2\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_i\right)\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p}), \quad i \in \widetilde{J} \\
\mathbf{H}_{i,j} = \frac{\partial^2 \ell}{\partial p_i \partial p_j} &= 4p_jp_i\,\text{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) - 2p_i(\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.j} - \boldsymbol{\mu}_{.i}^\top\boldsymbol{\Theta}\boldsymbol{\mu}_{.j} - \\
&\quad 2p_j(\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - 4p_ip_j(\boldsymbol{y}-\boldsymbol{\mu p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}(\boldsymbol{y}-\boldsymbol{\mu p}), \quad (i,j) \in \widetilde{J}^2, i \neq j
\end{aligned} \tag{10}$$

in which the coloured sections pair one by one with the corresponding coloured sections of the gradient, given in Equation (9).

Matrix calculus can largely ease the derivation of complex algebraic expressions, thus we remind in Appendices A.1 and A.2 relevant matrix properties and derivations. The numerical consistency of these derivatives was asserted with the **numDeriv** package, using the more stable Richardson's extrapolation ([For81]).

However, the explicit formulas for the gradient and the hessian matrix of the log-likelihood function, given in Equation (9) and Equation (10) respectively, do not take into account the simplex constraint assigned to the ratios. While some optimisation methods use heuristic methods to solve this problem, we consider alternatively a reparametrised version of the problem, detailed comprehensively in Appendix A.3.

3 Simulations

3.1 Simulation of a convolution of multivariate Gaussian mixtures

To assert numerically the relevance of accounting the correlation between expressed transcripts, we designed a simple toy example with two genes and two cell proportions. Hence, using the simplex constraint (Equation (2)), we only have to estimate one free unconstrained parameter, θ_1, and then uses the mapping function, defined in Appendix A.3 to recover the ratios in their original space.

We simulated the bulk mixture, $\boldsymbol{y} \in \mathcal{M}_{G\times N}$, for a set of artificial samples $N = 500$, with the following generative model:

- We have tested two levels of cellular ratios, one with equi-balanced proportions ($\boldsymbol{p} = (p_1, p_2 = 1 - p_1) = (\frac{1}{2}, \frac{1}{2})$ and one with highly unbalanced cell populations: $\boldsymbol{p} = (0.95, 0.05)$.
- Then, each purified transcriptomic profile is drawn from a multivariate Gaussian distribution. We compared two scenarios, playing on the mean distance of centroids, respectively $\mu_{.1} = (20, 22), \mu_{.2} = (22, 20)$ and $\mu_{.2} = (20, 40), \mu_{.2} = (40, 20)$) and building the covariance matrix, $\boldsymbol{\Sigma} \in \mathcal{M}_{2\times 2}$ by assuming equal individual variances for each gene (the diagonal terms of the covariance matrix, $\text{Diag}(\boldsymbol{\Sigma_1}) = \text{Diag}(\boldsymbol{\Sigma_1}) = \boldsymbol{I}_2$) but varying the pairwise correlation between gene 1 and gene 2, $\mathbb{C}\text{ov}\left[x_{1,2}\right]$, on the following set of values: $\{-0.8, -0.6, \ldots, 0.8\}$ for each of the cell population.
- As stated in Equation (1), we assume that the bulk mixture, $\mathbf{y}_{.i}$ could be directly reconstructed by summing up the individual cellular contributions weighted by their abundance, without additional noise.

3.2 Iterated optimisation

The extremum, and by extension the MLE, is a root of the gradient of the log-likelihood. However, in our generative framework, the inverse function cancelling the gradient of Equation Equation (8) is non-closed. Instead, iterated numerical optimisation algorithms that consider first or second-order approximations of the function to optimise are used to approximate the roots.

The *Levenberg-Marquardt (LM)* algorithm ([Lev44]) bridges the gap between between the steepest descent method (first-order) and the Newton-Raphson method (second-order) by inflating the diagonal terms of the Hessian matrix. Far from the endpoint, a second-order descent is favoured for its faster convergence pace, while the steepest approach is privileged close to the extremum since it allows careful refinement of the step size. Specially, we used the LM implementation of R package `marqLevAlg` to infer estimates of the cellular ratios from the bootstrap simulations ([Phi+21]). It notably includes an additional convergence criteria, the relative distance to the maximum (RDM), that sets apart extrema from spurious saddle points.

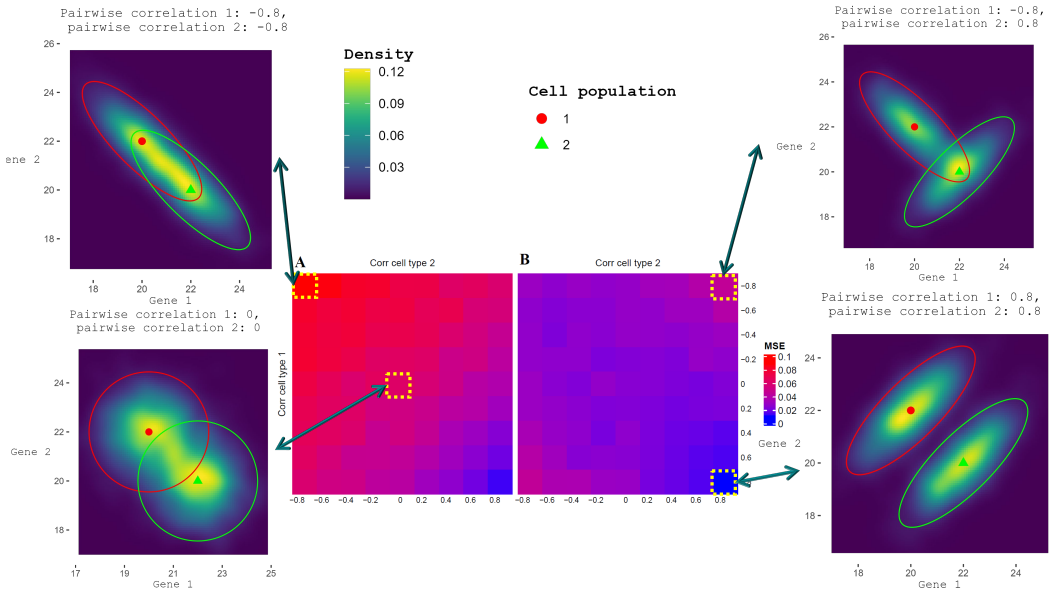


Figure 2. We used the package **ComplexHeatmap** to display the mean square error (MSE) of the estimated cell ratios, comparing the NNLS output, as implemented in the DeconRNASeq algorithm ([Gon13]), in Panel **A**, with our newly implemented DeCovarT algorithm, in Panel **B**. The lower the MSE, the least noisy and biased the estimates. In addition, we added the two-dimensional density plot for the intermediate scenario, for which each population is parameterised by a diagonal covariance matrix, and the most extreme scenarios (those with the highest correlation between genes). The ellipsoids represent for each cell population the 95% confidence region and the red spherical icon and the green triangular icon represent respectively the centroids (average expression of gene 1 and gene 2) of cell population 1 and cell population 2.

3.3 Results

We compared the performance of DeCovarT algorithm with the DeconRNASeq algorithm ([Gon13]).

Even with a limited toy example including two cell populations characterised only by two genes, we observe that the overlap was a good proxy of the quality of the estimation: the less the overlap between the two cell distributions, the better the quality of the estimation Figure 2.

The package used to generate the simulations and infer ratios from virtual or real biological mixtures with the DeCovarT algorithm is implemented on my personal Github account DeCovarT.

4 Perspectives

The new deconvolution algorithm that we implemented, DeCovarT, is the first one based on a multivariate generative model while enforcing explicitly the simplex constraint. Hence, it provides a strong basis to further derive statistical tests to assert whether the proportion of a given cell population differs significantly between two distinct biological conditions.

Extend the Simulation Framework To evaluate the biological and statistical interest of `DeCovarT`, we need to expand the simulation framework, by encompassing a larger number of cell types, genes, and testing the sensitivity of the model by voluntarily including noise in the benchmark evaluation.

The next phase of our evaluation involves real-world experiments, encompassing both blood and solid tumoral samples. To that end, we could start from the Kassandra benchmark, by [Zai22]. This large-scale project evaluates the performance of five established gold-standard and signature-based deconvolution algorithms, including EPIC [Rac17], CIBERSORT [New15], CIBERSORTx [New+19], quanTIseq [Fin19], and ABIS [Mon+19]. The evaluation involves deconvolving six publicly available datasets annotated with both flow cytometry and bulk RNA-seq expression.

Enhanced Inference and Integration of Co-Expression Networks All the popular differential gene expression analyses, such as *limma + voom* ([Rit+15]), *EdgeR* ([RMS10]) and *Deseq2* ([Var+16]), tend to overlook gene-gene interactions, comparing independently the expression between two conditions for each gene. The usual univariate approach of DGEAs additionally underlies the need of adjusting $p-$values, as numerous genes are examined simultaneously, and without accounting for interactions between them, the probability of observing false positives increases.

To account for correlations among observations, two consecutive papers, by [CL23] and [CCB22], present an innovative Bayesian framework which models proteomic expression across diverse biological conditions as multivariate Gaussian distributions. Insightful discussions with the main author, Marie Chion, suggest a straightforward extension of the method to transcriptomic expression, given the close relationship between two kinds of omics, both depicting counts.

While the methodology was originally designed to delineate differentially expressed genes between two conditions, the method can be readily extended to incorporate a *one-vs-all strategy*. This extension allows for the identification of markers specific to a particular cell population in comparison to all others. Furthermore, the generative model aligns closely with our deconvolution framework, leveraging the same distributions to describe cellular omic profiles. Alternatively, differential network approaches, such as INDEED, by [Zuo16], implement heuristic and dual-optimisation approaches, finding the sweet spot between maximising the mean differences between purified expression profiles and differentiating the neighbourhood network structure.

The gLasso algorithm used to derive the precision matrix associated to each purified cell profile is subjected to *parameter shrinkage*, like any penalty regularisation approach. Notably, in our setting, shrinkage tends to systematically underestimate the non-zero partial correlations of the precision matrix.

To mitigate this issue, one approach is to incorporate the *support* (indicating non-null inputs), derived from the gLasso output, into a conventional Maximum Likelihood Estimation (MLE) framework. The general concept is to utilising the true "zeros" to impose topological constraints on the final Gaussian Graphical Model (GGM). However, it's important to note that unless the undirected Markov network obtained from the gLasso output is a *chordal graph*, there is usually no straightforward mapping between the two topological spaces.

Finally, the inclusion of prior biological knowledge, such as the strength of relationships between transcription factors, retrieved from Protein-Protein Interaction (PPI) networks, can help reduce the exponential space of undirected graphs to explore.

Enhanced Inference and Integration of Co-Expression Networks All the methods outlined in Section 4 yield a subset of genes that distinguish a particular cell population from all others. However, when we combine these gene subsets, we often end up with a non-scalable signature matrix, presenting strong multicollinearity resulting from the redundancy between the gene markers identified.

To further refine the final set of genes able to delineating any cell population included in the signature matrix, *AutoGeneS*, by [Ali21], introduces a greedy genetic approach coupled with a dual optimisation approach [1]. Precisely, the loss function involves minimising inter-population correlation while simultaneously maximising the distance of the centroids.

We propose instead of this dual optimisation approach the minimisation of the *global overlap* between the concatenated distributions of the cellular profiles. Indeed, this metric not only captures in a single criterion the combined influence of mean inter-cluster distance and differential network structure in delineating purified cellular expression profiles, but supplies a straightforward score easy to interpret. The *overlap* metric precisely measures the shared probability mass and the degree of concurrence in probability densities. In simpler terms, it quantifies the global probability of incorrectly assigning an expression profile to the wrong cell subtype when utilising a maximum a posteriori approach, with the knowledge of each cellular profile's individual parameters.

[1] In standard approaches that rely on linear regression, the condition number serves as the gold-standard metric for assessing the level of precision of the linear model achievable with the design matrix

Joint Estimation of purified Expression Profiles and Cellular Ratios The generative model underlying the `DeCovarT` framework (Figure 1b) assumes that both the ratios and the purified cellular expression profiles are unobserved and need to be inferred from our model. However, we derived explicit formulas for the Gradient (eq. (9)) and Hessian (eq. (10)) of the associated log-likelihood function as if the purified expression profiles had been observed, by heuristically replacing the unknown and sample-specific purified expression profiles $\boldsymbol{X}_{.i}$ with their averaged counterparts $\boldsymbol{\mu}$. However, jointly optimising the cellular ratios and the purified expression profiles results in a non identifiable problem exhibiting an infinite number of solutions, without strong prior assumptions or regularisation of the unknown parameters to estimate. Finally, it's a highly intractable analytical task, and it is quite likely that no explicit form of the Gradient, nor the Hessian, could be derived.

We detail in Appendix B a Gibbs sampler to approximate the target distribution, here the joint value of the purified profiles and the cellular ratios. In addition, MCMC sampling allows for straightforward incorporation of prior knowledge, and streamlines the derivation of Maximum a Posteriori (MAP) estimates and *credible intervals*.

Precisely, by coupling Gibbs and Metropolis Hasting samplers, we ensure at each iteration that the estimated parameters adhered to the "balance condition", an essential property guaranteeing the convergence of MCMC chains to a stationary distribution identifiable to the target distribution.

References

[Lev44] Levenberg, Kenneth. "A method for the solution of certain non-linear problems in least squares". In: *Quarterly of Applied Mathematics* (1944). ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/10666. URL: https://www.ams.org/qam/1944-02-02/S0033-569X-1944-10666-0/.

[For81] Bengt Fornberg. "Numerical Differentiation of Analytic Functions". In: *ACM Trans. Math. Softw.* (1981). DOI: 10.1145/355972.355979. URL: https://doi.org/10.1145/355972.355979.

[RMS10] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data". In: *Bioinformatics* (Jan. 1, 2010). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: https://doi.org/10.1093/bioinformatics/btp616.

[Maz11] Mazumder, Rahul and Hastie, Trevor. "The Graphical Lasso: New Insights and Alternatives". In: *Electronic Journal of Statistics* (2011). DOI: 10.1214/12-EJS740.

[Gon13] Gong, Ting and Szustakowski, Joseph D. "DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data". In: *Bioinformatics (Oxford, England)* (2013). DOI: 10.1093/bioinformatics/btt090.

[New15] Newman, Aaron and Liu, Chih and others. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles". In: *Nature methods* (2015). DOI: 10.1038/nmeth.3337.

[Rit+15] Matthew E. Ritchie et al. "Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies". In: *Nucleic Acids Research* (Apr. 20, 2015). ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. URL: https://doi.org/10.1093/nar/gkv007.

[Var+16] Hugo Varet et al. "SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data". In: *PLOS ONE* (June 9, 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0157022. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157022.

[Zuo16] Zuo, Yiming and Cui, Yi and others. "INDEED: Integrated Differential Expression and Differential Network Analysis of Omic Data for Biomarker Discovery". In: *Methods (San Diego, Calif.)* (2016). DOI: 10.1016/j.ymeth.2016.08.015.

[Rac17] Racle, Julien and de Jonge, Kaat and others. "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data". In: *eLife* (2017). Ed. by Alfonso Valencia. DOI: 10.7554/eLife.26476.

[Fin19] Finotello, Francesca and Mayer, Clemens and others. "Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-seq Data". In: *Genome Medicine* (2019). DOI: 10.1186/s13073-019-0638-6.

[Mon+19] Gianni Monaco et al. "RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types". In: *Cell Reports* (2019). DOI: 10.1016/j.celrep.2019.01.041.

[New+19] Aaron M. Newman et al. "Determining cell type abundance and expression from bulk tissues with digital cytometry". In: *Nature Biotechnology* (2019). DOI: 10.1038/s41587-019-0114-2.

[Ali21] Aliee, Hananeh and Theis, Fabian J. "AutoGeneS: Automatic Gene Selection Using Multi-Objective Optimization for RNA-seq Deconvolution". In: *Cell Systems* (2021). DOI: 10.1016/j.cels.2021.05.006.

[Phi+21] Viviane Philipps et al. "Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg". In: *The R Journal* (2021). ISSN: 2073-4859. DOI: 10.32614/RJ-2021-089. URL: http://arxiv.org/abs/2009.03840.

[CCB22] Marie Chion, Christine Carapito, and Frédéric Bertrand. "Accounting for Multiple Imputation-Induced Variability for Differential Analysis in Mass Spectrometry-Based Label-Free Quantitative Proteomics". In: *PLOS Computational Biology* (Aug. 29, 2022). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010420. URL: http://arxiv.org/abs/2108.07086.

[Zai22] Zaitsev, Aleksandr and Chelushkin, Maksim and others. "Precise Reconstruction of the TME Using Bulk RNA-seq and a Machine Learning Algorithm Trained on Artificial Transcriptomes". In: *Cancer Cell* (2022). DOI: 10.1016/j.ccell.2022.07.006.

[CL23] Marie Chion and Arthur Leroy. "A Bayesian Framework for Multivariate Differential Analysis Accounting for Missing Data". In: (July 18, 2023). DOI: 10.48550/arXiv.2307.08975. arXiv: 2307.08975 [stat]. URL: http://arxiv.org/abs/2307.08975 (visited on 02/01/2024). preprint.

6.2 Conclusion

In a reduced toy example, involving two cell populations characterized solely by two genes, we first demonstrated that the level of mutual correlation between the two transcripts significantly influences the final precision of the estimated cellular ratios. We additionally exhibited that the loss of performance resulting from overlapping purified profiles was partially mitigated with `DeCovarT`, thanks to the explicit integration of underlying transcriptomic networks, compared to standard algorithm `DeconRNASeq` that relies on a quadratic optimisation approach. Moreover, the utilization of a generative model simplifies the development of statistical tests, allowing for a meaningful detection of significant changes in cell composition.

As detailed in Section 6.1, the current implementation version of `DeCovarT`, did not scale well with noisy datasets or unbalanced cellular composition. As briefly evoked in the manuscript, it turned out that the significant discrepancy observed between the estimated cellular ratios and the values used to simulate the artificial bulk mixtures proceeds essentially from the heuristic approach we adopted to integrate the unobserved and individual cellular expression profiles. Briefly, instead of considering them as another multidimensional parameter to estimate, in this first approach, we approximate sample-specific expression variations by the prior mean parameter.

As, the joint estimation of the purified expression profiles and the cellular ratios is challenging, we briefly introduce a Gibbs sampling approach in the perspectives of the paper to alleviate the analytical burden. Additionally, this novel modelling framework would provide a more comprehensive description of the transcriptomic variability, at the individual level.

6.3 Publication Outline

Adopting a network-based and multidimensional approach to model the bulk mixture impacts the whole deconvolution process, from the selection of marker genes to the derivation of statistical tests asserting a level of confidence to the cellular ratios estimated (chapter 5).

Accordingly, we plan to decompose the publication process in three stages. First, a methodological paper, in collaboration with [Chi18] and extending the findings from [AT21], would focus on a differential network approach to select the set of best discriminating genes to be included in the signature matrix. Subsequently, a forthcoming manuscript, extending the current arXiv preprint, will provide a comprehensive exposition of the fundamental theoretical concepts that underpin the development of the `DeCovarT` algorithm. We intend to submit this preliminary manuscript to a peer-reviewed journal renowned for its statistical acumen. `Computo`, by [Chi23], would completely fulfil these demands, since the editorial boarding focuses not only on the quality and originality of the statistical methods submitted, but also emphasises on the reproducibility of the results by constraining the submission of data necessary to repeat the simulations and enforcing efficient coding.

Furthermore, I have initiated a collaborative effort with Michael Rera and Savandara Besse to empirically validate the performance and robustness of the `DeCovarT` algorithm on a practical biological case study. Specifically, Michael Rera's research team is keen on applying the first cellular deconvolution method to Drosophila transcriptomic samples, with the ultimate goal of advancing their understanding of the causal mechanisms underlying the ageing process.

We conclude this thesis in the following section 6.3 by reviewing some avenues and ongoing endeavours to improve the biological applicability and statistical relevance of the models presented

in this thesis. We notably discuss the opportunities presented by the integration of orthogonal biological modalities, and suggest, on a side note, to generalise the models presented to non-Gaussian distributions that may offer a more accurate representation of the biological context.

Thesis overview

Conclusion Throughout this PhD manuscript, we adopted a data-driven approach to explore the biological drivers underlying the variability of biological mechanisms involved in the evolution and severity of diseases. In particular, we explored the relevance of two biological inputs, namely immune cell populations and transcriptomic expression profiles, to better understand the drivers of the intrinsic heterogeneity of living human organisms.

In addition, recently, the advent of cutting-edge methodologies like next-generation sequencing has led to an exponential growth in biomedical data. This staggering data volume has posed a challenge for human researchers in identifying meaningful biological patterns. Simultaneously, there has been a notable shift in the therapeutic paradigm, transitioning from "one-size-fits-all" treatments to Precision Medicine, theoretically customized to address each patient's unique biological profile.

In order to navigate the magnitude and intricacy of these extensive datasets, we conducted an extensive review of various statistical methodologies to unravel the biological determinants underlying the fluctuations observed in biologiocal activity and response. Most of them expand directly Gaussian distributions, precisely we exhibit the interest of mixtures models and convolution of independent variables to adjust to complex and multi-modal distributions that do not not seem normally distributed at first glance.

At a low level of granularity, the tropic context, such as disease state and tissue location, strongly impacts the variety of transcriptomic profiles across individuals. However, even within a cohort of patients affected by the same ailment, we observed strong heterogeneity of clinical factors, which often translates in varying response to clinical treatments. The unaccounted variability observed among patients afflicted by the same disease suggests the presence of an underlying latent factor influencing the progression of the disease. We demonstrated in Part II the interest of mixture models, and in particular Gaussian Mixture Models, for the unsupervised classification and stratification of heterogeneous individual profiles into endotypes.

We first started by reviewing a whole set of computational solutions, implemented as R packages, to infer gaussian-distributed clusters. We compared for the seven most popular ones their relevance in estimating the hidden parameters of GMMs, visualising the clustering inferred and asserting the quality of the estimation. We notably demonstrate the versatility of the `mclust` package, providing the end-users with a vast array of computational methods and visualisations to perform tasks as varied as initiating the cluster assignments, providing prior information in a Bayesian framework or extracting key variables for visualisation in high dimension. Interestingly, we also pinpointed the impact of initialisation and the characteristics of the mixture (overlap, number of components, ...) on the accuracy of the final estimate returned, exhibiting significant differences across packages.

We ended this section by a practical biological use case, in which we leveraged the `mclust` package to separate and classify patients from the same cohort affected by the Sjögren's primary

syndrome into endotypes. This data-driven, unsupervised approach, pioneered by the number of patients and omics integrated in the study, revealed notably the presence of four clearly distinct endotypes, characterised by specific molecular fingerprints. On top of hat, sensibility analyses revealed that most of the transcriptomic differences between subgroups of patients were actually driven by changes of the cell pool, as revealed by both physical, cytometry-based approaches and numerical deconvolution algorithms.

This finding led us in focusing on data-driven, numerical approaches, referred to as deconvolution algorithms, to estimate the cellular composition of heterogeneous biological samples in part III. They notably enable to retrospectively analyse historical bulk analysis, for which the original raw material is not available anymore and no cytometry analysis has been carried out.

We started to review a number of them in the last part of this PhD manuscript, focusing on partial, semi-supervised approaches relying on cell transcriptomic references. In brief, these signatures of purified cell population are composed of the averaged transcriptomic expression measures for a minimal set of highly discriminative genes. We showed however that not of these methods account for intrinsic co-expression networks within a homogeneous cell population.

Hence, we conclude our manuscript by the introduction of a new standalone deconvolution tool, `DeCovarT`. In opposition to gold-standard deconvolution approaches, such as `CIBERSORT` ([New+15]), we embraced a paradigmatic shift by adopting a systematic and connected approach of the deconvolution problem. To that end, we modelled the bulk transcriptome as a convolution of multi-dimensional variables, each characterising the distribution of a purified cell expression profile. Expanding the generative model to a multi-dimensional framework facilitates the incorporation of transcriptomic co-expression structures, encoded explicitly in the covariances of the individual cell profiles. This new feature, in turn, has been found to significantly improve the precision and robustness to noisiness of cellular ratio estimations.

Genes, environment, proteins, cell populations, and many other biological entities intertwine altogether and impact general well-being and treatment outcomes. Biological systems, far from being the mere sum of atomic entities acting independently, are better described by large networks connecting a variety of biological processes.

Similarly, only the cooperation between a variety of life-research fields are susceptible to unravel the causal mechanisms underlying the complexity of biological systems. I further highlight in next Section 6.3 the relevance of systematic approaches integrating several biological modalities in the field of computational medicine.

Perspectives We voluntary choose to restrain on Gaussian distributions and direct extensions of the to model the variability of the interactions occurring between biological features. Indeed, normally distributed datasets exhibit a number of interesting statistical properties, streamlining the derivation of complex analytical formulas and the inference of confidence intervals to assert the quality of the model. However, they perform quite badly in approximating real-world biological data when the assumptions of symmetric and bell-shaped distributions do not hold. In particular, they tend to underperform in the presence of outliers, high-dimensional datasets, or skewed or inflated distributions.

It would thus be interesting to compare the performance of the computational tools we implement and deploy with other kinds of distributions, closer to the reality. For instance, transcriptomic data is by nature counts, usually exhibiting strong zero-inflation and negative skewness ([MLR12]). Hence, mixtures or convolutions of Poisson log-Normal (PLN) or Negative Binomials ([RMS10]) are likely better tailored for describing mixtures or convolutions of raw bulk RNASeq datasets.

For instance, PLN distributions have been used in [CMR21] to propose a new Joint Species Distribution Models (JSDM) for studying the combined abundances of multiple species in ecological communities. We can predict that the versatility of PLN models would enable to easily extend that study to transcriptome, encoding gene co-expression dependencies in the covariance matrix instead of species relations.

It is important to acknowledge that the accuracy of the methods we have devised to unravel the intricacies of biological systems among patients, or even within tissues, is somewhat limited since they only leverage a single biological entity. For example, the patient stratification in chapter 4 is solely reliant on individual transcriptomic profiles, processed from bulk RNASeq techniques. Integrated methods aimed at combining diverse biological modalities. They are usually classified into *early*, *late*, or *intermediate* approaches ([Pic+21]). Early integration merges all datasets into a single comprehensive dataset for model learning, while late integration builds individual models for each dataset and then combines the resulting probabilistic frameworks.

On the other hand, intermediate approaches represent an hybrid strategy, aiming at deriving a model for each modality, strengthened by the input of other relevant biological annotations. An array of network-based approaches, such as Similarity Network Fusion (SNF, [Wan+14]), Non-Negative Matrix Tri-Factorization (NNMF, [Mal+19]) or spectral clustering ([Don+14]), have been precisely designed to uncover shared structures and similarities across multi-layered graphs, connecting several kinds of omics and clinical features.

However, striking the perfect balance between integrating diverse datasets, which may lack of shared biological mechanism, and creating a robust and insightful model, is a challenging task, especially for this kind of knowledge hypergraphs exhibiting large noises and numerous spurious correlations.

Studying omics and clinical datasets, even more so when combining distinct biological features together, generated from distinct medical centres, is highly challenging. This analytic complexity is further strengthened by the diversity of tools and methods to address roughly the same objectives. In particular, I should emphasize on the importance of data cleaning and preprocessing, coupled with clean documentation of the parameters of the algorithms used, to ensure the reproducibility and robustness of the statistical models. Indeed, the quality and abundance of datasets and the choice of the hyperparameters controlling the behaviour of the algorithm have a substantial impact on the performance of any model, even, and maybe more critical, on the most advanced approaches. As an illustrative example, we demonstrated that even minor alterations in the algorithmic implementations to mitigate numerical underflows strongly impact the outcome of

the benchmarked R packages (Chapter 3). These adjustments to the native EM algorithm, often poorly documented, yield substantial disparities in the accuracy and variance of the estimated parameters for Gaussian mixture estimations, particularly conspicuous in inherently complex distributions characterised by pronounced overlap or disequilibrium across clusters.

Adherence to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles is undeniably a pivotal strategy for addressing the persistent "reproducibility crisis" ([Bak16]), a recurrent issue that undermines the robustness of findings in the biomedical domain. This challenge is poignantly exemplified in the meta-analysis [SB16], an eye-opener in the specific field of mechanistic Boolean models that revealed that fewer than half of the results submitted to BioModels could be consistently reproduced. This issue is, in part, attributed to inadequacies in coding practices and data cleansing within the field of bioinformatics, coupled with lack of comprehensive documentation. In addition to adhering to universal guidelines, any statistical model developed for predicting disease outcomes, based on the molecular profiles of patients, should also adhere to the principles of *robustness* and *interpretability*. Robustness ensures that the model remains stable in the face of minor alterations of the assumptions of the model, and/or the presence of noisiness. An interpretable model enables biologists to readily discern the significance of the results, facilitating the identification of the factors underpinning the algorithm's outcomes. Moreover, the alignment with FAIR principles is likely to foster enhanced interdisciplinary collaboration among teams with diverse expertise. By compelling researchers to conduct their modelling endeavours in accordance with universal guidelines and language, it promotes a more cohesive and collaborative environment.

In order to achieve an universal modelling framework, shared by distinct end-users and integrating knowledge for multiple biological sources, we should highlight in particular two collaborative initiatives, achieved by a consortium of researchers guided by the same ideal of reproducible and reusable science. The SBML language, for Systems Biology Markup Language (see [Kea+20, Fig. 1]) is an unique attempt to describe with the same language seemingly unrelated modelling approaches. It notably encompasses descriptions of Boolean models, metabolic networks (such as flux balance analysis), or more closely related to my statistical background, a `distrib` add-on to store parameters of stochastic distributions. More directly involving and affecting me, Computo ([Chi23]) is an open, peer-reviewed journal developed in response to the reproducibility crisis, focusing not only on the quality and originality of the statistical methods submitted, but also emphasising on the reproducibility of results, including the submission of necessary data and efficient coding.

As a closing note, I truly believe that comprehending the intricacies of biological systems necessitates the adoption of a collaborative and interconnected approach, mirroring in doing so the complex interplay among the atomic units of a living organism that collectively contribute to its biological fitness.

Appendix A

Appendix of Chapter 1

Outline of an industrialised pipeline for analysing transcriptomic expression In this chapter, we detail a comprehensive pipeline to perform gene expression analysis in RNA-seq, encompassing the mapping of raw reads to downstream analysis. This pipeline involves the following steps:

- *Data cleansing*, better known as *data cleaning*, is the process of detecting and correcting corrupt records from an Expression matrix, precisely identifies incomplete or irrelevant parts of the data to gain biological insights and save memory storage. It also involves identifying *primary keys* consistently characterising a sample, and replacing ambiguous terms, such as erroneous gene notations or special non-ASCII character symbols that are not computer-readable on any operating system (report to Appendix A.1).

- *Preprocessing* refer to any task evaluating the quality of the collection of raw reads, which make up the RNA-Seq library, and composes the output of sequencing technologies (refer to Section 1.2.2). Preprocessing operations involve a variety of Quality Control and Quality Filtering, followed by Alignment to a reference genome(report to Appendix A.3).

- Different normalization and transformation functions enable to correct for technical variability, such as library size and sequencing depth and enforce assumptions required by the differential analysis models, respectively. It is usually coupled with a second stage of Quality Control to check the distribution of read counts, in order to ensure data quality post-normalization, and assess the similarity between samples, notably that samples with the same phenotype group together (report to Appendix A.1).

- Downstream analyses usually involve Differential Gene Expression Analysis (DGEA) (Appendix A.4.1) and Functional Pathway Analysis, in order to identify individual genes, or modules of genes, that constitute the fingerprint of a disease/phenotype condition.

- The final phase typically involves establishing a centralized repository to comprehensively document the entire analysis workflow and its primary outcomes. Visual report provides biologists with a succinct and clear overview of the methods employed and the main findings, which in turn simplifies the biological interpretation of the results.

  Experts can additionally evaluate the performance of the pipeline, by confirming the biological significance of the identified differentially expressed genes. For instance, when

working with tumour samples, the absence of observed dysregulations in the mechanisms associated with tumourigenesis can be a cause for concern.

Historically, omic analyses were primarily conducted using point-and-click and user-friendly software like OmicSoft [Li+14]. However, these tools came with a hefty price tag and offered limited access to core functionalities of the functions implemented.

In addition, these tools are not regularly updated, since they do usually dot not provide open-access resources, and depend on the releases and involvement of a limited team.

Ultimately, they do not comply FAIR principles: since their business model requires to pay on a regular basis fees, code is mostly internal, preventing from understanding the choices of implementation.

Consequently, our team of biostatisticians started using individual scripting languages like R, Python, or SAS (although SAS required an extensive license) to conduct omic analyses. While this approach provided greater flexibility, it also resulted in diverse outcomes and procedures, which often depended on the developer's skillset and personal preferences. The development of an integrated RNA-Seq pipeline was conceived to tackle this diversity, with the objective of producing consistent and reliable outputs.

Precisely, I contributed to the Development of the First Industrial RNA-Seq Pipeline as a member of a team of six statisticians at Servier company, all from diverse professional backgrounds, including biostatisticians focusing on biomarker discovery from transcriptomic data and bioinformaticians dedicated to raw sequence analysis (see Appendix A).

Advantages of Developing a Unified Repertoire: Standardising best practices in RNA-Seq analysis encompassed multiple objectives:

- As a team member, my primary goal was to collaborate with colleagues, some of whom I had never interacted with before, in order to establish a robust, consensus-driven, and efficient workflow for processing and analysing transcriptomic data within an industrial context. Simultaneously, I aimed to enhance my own skills in the fields of biostatistics, bioinformatics, and best practices for code sharing. Being part of this multidisciplinary team provided a unique opportunity to learn from experts with diverse backgrounds, thereby expanding my proficiency in a wide field of areas related to computational medicine.

- Additionally, this experience heightened my awareness of the importance of Findable, Accessible, Interoperable, Reusable (FAIR) practices, and I actively advocated for their implementation within my statistical team. Indeed, RNA-seq analysis is a complex and time-consuming process, necessitating the collaboration of multiple teams, which can make it susceptible to human and technical errors. Notably, the choice of tools and parameters may vary depending on the specific research question and dataset characteristics. Therefore, it is crucial to thoroughly document the analysis and identify universally applicable best practices to ensure the reproducibility of results across teams and projects.

  In the industrial context of the Servier environment, the creation of a centralised Github repository for storing all code snippets used in the analysis of transcriptomic data has had a profound impact. It has not only significantly improved the reproducibility of our data-processing workflow but has also resulted in substantial time and computational resource savings, ultimately leading to cost reductions. Moreover, it has increased interdisciplinary collaboration: indeed, one of our primary objectives was to compile the best practices in each key stage of the RNA-Seq workflow, and we achieved this purpose by aggregating enlighten recommendations from domain experts. We notably focused on developing versatile

differential analysis frameworks tailored for complex experimental designs and creating insightful visualisations (PCA, volcano plots, ...) that are both professional-looking and insightful.

- Transcriptomic data, offering detailed insights into gene expression, plays a pivotal role in advancing our understanding of complex biological processes and facilitating the development of personalized treatment strategies. This type of data is a treasure trove of information, actively harnessed by biologists, physicians, and biostatisticians to conduct research on multifaceted and heterogeneous diseases.

  Furthermore, as the cost of collecting and analysing transcriptomic data continues to decline, thanks to automated bioinformatic pipelines and the increasing prevalence of high-throughput RNA-Seq technologies (as discussed in Section 1.2.3 and Section 1.2.2), it is the the cleanest and most abundant sources of omics data at our disposal in the Servier ecosystem, accordingly, all the papers reported throughout this manuscript depend on transcriptomic datasets as raw material inputs, as referenced in Chapter 4, Appendix E, and Appendix F.3). And the first common task of these analyses usually involves rigorously process transcriptomic data and produce comprehensive reports detailing the methods, parameters, and key findings from our analyses.

- My personal specific interest with respect to my PhD project was to implement a robust and versatile transcriptomic routine, specially designed to retrieve the signature matrix of purified cell populations. This matrix is required as input for most partial deconvolution algorithms (see section 5.1). While the order and general principles of each step of the RNA-Seq pipeline are similar, collecting and retrieving a robust and versatile reference profile presents its own specific challenges:

  1. The reference signature matrix must enable flexible, accurate, and robust deconvolution across various tissues and biological conditions. Specific challenges arise due to multi-collinearity and noise resulting from the intrinsic complexity and variability of heterogeneous tissues. This complexity becomes more pronounced when the goal is to discriminate among a large pool of closely related cell populations at different abundance levels.
  2. To generate the most reproducible and informative signature matrix, it's often necessary to collect data from different studies, which may have been conducted on different sequencing platforms. Consequently, it becomes essential to evaluate and correct for blocking variables that introduce strong batch effects and can significantly reduce statistical power and sensitivity.

By pursuing these objectives, I not only contributed to the successful establishment of the RNA-Seq pipeline at Servier but also applied this pipeline in practice in 3-4 pre-clinical studies to which I was personally involved during my thesis. This experience proved to be mutually beneficial for both myself and the industrial team. In the upcoming Appendix A, I will delve into the guidelines we adhere to for the development of this industrial tool.

Modularity design of the pipeline This pipeline follows a modular design approach by dividing the overall workflow into separate, dedicated modules (in our use case, R packages). The key design principle was *uncoupling* the different tasks, as this makes each module easier to maintain, with one dedicated expert for the maintenance of an individual module. It also helps facilitates comprehension of the whole workflow while allowing straightforward access to

specific functions. For example, `boxplot` drawing is logically located in the visualisation module. Interestingly, this strong modularity mirrors the core principles and philosophy underlying the Linux system design [TD01], in which each function focuses on accomplishing perfectly well a single, well-defined task.

The most important packages, ordered by operational sequence and detailed further in this chapter, include:

1. The `bbcWrangling` package provides functions for data management and cleaning. These functions include trimming missing values, filtering samples based on phenotype features, or merging expression matrices together. They are specifically designed to handle the unique format of Bioconductor `ExpressionSet` objects. This module is complemented by the optional `bbcFetch` package, which is used to download, clean, and format datasets from online and public repositories, such as GEO or ArrayExpress, into the ExpressionSet format. Both are employed in Appendix A.1.

2. The `bbcPreprocessing` package encompasses all functions related to data transformation and normalization, ensuring that samples can be compared to each other across studies and integrated in downstream analyses (refer to Appendix A.3).

3. The `bbcViz` package gathers all functions implementing quality control representations, including those evaluating the impact of normalisation functions through boxplots and kernel plots. It also includes multi-dimensional projections such as PCA, t-SNE and UMAP, to ensure that samples with similar biological functions group together in lower-dimensional spaces. We utilised it to generate the various plots in Appendix A.3.4.

4. The `bbcSupervised` package houses all functions dedicated to downstream analyses, notably functional enrichment and Differential Gene Expression Analysis (DGEA) (refer to Appendix A.4).

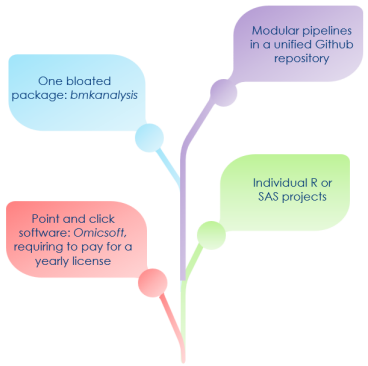
5. In addition to the core packages, we have implemented a wide ecosystem of wrapper and helper functions:

   - `bbcUtils`: This package contains short statistical or data management functions that are consistently used by most of the packages.
   - `bbcCorpo`: It includes corporate themes and templates for professional reports.
   - `bbcData`: This package stores toy example datasets as well as regularly updated gene databases.
   - `bbcVerse`: This is a wrapper utility package designed to load main packages without impacting the user and manage potential compatibility conflicts between modules or with the operating system.

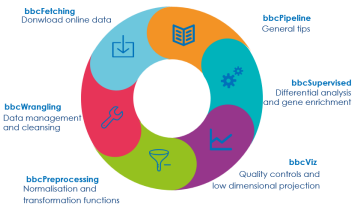
In conclusion, all these packages are accessible to any internal member of the Servier company via a centralised and secured Github repository through an individual token. We uphold the FAIR principles and enforce a comprehensive set of guidelines that must be adhered to before any package in development can be released as an industrial version on the *main* branch of each package.

Biological annotations Throughout this chapter, we employ the following abbreviations:

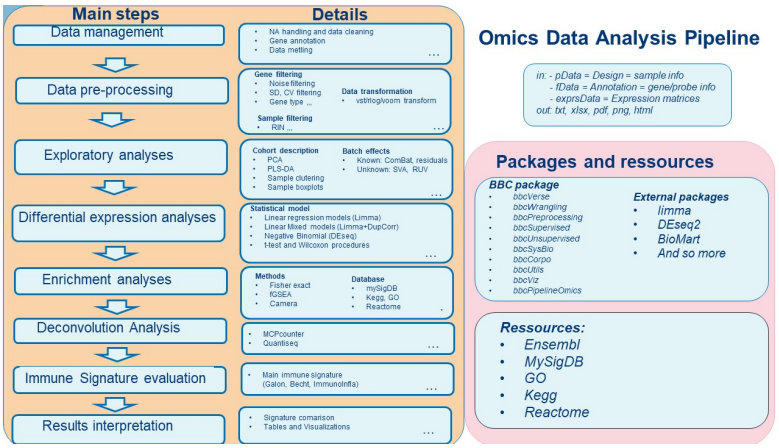
1. We use the following short names for cell populations:



(a) Brief historical outline conducting to the development of an unified industrial pipeline.We represent on this simplified schematic the main stages in the development of a standardised repertoire of functions dedicated to RNA-Seq analysis, starting from a corporate software to a modular Github folder, including the asynchronous and individual development of code snippets.



(b) Modularity and versatility: guidelines of our industrial RNA-Seq workflow.The design principle underlying the development of our RNA-Seq workflow, inspired from the best programming practices enforced by the Linux architecture team, consists of *uncoupling* pre-processing tasks, each dedicated to a specific functionality (quality control, visualisation, download online datasets, . . .).



(c) Comprehensive review of the operations implemented in our RNA-Seq workflow.The left panel details the main tasks integrated in our pipeline, as well as the tools and papers used to execute each of them. The top right panel details the main input and output required by our pipeline, namely the Bioconductor ExpressionSet object (we choose this object for its comprehensiveness and for the automatic process controls preventing any misused). The bottom right panel enumerates the core custom packages developed and integrated in our pipeline, along with the main available internal resources.

Figure A.1: Industrial R pipeline, taking as input the counts matrix returned by Appendix A.2.

- `t_cells` and `b_cells` to design subsets of lymphocytes, namely T cells characterised by a $CD3^+$ marker, and B cells characterised by a $CD19^+$ marker.
- `c_dc` and `p_dc`, for conventional (or classical) and Plasmacytoid dendritic cells, respectively
- `c_mo`, for conventional monocytes characterised by the $CD14^+$ marker
- `pmn` stands for Polymorphonuclear neutrophils, in opposition to Peripheral Blood Mononuclear Cell (PBMC)

See details about the features and biological functions of these immune cells in Section 2.1

2. We use the following short names for diseases:
   - `hc`, for healthy controls
   - `sle`, short for Systemic lupus erythematosus

A.1 Data import and cleaning

In this section, we aim to introduce the most widely used online repertoires housing transcriptomic data as well as describing the structure underlying the organisation of the datasets stored within these repositories. Additionally, we will delve into the utility functions employed to compress and clean the retrieved datasets, retaining only the most pertinent biological and technical factors and standardising gene, tissue, and cell population notations for enhanced consistency across projects.

A.1.1 Import relevant files

The majority of high-throughput transcriptomic datasets that are publicly accessible are deposited in the following two public repositories: *Gene Expression Omnibus (GEO)*, which is hosted by the NCBI organization, and *ArrayExpress*, hosted by the EMBL-EBI organization.

GEOs objects are stratified with the following hierarchical structure:

1. A *Platform* object details the general sequencing protocol and lists the probes or gene annotations. Besides, it gathers all the experiences that have been performed under this specific sequencing method. Platform ID follows the following notation: GPL, followed with an accession number.
2. A *Series* object identifies a set of Samples associated with the same biological experiment and additionally summarises phenotype features and the global design. It is identified with the GSExxx flag.
3. A *Sample* record, identified as GSMxxx, describes the conditions under which an individual Sample was handled, the manipulations it undergoes like the platform used, and the abundance measurement of each annotated transcript.

The second biggest source of public datasets is *ArrayExpress*, with differs from GEO with additional restrictions on the format of the datasets submitted, imposing to make them MIAME-compliant. Currently, 76,635 studies are represented in the ArrayExpress compendium. The general format is for each repository, a zipped MAGE-TAB document which splits itself into an *Investigation Description Format (IDF)* file, equivalent to the MIAME experimental data of the ExpressionSet object and describing top-level protocol experiences: `experimentData`, *Array*

Design Format (ADF), equivalent to the feature data: `Biobase::fData`, with the position (for microarray only) and the annotation of the measured transcripts, *Sample and Data Relationship Format (SDRF)*, equivalent to the phenotype data: `Biobase::pData` and eventually the raw and processed data files, that store transcriptomic expression.

While the packages `GEOquery`, [Dav22], [DM07], and `ArrayExpress`, [KES22], [Kau+09], have been designed to automatically query and fetch online databases, they are rather restricted to a specific format, unfortunately rarely met in practice. Depending on the level of precision required, along with general feature description files, we let the user to choose between raw or pre-processed data (generally, a tab-delimited file enumerating for each probe or identified transcript its total expression). In practice, one of the major bottlenecks is the absence of pre-processed expression data in the majority of RNASeq experiments, and the lack of standardisation of raw datasets. We have therefore attempted to partially resolve these limitations by respectively developing proprietary functions `bbcFetching::import_normalised_data` and `bbcFetching:::import_raw_files` for normalised and raw datasets, both functions attempting at first to parse local files, then fetch them online, and finally homogenises the output into an ExpressionSet object:

In addition, we display respectively the architecture of ArrayExpress and GEO databases[1], in Figure A.2.

A.1.2 Data wrangling with ExpressionSet

An ExpressionSet object is composed of an expression matrix, a gene annotation dataset and a phenotype dataset storing patient information, associated with general metadata stored in a MIAME object. In this section, we enumerate all the data formatting and quality check-ups performed to clean and leverage essential information from ExpressionSet:

1. Data format: We require that transcriptomic expression is stored in a matrix, and that the annotation sets are `data.frame` objects.

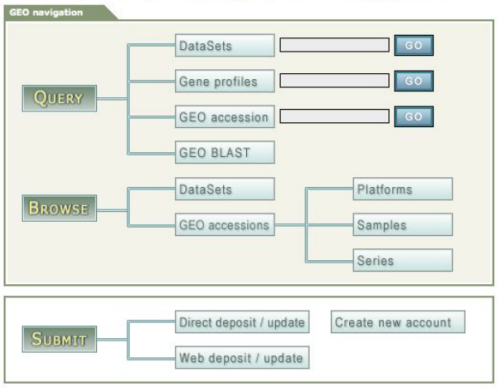
2. All elements composing the ExpressionSet must be documented with colnames and rownames, such that the colnames of the ExpressionSet match the rownames of the `pData` object (correspond to unique identification of samples) and its rownames match the rownames of the `fData` object (unique identification of transcripts). We recall the general structure of ExpressionSet objects as well as the operations required to manipulate them in Figure A.2.

3. Variable types: carefully prepare samples and features annotation data by formatting numerical variables in numeric format and converting textual variables as factors. When comparing several ExpressionSet objects, it is relevant to keep track of factor assignment, and homogenises them across batches. We marked unequivocally any misisng information using the "reserved" R variable `NA`, see Appendix A.1.2 for details.

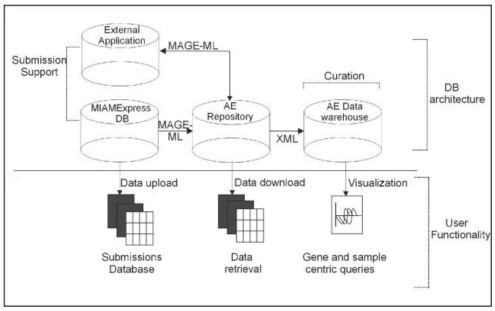
4. Gene annotation: this step ensures that gene ID used can be uniquely mapped to their corresponding most updated HGNC symbol. We can additionally filter out genes that are not involved in any of the biological functions of interest, see details in Appendix A.1.2.

5. (Optional) There is currently a strong lack of standard nomenclature to refer to cell types. We thus provide in Appendix A.1.2 helper functions to homogenise them across samples. Similarly, we may as well use ontologies to uniquely identify diseases or tissues.

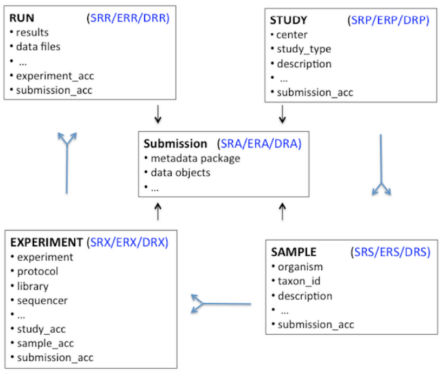
[1]This last repository is specifically dedicated to store high throughput sequencing experiments, and reveals its full potential as an unlimited and parallel data storage, at the raw read alignment level, along with NCBI GEO and EBI ArrayExpress databases. Currently, the R package SRAdb, [ZD22], [Zhu+13], is recommended for automatically querying, downloading and extracting alignment information



(a) Relation database and user API interface, from Wiki GEO

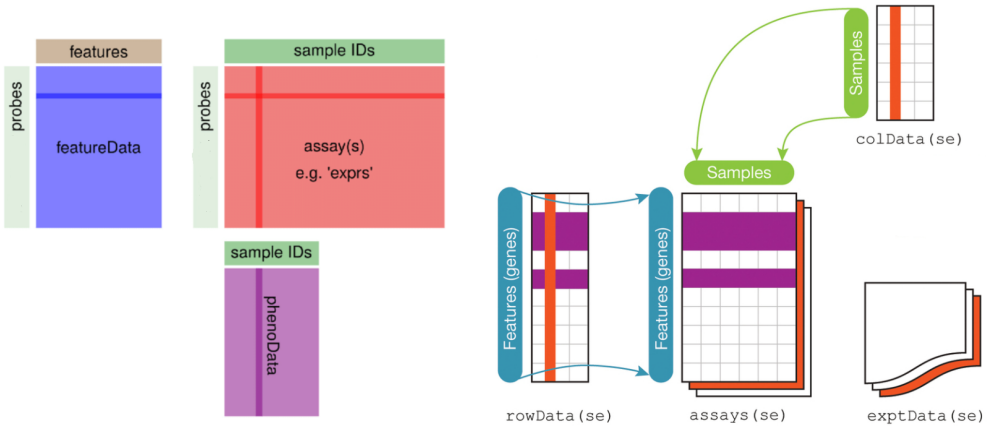


(b) The ArrayExpress architecture, with user functionality detailed at the bottom on its vignette.



(c) The graphical representation describing the entity relationships between the tables in SRAdb vignette

Figure A.2: ERD (Entity-Relationship Diagram), of GEO, ArrayExpress and SRA databases



(a) Structure of a Bioconductor's ExpressionSet object [Hub+15]–[KR18].

(b) The SummarizedExperiment object [Mor+22] is a generalisation of ExpressionSet. It is traditionally used to store the results of several experiences (provided at least some Gene IDs match). Besides, this format is required in some normalisation and differential analysis functions, notably for the *vst* normalisation applied by DeSeq2 Bioconductor package [LAH22].

Figure A.3: General structure of the objects most commonly used in R to store transcriptomic data

Sample annotation

Data cleansing is partly automated with `bbcWrangling::clean_ExpressionSet_names`:

1. First, simplify the identification colnames of objects `Biobase::pData` and `Biobase::fData`, mostly by complying them with the ruling names' syntax of R.

2. All non-ascii characters can be trimmed, however, perform this operation with care, since some of the characters of gene names may not comply with ASCII character encoding format.

3. The most useful function, removes any constant colname of object `Biobase::pData()`, and stores it as a `MIAME` object in slot `Biobase::experimentData`. We mean by *constant* any column that stores information common to all the samples, or patients, considered for the experiment, which can be typically the sequencing technology, the general experimental name or setting, or general information about laboratory and biologists contacts.

In addition to simplifying the manual curation of phenotype data, these management operations improve the identification of relevant ((e.g., phenotype condition) and confusing factors (e.g., batch replicates). Additionally, they optimise storage space utilization required to keep in memory the ExpressionSet object. For instance, the size taken by the phenotype dataset in memory hardware shifts from 322.8 Kb and 64, to 228 Kb and 26 on the cleaned ExpressionSet.

However, while part of the data curation is automated, there is usually an additional need of manual curation, and we displayed code snippets to perform data wrangling operations carried on purified cell expression sets:

- Rename columns, such that shared phenotype information, are labelled consistently across several Expression Sets, for instance the columns returning the cell assignment or the patient id
- Convert to factor any categorical variable which may contribute to biological or technical variability, since they will be required for batch correction and downstream analysis. Perform similarly for continuous values. For the last two experiences described, we identified three shared categorical variables of interest: patient assignment, cell type assignment and disease phenotype, and one continuous variable, age of the patient.
- Regarding categorical variables, you may attempt to merge closely related cell factors to gain statistical power and promote comparison across samples, a point further developed in Appendix A.1.2

In our example, we decided to concatenate closely related cell populations, by summing up their expressions within the same individual, based on our linear assumption of the reconstruction of the whole mixture:

```
# helper function to bind expression and phenotype feature
GSE137143_TPM_count_clean <- bbcWrangling::aggregate_by_pdata(
GSE137143_TPM_count_clean,
 col=c("sample_id"), sum)
```

We display respectively in Table A.1 and Table A.2, the relevant variables for phenotype description, along with the potential confusing technical factors. All these variables may contribute to the global transcriptomic variability across samples:

Table A.1: Visualise some phenotype characteristics of the patients from cohort GEO 149050

sample_id	cell_type	disease	age
GSM4489145	t_cells	hc	33.0
GSM4489146	t_cells	hc	41.0
GSM4489147	t_cells	hc	32.0
GSM4489148	**t_cells**	**hc**	**52.0**
GSM4489149	t_cells	hc	37.0
GSM4489150	t_cells	hc	24.0
GSM4489151	**t_cells**	**hc**	**58.0**
GSM4489152	t_cells	hc	50.0
GSM4489153	t_cells	hc	35.0
GSM4489154	t_cells	hc	36.0

(10 first lines / 288 lines)

Table A.2: Visualise some phenotype characteristics of the patients from cohort GEO 137143

sample_id	cell_type	disease	age
x13311_c_mo	**c_mo**	**hc**	
x13311_t_cells	**t_cells**	**hc**	
x13311_t_cells	**t_cells**	**hc**	
x22012_c_mo	**c_mo**	**hc**	
x43213_c_mo	c_mo	ms	34.0
x43213_t_cells	t_cells	ms	34.0
x43213_t_cells	t_cells	ms	34.0
x46913_c_mo	c_mo	ms	72.0

(8 first lines / 427 lines)

Gene annotation

One possibility to automatically update gene symbols to their respective modern gene nomenclature, typically the HUGO Gene Nomenclature Committee (HGNC) nomenclature, is to benefit from regularly cleaned online databases, which are available with R package `AnnotationDbi`.

However, we implemented our own set of annotation functions, that both extend significantly the bare features provided in R packages, while easing the interface with the specific R object `Biobase::ExpressionSet()`. The core function, `bbcPreprocessing::from_probe_to_gene`, is available externally for the regular user, and enables to automatically update genes to their newest format, performing the following steps:

1. (Optional) Unfortunately, most of the objects downloaded on GEO or ArrayExpress are not provided with recent, update gene nomenclature. In the worst case, the feature dataset is only composed of the `row.names` argument of the expression matrix, in that case, you may attempt to infer the type of nomenclature used. However, unfortunately, it may happen that the format used to name genes is exotic, or that expression is provided at a lower level than of the gene (typically, at the transcript isoform level for RNASeq, or at the probe level for microarray datasets). In that case, function `bbcPreprocessing::add_ExpressionSet_annotation()` retrieved automatically, from slot Biobase::annotation(), the gene feature annotation. For instance, GPLs (GEO Platforms) are often provided for microarray expression, or even made available as R packages of datasets, that can be queried using standard `AnnotationDbi` requests.

2. Function `bbcPreprocessing::get_genes_info` is called internally to update gene annotation. To uniformise the output of our analysis, we decided to force conversion to HGNC. However, any type of gene format among HGNC, Ensembl and ENTREZID can be provided as input, with its category manually filled in through argument `input_type`. The process joint is performed depends mostly on the nature of the input type:

- `ENTREZID` symbols can be directly matched (mapping between HGNC symbols and ENTREZID is 1-1), while the mapping between HGNC and ENSEMBL is of type one-to-many, requiring additional nesting operations. Indeed, an Ensembl stable ID consists of four compulsory parts: ENS(species)(object type)(identifier) and an optinal suffix indexing the (version), for example, ENSG00000141510.11 and the version index is likely to confuse most of automated naming tools.
- Mapping is more challenging for type `HGNCSYMBOL`. First of all, HGNC symbols are not all R syntaxic, and we commonly observed unwanted transformation of th original name (typically, when genes are used as `colnames` for datasets). Second, numerous *aliases* (old reference names, resulting generally from poor gene alignment) have been used as alternative names, some matching more than one current gene. For instance, the `OR4H6P` pseudogene, an old remnant of the olfactory superfamily, is known under 32 distinct names. In addition, we observe an *invariant* transformation (unique identification) under upper and punctuation trimming (replaced by a dash) transformation for column `HGNCSYMBOL`, and punctuation trimming alone for column `ALIAS`. Thus, from the original set of input genes, we first remove genes from the genes that match `HGNCSYMBOL` column under capitalisation and punctuation transformation, then remove genes that match `ALIAS` column, and displays those genes which have not been found in any of the `NCBI_gene` database.

3. Additional gene annotations, for 61538 unique HGNC symbols, are available in our regularly updated internal database `bbcData::NCBI_gene()`, among the following 12 features:

   - `HGNCSYMBOL` and its counterpart `ALIAS`, `ENTREZID` and `ENSEMBL` store gene updated correspondences, among the three most common gene nomenclatures. Of note, `HGNCSYMBOL` is the primary key of our table, uniquely identiyfing each row of our database, and there's a 1-1 match with `ENTREZID` symbol. General format for `ENTREZID` is the use of numbers, while the Ensembl nomenclature requires to precede the gene name with *ENSG*, followed by a series of number. They have been extensively used to identify genes in more than 70 species.
   - `GENENAME` and `GENEBIOTYPE` store respectively the specific and the general biological function of each transcript. 11 biological functions are available for the `GENEBIOTYPE` class, including `protein_coding`, tRNA(transfer RNA), rRNA(ribosomal RNA) that play an active role in the gene transcription and traduction phases and finally a wide ecosystem of microRNA subtypes, such as scRNA (small conditional RNA), snoRNA(small nucleolar RNA) and snRNA(small nuclear RNA). The role of microRNA in the regulation of gene expression has notably garnered extensive attention over the last few years [MVS14]-[CCZ16].
   - `MAP` locates the general position in the genome, detailling notably the chromosome name and arm. `GENESEQSTART` and `GENESEQEND` locate precisely the nucletoide position of the gene.
   - Finally, `TRANSCRIPTLENGTH` and `TRANSCRIPTNUMBER` return respectively the averaged size and the total known number of transcripts associated to the gene[2]. We detail in Appendix A.6 the protocol to fetch, aggregate and map distinct gene IDs conventions, and then how to populate them with additional general biological features.

[2]Remember from general biological introduction, Section 1.1.1 Post-transcriptional regulation, that an unique gene can give rise to several distinct transcripts, a phenomena known as *alternative splicing*

4. When mapped back to the ExpressionSet object, we performed the following series of data wrangling operations:

    - Any of the original genes that could not have been mapped to an updated HGNC-SYMBOL, as well as genes with a different biological function than the one listed in user-provided argument `gene_function` (by default, we discard any gene that does not code for a functional protein) are removed.
    - Any old gene that matches more than one recent HGNC symbol are also trimmed (typically, some aliases have been used to name recent HGNC symbols with completely distinct biological functions).
    - On the contrary, the expression of a whole set of genes that match the same recent HGNC symbol is aggregated, under the following default protocol: it is summed with RNASeq (since bulk sequencing technologies directly return the total number of counts observed in the sample), while it is averaged with microarray technology. For instance, in the Affymetrix technology, the expression of a RNA transcript is given by the number of complementary probe sequences, of 25 bases long each, it matches. However, this small size increases the risk of mismatch and it is thus common to design several probe sequences that target different regions of the genes of interest.
    - As a practical example, Table Table A.3, illustrates practical gene feature tidying operations of the function `bbcPreprocessing::from_probe_to_gene`, applied on the raw ExpressionSet of the GSE149050 study. Gene *5_8S_rRNA* (ribosomal RNA), alternatively known as ENSG00000276871, is removed since it is not associated to any known HGNC symbol. Old aliases that are mapped to more than one updated gene, such as *CH507-154B10.1* (not assigned to any known biological function and associated to more than three distinct locations in the genome), or *DEC1* have thus been removed. In the former two cases, the paired HGNC symbols even correspond to genes with distinct biological functions, with only one coding actually for a protein. On the contrary, old aliases that can be unequivocally associated to one known HGNC symbol are conserved, such as *AAED1*, mapped to gene *PRXL2C*, both involved in the glycolysis pathway.

Table A.3: Feature table illustrating some difficulties handled automatically by corporate function, to handle one-to-many or one-to-null gene mapping. We additionally highlight genes associated with more than one ENSEMBL ID.

original_input	HGNCSYMBOL	ENTREZID	GENEBIOTYPE	GENELENGTH
5_8S_rRNA				
7SK	**RN7SK**	**125050**	**snRNA**	**328.0**
AAED1	PRXL2C	195827	protein_coding	15,741.0
AIM1	CRYBG1	202	protein_coding	211,301.0
AIM1	AURKB	9212	protein_coding	5,868.0
AIM1	**SLC45A2**	**51151**	**protein_coding**	**24,980.0**
CH507-154B10.1	LOC102724701	102724701		
CH507-154B10.1	LOC105379499	105379499		
CH507-154B10.1	LOC107987290	107987290		
DEC1	BHLHE40	8553	protein_coding	5,887.0
DEC1	DELEC1	50514	lncRNA	551,677.0
GGTA1P	GGTA1	2681	protein_coding	55,037.0
GGTA1P	GGTA2P	121328	processed_pseudogene	1,125.0

(13 first lines / 13 lines)

```
GSE149050_raw_count_clean <-
  bbcPreprocessing::from_probe_to_gene(GSE149050_raw_count_clean,
  microarray_type = FALSE
)
# add explicitly a column with some annotation
Biobase::fData(GSE149050_raw_count_clean) <-
  Biobase::fData(GSE149050_raw_count_clean) %>%
```

```
  dplyr::mutate(Genes = row.names(.))
```

Cell type annotation

It turns out that the nomenclature for naming cell populations is inconsistent between the pathologists and haematologists, the former focusing on the functionality and the latter on surface markers. In addition, tabular structures, complying with the DBMS, for database management system, format (an unit in the table is necessarily atomic) are not tailored to store efficiently complex lineage relationships between distinct entities.

To that end, noSQL approaches, notably graph approaches, are way better to store interconnected and heterogeneous relations across variables. In addition, a whole array of graph-search algorithms can be leveraged to perform valuable mining operations, such as retrieving automatically all the cell lines descending from an overarching cell population, or, applied to cellular deconvolution (Section 5.1), ease the computation of ancestral cell lines, summing the individual ratios of descending cell populations.

We exemplified this concept with the Kassandra project [Zai+22]. This initiative gathered a compendium of more than 212 datasets, with 17 distinct cell annotations. We displayed in Table A.4 canonical purified datasets of cell populations collected by the consortium. However, even this highly pre-processed database collection, the cell types concatenated together are not consistent in terms of cell lineage, hence the interest of automate the annotation of cell populations.

Table A.4: A quick summary of the datasets collected by the Kassandra algorithm, for the top 8 databases containing the most samples.

Array accession	Samples	Num cell types	Cell types
GSE133822	219	3	CD4_T_cells, CD8_T_cells, Monocytes
GSE103844	178	1	CD4_T_cells
GSE104744	129	3	Monocytes, CD8_T_cells, CD4_T_cells
GSE129829	89	1	CD4_T_cells
GSE60424	80	6	Neutrophils, CD4_T_cells, Monocytes, CD8_T_cells, Non_plasma_B_cells, NK_cells
GSE124073	73	1	Monocytes
GSE114065	71	1	CD4_T_cells
GSE117970	60	1	Monocytes

(8 first lines / 212 lines)

To simplify and standardise in an automated manner the mapping from the original Kassandra cell type annotation to updated cell ontologies, we benefit from the features implemented by the suite of R packages `ontologyX` [GRT17a], specially tailored for working with biological ontological datasets, and composed of three main compartments:

- *ontologyIndex* [Gre22], [GRT17b] can read in arbitrary ontologies and converts them as R objects. Besides, it provides a set of highly-relevant wrapper functions to prune complex graph ontologies, or perfom complex graph queries.
- *ontologyPlot* [Gre21a] enables visualisation of ontological terms and ontological annotation with a wide variety of graphical options.
- *ontologySimilarity* [Gre21b] facilitates fast semantic comparison across multiple ontological objects, including assessment of statistical significance.
- *ontoProc* [Car22a] is an extension of the three previously cited packages, specially designed to work with biological ontologies. It provides base commands to annotate cell or tissue ontologies, as well as wrapper functions of `ontologyX` packages, for example `ontoProc::onto_plot2` to visualise quickly cell ontologies.

We illustrate the increased reproducibility power provided by these tools, providing useful commands to identify and map cell terms, especially when they slightly differ from the ones provided by default in the cell ontology, as well as commands to handle more easily the graph, for instance enforcing the directed graph to be a tree and performing standard network queries, such as retrieving the set of siblings, ancestors, descendants and first-order relatives for a given node:

```
### load cell ontology
co <- ontoProc::getCellOnto()

### get correspondence to a specific term, when there is no direct match
monocytes_matches <- ontoProc::liberalMap(c("Monocytes"), co, useAgrep = TRUE,
                            ignore.case=T) #36 possible matches are returned

### from unique ontology ID key, retrieve the scientific, readable term
ontologyIndex::get_term_property(ontology=co, property="name",
                                 term="CL:0000576", as_names=F)
#> CL:0000576
#> "monocyte"

### get all children (direct-link), from monocyte lineage (5 children returned)
ontoProc::children_TAG(Tagstring = "CL:0000576", co)@ontoTags
#>  CL:00005763  CL:00005764  CL:00005761  CL:00005765  CL:00005762
#> "CL:0001022" "CL:0001054" "CL:0000860" "CL:0002393" "CL:0000875"

### rebuild Kassandra cell ontology

# get all ancestors of Kassandra

ancestors_kassandra <- co$id[ontologyIndex::get_ancestors

(co, updated_kassandra_annotations$ontoid)]

# get the root (leukocyte = CL:0000738 is the closest common ancestor)
ancestral_root <- co$id[ontologyIndex::get_ancestors(co, c("CL:0000738"))]

# prune the cell line, by removing spurious ancestral terms
# minimal_set enables to remove spurious or redundant terms
cell_ids <- setdiff(ancestors_kassandra, ancestral_root) %>%
                      ontologyIndex::minimal_set(co, .)
# ontoProc::onto_plot2(co, cell_ids, cex = 0.8) plot the associated lineage
```

We represent in Table A.5 the final mapping between the original Kassandra cell notations and approved cell ontology terms:

Table A.5: This table displays the mapping between the original cell type notations of the 17 cell populations identified within the Kassandra repertoire, Cibersort terms, and their corresponding usual notation in the haematologist and pathologist community. In addition, the primary ID key, identifying each cell term with an unique and computer-readable notation, is reported in the last column `ontoid`, and consists of upper-case prefix `CL` (for cell lineage) followed by an unique Arabic numeral ID. We highlight the rows storing cell populations that we were not able to unequivocally map to an unique ontology term.

old_annotation	cell_type	ontoid	cibersort_mapping
B_cells	**B cell**	**CL:0000236**	**B cell**
T_cells	**T cell**	**CL:0000084**	**T cell**
CD4_T_cells	CD4-positive, alpha-beta T cell	CL:0000624	T cell
CD8_T_cells	CD8-positive, alpha-beta T cell	CL:0000625	T cell
Classical_monocytes	classical monocyte	CL:0000860	monocyte
Eosinophils	eosinophil	CL:0000771	eosinophil
Basophils	basophil	CL:0000767	basophil
memory_B_cells	memory B cell	CL:0000787	B cell
Monocytes	monocyte	CL:0000576	monocyte
Naive_B_cells	naive B cell	CL:0000788	B cell
NK_cells	natural killer cell	CL:0000623	natural killer cell
Neutrophils	neutrophil	CL:0000775	neutrophil
Non_classical_monocytes	non-classical monocyte	CL:0000875	monocyte
Non_plasma_B_cells	**mature B cell**	**CL:0000785**	**B cell**
Plasma_B_cells	plasma cell	CL:0000786	B cell
PDC	plasmacytoid dendritic cell	CL:0000784	Dendritic cell
Granulocytes	**granulocyte**	**CL:0000094**	**neutrophil**

(17 first lines / 17 lines)

In Figure A.4, we present the finalised cell lineage tree, highlighting the overlapping cell populations from the Kassandra signature:

While we were able to find a 1-1 mapping for most of the cell type annotated by Kassandra, we were unable to find any direct correspondence to design `Non_plasma_B_cells`. Instead, we choose arbitrary to refer to them as `mature B cells`, but at least two other possible choices, illustrated in Figure Figure A.5 are likely valid:

- The most straightforward choice would have been to select a sibling term, at the same lineage hierarchical level, namely select `B-cell, C19+` term (most of B cells are unequivocally annotated with the so-called CD19 marker). However, we observe that among the first-order descendants of B-cell, $CD19^{+}$, only the mature B cell population is likely to be found in blood tissues, since the other described B cell stem lines are precursors, more likely to

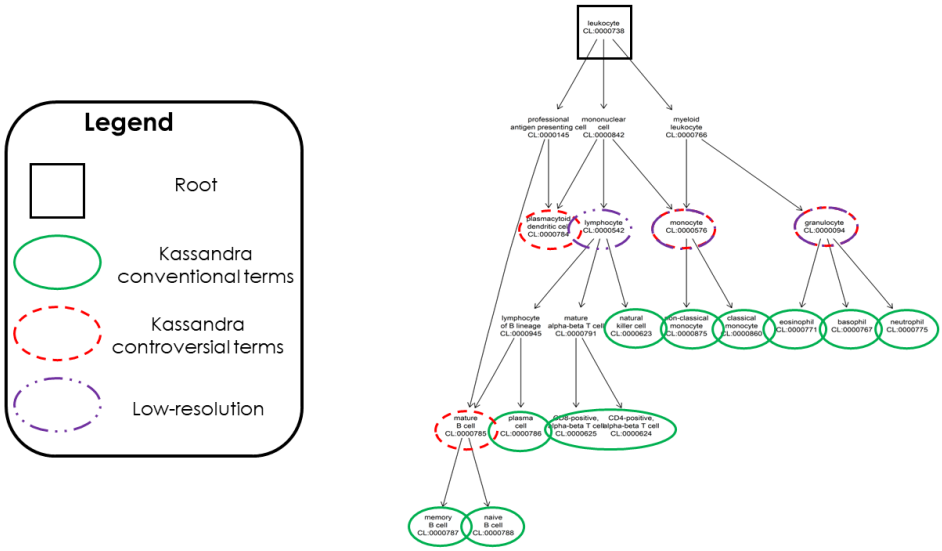


Figure A.4: Cell lineage tree of the cell types included in the Kassandra fingerprint

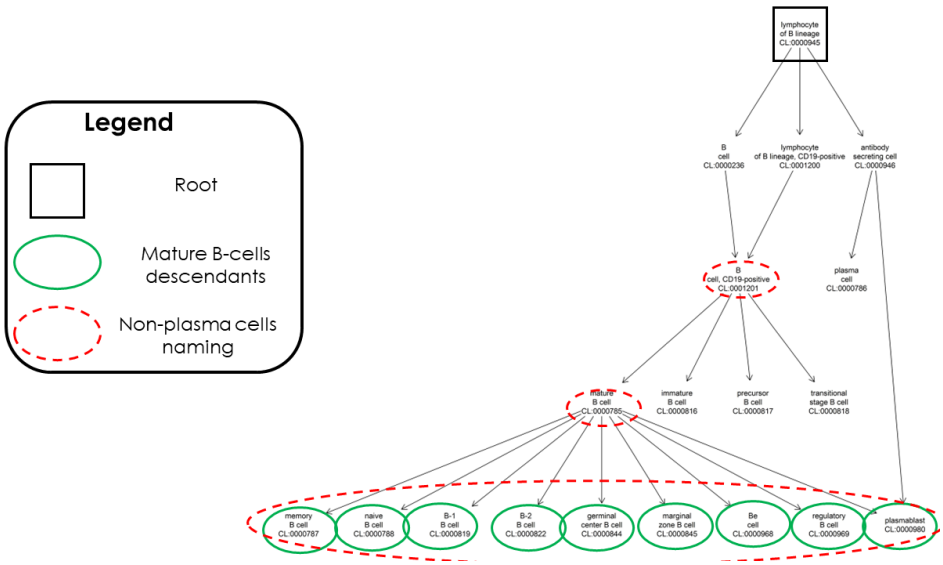


Figure A.5: Descendants up to third lineage of B cell clones, including all terms of tke Kassandra annotaiton database relating to B cells

prevail in bone marrow.

- Alternatively, it would have been possible to group all B cells that are not plasma cells under a single, overarching term by merging all lineages of B cells exhibiting distinct terminal biological functions, such as *naive B-cells*, *memory B-cells* or *regulatory B-cells* (see glossary term B-cell and Section 2.1.2).

A.2 Pre-processing of raw counts, using a Nextflow pipeline

Pre-processing, from raw reads to quantifying RNA-Seq counts through mapping to a reference genome, is particularly computationally intensive and usually requires powerful and high-scalable dedicated algorithms, built directly on assembly on in C++. For these reasons, the R environment, highly versatile and easy to handle for beginner programmers, is not particularly tailored to handle such operations (see Appendix A.2).

On the contrary, the programming framework Nextflow is designed for building and executing *data-intensive*, *scalable*, and reproducible computational pipelines, combining in an unified repertoire computational tools written in different programming scripts. To that end, it provides a domain-specific language (DSL) that simplifies the creation and management of complex workflow.

Accordingly, Nextflow offers several key benefits for implementing a pipeline to analyse raw reads, ensuring reproducibility (the entire pipeline's configuration, including software dependencies and versions, are stored in a computer-readable format), scalability (native parallel and distributed implementation), versatility (execution on a variety of computing infrastructures, including coupling local cluster with cloud computing platforms) and flexibility (same modular approach discussed in Appendix A underlies the Nextflow architecture design), enabling bioinformaticians to leverage their preferred assembly and mapping tools. This flexibility is highly valuable in RNA-seq analysis, where different tools and algorithms may be required for preprocessing, alignment, quantification, and downstream analysis, with regard to the nature of the platform and the quality of the library (single-end or paired-end, percentage of adapter contamination, ...). Finally, Nextflow provides the end-users with a variety of diagnostic tools to evaluate errors or warnings occurring throughout the pipeline process and controlling the individual computational cost of each task (see Appendix A.2).

To that end, Servier's bioinformatics team implemented its own custom Nextflow framework, `RNAExp`, using its own favoured alignment and mapping tools. I will not delve into comprehensive details on the hyper-parameters and methods implementation choices, since I was not personally involved in the development of this computational tool and since most of the transcriptomic datasets I studied were already mapped to a human reference genome and formatted as raw counts. However, I outline in Appendix A.2, the key tasks implemented in this Nextflow workflow.

Nextflow exhibits however some limitations: it has a strong learning curve, especially for users who are new to workflow management systems, the DSL syntax of Nextflow can indeed be particularly challenging for beginners, as it may appear obscure and unfamiliar. Developing complex pipelines dealing with intricate version dependencies or advanced workflow patterns (notably those requiring human inspection or cyclic loops) in Nextflow can be challenging. And the versatility and flexibility of the tool entails hard time in debugging and troubleshooting, particularly when dealing with issues across multiple stages or parallel executions. Accordingly, error handling strategies are essential to facilitate the resolution of a wide variety of exceptions

returned by a whole ecosystem of tools pre-processing raw reads. Ultimately, since Nextflow is relatively recent in the history of computational development, it may still have fewer resources and community support compared to more established programming framework, leading to fewer readily available solutions to specific problems.

Interestingly, [Ewe+20] discusses about a similar implementation of our Nextflow pipeline, namely the *nf-core/rnaseq* project, which provides an unique suite of testing and automation tools with comprehensive documentation for both developers and end users for assembling and mapping raw RNA-Seq reads. Notably, this framework project adheres to all nf-core guidelines, homogenising the coding style and increasing the reproducibility and interoperability across multiple operating systems and software versions.

A.3 Quantification of Gene Expression

In this section, we will provide a concise overview of various multi-faceted metrics used to globally assess the quality of a given RNA-Seq sample. Subsequently, we will delve into semi-supervised techniques designed to eliminate background noise and identify genes that do not exhibit consistent expression patterns across samples. Finally, we will conclude this section by presenting a comprehensive array of methods dedicated to correct for technical artefacts and ensuring the comparability of RNA-Seq distributions across different samples and studies.

A.3.1 Sample filtering

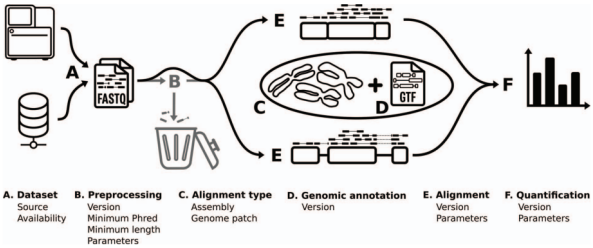
It is common to find individual RIN annotations in the phenotype dataset. The RNA Integrity Number (RIN) is a quality metric used in RNA-Seq experiments to assess the integrity of RNA samples, providing insights on the level of degradation or damage that RNA molecules have undergone, and thus can be used as a proxy of the accuracy and reliability of gene expression measurements.

The RIN quality score is computed using an automated electrophoresis system, such as the Agilent Bioanalyzer or TapeStation, by aggregating several measures: the ratio of 28S ribosomal RNA (rRNA) to 18S rRNA, the presence of degradation products, and the overall shape of the RNA profile. Samples with high RIN scores are indicative of intact RNA, while lower scores suggest RNA degradation. Accordingly, the commonly used RIN threshold for sample inclusion in RNA-Seq studies is typically set above seven, or eight a for more stringent quality control.

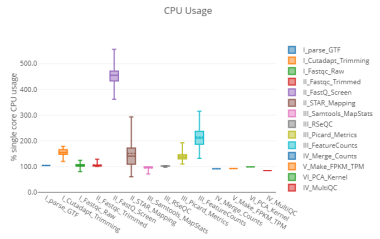
However, it is recommended to combine this metric in a broader quality control process, for example by projecting samples in low dimension ((see Appendix A.3.4 or Appendix A.3.4) or compute Heat map distance matrices (see Appendix A.3.4) to confirm their status of outliers. Indeed, there is no universal consensus on the RIN threshold to apply, and RIN calculation is not totally insensitive to the choice of the Electropherogram software. Ultimately, this metric does not provide information about potential infiltrates that may affect RNA-Seq quality and is not adjusted to evaluate small RNA-Seq data, usually coming along in lower quality.

A.3.2 Gene filtering

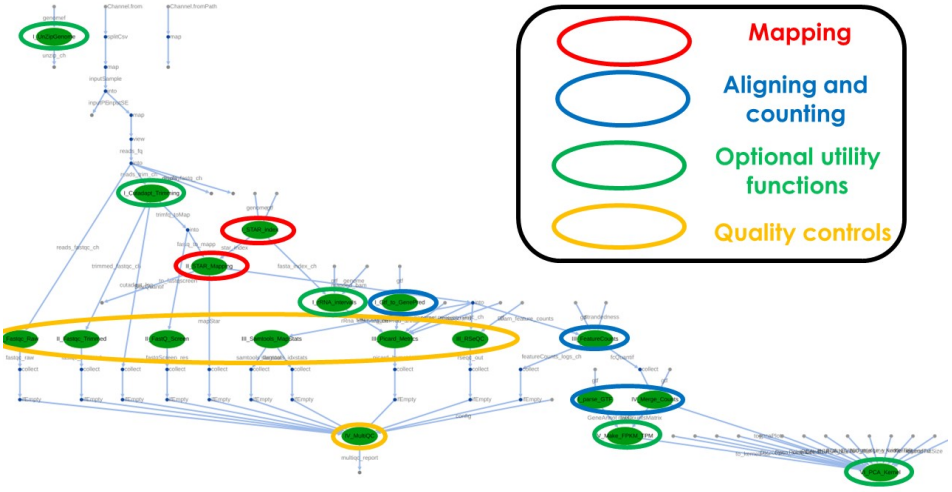
Discarding background noise in transcriptomic expression data is critical for the statistical power of downstream analyses. Indeed, removing genes with low expression levels that not provide meaningful biological insights, increase the overall Signal to Noise Ratio (SNR) of



(a) Didactic illustration of the RNA-seq bioinformatics pipeline..The RNA-Seq pipeline is usually split into generation of the raw reads RNA-Seq library (through amplification followed by sequencing, refer to Section 1.2.2), followed by quality controls and trimming of adapater sequences (see Section 1.2.2). Then, the subset of high-quality reads is mapped on the reference genome (see Section 1.2.2), and the total number of reads matching the features of interest (such as protein-coding genes) are quantified and returned as a matrix of raw counts (see Section 1.2.2). Illustration reproduced from [Sim+21, Fig. 1]. Comprehensive schematics are also reported in Figure 1.4.



(b) Total Computational Cost per Operation Implemented in the Nextflow Pipeline.On the x-axis, you can see the whole set of pre-processing tasks implemented in the pipeline. The y-axis represents the percentage of CPU utilization. It's worth noting that the percentage may exceed 100% since Nextflow allows for the implementation of highly-parallelised operations.



(c) Tree graph outlining the sequence of operations within our industrial Nextflow RNA-seq bioinformatics pipeline.This pipeline encompasses all pre-preocessing tasks, starting from quality control and the removal of erroneous bases in the raw read sequences, till quantification of transcript counts, including the alignment and mapping of reads to a human reference genome. It's worth noting that the preprocessing steps need not be executed sequentially, as the Nextflow distributed implementation allows for the concurrent execution of multiple preprocessing tasks.

Figure A.6: Industrial Nextflow framework for processing raw RNA-Seq reads

the transcriptomic dataset and automatically contribute to better identification of genes truly differentially expressed across cohorts.

Most of the methods used to perform this filtering assume that transcriptomic expression distributions exhibit a bimodal distribution, the first peak describing the background noise, and the second peak assumed to describe expressed genes and correspond to relevant biological signal. They traditionally return a threshold expression value, below which genes are considered unexpressed and filtered out, the challenge being to identify this threshold in a robust and consistent manner across samples and studies (alternatively, the Signal to Noise Ratio (SNR) or inter-variability score of a given gene across samples can be leveraged as a threshold, since genes with low variance are often considered uninformative).

The tools implemented for filtering Genes are further classified into manual-based filtering methods (simpler to implement, computationally efficient, but usually requires prior expert knowledge and usually does not consider the nature of the data distribution) and unsupervised, data-driven filtering methods, that retrieve automatically the parameters of the bimodal Distribution (requires more sophisticated statistical tools and substantial computational resources, however, based on a sound theoretical framework, they can effectively control the false positives/negatives ratio). These machine-learning methods subsequently use these parameters to consistently classify genes as "expressed" or "unexpressed".

Among the data-driven approaches, zero-inflated models effectively account for the commonly observed drop-out event, in which a significant proportion of genes have zero expression across all samples. Such methods are better tailored for scRNA-Seq analysis that usually exhibit a lower library depth compared to more traditional sequencing approaches, and increase the detection of lowly expressed genes with meaningful biological significance.

Up to now, we mostly employed existing, heuristic methods, or manual observation of the distribution, to infer the threshold setting apart background noise from truly expressed genes. Among them, we mostly capitalise on:

- We employed the zFPKM method first reported in [Har+13], to identify active genes from background noise. The authors of the study validate their conclusions on the ENCODE project and additionally show that lowly-expressed genes are usually associated with repressed chromatin.

  Briefly, the idea underlying the zFPKM normalisation is to mirror the half-Gaussian curve to the right half of the main peak of the bimodal distribution (assume to capture the true biological insights) to a full Gaussian distribution, of parameters (μ, σ). To that end, the distribution on the $\log_2$ FPKM expression values space is approximated by a kernel density estimation (kde) method, from which we derive the required parameters: mean expression μ as the kde maximum and the standard deviation, $\sigma = \frac{\mathbb{E}[\boldsymbol{X} \geq \mu] - \mu}{\sqrt{2\pi}}$, using the statistical method of moments that links the conditional expected mean of a sample following a Gaussian distribution to its mean and variance (see Gaussian distribution, section Moments).

  However, this paper does not discuss the specific limitations of the zFPKM normalisation method proposed, such as its performance in different experimental conditions, with distinct normalisation methods (recall from Appendix A.3.3 that FPKM or RPKM are not particularly advised for DGEA), or its sensitivity to outliers, limiting the generalizability of their conclusions to other datasets, experimental settings or pipelines. On a statistical point of view, the method is not really satisfactory, since a strong first peak noise may significantly bias the estimation of the variance and the mean of the second mode. Without

explicitly considering a mixture of two distributions, the method does not provide statistical evaluation of the probability that each gene is truly expressed in the mixture.

- To generalise the approach described before to any kind of normalisation, we developed custom filtering algorithms tailored to the most popular RNA-Seq normalization methods, before and after the $\log_2$ transformation. These methods are all variants of zFPKM, aiming to transform the original data to approximate a symmetric, bell-shaped Normal distribution. However, we departed from the conventional zFPKM framework by symmetrically adjusting the left half of the first peak (and not anymore the right half of the second peak). Subsequently, we extracted the maximum of each peak using the same kernel density estimation (kde) method.

  We then initiated an EM algorithm with the corresponding rough estimates, which was employed to simultaneously infer the parameters (mean and standard deviation) of both modes, assuming each follows ahead a Gaussian distribution. Interestingly, this approach bears some resemblance to the REBMIX algorithm ([NF11]). Consequently, we were able to precisely evaluate the uncertainty for each gene to be classified as noise or truly expressed.

  However, it's important to note that this method comes with several theoretical limitations. By altering the original distribution, we tend to overestimate the dispersion of the first peak, artificially increasing the overlap between the two clusters. Furthermore, we observed in a wide range of simulations that even after imposing this artificial symmetrisation of the distribution, the EM algorithm, as implemented in the 'mixtools' package, struggles to converge. Code snippets illustrating some of these functions, with regard to the nature of the transcriptomic dataset, are displayed hereafter:

```
# This function is applied on the distribution of raw counts
threshold_GSE149050 <- bbcPreprocessing::estimate_cutoff_lowcounts(
  GSE149050_raw_count_clean %>%
    Biobase::exprs())

# This function is applied with TPM counts
threshold_GSE137143 <- bbcPreprocessing::estimate_cutoff_lowcounts_norm(
  GSE137143_TPM_count_clean %>% Biobase::exprs())
```

After automatically determining a global threshold (typically set as the 0.95 quantile of the first peak of the bimodal distribution), biostatisticians must evaluate the minimal proportion of samples in which a gene surpasses this threshold, to be considered as expressed. In this framework, we employ a rigorous filtering criterion, retaining only those genes that exhibit expression levels exceeding the threshold in a minimum of 30% of the entire cohort.

```
NSample <- 0.3 * ncol(GSE149050_raw_count_clean) %>% round()
threshold_GSE149050 <- 7
GSE149050_raw_count_filtered <-
  bbcPreprocessing::filter_background(GSE149050_raw_count_clean,
                                      filter=threshold_GSE149050, NSample)
#> [1] "Start : 19121"
#> [1] "End : 12265"
#> [1] "Diff : 6856"
```

To conclude, filtering genes enhance the signal-to-noise ratio in downstream analyses and leads to faster and more efficient analyses by reducing the size of expression matrices. Accordingly, such methods increase the statistical power of Differential Gene Expression Analysis (DGEA), by removing uninformative genes and alleviating multiple test problem. Anecdotally, it simplifies multi-dimensional visualisation by enhancing biologically insightful expression patterns.

Nonetheless, the process of eliminating background noise must be executed with utmost caution as it has the potential to exclude genes that, while lowly expressed, might hold significant biological relevance, and the choice of the filtering method should align with the specific goals followed by downstream analyses and characteristics of the dataset.

Unfortunately, there is no universally agreed-upon method that outperforms all others in accurately and robustly estimating the background noise of transcriptomic datasets. The heuristic methods we have developed for this purpose lack strong theoretical statistical foundations, aren't versatile enough to handle all types of bimodal distributions, and do not fully capture the discrete count nature of RNA-Seq data nor the inherent positive constraints associated with transcriptomic expression.

To address the lack of flexibility, we are currently implementing non-parametric mixture-based models. To better reflect the discrete nature of RNA-Seq data and the positive constraints, we are considering probabilistic distributions, such as Negative Binomials or zero-inflated log-Normal distributions, that inherently accommodate these characteristics. For more details, please refer to Section 3.3.

A.3.3 Normalization and transformation

This step encompasses normalisation and transformation applied on RNA-Seq. Normalisation is applied to correct for variability in library size and sequencing depth, thus enabling to make samples comparable, while transformation functions are rather employed to guarantee that the distributions of RNA-Seq counts comply with the assumptions of downstream analyses, notably parametric methods (see Appendix A.4).

Information: Introduction to straightforward Normalisation methods

Conventional normalisation functions to correct for technical artefacts encompass:

- RPKM (Reads Per Kilobase Million) and FPKM (Fragments Per Kilobase Million) are both considered Library Size Normalization techniques. RPKM was designed for single-end RNA-seq, whereas FPKM was developed for paired-end RNA-seq, where each read is typically duplicated. In paired-end sequencing, it's possible for one read in a pair to fail to map to the reference genome. Consequently, FPKM accounts for the fact that two reads may map to a single fragment. Precisely, RPKM is computed as Equation (A.1), inserting the notations introduced in Definition C.1.8:

$$\text{RPKM}(y_{gi}) = \frac{\text{Number of Reads Mapped to the Gene} := \sum_{l=1}^{L_i} t_{li}\, \mathbb{1}_{t_{li}=g}}{\left(\frac{\text{Gene Length in Kilobases}}{1000}\right) \times \left(\frac{\text{Total Number of Mapped Reads}}{1{,}000{,}000}\right)} \tag{A.1}$$

  , with g indexing the gene, i indexing the sample, l the read and $L_i = \sum_{g=1}^{G} y_{gi}$ the library length of a given sample, namely the number of reads mapped to the reference genome. Both methods can correct for library depth and gene length, and should be used for in-sample comparison between genes of the expression levels rather than differential expression analysis.

- TPM, for Transcripts Per Million, is a Total Count Scaling method. TPM is calculated similarly to RPKM, both methods only diverge by the order of operations: in TPM, you normalise for gene length first, and then for sequencing depth. However, the effects of these operations are fundamental: with TPM, the total sum in each sample is the same (normalised to equal one million of reads), enabling straightforward comparisons of the proportion of reads mapped across samples. In contrast, with RPKM and FPKM, the sum of the normalised reads in each sample may be different, and this makes it harder to compare samples directly. Accordingly, TPM is often preferred when quantifying gene expression levels and in Differential Gene Expression Analysis (DGEA) analysis, especially for comparing the expression levels of long transcripts.

- Quantile Normalisation aligns the quantiles of expression distributions across samples, thus rendering distributions comparable between samples, and is agnostic to the sequencing technique, hence suitable for both microarray and RNA-Seq data. However, it does not account for differences in library size and should not be employed for differential expression analysis.

The choice of normalization method depends on the specific goals of the analysis and the characteristics of the RNA-Seq dataset. Differential Gene Expression Analysis (DGEA), such as like DESeq2 or limma framework, typically require more sophisticated normalisation techniques, while simple methods, like RPKM or TPM, may be suitable for general expression quantification. To better understand these concepts, we refer the interested reader to [Blo15].

Indeed, [Zha+21] compares the performance and statistical power of TPM, FPKM, and Normalized Counts approaches for the Analysis of RNA-seq Data from Patient-Derived Models

(precisely patient-derived xenograft, which better reproduce true biological conditions compared to pure in-vitro cell lines models). They notably found that transformed counts, such as VST or RLR in DESeq2 or TMM in EdgeR, described hereafter, showed the lowest coefficient of variation (CV) and highest intraclass correlation (ICC) values, otherwise the highest Signal to Noise Ratio (SNR) across biological conditions, and so likely the best classification predictability using clustering approaches. Overall, the simulations they carried on support the use of transformed counts for downstream analyses of RNA-seq data across samples, highlighting their accuracy and consistency in inter-sample comparisons.

Warning: Limitations of straightforward normalisation methods

It's worth noting that all the normalisation methods described previously do not stabilise the variance of the count data, whereas genes with very low read counts tend to exhibit, on average, a higher SNR. We detail hereafter some transformation methods that account for the intrinsic variability of gene expressions, and in key Heteroskedasticity the main causes for this phenomena.
In addition, none of the aforementioned methods account for compositional biases in RNA-Seq data, resulting from the nature of the sequence itself. The TEMT algorithm, natively implemented for deconvolution purposes, features an additional normalisation of the counts, based on the composition of the sequence ([LX13]).

Information: Introduction to Complex normalisation methods

We enumerate hereafter standard transformations, ordered by increasing computational and statistical complexity:

- The RLE (for Relative Log Expression) normalisation is also implemented in the DESeq2 package [And+13], and is computed as follows Equation (A.2):

$$\mathrm{RLE}(x_{g,i}) = \log_2\left(\frac{x_{g,i}}{\mathrm{med}(\boldsymbol{x}_{g.})}\right) \tag{A.2}$$

  where $\mathrm{med}(\boldsymbol{x}_{g.})$ is the median count of gene g across all N samples.

- DESeq2 Normalisation, also known as "median of ratios" [LHA14], is given by Equation (A.3):

$$\mathrm{DESeq2}(y_{g,i}) = \log_2\left(\frac{y_{g,i}}{s_i}\right) \tag{A.3}$$

  The size factor s_i, specific to each sample, is given by Equation (A.4)

$$s_i = \frac{\sum_{g=1}^{G} y_{g,i}}{\mathrm{med}(\sum_{g=1}^{G} y_{g,1}, \ldots, \sum_{g=1}^{G} y_{g,N})} \tag{A.4}$$

  and $\mathrm{Med}(\sum_{g=1}^{G} y_{g1}, \ldots, \sum_{g=1}^{G} y_{gN})$ the median of total counts across all samples, and is computed assuming a Negative Binomial (NB) distribution.

- The Variance Stabilising Transformation (VST) transformation, described in [AH10], is close in principle to Equation (A.3), replacing the median operator of the scaling factor s_i (Equation (A.4)) by the mean of counts for that gene across all samples.

  Both methods assume Negative Binomial (NB) distributions to compute the scaling factor for each gene. NBs simplify the modelling of the count distribution characteristics, such as the library size and RNA composition, in the modelling process. In addition, the VST transforms the raw counts data into a space where the variance is approximately independent of the mean, making it suitable for statistical modelling and visualisation.

- TMM, for Trimmed Mean of M-values , computes the trimmed mean of log-fold changes for each gene, as given in Equation (A.5).

$$\mathrm{TMM}\,(y_{gi}) = s_i \times \frac{\sum_{g=1}^{G} w_{gi} y_{gi}}{\sum_{g=1}^{G} w_{gi} l_g} \tag{A.5}$$

  where s_i is the same scaling factor reported in Equation (A.4), w_g is the weight assigned to each gene g, and l_g is the *effective gene length*. This method is notably used by the EdgeR differential analysis framework [RMS10].

To put it into a nutshell, TMM is better suited for between-sample comparisons, and VST for stabilising the variance, rendering these methods tailored for DGEA, by enforcing the

homoscedascity constraint. On the contrary, RLM is primarily used for quality control and visualisation purposes, assessing whether the quantified counts within each sample behave as expected.

However, [Maz16] suggests that all these transformation functions (TMM from edgeR, RLE from DESeq2, and MRN (Median Ratio Normalization, first reported in [Maz+13]), another transformation function, are closely related and further demonstrates that the computation of the relative size of transcriptome, s_i, or even the normalised counts are strictly equal in *simple experimental designs*, such as the "two-conditions-without-replicates" (referring to *replicate* as repeated measures of the same tissue, in the same individual) scenario illustrated in Appendix A.4.1. Aligning with these theoretical properties, [Maz16] has shown with numerical simulations on a simple experimental design, which involves only two conditions and no replicates, that any of the normalization methods described before yield similar estimates.

We present our wrapper custom functions that apply VST transformation directly on an ExpressionSet object:

```
# Creation of a DESeq object
dds <- DESeq2::DESeqDataSetFromMatrix(
  countData = Biobase::exprs(GSE149050_raw_count_filtered),
  colData = Biobase::pData(GSE149050_raw_count_filtered),
  design = ~ 1 + cell_type)

# Estimation of vst normalized expression matrix
counts_vst <- SummarizedExperiment::assay(dds %>% DESeq2::vst(blind = F))

# Update the ExpressionSet with vst normalised expression
GSE149050_vst_filtered <- bbcWrangling::update_ExpressionSet_object(
  ExpressionSet = GSE149050_raw_count_filtered,
  expression_data = counts_vst)
```

Warning: Limits of normalisation/transformation methods

It should be noted, however, that none of these methods are well-suited for datasets characterised by a substantial proportion of differentially expressed genes, since the majority of these methods assume that the expression of most genes remains unaffected by changes in the phenotype or experimental conditions.
In addition, it is recommended to apply all these methods in the raw count space to achieve optimal performance. Regarding this issue and the strong assumptions underlying these transformations, all these methods usually come along with their custom differential analysis pipeline. Indeed, the raw count distributions are not Normal-shaped, deterring for instance from using the conventional `limma` package which considers that the Gauss-Markov assumptions (see Theorem C.1.2).

To conclude, the choice of the normalisation and transformation functions depend on the objectives and the nature of the study. Study [Dil+13], under the Statomique initiative, compares the performance of the seven most employed normalisation methods for RNA-seq differential

analysis and demonstrates that DESeq [AH12] and TMM [RMS10] are the more accurate methods, acknowledged by robust reconstruction and correction of a set of 30 Housekeeping genes genes. Indeed, fundamental genes are anticipated to exhibit similar expression levels across samples, irrespective of the underlying biological conditions. Consequently, we expect precise normalization methods to consistently yield estimates for these genes, regardless of variations in library size or specific compositional biases.

A.3.4 Quality control and data exploration

Once these filtering and normalisation operations have been performed, it is common to perform quality controls using a variety of visualisations to ensure their correctness and interest in enhancing biological insights. We further decompose these methods into univariate quality plots to control the distribution of Counts (see Appendix A.3.4, and *Multidimensional Scaling* (MDS, see Appendix A.3.4) Plot to assess the similarity between samples, and verifies that samples associated to the same phenotype are indeed clustered together. For a comprehensive comparison of best practices to verify the integrity of RNA-Seq analysis, we kindly refer the reader to the following review papers: [Sch+23], [Col+21], [Cor+18] and [Su+14].

Control the distribution of read counts

Univariate visualisations, such as kernel plots (Appendix A.3.4), or boxplots (Appendix A.3.4), help bioinformaticians to verify the effectiveness of normalisation methods, reported in Appendix A.3.3, on the correction of technical artefacts at a low resolution level, by providing a quick overview of the distribution of gene expression values across samples.

Boxplots Complementary to density plots reported in Appendix A.3.4, we display in Figure A.7 the boxplot distributions concatenated per cell type, plotted with the internal function `bbcViz::draw_boxplot`

Conclusion: Boxplot insights

Boxplots can reveal batch effects or systematic variations between groups of samples. Unwanted batch effects can lead to differences in the central tendency (median) and spread (interquartile range) of expression values between groups. They highlight outliers samples that deviate significantly from the rest of the distribution, suggesting a sample quality issue, such as contaminations or human-made errors. Finally, by comparing boxplots of gene expression values before and after normalization, biologists can easily assess whether the normalization process has effectively reduced variability between distributions.
Paired comparison reported in Figure A.7 confirm the absence of outlier samples, and confirms that the VST normalisation has effectively homogenised the variability across samples.

Kernel plots We report the kernel plots, before and after gene filtering respectively, with the custom function `bbcViz::draw_kernel`, in Figure A.8:

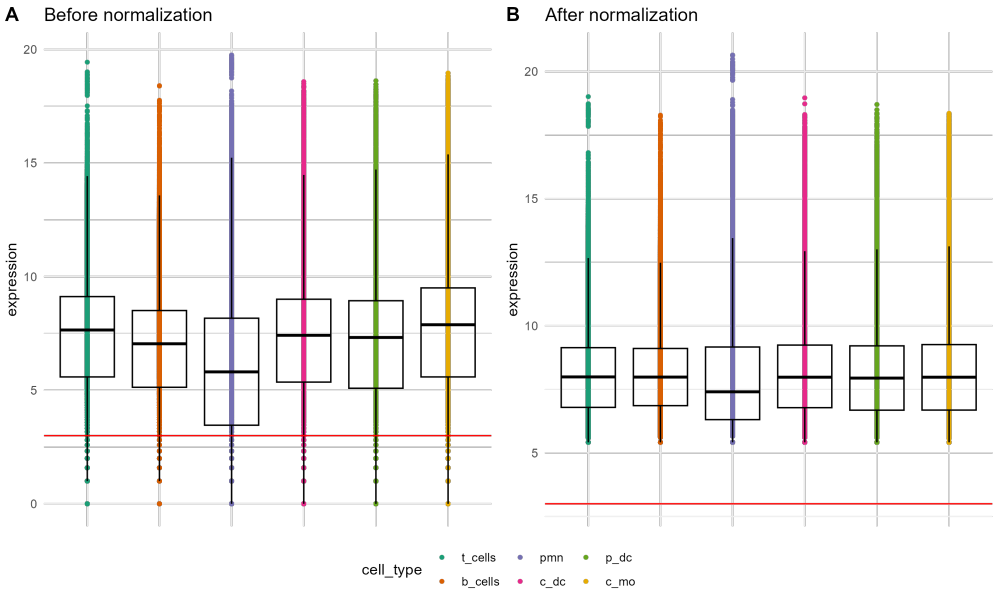
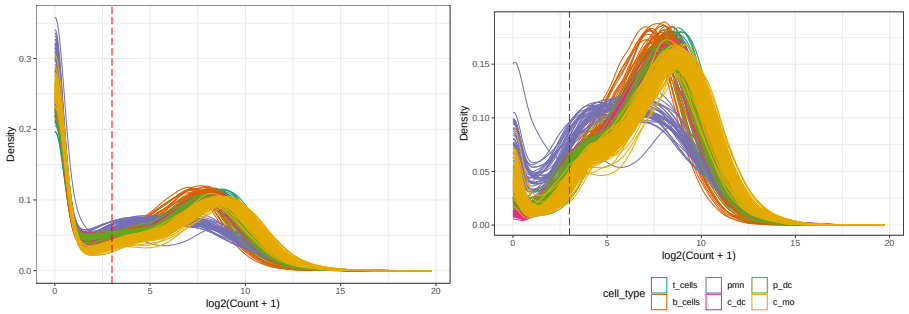


Figure A.7: Boxplot representations of transcriptomic expression aggregated per cell type are shown before (Subfigure A) and after normalization (Subfigure B). The x-axis represents the cell population, while the y-axis represents transcriptomic expression, after applying a log_2 transformation.



(a) Before filtering **(b)** After filtering

Figure A.8: Density distribution of the gene counts, after applying $\log_2$ transformation and removing null counts

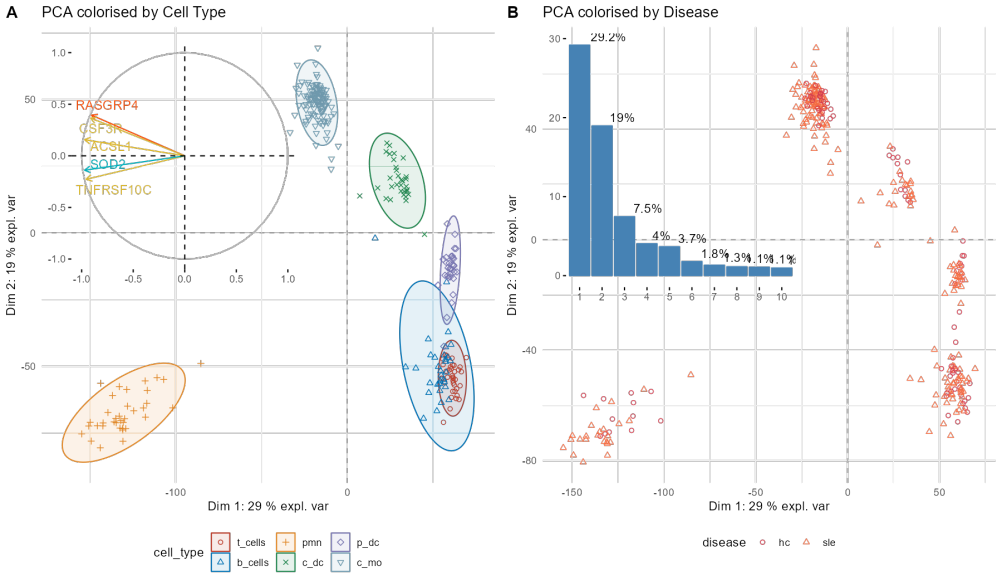


Figure A.9: In the left panel, the PCA projection is colourized by cell population, while in the right panel, the same projection is colourized by disease. For abbreviations used in this representation, report to Appendix A.

Conclusion: Kernel plot insights

Kernel Density Plots provide more details of data distribution compared to boxplots, by returning the empirical probability density function of the data. Accordingly, Kernel density plots can reveal bimodal (or more) distributions, which might be indicative of subpopulations within the samples and can help assess whether the data follows a Normal, bell-shaped distribution.
Paired comparison reported in Figure A.8 confirms the threshold used to filter background genes, reported in Appendix A.3.2, followed by a log_2 normalisation, effectively turns the bimodal distribution observed into an unimodal one (removal of the peak related to the background noise), and enforces that the transcript distributions approximate Gaussian distributions.

Multidimensional projections to identify patterns of expression

Principal component analysis (PCA) *Principal component analysis* (PCA) allows a representation of the samples in a low dimensional space estimated considering all genes' expressions (after background genes filtering). We compute the matrix resulting from PCA projection using internal function `bbcUnsupervised::compute_pca`. The projection of individual samples from the GEO149050 cohort onto the two-dimensional space described by the two largest eigenvectors is depicted in Figure A.9:

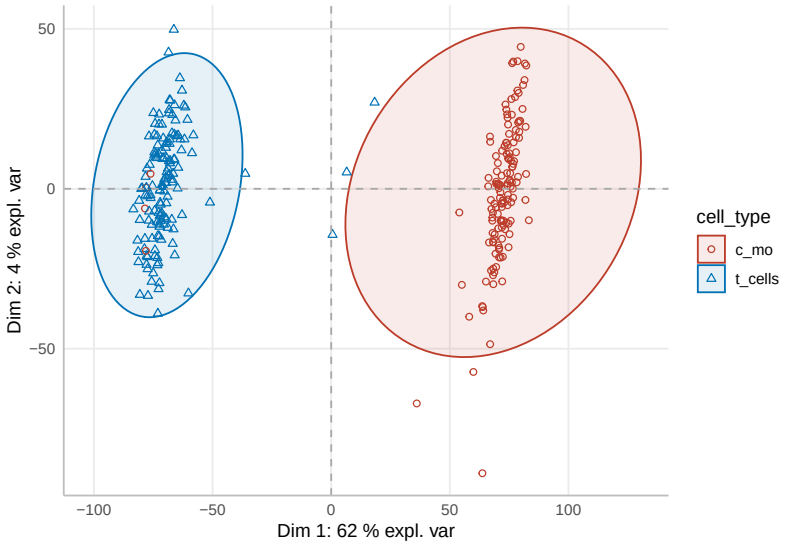


Figure A.10: PCA projection for cohort GEO137143, colourised by cell type. For abbreviations used in this representation, report to Appendix A

Conclusion: Evaluation of PCA projections

- Most of the cell populations are well clustered, on the simple basis of transcriptomic expression, and notably the PMN group expression clearly differentiates from the other cell types. Expected, the transcriptomic expression of B cells and T cells is relatively close.
- The general phenotype (disease, whether the sample proceeds from a patient suffering from SLE, or an healthy control) appears to have a negligible impact on the final transcriptomic profile.

Briefly, we represent in Figure A.10 the resulting pca projection for study GEO 137143, after performing the same normalisation and pre-preprocessing steps as described before:

Warning: Detection of outliers through PCA projection

We note at least five samples which seem to be clearly outliers, maybe resulting from technician wrong annotation of the samples.

Partial least squares - discriminant analysis (PLS-DA)

Information: Introduction to Partial least squares

Partial least squares - discriminant analysis (PLS-DA) allows a representation of the samples in a low dimensional space estimated considering genes that are selected as the most discriminant between groups of a variable of interest.

PLS-DA is computed with function `bbcSupervised::compute_plsda` and visually projected in a two dimensional space with `bbcViz::draw_plsda_individuals`. We displayed in Figure A.11 the corresponding PLS-DA bi-dimensional projections, enabling to quickly identify whether we are able or not to discriminate cell populations on the one hand, phenotype origin on the other.

Heat map and sample clustering Unsupervised clustering methods were employed to assess whether the normalization functions applied in Appendix A.3.3 keep on grouping the phenotypic contrasts of interest together. We utilized the Agglomerative hierarchical clustering method, implemented by the R `mclust::hc` function from the `mclust` package [Scr+16] to compute the similarity distance matrix. Subsequently, we employed the pheatmap package to generate the heat map shown in Figure A.12.

Information: PCA exploration preliminary conclusions

Unsupervised clustering tends to confirm that cell types are perfectly separated, once accounted in the design, with in particular, a strong dichotomy in terms of expression profiles between Polymorphonuclear Neutrophils (PMN) and Peripheral Blood Mononuclear Cell (PBMC) cell populations. On the other hand, discriminating disease phenotypes within a given cell population, namely identifying cell populations extracted from normal tissues from those displaying a disease phenotype is a comparably much harder task.

A.3.5 Batch effect correction

A *blocking factor* is a categorical variable, that is not of primary interest in the main experimental objectives. Identifying blocking factors is crucial to control for unwanted sources of variation that could otherwise confound the analysis. The most common sources of undesired variability proceed from *batch effects*, which refers to the variation introduced during sample processing or sequencing from the use of different equipment or reagents.

When unwanted sources of variation are not accounted for, they can introduce noise into the analysis, making it harder to spot true differences between biological conditions. Hence, blocking these confusing factors can increase the statistical power of your analysis and contribute to identify meaningful differences.

Two complementary strategies can effectively mitigate batch effects in the analysis: the first strategy involves directly incorporating batch information into the design of the differential analysis, as discussed in Appendix A.4.1. Alternatively, the second approach focuses on correcting technical variations before conducting the subsequent downstream analyses.

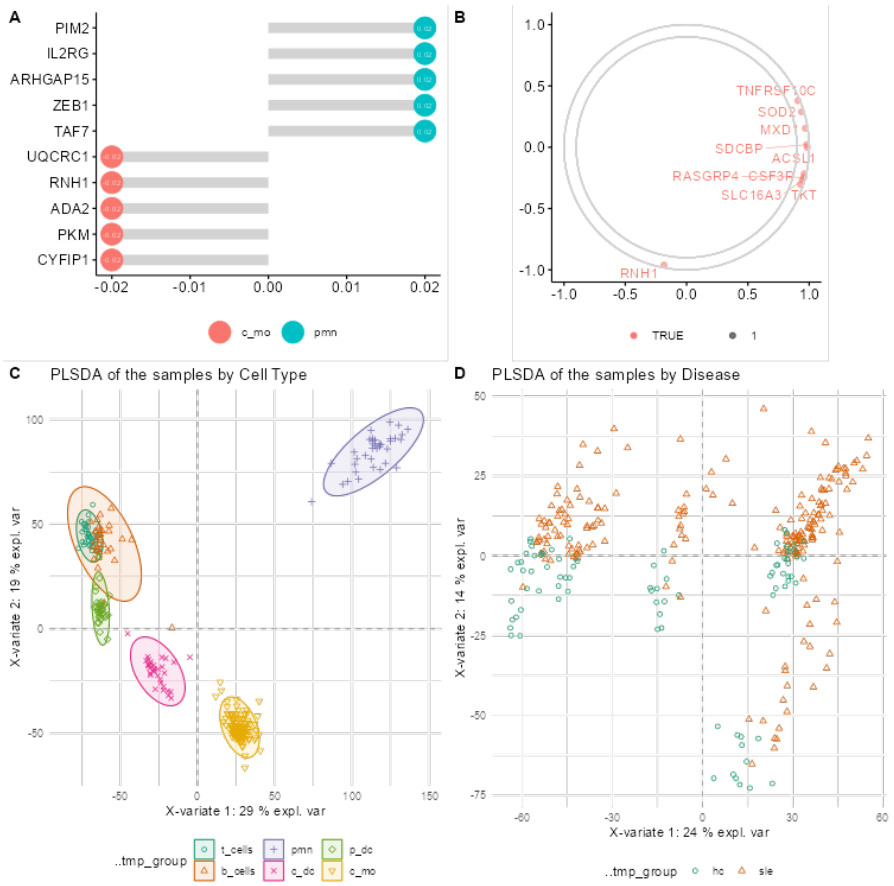


Figure A.11: In Panel A, the first 10 *loadings*, representing the 10 genes that contribute the most to the prediction of the six cell subsets included in the PLS-DA analysis, are displayed. The x-axis indicates the log2 fold-change value (lFC) after data standardization. Notably, the most distinguishable populations are monocytes (in red) and PMN (in blue). The correlation circle, reported in Panel B, of the associated PLS-DA projection, restrained to the top 10 genes with the highest loading, is an interesting complementary representation. This visualization illustrates how strongly these genes are correlated with the first two principal components. The module of the gene vectors denotes their total contribution to the *cross-covariance* matrix, and the angle with each axis is the degree of cosine correlation with respect to the first two loadings. PLS-DA projections are reported for Panel C and for Panel D, colourized by cell type and disease, respectively. For abbreviations used in this representation, report to Appendix A

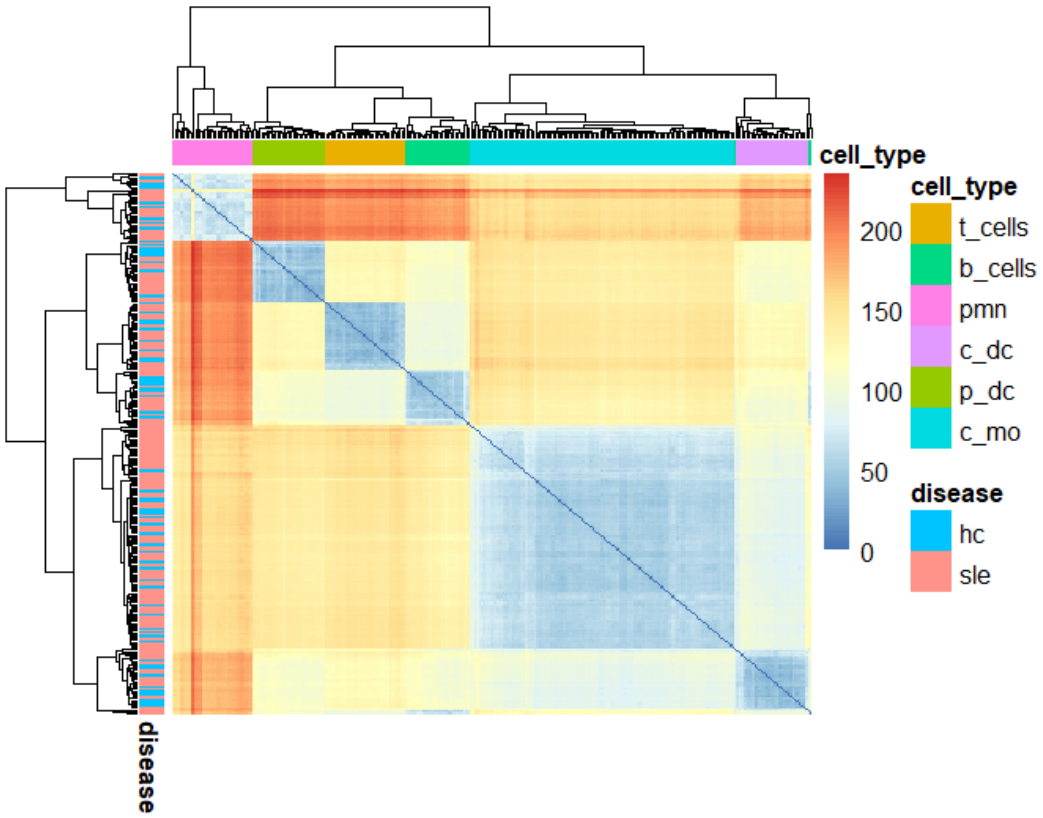


Figure A.12: Heat map followed by hierarchical clustering, showing explicitly that the profile of all listed cell populations clearly discriminate.

We illustrate this concept by applying the supervised *ComBat* (appendix A.3.5) method to the expression matrix resulting from merging the two separate cohorts, GSE149050 and GSE137143, for the two cell populations shared by both datasets: monocytes and PMN. We demonstrate with PCA visualisations that the batch correction effectively addresses some of the variations introduced by differences of sequencing technology and protocols.

Combat

To apply Combat correction, a parametric empirical Bayes framework for originally adjusting batch expression across microarray datasets (see [JLR07]), we need at least one observation for each factorial combination of categorical variables. Indeed, if this assumption is violated, there is no reference anymore to set apart explicitly the noise resulting from the technical bias from the expected biological variability, of interest. That's why we use the helper function `bbcWrangling::filter_samples_from_expr` to select genes and factors (concerning cell populations, only monocytes and T cells populations are shared in both experiences) present in both conditions (failure to enforce so results in a singular, degenerate problem).

We use the supervised `sva::ComBat` [Lee+22] to apply ComBat correction. This function is versatile, allowing users to specify both the design of the experience (the expected biological variability, that should not be impacted by the correction process) and the technical indicator variable. The resulting PCA from the aggregation of both GEO datasets, GSE149050 and GSE137143, before and after batch removal, is pictured in Figure A.13

Conclusion: Graphical evaluation of Combat normalisation

By comparing the PCA plots before and after batch correction, biologists can visually assess whether the batch effect has been mitigated. Successful batch effect reduction is indicated by bringing closer previously distant samples from different batches. Indeed, the key paradigm that underlies batch correction is to decrease the influence of technical bias as a source of variation, while optimising the separation between samples reflecting biological variation.
We can observe in Figure A.13 that the ComBat correction only partially mitigated the batch effects. Samples from the same cell population are now closer to each other, but we can still distinguish different cohorts within the same cell subtype.

Surrogate Variable Analysis (SVA)

To conclude that section, note that unsupervised methods can alternatively be used to identify and account for hidden sources of variability in high-dimensional datasets, by reproducing and estimating their effect using latent variables.

Figure A.13: PCA projection before (left column) and after (right column) applying ComBat correction to assess how much the observed batch effect between the two cohorts, GSE149050 and GSE137143, has been mitigated. Each row in the panel, from top to bottom, is colour-coded by technical batch (identified as the cohort), disease, and cell type.

Warning: Surrogate Variable Analysis (SVA)

SVA is a statistical method, mostly used in a high-dimensional framework. By removing the effects of undesired variability and controlling false positives by amplifying the true biological signals from unwanted sources of variation, SVA can enhance the statistical power of downstream analyses and make them more robust.
We refer to these factors that capture unmodelled sources of variability as *surrogate variables* (also known as hidden or latent variables) and in particular, SVA utilizes a singular value decomposition (SVD) to estimate them from the covariance structure of the data.

However, we should highlight that these unsupervised methods for addressing non modelled variability in high-dimensional data may not effectively capture all sources of unmodelled variability, leading to suboptimal correction, and are not straightforward to interpret. Indeed, without domain-specific knowledge to make sense of the identified hidden sources of variability, it is usually infeasible to interpret whether the detected surrogate variables are biologically meaningful ("wanted" insight, [Jin+21], [Law+20]), such as variation associated with a specific cell type or stem from technical causes ("unwanted" variation), such as batch effect ([HS21], [Kas+22a]) .

SVA also comes along with its own limitations: this projection method assumes that non modelled variability follows the Gaussian framework reported in Theorem C.1.2, or at least is linearly related to the observed data. SVA, is computationally intensive and time-consuming when applied to high-dimensional data, can be computationally intensive and time-consuming, but on the other hand, it tends to work best with large sample sizes. Finally, optimising the hyper-parameters, such as the number of surrogate variables to estimate, is critical for its performance.

A.4 Downstream analyses

In this section, within the realm of downstream analyses handling transcriptomic data, we focus on Differential Gene Expression Analysis (DGEA), since these methods are widely popular among biologists for their straightforward interpretability.

For an exhaustive examination of enrichment-based methods, which constitute the second category of downstream methods, we kindly direct interested readers to refer to Section 5.1, in which we applied these methods to identify pathways and marker genes that characterize distinct cell populations.

A.4.1 Differential expression analysis

Differential Gene Expression Analysis (DGEA) aim to test whether the difference of the mean expression of a given gene between two (or more) biological conditions (often control versus disease, or drug 1 outcome vs drug 2) is statistically significant [Smy+22].

It usually involves the following steps:

1. First, build a *design matrix* that describes the experimental design and sample groups, possibly including any covariates or confusing variables.

2. Apply one of the most popular statistical tools employed for differential analysis, namely DESeq2, edgeR, or limma-voom, to identify genes that are significantly differentially expressed between groups. Precisely, each gene comparison undergoes a hypothesis test, where the alternative to the null hypothesis concludes to a significant difference in gene expression between biological conditions. Alternatively, you may employ any of the parametric or unparametric method listed in Appendix A.4.1: they are easier to implement and agnostic to the nature of the dataset compared, but this flexibility and robustness comes at the expense of statistical power and precision (especially when the assumptions underlying the theoretical application of counts-based models, such as DeSeq2, or Gaussian-distributed, such as limma, hold [Jea+10]).

3. Conducting numerous statistical tests simultaneously, on thousands of genes, to determine which are differentially expressed between conditions, requires to deal with "multiple testing" problem. Indeed, with thousands of tests performed independently, the probability of observing false positives (genes that appear differentially expressed by chance) becomes significant.

   To address the multiple testing problem, researchers commonly adjust the p-values obtained, using methods designed to either control the family-wise error rate (FWER) or the false discovery rate (FDR). FDR control methods focus on controlling the expected proportion of false positives among the set of differentially expressed genes, while FWER control methods aim to control the total probability of making at least one Type I error across all conducted tests. FWER control methods, such as the Bonferroni Correction, are more conservative, thereby better fitted when stringent control over Type I errors is critical, while FDR methods, like the Benjamini-Hochberg procedure, are more flexible, and often preferred when finding the sweet sport between the ability to detect true positives while controlling the rate of false discoveries.

4. Define a fold change (FC) and $p-$ value threshold to identify genes with biologically meaningful changes

5. Generate visualisation plots, such as volcano plots, heatmaps, concordance plots, and interaction plots to visualise the set of differentially expressed genes and apprehend intricate co-expression patterns.

From scratch

Without assumption on the distribution of the transcripts, or without using any non base R package, the easiest way to compare transcriptomic expression is certainly by resorting to tests provided in `stats` package:

- t-test should be used for normally distributed variables
- Wilcoxon and Kruskall-Wallis are non-parametric tests, particularly relevant for small samples, that make no assumption on the distribution of the data, the first one being tailored to compare two groups, while the second one is used to determine whether a significant change occurs globally across multiple conditions.

```
# Contrasts
contr_matrix <- c("sle//hc")

# Perform group comparison with the chosen test ("t", "kruskal")

# For two contrasts
GC_res <- dea_univariate(
  ExpressionSet = GSE149050_vst_filtered,
  var = "disease",
  contrasts = contr_matrix,
  test = "t",
  correction = "BH")

# Across multiple conditions, no need to provide contrast
GC_res <- dea_univariate(
  ExpressionSet = GSE149050_vst_filtered,
  var = "cell_type",
  test = "kruskal",
  correction = "BH")
```

DESeq2 framework

Information: Design and contrasts in differential analysis with DESeq

- First, we consider a simple linear regression model in which the cell type origin is the unique source of variability in transcriptomic expression across samples.
- We use the Bioconductor package DESeq([AH12]) to perform the differential expression analysis, in which counts are modelled using Negative Binomial distribution and the regression model fitting using a Generalized Linear Model framework.
- Fold-change (FC) and adjusted p-values (FDR of Benjamini-Hochberg correction) are used to identify differentially expressed genes (DEG).
- A common standard for identifying genes that are significantly differentially expressed uses the following dual criteria: an absolute fold-change greater than 1.3, $|FC| > 1.3$, and an adjusted p-value below 0.05, $p_{\text{adj}} < 0.05$, used as a proxy to control for the false discovery rate [Dey+22].

The following script performs DESeq analysis and returns a tibble with, for each contrast, the table of DEGs (differentially expressed genes):

```
# common biological thresholds
t_pvalue <- 0.05; t_FC <- 2
```

```
# Model without covariate
model <- formula(~0 + cell_type)

# Contrasts (two possibilities to create them)
contr_list <- list(
  c("cell_type", "t_cells", "b_cells"),
  c("cell_type","p_dc", "c_dc"))

# Perform differential expression analysis with DESeq
DEA_res_seq <- bbcSupervised::dea_deseq(
  ExpressionSet_object = GSE149050_raw_count_filtered,
  model = model,
  contr_list = contr_list,
  feature_colname = "Genes")

# Add the tables of DEGs to the results
DEA_res_seq <- DEA_res_seq %>% mutate(DEgenes = purrr::map(
  data,
  ~ .x %>% bbcSupervised::subset_deg(., "FC", t_FC,
  "FDR", t_pvalue,
  order = "FC")
))
```

Limma framework

Information: Design and contrasts in differential analysis with limma

- We consider in first intention the same linear regression framework reported in Appendix A.4.1, including only one discrete explanatory variable. We refer to this configuration as *fixed*, since we observe all the possible modalities taken by the covariate`cell_type`.
- In that section, we consider the R package `limma` to perform the differential expression analysis
- Normalized expressions are modelled through a Normal distribution and fitted using a Linear Model. A moderated t-statistics is then derived through a Bayesian estimation of the variance.
- Fold-change (FC) and adjusted p-values (FDR of Benjamini-Hochberg correction) are used to identify differentially expressed genes (DEG)

```
# common biological thresholds
t_pvalue <- 0.05; t_FC <- 2
```

```
# We do not consider the presence of intercept
# Instead, we compute the averaged expression for each cell type
model <- model.matrix(~ 0 + cell_type, data = GSE149050_raw_count_filtered)
colnames(model) <- GSE149050_pheno_data %>%
  pull("cell_type") %>% unique ()

# Contrasts (as a proof of concept, we consider only two cell contrasts,
# in order to separate closely related cell populations)
# makeContrasts function enables to precisely compute the contrast of interest,
# even when the resulting design is singular
contr_matrix <- limma::makeContrasts(
  T_vs_Bcell = t_cells - b_cells,
  cDC_vs_pDC = p_dc - c_dc,
  levels = colnames(model))

# Perform count data normalization to apply Limma to be tested
data_normalized <- bbcSupervised::dea_limma_normalization(
  ExpressionSet_object = GSE149050_raw_count_filtered, model = model)

# Perform differential expression analysis with Limma
DEA_res <- bbcSupervised::dea_limma(
  ExpressionSet_object = GSE149050_raw_count_filtered,
  data_norm = data_normalized,
  model = model,
  contr_matrix = contr_matrix,
  feature_colname = "Genes")

# Add the tables of DEGs to the results
DEA_res <- DEA_res %>% dplyr::mutate(DEgenes = purrr::map(
  data,
  ~ .x %>% bbcSupervised::subset_deg("FC", t_FC, "FDR", t_pvalue, order = "FC")
))
```

Instead, we could have considered a *linear-mixed* model, including in the experimental design both random and fixed variables (in that context, we assume that `cell_type` has a *fixed* effect, while `patient_id` has a *random* effect, since the patients included in the clinical trial are likely to constitute only a sample of the whole population). An example of such a design is proposed in the code hereafter:

```
# Linear-mixed model specification
model <- model.matrix(~ 0 + cell_type + patient_id,

data = GSE149050_raw_count_filtered)

# Perform count data normalization to apply limma
data_mixed_normalized <- bbcSupervised::dea_limma_normalization(
```

```
  ExpressionSet_object = GSE149050_raw_count_filtered,
  model = model,
  random = TRUE,
  var_random = "patient_id")

# Perform differential expression analysis with Limma
DEA_mixed_res <- bbcSupervised::dea_limma(
    ExpressionSet_object = GSE149050_raw_count_filtered,
    data_norm = data_random_normalized,
    model = model,
    contr_matrix = contr_matrix,
    random = TRUE,
    var_random = "patient_id",
    feature_colname = "Genes")
```

Visualisations

DEGs table We displayed in Table A.6 the total number of genes identified as DEGs, as a relation of the prior thresholds established, between C(lassical) and P(lasmacytoid) cells subsets[3].

Table A.6: Total number of DEGs identified, for the following respective thresholds, FC: $\{1.3, 1.5, 2\}$, and adjusted p-values: $\{0.05, 0.1\}$.

pvalue threshold	0.05			0.1		
foldchange	Total	Up	Down	Total	Up	Down
1.3	5,397	2,480	2,917	5,400	2,481	2,919
1.5	3,659	1,619	2,040	3,659	1,619	2,040
2.0	1,883	781	1,102	1,883	781	1,102

and we displayed the top 10 most differentially expressed genes in Table A.7, ordered by decreasing absolute value of log2 Fold change:

[3]Second contrast studied can be observed as well on the HTML report

Table A.7: Table of differentially expressed genes, showing for each of them their respective p-value and fold change.

Genes	log2FC_T_vs_Bcell	log2FC_cDC_vs_pDC	FDR_T_vs_Bcell	FDR_cDC_vs_pDC
JCHAIN	-7.7	8.4	0.0	0.0
MS4A1	-7.5	0.2	0.0	0.2
BANK1	-7.1	0.0	0.0	0.8
TCL1A	-6.7	7.3	0.0	0.0
IL7R	6.5	-0.3	0.0	0.1
CD3D	6.2	0.0	0.0	0.8
CD3E	6.1	1.4	0.0	0.0
CD19	-6.1	-0.0	0.0	0.8
CD22	-5.9	-0.5	0.0	0.0
HLA-DRA	-5.9	-1.9	0.0	0.0

(10 first lines / 10 lines)

Conclusion: Analysis of DEGs table

- Even being stringent on these thresholds, by considering a high fold-change of 2 and a small adjusted $p-$ value of 0.05, we still observe a great number of genes identified as DEGs (1883 in total, with 781 up-regulated in cDC in comparison with pDC and so 1101 down-regulated)
- Similarly, the number of identified DEGs between B cells and T cells is significant, with 525 up-regulated genes identified in T cells in comparison with B cells and 687 down-regulated)

P-value distribution For the two-described contrasts, we plot in Fig Figure A.14 the density distribution and histograms of the computed $p-$ values.

Conclusion: Sound evaluation of P-values distributions

P-value distributions for both studied cell contrasts display strong signal, since the distribution has a strong left heavy tail, implying that many genes are significantly differentially expressed (on average, if the assumption $H0$ of null differential expression between two cell types hold, you would expect an uniform distribution between 0 and 1, with on average, 5% of the genes considered as differentially expressed, with a p-value below 0.05).

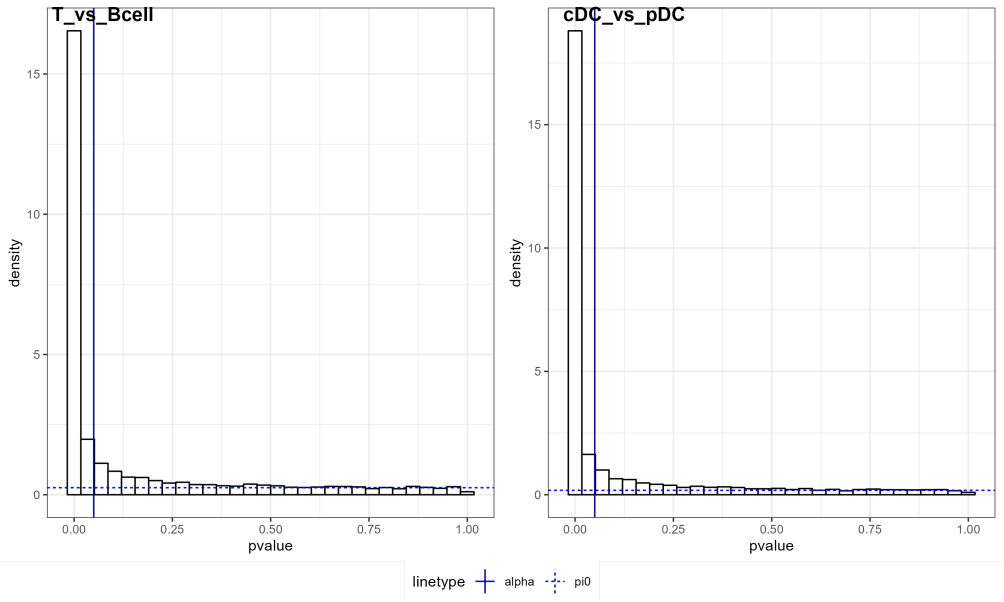
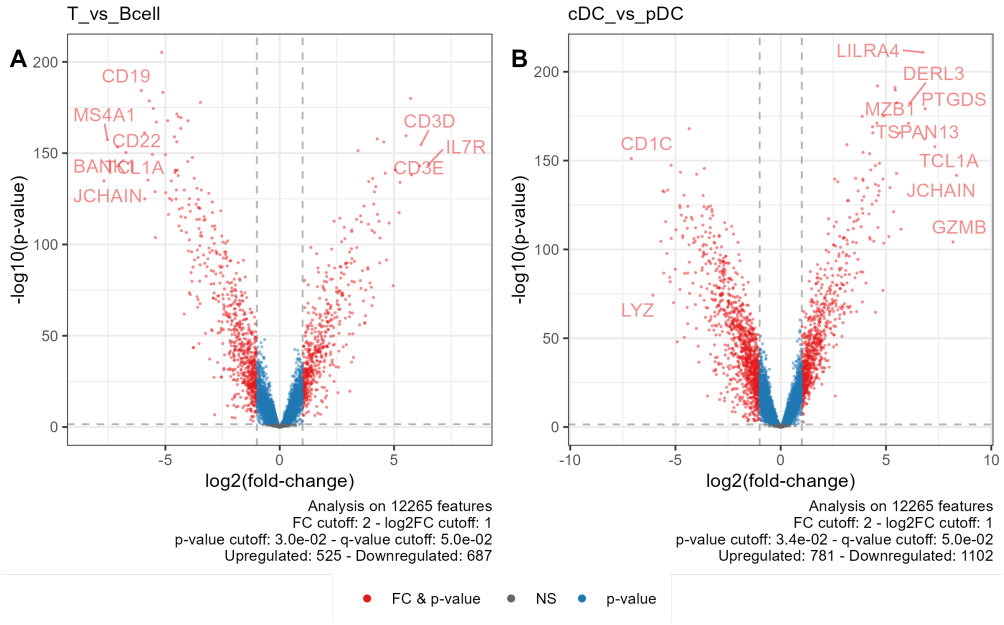


Figure A.14: P-value histogram distribution

Volcano plots



Concordance plots One way to evaluate the quality of the batch correction, reported in Appendix A.3.5, is to compare whether the fold changes, or the related p-values, for the same biological contrast, are similar across several cohorts. Visually, this would correspond in the perfect scenario to an unidimensional manifold scatter plot (fold change values in that context

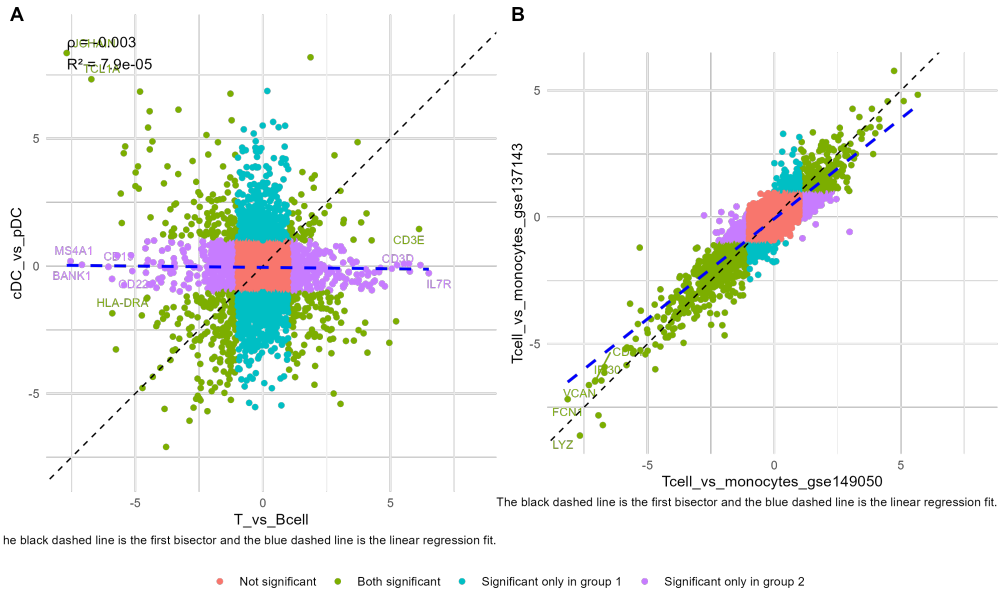


Figure A.15: As a proof of concept, we displayed first to the left the scatter plot between two conditions highly dissimilar (subfig A) and then between the two same biological contraststst, but processed in distinct batch conditions (subfig B).

should all line up along a straight line of slope 1 and of null intercept), we displayed as a toy example two contrast comparisons in Fig Figure A.15

More interestingly, it can be interesting to compare the set of genes that are differentially expressed between monocytes and T cells, for each batch and for the concatenated expression matrix. It can also be interesting to evaluate the global impact of the batch effect on the transcriptomic expression. We supply in the following snippet of code instructions to generate the corresponding contrasts of interest

```
# add dummy variable combining all categorical modalities
Biobase::pData(combined_ExpressionSet) <-

 Biobase::pData(combined_ExpressionSet) %>%
  tidyr::unite(col="combined_factor", platform, cell_type, remove = FALSE)

# First the design
design_interaction <- model.matrix(
  ~ 0 + combined_factor, data = combined_ExpressionSet)
colnames(design_interaction) <- c("GEO149_M", "GEO149_T", "GEO137_M", "GEO137_T")

# Then, compute the contrasts (coefficients of interest)
contr_matrix_interaction <- limma::makeContrasts(
  TvsMinGEO149 = GEO149_T - GEO149_M,
  TvsMinGEO137 = GEO137_T - GEO137_M,
  Diff = (GEO137_T - GEO137_M) - (GEO149_T - GEO149_M),
```

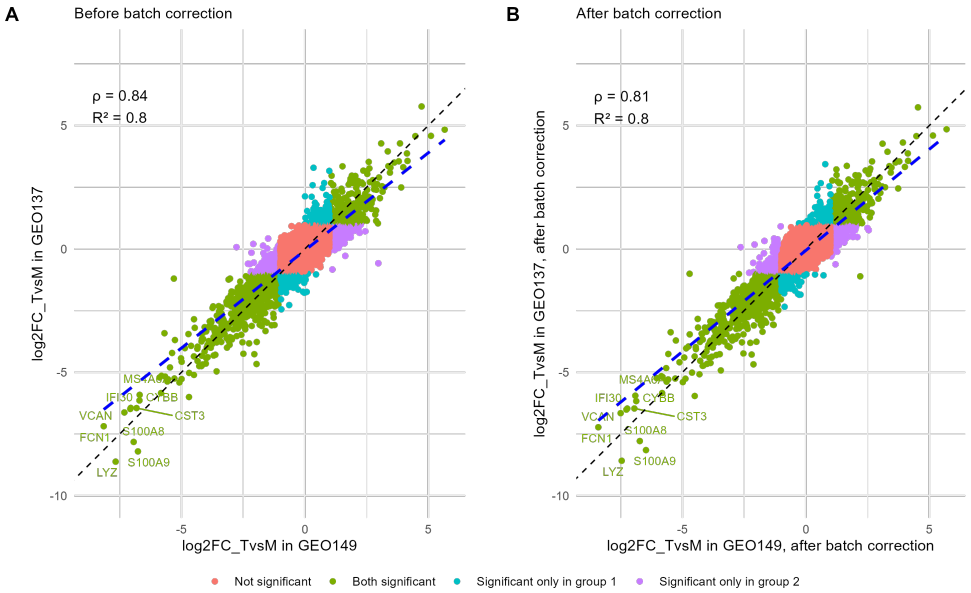


Figure A.16: Correspondance plot of fold-change values, before and after batch correction, comparing monocytes and T cells expression

```
  levels = design_interaction)

# Finally, perform differential expression analysis with Limma
DEA_uncorrected <- bbcSupervised::dea_limma(
  ExpressionSet_object = combined_ExpressionSet,
  model = design_interaction,
  contr_matrix = contr_matrix_interaction,
  feature_colname = "Genes")

Biobase::pData(combined_ExpressionSet_combat) <-
 Biobase::pData(combined_ExpressionSet_combat) %>%
  tidyr::unite(col="combined_factor", platform, cell_type, remove = FALSE)
DEA_corrected <- bbcSupervised::dea_limma(
  ExpressionSet_object = combined_ExpressionSet_combat,
  model = design_interaction,
  contr_matrix = contr_matrix_interaction,
  feature_colname = "Genes")
```

The impact of batch correction effect, as the difference between the averaged fold change values, is computed in Figure A.16

Heatmaps

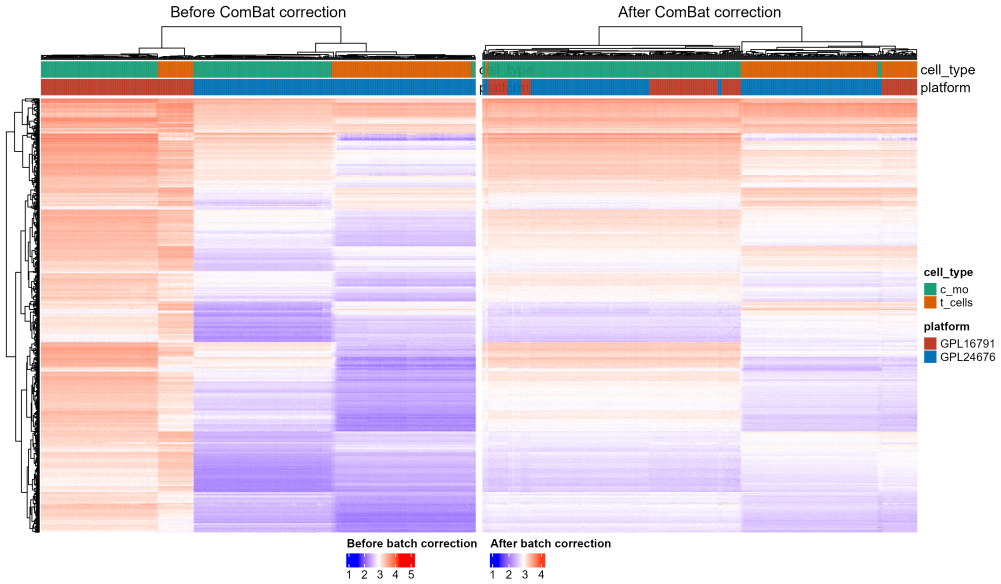


Figure A.17: Heatmaps (left: before applying batch correction, right: after applying batch correction) of the $\log_2$-transformed transcriptomic expression resulting from the merging of studies GSE149050 and GSE137143. For the sake of comprehensiveness, we employed the Bioconductor ComplexHeatmap package to visualise the expression matrices, instead of the `pheatmap` package used in Appendix A.3.4.

Profile plots The purpose of *profile plots* is to visualize the potential presence of any interaction for each subset of the universe of possible categorical combinations, in a multivariate linear approach including more than one discrete variable. An interaction between two factors occurs when the effect of one of the factors on the variable to be explained varies according to the levels of the other factor.

The interaction plot can be visualized with the R function `stats::interaction.plot` or with our custom ggplot-like function `bbcViz::draw_profile`. In Figure A.18 , we present the interaction plot of the two genes that exhibit the greatest variability between the two cohorts, GSE149050 and GSE137143.

Venn diagram Venn diagrams are standard representation to compare enriched sets of genes across biological conditions or cell types (Figure A.19):

For the interested reader, we refer to [Gau21, Chapter 1, Section 3] for a detailed review of the statistical framework and assumptions underlying the top 3 most popular differential methods, namely limma, EdgeR and DESeq2 (in addition, the PhD manuscript introduces its own DGEA method, described hereafter), and to [Gau21, Chapter 2, Section 2] and [Bic20, Chapter 2, Section 2] for a comprehensive review of family of methods dealing with multiple testing problem, occurring when performing simultaneous comparisons of independent genes across two biological conditions.

To conclude, [CDL17] demonstrates the better performance achieved by `NOIseq` [Tar+11], `DESeq2` [LHA14] and `limma+voom` [Law+14] methods, consistent with findings from [Rit+15] compared to a compendium of six mapping methods combined with nine differential expression analysis frameworks and evaluated against qRT-PCR data. They additionally demonstrate that mapping methods have minimal impact on the final DEGs analysis, and the combination of

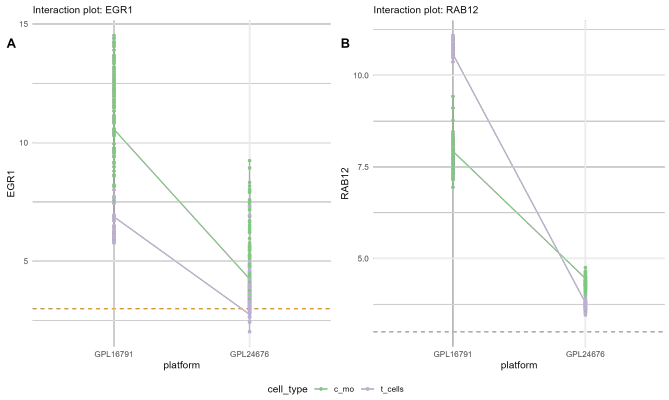


Figure A.18: Interaction plot, showing the difference of averaged fold change for the two genes displaying the most dissimilar contrast, before any batch correction is applied.

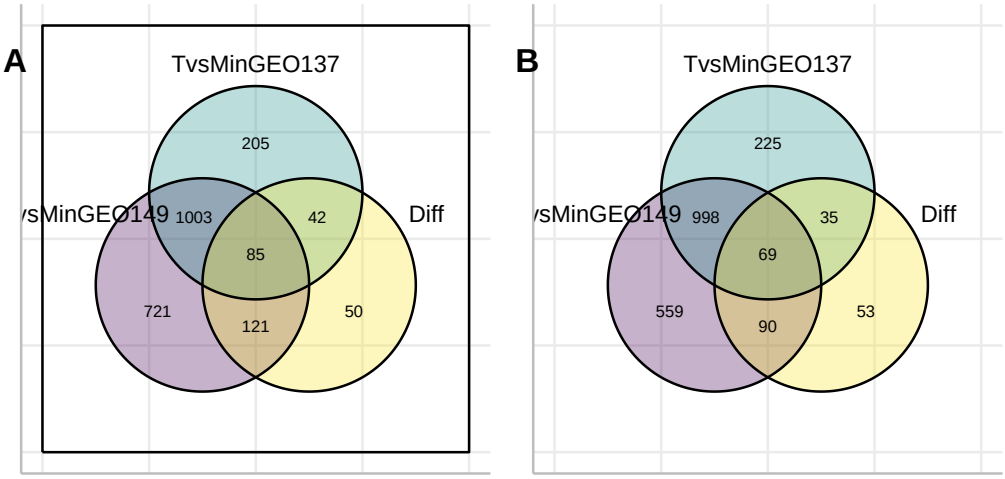


Figure A.19: Compare the set of genes considered as differentially expressed in T cells with respect to monocytes, before and after Combat batch correction

different methods can produce more accurate results, this consensual option being included in the available software, consexpression, implemented on Github by the Bioscience team.

[Hej+22] introduced recently the dearseq DGEA method, that implements a robust statistical test, accounting for data heteroscedasticity, and which was released as a Bioconductor package. This method outperforms other differential analyses methods in a number of virtual simulation experiences, and notably the agnostic and non-parametric Wilcoxon rank-sum test. Ultimately, the paper recalls the intrinsic difficulty of generating insightful and biological relevant numerical simulations, highlighting a major flaw in the data generation process of previous study [Li+22], which led to incorrect comparisons of the performance of differential expression analysis.

A.4.2 Multi-level classification of cell populations

Intrinsic complexity of multi-classification prediction In the previous section, the design matrix was straightforward to build, since we've only compared two pairwise levels, such as phenotypes or cell type. Nevertheless, in a multivariate contrast that involves comparing multiple conditions, the task is much more challenging.

Indeed, while binary classification is a common machine learning problem, multi-class classification is inherently more complex due to the combinatorial explosion of possible outcomes to consider (time- and memory-consuming), and the commonly observed pattern of increased overlap when increasing the number of classes (especially when two classes exhibit similar characteristics, or the dataset is strongly *imbalanced*). Out of this exponential explosion of computational cost, most of the machine learning algorithms have been designed for binary classification and do not scale well with increasing number of levels to predict.

Furthermore, interpreting and visualising the outputs and key features driving the classification decisions is usually intricate, as it involves understanding how the model makes decisions across multiple classes. In particular, it becomes more complex to evaluate globally the performance of the algorithm (accuracy is commonly used in binary classification, while multi-class problems often require specialized metrics such as confusion matrices). Finally, the risk of overfitting increases with the number of classes, especially with small or imbalanced datasets, even though this issue can be partly alleviated by a whole realm of regularisation techniques.

The conventional approach when tackling multi-class classification involves beginning with standard algorithms initially designed for binary classification and subsequently adapting them to handle multi-class problems through feature engineering methods. Among them, the "one-hot encoding", using a dummy variable to decompose the complex multi-class problem into a series of simpler binary classifications (see details in Definition A.4.1), is certainly the most popular.

Definition A.4.1: One-vs-all (OvA) strategy

The OvA strategy, also known as "one-vs-rest", is a commonly used technique in machine learning for multi-level classification problems. This strategy consists of transforming a complex multi-class classification problem into several binary classification problems, for which most of the classification techniques have natively been developed. Applied to our deconvolution problem, for each of the J cell populations, you train a binary classifier to distinguish a given cell population j from all others. While this approach is straightforward, interpretable and can be easily deployed with any binary classifier, it is sensitive to class imbalance *and performs badly when there is significant overlap between two closely related classes.*

Practical use case: identify gene markers of a cell population Let's delve into a practical example of a complex multi-class problem, as outlined in Appendix A.4.2. In this analysis, we employ the "limma + voom" statistical framework, with the objective of identifying the smallest subset of genes that distinctly define a specific cell population when compared to all others.

In practice, the one-hot encoding (Definition A.4.1) can be implemented in both versions, with distinct statistical properties:

- The most straightforward strategy consists of creating a "dummy" indicator variable for each cell population: it equals 1 when the sample belongs to the cell population of interest and 0 otherwise. However, this strategy may mask individual differences between cell populations, so it's not recommended when the profile of the remaining cell populations strongly diverges (risk of losing relevant biological information), or when the dataset is imbalanced (which is an actual issue in our use case, as illustrated in our frequency table in Table A.1).
- Alternatively, we can evaluate the significance of the difference between the cell population of interest and the averaged expression of all other cell types (assuming equal contribution of each cell population, denoted as the common denominator). This strategy is recommended when the groups being compared are heterogeneous.

For a detailed explanation of the differences between these two approaches and their respective pros and cons, we recommend the Bioconductor discussion thread Limma: Contrasts comparing one factor to multiple others.

We display hereafter a code snippet to automatically build the associated design contrast for GEO study 149035, combing the power of `limma` package with our user-friendly differential functions:

```
### Retrieve the minimal set of coefficients
# Compute the averaged expression of each cell type,
# without intercept or reference cell line
model <- model.matrix(~ 0 + cell_type,
                      data = GSE149050_raw_count_filtered)
colnames(model) <- GSE149050_pheno_data %>%
pull("cell_type") %>% unique ()
# Perform vst data normalization to apply limma,
# enforcing its gaussian-distributed assumptions
data_normalized <- bbcSupervised::dea_limma_normalization(
  eset_object = GSE149050_raw_count_filtered, model = model)

### Build the contrasts of interest
# the number of contrasts to be compared against
num_cells <- length(cell_type_labels)

# build the contrast programmatically for each cell population
contr_matrix <- purrr::map(cell_type_labels, function(.x) {
  # identify the other cell types to be aggregated
  other_celltypes <- setdiff(cell_type_labels, .x)

  # contrast of the cell of interest, vs all-others
```

```
  contr_matrix_per_cell <- limma::makeContrasts(
    contrasts= paste0(.x, "-",
                      "(", paste(other_celltypes, collapse = " + "), ")/",
                      num_cells -1),
    levels = colnames(model))
  return(contr_matrix_per_cell %>% as.data.frame())
}) %>% purrr::list_cbind() %>% as.matrix() # combine all contrasts together
colnames(contr_matrix) <- paste0(cell_type_labels, "_Vs_All")

### Perform differential expression analysis
DEA_res_all_contrasts <- bbcSupervised::dea_limma(
  eset_object = GSE149050_raw_count_filtered,
  data_norm = data_normalized,
  model = model,
  contr_matrix = contr_matrix,
  feature_colname = "Genes")
```

It can be interesting now to represent marker genes that are shared across several populations, however, Venn diagrams are not anymore adapted with multiple conditions. Indeed, the total number of subsets, grows exponentially with the number of pairwise conditions to compare: with J the number of cell types, there are possibly $2^J - 1$ non-empty intersection sets to consider.

Upset plots, using R function `ComplexUpset::upset` [Kra20], displays an alternative, user-friendly graphical representation. To generate the global table of indicator assignments, we design the `bbcUtils::binary_membership` function.

Information: Upset plots: interest

The upset plot consists of three panels:

- The *Set size* panel, located in the bottom left corner, displays the size of each set being compared (i.e., the number of genes identified as differentially expressed for al six cell populations).
- The *Intersection size* panel, positioned in the top right corner, indicates the size of each possible subset. The total number of subsets considered is equal to the combinatorial enumeration of partial permutations on J items, and the cardinality for each of them is given by the intersection of the cell populations included. Setting the "group_by" argument to true orders them by cell population and then by decreasing cardinality.
- Lastly, the *group* panel provides a customizable representation of the groups being compared.

We illustrate this concept in Figure A.20.

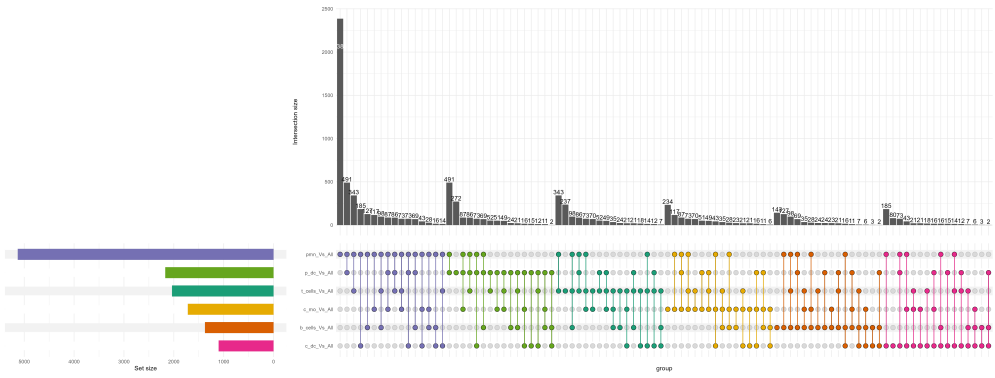


Figure A.20: Upset plot of the overlapping subsets of genes identified as DEGs, in study GEO149050.

Conclusion: Upset plots: conclusions

From Figure A.20, we can clearly see that the PMN group stands out as the most dissimilar. This aligns with our earlier observations in the PCA projection (refer to Appendix A.3.4), where we observe that the PMN cluster exhibited the highest contribution to the overall variability in transcriptomic expression and appeared the most distant from all other clusters.
Far behind, the second most unique profile is composed of the conventional (or classical) dendritic cells (`c_DC`).

From Figure A.20, the union of DEGs, `length(Reduce(union, list_venn))=6712`, is still too large to perform efficient and robust deconvolution. In linear regression framework, it is thus typical to refine the resulting *signature matrix* by optimising the condition number of the union of the merged gene sets (see Section 5.1 for details).

However, this approach, including other popular linear regression frameworks, such as EdgeR and Deseq2, tend to overlook gene-gene interactions. Instead, they perform statistical tests independently for each individual gene.

A.5 Conclusions and perspectives

To conclude that chapter illustrating an example of integrated and user-friendly RNA-Seq pipeline, I review some other freely available public initiatives which have been implemented.

For instance, the SARTools R pipeline, described in [Var+16] is designed for conducting seamlessly integrating differential analysis of RNA-Seq count data, using either DESeq2 [And+13] or EdgeR [RMS10] implementations. It is crafted to facilitate user-friendly multi-class comparisons of a single biological factor, and allows the inclusion of potentially confounding variables such as batch effects or sample pairing. One of the primary objectives of SARTools is indeed to provide access to the core functionalities of these analysis packages while preventing users, including those with limited experience, from misusing certain features. Additionally, the package offers a range of diagnostic plots for systematic quality control and hypothesis testing at critical stages

of the analysis workflow. Furthermore, SARTools automates the generation of a comprehensive HTML report that summarize all downstream analysis and quality control outputs. This report can be utilised as a centralized repository, documenting the entire analysis process, including parameter values and versions of R packages employed. Such an approach significantly enhances the reproducibility of experiments, promotes method standardization, and can lead to reductions in both time and cost.

[Cor+20] and [Con+16] provide bioinformaticians with comprehensive guidelines detailing the advantages and disadvantages of RNA-Seq methods and procedures, with the purpose of simplifying the selection of the most appropriate algorithms and pipelines with respect to the objectives and nature of the study.

Precisely, [Cor+20] tested the performance of 192 RNA-seq pipelines using different combinations of algorithms for trimming, alignment, counting, and normalisation, evaluated against two independent cell lines, and assessed the differential gene expression performance of 17 methods, validated against qRT-PCR. The respective precision, reproducibility and accuracy in quantifying raw gene expression and differential gene expression are reported in [Cor+20, Table 1], in which we observe that the counting algorithm HTSeq ([APH15]) coupled with the TMM [RMS10] normalisation and either the RUM [Gra+11], STAR [Dob+13] and TopHat2 [Kim+13] alignment algorithms exhibit the best accuracy and precision performance. They additionally conclude the absence of significant impact resulting from trimming algorithms. Regarding differential analyses frameworks, reported on [Cor+20, Fig. 7], they conclude that the most robust and reproducible method was `limma` with "trend" normalisation [Rit+15] followed respectively by `baySeq` [HK10], `limma` with "voom" normalisation [Law+14] and `edgeR` GLM (for generalised linear model, [RMS10]).

While [Cor+20] evaluate quantitatively the performance of several RNA-Seq pipelines, [Con+16] focuses on describing qualitatively the major steps involved in RNA-seq data analysis along with challenges associated, including experimental design, quality control, read alignment, quantification of gene and transcript levels, visualization, differential gene expression, and optional operations, depending on the objectives of the study, such as alternative splicing, functional analysis, gene fusion detection, and eQTL mapping. It notably addresses the interest, and the best guidelines to introduce control samples and a randomised sample processing, as well as advices to sequence error-free runs.

It also discusses the integration of RNA-seq with other functional genomics techniques, to connect gene expression regulation with molecular, physiological and functional annotations.

Although I did not delve into details about other technologies for analysing transcriptomic data, such as microarrays or q(for quantitative)PCR, the general principles, order of steps, and objectives outlined in Appendix A to analyse end-to-end this type of data hold.

In essence, all methods dealing with transcriptomic data must go through a series of pre-processing steps. These steps begin with an assessment of the overall quality of the raw data (as discussed in Appendix A.2). Then, gene counts aggregation and normalisation processes enable quantitative comparisons between samples, as described in Appendix A.3. Eventually, the Variance Stabilizing Transformation (VST) normalization method, which coerces datasets into following Gaussian distributions, simplifies the application of DGEA with the `limma` package, regardless of the sequencing platform (refer to Appendix A.4.1).

For a comprehensive comparison of the advantages and disadvantages of RNA-Seq versus microarray, please refer to Section 1.2.2. Additionally, for a detailed and didactic introduction to microarray processing, applied to Affymetrix microchips, please see the work by [KR18].

A.6 Appendix A: Gene notations

A.6.1 Gene terminologies

The need for a standardised convention to refer to human genes is essential, advocating the development of universal gene naming protocols. Practical examples, where the use of symbol aliases instead of their updated HGNC symbol, led to wasted clinical experiences and more worryingly to harmful and erroneous medical recommendations, are reviewed in [Bra+21]. Consider for instance the confusion involving the two unrelated genes currently approved as SF3B3 (splicing factor 3b subunit 3, plays a role in activating the immune system, notably in triggering inflammatory bowel condition of the Crohn's disease [Gon+20]) and SAP130 (Sin3A associated protein 130, a co-repressor protein associated with histone deacetylases, [FYA03]), that both encode proteins that are around the same molecular weight. It turns out that numerous papers use SAP130 as an alias to the HGNC approved name for SF3B3, starting from a 2013 publication, that referred to SAP130, a subunit of histone deacetylase, when they were actually studying SF3B3 [Suz+13] Antibody companies then wrongly attached this article to products for both the SAP130 and SF3B3 genes, which finally leads to misidentification of an identified biomarker in [Liu+20] study, in which they survey the activity of the SAP130 protein rather than the SF3B3-encoded protein.

We decided to use HGNC symbol as the primary key of our database, namely the unique identifier index of each row of our table. We chose HGNC convention since is the only internationally appointed authority for providing standardised guidelines for naming any gene, from protein-coding genes to pseudogenes [Bru+20]-[Yat+17]. In addition, the committee focused on ensuring consistency and clarity in research publications, by requiring that each entry matches a DNA segment with proved phenotype or function. Starting from 1979, more than 40 000 carefully curated human loci inputs can be retrieved from the HGNC database, half of them coding for proteins. To alleviate common confusion regarding Human Genome Organization (HUGO) and HUGO Gene Nomenclature Committee (HGNC) terms, note that the HGNC working group is one of the components of the larger initiative led by the overarching HUGO organisation. HUGO oversees various aspects of human genomics research that extends beyond gene nomenclature. Its general purpose is to provide a platform for any initiative related to human genomics, not only including gene naming, but also promoting genomics education, supporting genomic data sharing, and developing global coordination in genomics research [Lee+21].

It appears than only eight HGNC symbols (retrieved with key `SYMBOL`) matched ambiguously more than an unique ENTREZID symbol. Once removed, by keeping for each of them the most general annotation, we successfully generate a `1-1` mapping for both `ENTREZID` and chromosome name. On the contrary, one HGNC symbol is likely to match more than several old aliases (`ALIAS`) still commonly used by biologists, or more than one `ENSEMBL` symbol.

Ultimately, contrary to `ENTREZID` nomenclature, `HGNC` names provide insights on the biological function of the gene they referred to [Bru+20], Table 1, while `ENTREZID` entries are just ordered integer indexes. Delving into details, all the members of a gene family (genes grouped together based on homology, shared phenotype characteristic or membership of a protein complex) are usually designated by the same root symbol, followed by an Arabic numeral for unique identification (for example, KLF1, KLF2 and KLF3 are all antibodies). *Pseudogenes* (genetic sequence incapable of producing a functional protein product but homologous to a functional gene) are named after their ancestral parent gene, and *non-coding RNA genes* are prefixed according to their RNA type (miRNA, snRNA, snoRNA for instance, see [Bru+20], Table 2 and [WB11] for details). We should note however that a consensual nomenclature lacks for long non-coding RNAs (lncRNAs,

$>$ 200) nucleotides) and the HGNC can not be used for naming *isoforms* (alternate transcripts or splice variants, see Section 1.1.1).

Contrary to HGNC guidelines, ENSEMBL protocol, which names a gene after a specific localisation and mapping, does not prevent from naming mutated genes, such as an alternative-sequence variant or a by-product resulting from gene translocation or fusion. Precisely, ENSEMBL naming [Bir+04] is a partly unsupervised process, summarised in [Ake+16, Fig. 3] in which gene sequences are mapped onto genome assemblies and uniquely indexed through their genomic coordinates, which are not always manually curated, while the precise localisation in the genome might differ from an individual to another.

Hence, most of the many-to-one mappings between HGNC and ENTREZID can be explained mostly by the presence of *haplotypes* (regions of the genome which are known under two or more versions, possibly vastly differing across individuals) or duplicated genes, with distinct *loci* (genome localisations). For instance, `AGPAT1` is mapped to 9 distinct ENSEMBL sites, each corresponding to a specific haplotype, see forums Why am I getting different ensembl gene ids for a given gene symbol?, Alternative sequence gene Ensembl ID, How to deal with the case that one gene symbol matches multiple ensembl ids? and illustrative haplotype video for details.

However, the advantage of Ensembl database lies in the documentation of a cluster of related spliced transcripts (ENST...) with overlapping coding sequences[4].

To conclude that part, in order to minimise gene symbol confusion, the HGNC recommends to pair each HGNC symbol with its unique, most consensual ID, since they are less proned to nomenclature changes, being directly associated with the gene sequence. Another way to minimise confusion between approved and alias symbols is by displaying *curator notes* on the approved gene symbols that are used as well as alias for another gene or when multiple genes share the same alias [5]. Finally, the HGNC multi-symbol checker tool is another fast and easy way to check that you are using HGNC-approved gene symbols in your manuscripts.

A.6.2 Automated methods for gene annotation

Let's step in programmatic details: to build automatically our own, regularly updated `NCBI_gene` database, we first start by loading `org.Hs.eg.db` [Car22b], a genome annotation database for Human, together with the wrapper Bioconductor package `AnnotationDbi` [Pag+23], an user-friendly interface using SQLite-based conventions to query omics databases[6].

```
library(org.Hs.eg.db); library(AnnotationDbi)
symbol_humans <- AnnotationDbi::keys(org.Hs.eg.db, keytype="SYMBOL")
NCBI_gene <- AnnotationDbi::select(org.Hs.eg.db,
keys=symbol_humans,
                                   columns=c("SYMBOL", "GENENAME","ENTREZID"),
                                   keytype="SYMBOL")
# remove duplicated genes (when HGCN SYMBOLs match more than one ENTREZID)
```

[4] Indeed, transcripts that belong to the same gene may vastly differ in transcription start and end sites, as well as the sequence of exons, and can give rise to very different proteins, see biological introduction, Section 1.1.1 Post-transcriptional regulation.

[5] More than 450 approved gene symbols match aliases for different genes

[6] Alternatively, you may leverage [Rai17] database, another `AnnotationDbi` object, however, keep in mind that Ensembl databases are less curated and homogenised than ENTREZID or HUGO-based databases. A wrapper function to `AnnotationDbi::select`, `clusterProfiler::bitr()` is also available in `clusterProfiler` package [Yu22]–[Yu+12]

```
duplicated_genes <- unique(NCBI_gene$SYMBOL[duplicated(NCBI_gene$SYMBOL)])
NCBI_gene <- NCBI_gene %>% dplyr::filter(!SYMBOL %in% duplicated_genes)

# get chromosome localisation
NCBI_chrom_location <- AnnotationDbi::select(org.Hs.eg.db, keys=symbol_humans,
                                             columns=c("SYMBOL","MAP"),
                                             keytype="SYMBOL")

NCBI_gene <- NCBI_gene %>% dplyr::inner_join(NCBI_chrom_location, by="SYMBOL")
# get aliases for each symbol
NCBI_aliases <- AnnotationDbi::select(org.Hs.eg.db, keys=symbol_humans,
                                      columns=c("SYMBOL","ALIAS"),
                                      keytype="SYMBOL") %>%
  tidyr::chop(cols=c("ALIAS"))
NCBI_gene <- NCBI_gene %>% dplyr::inner_join(NCBI_aliases, by="SYMBOL")

#get ensembl for each symbol
NCBI_ensembl <- AnnotationDbi::select(org.Hs.eg.db, keys=symbol_humans,
                                      columns=c("SYMBOL","ENSEMBL"),
                                      keytype="SYMBOL") %>%
  tidyr::chop(cols=c("ENSEMBL"))

NCBI_gene <- NCBI_gene %>%
  dplyr::inner_join(NCBI_ensembl, by="SYMBOL") %>%
  dplyr::rename(hgnc_symbol = SYMBOL)
```

In a second time, we use `biomaRt` [Dur+05]–[DH22] package to access detailed gene annotations, such as the most likely biological function assigned to the transcript, or its precise localisation in the genome. To do so, we first download the database storing Homo sapiens annotation with function `biomaRt::useEnsembl`, then retrieve biological function and precise nucleotide/mapping locations in the human genome with `biomaRt::getBM` function and respectively keys `gene_biotype`, `start_position` and `end_position`[7]. Finally, we use the key `cdna` to retrieve all the known RNA transcripts of a given HGNC gene[8].

```
# download associated database
human_ensembl <- biomaRt::useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")

# get biological types
gene_biologic_functions <- biomaRt::getBM(
```

[7]Since there is no `1-1` mapping between biological function and HGNC symbol, we only keep the first entry in the list, except when coding for a gene. Similarly, there is no unique localisation of the start and the end of the gene sequence, thus, we compute instead the mean of all known positions. From these values, we can then easily deduce the `GENELENGTH` for each gene, a required entry for most RNASeq normalisation methods: GENESEQEND − GENESEQSTART + 1

[8]Since this operation might generate memory overflow due to the size taken by the complementary sequences, and since we are only interested in getting the number and the average size of transcripts, it is recommended to run this last operation in parallel, or sequentially process each transcript.

```
attributes= c("gene_biotype", "hgnc_symbol"),
filters=c("hgnc_symbol"),
values = NCBI_gene %>% dplyr::pull(hgnc_symbol),
mart=human_ensembl) %>% dplyr::distinct(hgnc_symbol, .keep_all=TRUE)

# get gene positions
gene_positions <- biomaRt::getBM(attributes= c("start_position", "end_position",
"hgnc_symbol"),
filters=c("hgnc_symbol"),
                                 values = NCBI_gene %>% dplyr::pull(hgnc_symbol),
                                 mart=human_ensembl) %>%
  dplyr::group_by(hgnc_symbol) %>%
  dplyr::summarise(GENESEQSTART = round(mean(end_position)),
                   GENESEQEND = round(mean(start_position)),
                   GENELENGTH = end_position - start_position + 1)

# transcript number and size
transcript_base <- biomaRt::getBM(
attributes= c("hgnc_symbol", "cdna"),
filters=c("hgnc_symbol"),
                                  values = NCBI_gene %>% dplyr::pull(hgnc_symbol),
                                  mart=human_ensembl) %>%
  dplyr::group_by(hgnc_symbol) %>%
  dplyr::summarise(TRANSCRIPTLENGTH = round(mean(nchar(cdna))),
TRANSCRIPTNUMBER = dplyr::n()))


# Populate NCBI database with additional feature informations
NCBI_gene <- NCBI_gene %>%
  dplyr::left_join(gene_biologic_functions, by = "hgnc_symbol") %>%
  dplyr::rename(GENEBIOTYPE = gene_biotype)
NCBI_gene <- NCBI_gene %>% dplyr::left_join(gene_positions, by = "hgnc_symbol")
NCBI_gene <- NCBI_gene %>% dplyr::left_join(transcript_base, by = "hgnc_symbol")
```

In summary, we use `AnnotationDbi` to match HGNC symbols with Ensembl ones, and then uses the `biomaRt` package to retrieve the biological function and the known transcripts. Additional gene feature annotations, such as pathway assignment (GO terms), can also be retrieved using `biomaRt` or `clusterProfiler`, see biomaRt vignette, however all these packages depends on Bioconductor annotation packages (datasets stored and documented as R objects), limiting the number of organisms available and regular updating (every day for NCBI databases, against every six months for Bioconductor package). Additionally, requiring download of the entire Ensembl or HUGO database each time you perform a query is cumbersome, hence several R packages alternatives have been developed to partly handle these limitations:

- *gProfileR* [RKA19] is a R API to the web server, unfortunately it only supports one-to-one conversion, preventing for instance to convert gene symbols to Ensembl.
- *UniprotR* [SM22], [Sou+20] is a R API to the *Uniprot* protein database, which supports

almost all species. However, being a protein database implies that it can only be used to fetch protein-coding genes.

- *genekitr* [Liu23] is a wrapper package that aims at combining both Ensembl and NCBI databases to provide comprehensive gene annotations, while optimising running speed, all these operations performed using `genekitr::transId` function. The features enabled by this package are closely related to the ones implemented in our own internal annotation function `bbcPreprocessing::from_probe_to_gene`, enabling for instance to set apart HGNC from `ALIAS` symbols, or to return simultaneously several gene nomenclatures. In addition to our own implementation, it allows through `genekitr::plotVenn` to quickly identify overlapping and unmatched gene sets between several databases.

- On the contrary, we strongly deter from using functions `limma::alias2Symbol`, a helper function from `limma` package, since it does not preserve the original order and does not explicitly detail discriminate input genes mapping several standard HGNC symbols from ones that could have been formally identified and `Seurat::GeneSymbolThesarus`, since its speed is slow and suffers from the same limits as `limma::alias2Symbol` function.

Appendix B

Appendix of Article 1

Supplementary Notes on Gaussian Mixture Models in R

Bastien Chassagnol* Antoine Bichat† Cheïma Boudjeniba‡ Pierre-Henri Wuillemin§
Mickaël Guedj¶ David Gohel‖ Gregory Nuel** Etienne Becht††

For the sake of readibility, we display in Table 1 the general configuration used to run all the benchmarks tested.

Table 1: Global options shared by all the benchmarked packages.

Initialisation methods	Algorithms	Criterion threshold	Maximal iterations	Number of observations
midrule hc, kmeans, small EM,rebmix, quantiles, random	EM R, Rmixmod, bgmm, mclust, flexmix, EMCluster, mixtools, GMKMCharlie	10^{-6}	1000	100, 200, 500, 1000, 2000, 5000, 10000

Furthermore, the code snippets, data, and figure subfolders required for replicating the figures documented in this supplementary material are readily accessible through the following public GitHub repository: GMM_appendix.

Appendix A: In-depth statistical elements about parameters estimation in GMMs

Application of the EM algorithm to GMMs

While solving Equation (10) to retrieve the MLE estimates in the M-step of the EM algorithm, we have to enforce the non-negativity and sum-to-one constraint of the mixture models (Equation (2)). This is enabled by the *Lagrange multipliers* tip, which consists in practice to add the equality constraint over the parameters to estimate, here $-\lambda(\sum_{j=1}^{k} p_j - 1)$, to the function to be optimised (Walsh 1975).

The evaluation of the roots of the derivative of the auxiliary function (see Equation (10)) at the parameter p_j with the additional unit simplex constraint (2) allows to readily compute a MLE estimate of the ratios, valid for any finite mixture model (Equation (1)):

$$\hat{p_j} = \frac{\sum_{i=1}^{n} \eta_i(j)}{n} \tag{1}$$

Additionally, we restrained in both the univariate and multivariate settings to the fully *unconstrained parametrisation*, in which each component follows its own parametric distribution. The general derivative of the auxiliary function with respect to each component parametric distribution ζ_j, is given by Equation (2)[1]:

*LPSM, Sorbonne Université, bastien\protect_chassagnol@laposte.net
†Les Laboratoires Servier, IRIS
‡Les Laboratoires Servier, IDRS
§LIP6, Sorbonne Université
¶Les Laboratoires Servier, IRIS
‖ArData
**LPSM, Sorbonne Université
††Les Laboratoires Servier, IRIS, etienne.becht@polytechnique.edu
[1]It is equivalent to compute the MLE of a sample following distribution f_{ζ_j} weighted by the vector of posterior probabilities.

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \zeta_j} = \sum_{i=1}^{n} \eta_i(j) \frac{\partial \log(f_{\zeta_j}(X_i|S_i = j))}{\partial \zeta_j} \tag{2}$$

Accordingly, if a closed form for the computation of the MLE in supervised cases is known (and fortunately this is the case for both the univariate and multivariate Gaussian distributions), the computation of the maximum of the auxiliary function can be readily calculated.

Plug-in the corresponding parametric distribution in the auxiliary function (10) yields the following formula for the univariate GMM (Equation (3)):

$$Q(\theta|\hat{\theta}_{q-1}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \eta_i(j) \left(\log(p_j) - \log(\sigma_j) - \frac{(X_i - \mu_j)^2}{2\sigma_j^2} \right) + K \tag{3}$$

and Equation (4) for the multivariate GMM:

$$Q(\theta|\hat{\theta}_{q-1}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \eta_i(j) \left[\log(p_j) - \frac{1}{2} \left(\log(\det(\boldsymbol{\Sigma}_j)) + (x_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (x_i - \boldsymbol{\mu}_j) \right) \right] + K \tag{4}$$

K is a constant with respective values of $\frac{-nD\log(2\pi)}{2}$ and $\frac{-n\log(2\pi)}{2}$ in the univariate and multivariate setting.

In the univariate setting, the individual MLE mean μ_j, and variance, σ_j, estimates are readily available (Equations (5) - (6)):

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \mu_j} = 0 \Leftrightarrow \mu_j = \frac{\sum_{i=1}^{n} \eta_i(j) X_i}{\sum_{i=1}^{n} \eta_i(j)} \tag{5}$$

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \sigma_j} = 0 \Leftrightarrow \sigma_j^2 = \frac{\sum_{i=1}^{n} \eta_i(j)(x_i - \mu_j)^2}{\sum_{i=1}^{n} \eta_i(j)} \tag{6}$$

Before finding the optimum of the auxiliary function in the multivariate setting, we remind the interested reader of some relevant calculus formulas below:

Transpose matrix properties

a. $\det(p\boldsymbol{A}) = p^G \det(\boldsymbol{A})$

b. $\det(A^{-1}) = \frac{1}{\det(A)}$

c. $\left(\boldsymbol{A}^{-1}\right)^\top = \boldsymbol{A}^{-1}$ [a]

[a] when $\boldsymbol{A}$ is itself symmetric, as by definition, $\boldsymbol{A}^\top = \boldsymbol{A}$

Matrix calculus

Given a symmetric matrix $\boldsymbol{A}$ of full rank D and two vectors $\boldsymbol{x}$ and $\boldsymbol{\mu}$ of size D, the following derivative properties hold:

a. $\frac{\partial \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}^\top$

b. $\frac{\partial (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{A} (\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -2\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{\mu})$

c. $\frac{\partial \log(\det(\boldsymbol{A}))}{\partial \boldsymbol{A}^{-1}} = -\boldsymbol{A}$ [a]

[a] Other matrix calculus formulas and notations are available on Matrix calculus and demonstration details from *The Matrix Cookbook* (Petersen and Pedersen 2008).

Using the calculus formulas derived in the previous boxes, a closed form for the MLE estimate of the mean, $\boldsymbol{\mu}_j$, and covariance, $\boldsymbol{\Sigma}_j$, is readily computed (see Equations (7) - (8)):

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \boldsymbol{\mu}_j} = \sum_{i=1}^{n} \eta_i(j) \boldsymbol{\Sigma}_j^{-1}(x_i - \boldsymbol{\mu}_j) = 0 \Leftrightarrow \mu_j = \frac{\sum_{i=1}^{n} \eta_i(j) \boldsymbol{x}_i}{\sum_{i=1}^{n} \eta_i(j)} \tag{7}$$

$$\frac{\partial Q(\theta|\hat{\theta}_{q-1})}{\partial \boldsymbol{\Sigma}_j^{-1}} = \frac{1}{2} \sum_{i=1}^{n} \eta_i(j) \left[\boldsymbol{\Sigma}_j - (x_i - \boldsymbol{\mu}_j)(x_i - \boldsymbol{\mu}_j)^\top \right] = 0 \Leftrightarrow \boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{n} \eta_i(j)(x_i - \boldsymbol{\mu}_j)(x_i - \boldsymbol{\mu}_j)^\top}{\sum_{i=1}^{n} \eta_i(j)} \tag{8}$$

Explicitly optimising the equations ((3)-(4)) yield the following MLE parameters in both the univariate and multivariate settings (Table 2), as detailed in (Leytham 1984; Redner and Walker 1984):

Table 2: An overview of the practical implementation of the EM algorithm in GMMs.

	Univariate GMM	**Multivariate GMM**
E-step	$\eta_i(j) = \frac{\hat{p}_j^q \mathcal{N}(x_i \vert \hat{\mu}_j^q, \hat{\sigma}_j^q)}{\sum_{j=1}^{k} \hat{p}_j^q \mathcal{N}(x_i \vert \hat{\mu}_j^q, \hat{\sigma}_j^q)}$	$\eta_i(j) = \frac{\hat{p}_j^q \mathcal{N}_D(\boldsymbol{x}_i \vert \hat{\boldsymbol{\mu}}_j^q, \hat{\boldsymbol{\Sigma}}_j^q)}{\sum_{j=1}^{k} \hat{p}_j^q \mathcal{N}_D(\boldsymbol{x}_i \vert \hat{\boldsymbol{\mu}}_j^q, \hat{\boldsymbol{\Sigma}}_j^q)}$
Ratios estimation	$\hat{p}_j^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j)}{n}$	$\hat{p}_j^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j)}{n}$
Mean estimation	$\hat{\mu}_j^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j) x_i}{\sum_{i=1}^{n} \eta_i(j)}$	$\hat{\boldsymbol{\mu}}_j^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j) \boldsymbol{x}_i}{\sum_{i=1}^{n} \eta_i(j)}$
(Co)Variance estimation	$\left(\hat{\sigma}_j^2\right)^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j) \left(x_i - \hat{\mu}_j^{q+1}\right)^2}{\sum_{i=1}^{n} \eta_i(j)}$	$\left(\hat{\boldsymbol{\Sigma}}_j^2\right)^{q+1} = \frac{\sum_{i=1}^{n} \eta_i(j) \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j^{q+1}\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j^{q+1}\right)^\top}{\sum_{i=1}^{n} \eta_i(j)}$

In both cases, obtaining the parameters of each component's parametric distribution turn to be equivalent to the computation of the mean and variance of a weighted sample, which can be computed in R with `stats::weighted.mean` and `stats::cov.wt` functions[2]. Importantly, the value of the mapping function only depends on the set of the observations X, but does not depend on the parameter to estimate θ. Indeed, the statistic computed by the EM algorithm is sufficient, which is one of its main advantages.

The complete code associated to our R implementation is implemented respectively with `enmix_univariate` and `enmix_bivariate` for the univariate and multivariate setting, available on GitHub at RGMMBench, as well as the programs used to generate the several plots and tables of the article. We additionally made two choices not clearly set in the literature:

- The algorithm stops when when the absolute difference between consecutive log-likelihoods falls below a user-defined threshold *epsilon*, with a maximal number of *itmax* iterations allowed to reach this convergence.
- In order to avoid numerical underflows resulting in inconsistent ratios, of type 0/0, we rely on the fact that Gaussian distributions belong to the exponential family to log-rescale our observations and compute efficiently the posterior probabilities in the E-step of the EM algorithm. First, to avoid null values for highly unlikely observations, those far from the centroids, we use the log attribute of `stats::dnorm` and `mvtnorm::dmvnorm` functions, see Equation (9):

$$\begin{aligned} \ell(\theta|x) &= \log(\sum_{j=1}^{k} p_j f\zeta_j(x)) \\ &= \log\left(\exp\left[\log(p_j) + \log(f\zeta_j(x))\right]\right) \end{aligned} \tag{9}$$

Second, we rewrite our sum of exponentials, the one enclosed into the log, to use the Taylor' series of $\log(1+x)$, with $|x| \ll 1$, see Equation (10):

[2]We assign "ML" to the argument *method* to get the biased but true MLE estimate of the covariance

$$\begin{aligned}\log\left(\sum_{j=1}^{k} e^{a_j}\right) &= \log\left(\exp(a'_j) \times \left[1 + \sum_{j \neq j'} \exp\left(\frac{a_j}{a_{j'}}\right)\right]\right) \\ &= a_{j'} + \text{log1p}\left(\sum_{j \neq j'} \exp\left(\frac{a_j}{a_{j'}}\right)\right), \text{ with } j' = \arg\max_{\forall j \in \{1,\dots,k\}} (e^{a_j})\end{aligned} \tag{10}$$

with `log1p` the R function dedicated for this Taylor's development. The posterior probabilities are then given by Equation (11):

$$\log\left(\mathbb{P}_\theta(S = j|X = x)\right) = \log(p_j) + \log(f_{\zeta_j}) - \ell(\theta|x) \tag{11}$$

- We stop the algorithm early when the estimates are trapped in the boundaries of the parameter space, typically when the ratio of a component or its associated variance tends to zero. This case rarely occurs in our simulations: once in univariate and never in multivariate.

Parsimonious parametrisation of multivariate GMMs

Parsimonious parametrisation of GMMs models are provided by the following *eigenvalue* factorisation of the covariance matrix (Equation (12)):

$$\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{Q}_j \boldsymbol{D}_j \boldsymbol{Q}_j^\top \tag{12}$$

with $\lambda_j = \det(\boldsymbol{\Sigma}_j)^{\frac{1}{D}}$ a scalar proportional to the total volume of the ellipsoid (or area in bi-dimensional setting), $\boldsymbol{D}_j$ a diagonal matrix storing the eigenvalues normalised such that $|\boldsymbol{D}_j| = 1$[3] and $\boldsymbol{Q}_j$ a $\mathcal{M}_D(\mathbb{R})$ *orthogonal matrix* whose columns are D linearly independent *eigenvectors* generating an orthonormal basis in $\mathbb{R}^D$ while $\boldsymbol{Q}_j^\top$ is its corresponding transpose matrix. The existence of the decomposition is guaranteed by the positive definiteness constraint over the covariance matrix while the orthogonality of $\boldsymbol{Q}_j$ results from its symmetry. When the matrix to factorise is positive-definite and symmetric, we also refer to it as *spectral decomposition*, a special case of *eigendecomposition*.

Each of these matrices can be constrained to be equal or variable across clusters, hence this decomposition reveals 14 possible models with different geometric characteristics, namely:

- two models with the *spherical family*, for which only λ_j is used to control the *isotropic* (same radius in any dimension) volume of each component of the corresponding distribution structure
- four models with the *diagonal family*, using λ_j with possibly distinct diagonal elements and $\boldsymbol{D}_j$ to specify the shape of the density contours. In that context, $\boldsymbol{Q}_j$ is henceforth a permutation matrix, whose inputs are only zeros and an unique one per row.
- eight models with the *general family*, using additionally $\boldsymbol{Q}_j$ to determine the orientation of the main axes of the ellipsoids. Indeed, in the last two families described, this matrix was equal to the identity, hence the axis of the ellipsoids were aligned with the standard $\mathbb{R}^D$ basis.

We detail the main characteristics of the 14 parametrisations (28 if we add for each model the equiproportional hypothesis) in Table (3):

- The first column describes in general and understandable terms each parametrisation, with I meaning invariant (alternatively, not used in the parametrisation), E means equal and V variable while the second column matches the corresponding matrix decomposition of the covariance matrix. These 14 models are all

[3] Langrognet et al. (2021) enforces an additional but, in our opinion, superfluous constraint that the eigen values are sorted by decreasing order

Table 3: The 14 canonical parametrisations of the within-group covariance matrix $\boldsymbol{\Sigma_j}$ with the corresponding geometric representations.

Model	**Notation**	**Family**	**M-step**	**Number of parameters**	**Representation**
EII	$[\lambda I]$	Spherical	CF	$\alpha + 1$	
VII	$[\lambda_j I]$	Spherical	CF	$\alpha + k$	
EEI	$[\lambda D]$	Diagonal	CF	$\alpha + d$	
VEI	$[\lambda_j D]$	Diagonal	IP	$\alpha + d + k - 1$	
EVI	$[\lambda D_j]$	Diagonal	CF	$\alpha + kd - k + 1$	
VVI	$[\lambda_j D_j]$	Diagonal	CF	$\alpha + kd$	
EEE	$[\lambda Q D Q^\top]$	Ellipsoidal	CF	$\alpha + \beta$	
EVE	$[\lambda Q D_j Q^\top]$	Ellipsoidal	IP	$\alpha + \beta$	
VEE	$[\lambda_j Q D Q^\top]$	Ellipsoidal	IP	$\alpha + \beta + (k-1)(d-1)$	
VVE	$[\lambda_j Q D_j Q^\top]$	Ellipsoidal	IP	$\alpha + \beta + d(k-1)$	
EEV	$[\lambda Q_j D Q_j^\top]$	Ellipsoidal	CF	$\alpha + k\beta - d(k-1)$	
VEV	$[\lambda_j Q_j D Q_j^\top]$	Ellipsoidal	IP	$\alpha + k\beta - (k-1)(d-1)$	
EVV	$[\lambda Q_j D_j Q_j^\top]$	Ellipsoidal	CF	$\alpha + k\beta - k + 1$	
VVV	$[\lambda_j Q_j D_j Q_j^\top]$	Ellipsoidal	CF	$\alpha + k\beta$	

included in one of the three super-families: spherical, diagonal and ellipsoidal listed before. As an example, the model VEI has variable volumes λ_j in relation with the cluster, however shares same general shape (as we can note on the Representations, all isodensities are distributed along the x-axis) and invariant directions (in other words, the transition matrix is the identity matrix, entailing that all scatter plots are aligned with the Cartesian coordinate axes).

- Varying the volume λ_j, given a fixed $\boldsymbol{Q}$ and $\boldsymbol{D}$, amounts to an *enlargement* (when all dimensions of a figure are changed in the same scale, also referred to as *isotropic* transformation), varying the eigenvectors $\boldsymbol{Q}_j$, given a fixed volume λ_j and $\boldsymbol{D}$ is equivalent to a rotation and finally varying the diagonal matrix $\boldsymbol{D}_j$, given the other parameters of Equation (12) are fixed, results in a *distortion* of the representation.
- CF means that the M-step is in closed form while IP entails that the M-step is iterative.
- The number of parameters enumerates the *degrees of freedom*, namely the number of parameters to truly estimate once the sum-to-one constraint is enforced (Equation (2)). In detail, k is the number of components of the GMM model, D its dimension, $\alpha = kD + k - 1$ is the number of parameters required to identify the mean vector of each component (kD) and the ratios $k - 1$ and $\beta = \frac{D(D+1)}{2}$ the number of covariance terms to estimate for a given component (D variance diagonal terms, the remaining terms being the pairwise symmetric covariance terms between the features). Note that the complexity of the covariance matrix in the fully unconstrained model (Model VVV) grows linearly with the number of components while exploding in the order $\mathcal{O}(D)$ with the number of dimensions. Meantime, the complexity of the parametrisation with the homoscedastic spherical family (Model EII) is constant.
- Last column displays the 14 most common GMMs parametrisations, by plotting the ellipses and centroids of a three components bivariate GMM parametrised by the mean vector and covariance of each component. For any additional detail, we refer the interested reader to **mclust** (Scrucca et al. 2016) and **Rmixmod** (Langrognet et al. 2021) vignettes for a general introduction to GMMs and to (Banfield and Raftery 1993; Celeux and Govaert 1992; Browne and McNicholas 2014) for the closed formulas of the models.

Parameters estimation in a high-dimensional context

However, while parsimonious representations can largely reduce the computational burden, none of them in the general family is able to handle degenerate cases where the number of features, D, exceeds the number of observations n. Likewise situations, when the number of features is consequent, are referred to as high-dimensional, raising the well-known issue of the "curse of the dimensionality". Two distinct approaches have been developed in the literature to handle these degenerate cases:

- The most naive approach aims to eliminate the least informative variables by applying a strong Lasso-type penalty on the parameters to be estimated. We only came across such an approach twice among the reviewed R packages, in the specific context of regressions of mixtures (see**RobMixReg** and **fmerPack** packages).
- The second category includes a larger diversity of methods, all inspired from the *factor analysis* approach whose paradigm is to consider that all the D features used to describe the observations can be spanned in a smaller subspace without lose of information. Precisely, the factor analysis theory describes the variability among observed and correlated variables by a substantial lower number of unobserved variables called *factors* or *latent variables*. In practice, for a given component j, the diagonal matrix storing the eigenvalues is decomposed into two-blocks. The first upper-right diagonal block, assumed generally of dimension $d_j \ll D$, stores the largest d_j eigenvalues and model the variance of the actual data of component j while the lower-left diagonal block, of dimension $D - d_j$, stores an unique parameter that can be interpreted as the variance of the residual error terms, constrained to be strictly inferior to the lowest variability of the informative variables. The dimension d_j can be considered as the intrinsic dimension of the latent subspace of cluster j spanned by the first d_j eigenvectors of $\boldsymbol{Q}_j$[4].

[4]Starting from eigen-decomposition described in (Equation (12)), this approach is equivalent to consider only the d_j largest eigenvalues resulting from the decomposition and sets the others to null.

When the sub dimension d_j is known, a closed version is generally available for the M-step of the EM algorithm, however d_j is itself an hyperparameter to estimate. Though, (Bouveyron, Celeux, and Girard 2011) has shown that a classical Cattell's scree-test could be used to asymptotically estimate the intrinsic dimension of each cluster. Compared to the previous approach, this method has a strong theoretical background and strong impact on the running times performance.

Taking a concrete use case from the help documentation of the package **HDclassif**, it enabled to cluster a dataset of 10 classes with 130 observations overall and described in a 1024-dimensional space (consider the famous machine-learning digit recognition problem). Variants of these approaches have been developed in the following packages: **HDclassif**, **fabMix**, **EMMIXmfa** and **pgmm**. We refer the interested reader to the educational vignette of **HDclassif**: HDclassif and papers (Paul David McNicholas and Murphy 2008; P. D. McNicholas et al. 2010; Paul D. McNicholas and Murphy 2010).

Historically, the first mention of a probabilistic framework with an application to dimension reduction in the context of finite mixture models goes back to Tipping and Bishop (1999), based on principal component analysis. G. J. McLachlan, Peel, and Bean (2003) and Mclachlan and Peel (2000) extend this original model by postulating that the distribution of the data within any latent class could be described using the tools of the factor analysis field[5] Finally, building on the parsimonious parametrisations already theorised for GMMs (see previous section) , Paul David McNicholas and Murphy (2008), P. D. McNicholas et al. (2010) and Bouveyron, Girard, and SCHMID (2007) proposed a variety of constraints, but this time directly defined on the projected subspace. Since all methods based on factor analysis provide a transition matrix, using the two or three most informative eigen values and their associated eigen vectors in order to project the dataset on a smaller subspace provides a simple visualisation tool for representing high dimensional datasets. However, this method may is not suitable for unravelling the clustering structure. Instead, *the GMMDR method*, first proposed by Scrucca (2010) and implemented in the `MclustDR` function, from **mclust** package, aims at recovering the subspace that best captures the underlying latent clustering structure (we notably expect invariance of the global overlap in the sampling space and the corresponding projected subspace). More precisely, the main objective of the GMMDR technique is to infer the global *change-of-basis matrix* $\boldsymbol{Q}$ that minimises the differences in the a posteriori probabilities of assigning each observation i to a given cluster s_i, knowing the value of the vector of observed covariates $\boldsymbol{x}_i$. Namely, we are looking for the orientation matrix $\boldsymbol{Q}$ that maximally ensures the following objective (Eq. (13)):

$$\hat{\boldsymbol{Q}} = \arg\max_{\boldsymbol{Q}} \left(\mathbb{P}_\theta(S_i = j|\boldsymbol{X} = \boldsymbol{x}_i) = \mathbb{P}_\theta(S_i = j|\boldsymbol{X}\boldsymbol{Q})\right) \text{ such that } S \perp \boldsymbol{X}|\boldsymbol{X}\boldsymbol{Q} \tag{13}$$

This procedure itself derives from the *sliced inverse regression* algorithm (K.-C. Li 1991), but instead of conditioning on the known response variable, GMMDR conditions on the estimated MAP cluster assignments. Since the solution returned by the following optimization problem is not unique, we generally constrain the projection matrix to be orthonormal (any of the vectors forming the basis are pairwise orthogonal, and individually of norm 1).

Model selection

When comparing several models with several number of components or parametrisations, the likelihood is uninformative as it can be arbitrarily minimised by increasing the complexity of the model or adding components. it is then necessary to penalise for complexity when comparing them. The general form of the penalty metric, *GIC* (for generalised information criteria), is given by Equation (14):

$$\text{GIC}(\theta) = \underbrace{p(\theta)}_{\text{penalty term}} - \underbrace{2\ell(X|\theta)}_{\text{log-likelihood of the model}} \tag{14}$$

[5]Although principal component analysis and factor analysis are closely related, we can differentiate both approaches by their differing objective: while PCA seeks to capture the overall variability of the dataset, factor analysis focuses on describing the intra-variability between covariates. In practice, the differences between the two approaches are minor, we can notably show that the output of PCA is one of the solutions suggested by standard factor analysis.

Among them, we set apart scores focused on selecting selecting the right number of parameters and components, namely the *degrees of freedom* (d.o.f.) of the model ($3k-1$ parameters for the univariate unconstrained GMM), and those focusing on retrieving readable clusters.

In the first category, the *AIC* (Akaike information criterion) (Schwarz 1978) is a *minimax-rate optimal* (score that minimises the risk in the worst case) but inconsistent metric (Yang 2005) , proned to overestimate the true number of components. *BIC* (Bayesian Information Criterion), and CAIC (consistent AIC), accounting for both the number of parameters and the sample size, are consistent metrics. Finally, the *MDL*(Minimum Description Length) criterion accounts for the number of parameters, sample size and number of components. Its core objective differs from the others as it aims at reducing the amount of code to encode both parameters and observations but is practically close to the BIC metric. A thorough description of these scores, with their formulas and theoretical properties, can be found in Fonseca (2008), Celeux, Fruewirth-Schnatter, and Robert (2018).

In the second category, the most commonly implemented is the *ICL (integrated complete-data likelihood)*, a BIC criterion with an additional entropy penalty (G. McLachlan and Peel 2000). As opposed to *BIC*, the entropy term reduces the number of components to a well-separated and readable clustering. Hence, it tends to underestimate their true number when components are overlapping. Alternative similar metrics are the *CLC* (Classification Likelihood Criterion), *AWE* (Approximate Weight of Evidence) and *NEC* (Normalised Entropy Criterion) metrics (Bacci, Pandolfi, and Pennoni 2012). The several metrics implemented by the reviewed packages are listed in Table 2.

The *Likelihood-ratio test* (LRTS) can also be used to compare *nested models*, with additional advantage to possibly derive a p-value yielding the probability that a complex model (with more components) should preferentially be used over a simpler one. Traditionally, common process is to add one component after the other, until hypothesis H_0 can not be rejected anymore. Under standard regularity conditions of Cramer's theorem, Wilk's theorem states that the Likelihood Ratio distribution follows asymptotically a χ^2 distribution, but unfortunately these conditions are not met in mixture models (G. McLachlan and Peel 2000). To counterbalance it, bootstrap inference (G. McLachlan and Peel 2000) is often used to derive an empirical distribution of the Likelihood Ratio.

Derivation of confidence intervals in GMMs

Punctual estimation, with a single estimate $\hat{\theta}$ for a given n-sample, is not enough to evaluate the performance of a specific method, as drawing another n-sample using the same parameters is likely to lead to a different distribution and estimation of $\hat{\theta}$. Instead, it can be interesting to retrieve the distribution or at least the variability of the estimated parameters, which can reveal useful to derive confidence intervals. However, obtaining the distribution or even an asymptotic approximation of the distribution of the parameters is not feasible in practice with mixture models (G. McLachlan and Peel 2000). Hence, most authors recommend to use bootstrap methods for the generation of confidence intervals, as suggested in (Efron and Tibshirani 1993; Basford et al. 1997).

Bootstrap distributions of the parameters are generally retrieved via *empirical* or *parametric* bootstrap, both available in the **mclust** package. In the *empirical* or *non-parametric* bootstrap Jaki et al. (2018), we draw iteratively N samples of size n with replacement from the original observed variable $x_{1:n}$. In the *parametric* bootstrap, N simulations are built from the parameter estimated with the available observations of X, via the EM algorithm or any method used for parameter estimation. In both cases, we obtain an empiric distribution of the parameter estimate: $\hat{\theta}_{1:N} = (\hat{\theta}^1, \dots, \hat{\theta}^N)$. Sample mean and standard deviation (SD) of this empirical distribution can be used to retrieve an asymptotic estimate of the variability of the parameter estimate $\hat{\theta}$, the bias or the MSE of the parameter estimates. To get unbiased estimates of the true standard deviation and mean of the estimates, it is of common practice to compute the empirical covariance matrix of the sample $\text{cov}[\hat{\theta}] = \frac{\sum_{j=1}^{N}(\hat{\theta_j}-\mathbb{E}[\hat{\theta}])(\hat{\theta_j}-\mathbb{E}[\hat{\theta}])^{\mathrm{T}}}{N-1}$, the square roots of its diagonal terms corresponding to the empiric SDs. Symmetric $1-\alpha$ asymptotic confidence intervals using the Central Limit Theorem (CLT) can then be simply derived Equation (15):

$$\mathbb{E}[\hat{\theta}_t] \pm \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \sqrt{\text{var}(\hat{\theta}_t}, \quad \forall t \in \{1, \dots, 3k\} \tag{15}$$

with $z_{1-\frac{\alpha}{2}}$ the $1-\frac{\alpha}{2}$ quantile of the standard Gaussian distribution.

If computing the covariance matrix is not possible analytically, it can be approximated by the expected Fisher Information Matrix $\mathcal{I}_{\text{exp}}(\theta)$ (FIM), given by Equation (16):

$$[\mathcal{I}_{\text{exp}}(\theta)]_{1 \leq i \leq 3k, 1 \leq j \leq 3k} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\theta|X)\right] \tag{16}$$

Indeed, the Cramér-Rao theorem states that the diagonal elements of the inverse of the FIM are upper bounded by the variability of the parameters: $\text{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}$. This implies that the ratio between inverse of the FIM and the variance $e(\hat{\theta}) = \frac{\mathcal{I}(\hat{\theta})^{-1}}{\text{var}(\hat{\theta})}$ converges to 1, using the asymptotic efficiency of the MLE estimate of GMMs.

Unfortunately, the computation of the expected FIM is still a hard task. Hence it is generally replaced by the observed FIM, the negative of the Hessian matrix of the incomplete log-likelihood function: $\mathcal{I}_{\text{obs}}(\theta) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\theta|X)$. Exact general formulas are provided for the univariate case in Louis (1982) and for the multivariate case in Oakes (1999). Yet, it has to be noted that the expected FIM generally outperforms the observed FIM in estimating the covariance matrix of the MLE (X. Cao and Spall 2012).

However all these methods require to compute second derivatives of the log-likelihood leading to some disadvantages from a computational point of view. More recently, L. Meng (2016) and Delattre and Kuhn (2019) proposed an accelerated algorithm requiring only computation of first order derivatives. A similar alternative is implemented in the **mixsmsn** package (Prates, Lachos, and Cabral 2021): `mixsmsn::im.smsn`, in which the Hessian matrix is approximated by the cross-product of the gradient of the log-likelihood Equation (17):

$$\mathcal{I}_{\text{obs}}(\theta) \approx -\frac{\partial \log(\ell(\theta|X))}{\partial\theta} \frac{\partial \log(\ell(\theta|X))}{\partial\theta}^T \tag{17}$$

according to an idea developed in paper Basford et al. (1997). For a more general introduction to Gaussian mixtures, including other models and parametrisations in the multivariate case, we refer the reader to the reference book *Gaussian parsimonious clustering models* Celeux and Govaert (1992).

An analytic formula of the overlap for univariate Gaussian mixtures

From an analytic point of view, the overlap between k components of variable X is given by Equation (18):

$$\text{OVL}(X) = 1 - \int_{\mathbb{R}} \max_j (p_j \varphi_{\zeta_j}(x)) dx \tag{18}$$

The 1 in Equation (18) corresponds to the integration of probability $f_\theta(X)$ distribution over its domain. The second part is the area under the curve of the component density function maximised on $\mathbb{R}$, with j the index of the component maximised at that point. It should be noted that the definition used here for the overlap is closely related to the definition of the *false clustering rate* (FCR) (Marandon et al. 2022).

Equation (18) simplifies for a two component mixture distribution to Equation (19):

$$\text{OVL}(X) = \int_{\mathbb{R}} \min\left(p_1 \varphi_{\zeta_1}(x), p_2 \varphi_{\zeta_2} dx(x)\right) \tag{19}$$

From a probabilistic point of view, we can rewrite Equation (19) as the overall probability of assigning a wrong label to a given observation. With two components, this simply decomposes as the sum of the probability of

mistakenly assigning an observation from component 2 to component 1 and the probability of assigning an observation from component 1 to component 2 Equation (20):

$$\begin{aligned}\text{OVL}(1,2) &= \text{OVL}(1|2) + \text{OVL}(2|1) \\ &= \mathbb{P}\left(p_1\varphi(X,\mu_1,\sigma_1) \leq p_2\varphi(X,\mu_2,\sigma_2)\right) + \mathbb{P}\left(p_2\varphi(X,\mu_2,\sigma_2) \leq p_1\varphi(X,\mu_1,\sigma_1)\right) \\ &= \int_{\mathbb{R}} p_1\varphi_{\zeta_1}(x)1_{p_1\varphi_{\zeta_1} \leq p_2\varphi_{\zeta_2}}dx + \int_{\mathbb{R}} p_2\varphi_{\zeta_2}(x)1_{p_2\varphi_{\zeta_2} \leq p_1\varphi_{\zeta_1}}dx\end{aligned} \tag{20}$$

We illustrate the computation of the overlap in some hard-hitting cases below, showing relation between the level of entropy and the individual standard deviations with the overlap measured in Figure 1. Means of component 1 and 2 are 5.28 and 8.45. Panels A and C correspond to balanced classes, while in panel B and D, class 1 is more abundant with a frequency of 0.9. Finally, in panels A and B, the variance of component 1 is smaller than the variance of component 2 with respective SDs of 1 and 3 and reciprocally for panels B and D. Interestingly, in panel D, using the MAP as defined in Equation (21), all observations issued from class 2 are wrongly assigned to class 1.

$$\eta_i(j) := \mathbb{P}_\theta(S_i = j|X_i = x_i) \tag{21}$$

The red area corresponds to the probability of misclassifying component 1 as component 2, while the green area corresponds to the probability of misclassifying component 2 as component 1. Total overlap is since the sum of red and green area, in Figure 1.

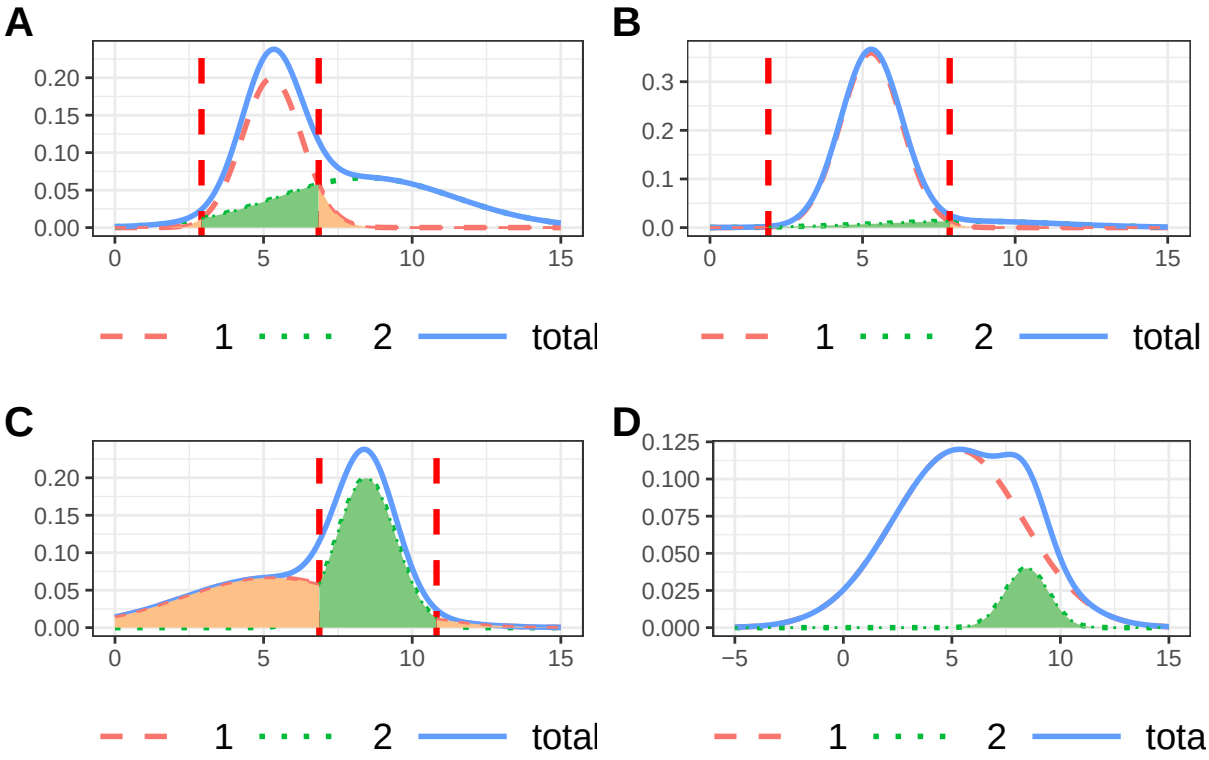


Figure 1: Illustration of the overlaps between a two-components GMM. Density function of component 1 is given by the red line, its of component 2 by the green line, and total density function $f_\theta(X)$ is represented in blue. The total overlap is given by the sum of the green and red areas.

There are two intersection points, x_1 and x_2 , with $\mu_1 < \mu_2$ when solving equation Equation (22):

$$p_1\varphi(x,\mu_1,\sigma_1) = p_2\varphi(x,\mu_2,\sigma_2) \tag{22}$$

in following case: if $\sigma_2 > \sigma_1$, then we must have $p_1 > \frac{\sigma_1}{\sigma_1+\sigma_2}$, else if $\sigma_2 < \sigma_1$, then $p_1 < \frac{\sigma_1}{\sigma_1+\sigma_2}$. In that case, they are given by following formula Equation (23):

$$(x_1, x_2) = \left(\frac{\sigma_1^2\mu_2 - \sigma_2^2\mu_1 \pm \sigma_1\sigma_2\sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_2^2-\sigma_1^2)\left[\log(\frac{p1}{p2}) + \log(\frac{\sigma_2}{\sigma_1})\right]}}{\sigma_1^2 - \sigma_2^2} \right) \tag{23}$$

Again, sign of term A and order of the roots yield two several cases, depending whether σ_1 is greater or not than σ_2. Both situations with unbalanced classes are illustrated in panel B and D on Figure 1:

- When $\sigma_1 < \sigma_2$, then $x_2 < x_1$ and $p_1\varphi(x, \mu_1, \sigma_1) < p_2\varphi(x, \mu_2, \sigma_2)$ on interval $[x_2, x_1]$. Hence, total overlap is given by Equation (24):

$$\text{OVL}(1,2) = p_1\left(\Phi(x_2, \mu_1, \sigma_1) + 1 - \Phi(x_1, \mu_1, \sigma_1)\right) + p_2\left(\Phi(x_1, \mu_2, \sigma_2) - \Phi(x_2, \mu_2, \sigma_2)\right) \tag{24}$$

- When $\sigma_1 > \sigma_2$, then $x_1 < x_2$ and $p_1\varphi(x, \mu_1, \sigma_1) < p_2\varphi(x, \mu_2, \sigma_2)$ on interval $[x_1, x_2]$. Hence, total overlap is given by Equation (25):

$$\text{OVL}(1,2) = p_2\left(\Phi(x_1, \mu_2, \sigma_2) + 1 - \Phi(x_2, \mu_2, \sigma_2)\right) + p_1\left(\Phi(x_2, \mu_1, \sigma_1) - \Phi(x_1, \mu_1, \sigma_1)\right) \tag{25}$$

An interesting result is obtained with the homoscedascity and balanced classes' assumptions of the k-means algorithm. There is only one intersection point in that case: $x_c = \frac{\mu_1+\mu_2}{2}$, that is simply the centre of the segment bounded by the means of the two components. The overlap is simply then $\text{OVL}(1,2) = 2\Phi(-\frac{|\mu_1-\mu_2|}{2\sigma})$.

To our knowledge, no closed formula has been determined returning the overlap generalised to more than two components (combinatorial set of inequations to solve), in the unconstrained multivariate setting (cubic equation to solve in bi-dimensional space). Indeed, even restraining the study to the bivariate setting (the calculation of the OVL then amounts to estimating the zone of intersection between two ellipses), the exact computation of the OVL involves multiple integration and the algebraic resolution of a quartic equation. A first step is provided by (Alberich-Carramiñana, Elizalde, and Thomas 2017), stating algebraic conditions for the existence of an intersection region and computing where applicable a closed formula of the OVL between two coplanar ellipses.

Accordingly, only stochastic approximations, relying on randomised algorithms, such as the Monte-Carlo integration with a rejection technique (knowing that the total area under the curve is normalised to one, we randomly simulate observations and the ratio of the number of observations falling in the intersection area is then used as a proxy of the overlap), are available so far (Maitra and Melnykov 2010; Pastore and Calcagnì 2019; Nowakowska, Koronacki, and Lipovetsky 2014).

Appendix B: Extensions of the EM algorithm to overcome its limitations

Two main alternatives were developed in parallel to the EM algorithm and are implemented in some of the reviewed packages: the CEM and the SEM algorithm. However, they do not have its theoretical properties, especially guarantee of the consistency of the algorithm.

The M-step of the *classification EM* (CEM) algorithm (Biernacki, Celeux, and Govaert 2000) maximises a function where each observation was assigned to the maximum a posteriori (MAP) estimate Equation (21). It generalises the well-known k-means algorithm making no assumption of homoscedascity or equibalanced clusters. Its main drawback is to not take into account uncertainty of the cluster assignment, inducing *inconsistency* of the algorithm (G. McLachlan and Peel 2000). EM*, referred in Kurban, Jenne, and Dalkilic (2017) and

implemented in the **DCEM** package, is a faster implementation of the CEM algorithm, with roughly a twice smaller complexity. To do so, only the posterior distributions associated to the lower half of the most uncertainly assigned observations are re-computed in the E-step of the EM-algorithm. This normally avoids to recompute data that is unlikely to change of cluster attribution from an iteration to another. However, the higher speed of this algorithm has not been theoretically proven, as the gain of running time per iteration of the algorithm may be alleviated by a greater number of steps to reach the convergence.

The *Stochastic EM* (SEM) replaces the MAP value for S in the E-step of the CEM algorithm by a random draw (or N of them in the N- variant of the algorithm) of the posterior distribution $\mathbb{P}_\theta(S|X)$. As this algorithm does not converge to a unique solution, but rather oscillates around a local maximum, the estimation is usually performed by averaging the late estimated values while ignoring the first estimates from the *burn-in phase*. A theoretical description of these algorithms, with discussion on their convergence properties, is detailed in Celeux and Govaert (1992). SEM algorithm has also a relatively faster convergence than EM algorithm but it is more proned to be trapped in a local maximum or to remove a component. Increasing the number of draws N may alleviate this issue, but at the extent of computational performances.

A wide variety of fast algorithms derived from the EM algorithm have been developed. cwEMM (component-wise EM algorithm), described in Celeux, Chrétien, and Forbes (2012), is a variation of the EM algorithm aiming at speeding up its convergence. The M-step at each iteration is only performed for one of the components $\theta_j = (p_j, \mu_j, \sigma_j)$, implying that the parameters of a given component are estimated sequentially rather than simultaneously. The theory behind relies on a *Gauss-Seidel* scheme and was first used by the SAGE algorithm. However, the constraints on the proportions set in Equation (2) are only guaranteed if the algorithm converges. Additionally, faster convergence is not theoretically proven for any situation. A list of general acceleration methods for the EM algorithm, not specific to GMMs, is available on **turboEM** (Bobb and Varadhan 2021).

Other EM-inspired algorithms focus on counterbalancing the main limitations of the EM algorithm. The *Variational Bayesian EM* (VBEM) algorithm performs a Bayesian estimation of the parameters. Indeed, the large space of all possible parameter estimates Θ can be hard to explore and the usual initialisation methods are uninformative, not taking into account expert recommendations. VBEM uses these prior assumptions on the parameters' distribution $\mathbb{P}(\theta)$ to optimise the posterior distribution $\mathbb{P}(\theta|X)$, based on Bayes' rule. Direct determination of the Bayesian posterior law of the parameters is generally an intractable problem, hence Variational Bayes only maximises an approximation of the true posterior, assuming that the parameters can be partitioned in independent distributions. This hypothesis is known as *mean-field approximation* (Murphy 2012).

The minimum message length (MML) EM algorithm, implemented in the **GMKMcharlie** package, is a completely unsupervised algorithm as it does not require any prior selection of the number of components (Figueiredo and Jain 2002) , by dealing explicitly with the possibility of discarding a component during the iteration. To do so, the selection criteria for the number of components is directly included in the optimisation procedure. However, its implementation is close from a Bayesian estimation of the parameters of the model, setting a non-informative Dirichlet prior distribution on the ratios and the higher expected performances of the algorithm are not demonstrated on real use cases (Figueiredo and Jain 2002).

The Expectation/Conditional Maximisation (ECM) (MENG and Rubin 1993) belongs to the super-family of GEM (general EM) algorithms, generally used when the maximisation of the auxiliary function yields a non-closed form to solve. To do so, the ECM algorithm replaces the intractable M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps (instead of inferring all parameters at once, the conditional step retrieves a set of optimal parameters, conditioned by the current value of the others). ECM is for instance used with GMMs including an additional linear constraint on the means of the components, as provided by the **mixtools** package. As documented in Table 2, **EMMIXmfa** implements an extension of the ECM, termed alternating expectation–conditional maximization (AECM) algorithm (X.-L. Meng and Van Dyk 1997), and which can be used to reduce the computational burden of estimating the parameters of mixtures of factor analysers. The AECM algorithm is an extension of the ECM algorithm that allows the complete data used for estimation to differ on each CM-step (generally, in order to speed the computation, by selecting a subset of the most leveraged observations). GEM algorithms share the same asymptotic theoretical properties of the

EM algorithm, especially the local consistency of the estimates returned.

A small simulation to evaluate the impact of outliers

Classical methods used for the parameters' estimation, especially the maximum likelihood estimation (MLE), are sensitive to the presence of outliers. A naive solution consists in assigning null weights to observations suspected to be outliers, so that they do not contribute [6]. Trimming aberrant observations from the distribution is justified theoretically by the principle of the *spurious outlier model* (Gallegos and Ritter 2005). However, this method is quite stringent, requiring human expertise or the use of general outlier detection tools not necessarily adapted to GMM estimation.

Two general approaches for dealing with outliers with a well-defined theoretical background are the *outliers mixture modelling* and the *trimming approach*. *Outliers mixture modelling* integrate an additional component accounting for the outliers in the distribution. Notably, the **mclust** (Fraley, Raftery, and Scrucca 2022) and **otrimle** (Coretto and Hennig 2021) packages use an improper uniform distribution to model the distribution of outliers. Unlike **mclust**, the **otrimle** package does not require the user to set in advance the proportion of outliers in the mixture (Coretto and Hennig 2016). As opposed, in the *trimming approach*, outliers are first removed before the complete estimation of parameters. Such methods are implemented in **tclust** (Iscar, Escudero, and Fritz 2022) and **oclust** (Clark and McNicholas 2019) packages.

tclust (Iscar, Escudero, and Fritz 2022) uses a robust constrained clustering method, where the user has to set an upper threshold to the ratio between the highest and the lowest variability among all components and a trimming ratio α. It extends the work of García-Escudero et al. (2008), with released constraints on the Gaussian distribution. First, the maximal degree of affinity, defined in Equation (26):

$$D(x_i|\theta) = \max_j \left(p_j \varphi_{\zeta_j}(x_i)\right) \tag{26}$$

is computed for each observation x_i, and corresponds for each point to the maximum probability to observe it in the distribution, given parameter θ. Then, α observations the least likely to be observed are trimmed for the estimation of the parameters. When we reach convergence of the estimated parameter and there is no change in the outliers identification from one iteration to another, the iterative algorithm stops. The use of constraints is an additional feature that avoids building over-dispersed or unbalanced clusters, the highest constraint of a ratio of 1 yielding clusters with equal sizes. However, the identification of an observation as aberrant is highly dependant on the variability constraint and the determination of these two hyperparameters is complex and highly dependant on the shape of the distribution. Additionally, a CEM algorithm is used to retrieve the parameters and the proportion of outliers, for which the MLE, in contrast to the EM algorithm, is not asymptotically consistent nor efficient.

Unlike **tclust**, **oclust** (Clark and McNicholas 2019) both retrieves the proportion of outliers and identifies them. To do so, it compares the complete log-likelihood of the mixture $\ell(\theta|X)$ with its value removing one observation $\ell(\theta|X \setminus X_i)$, for all observations. Observations are iteratively removed, based on the assumption that the Kullback-Leibler divergence between the original log-likelihood and the trimmed log-likelihood $\mathrm{KL}\left(\ell(\theta|X)||\ell(\theta|X \setminus X_i)\right)$ follows a Beta distribution. At each step, the observation that maximises the Kullback-Leibler divergence at a statistically significant threshold is removed. The algorithm stops trimming outliers, when this measure is not anymore statistically significant. However, the assumption of a Beta distribution only holds asymptotically and with non-overlapping clusters.

To integrate the impact of outliers in the estimation, we simulated a two-components GMM with well-separated and balanced clusters. The outliers distribution, corresponding to the additional noise component, was retrieved by randomly selecting *prop.outliers* points out of the total number of observations and drew their values from

[6] The use of weighted distributions has more general applications. It can be used to deal with a component distribution that does not fit exactly a Gaussian shape. For instance, to deal with heavy tail distributions, more weight can be given to central components and less weight to the tails.

an uniform distribution bounded by an interval five times as big as the 0.05 and 0.95 quantiles of $f_\theta(X)$. All estimates were obtained comparing the five reviewed initialisation methods, except with **otrimle** which has its own hierarchical clustering initialisation method.

The slowest package is **otrimle**, most of the time being taken by the initialisation step where proportion and identification of the outliers is performed. Running times of the other packages are generally not impacted by the presence of outliers.

Most of the reviewed packages, except the **bgmm** package, are not impacted by the choice of initialisation method. Additionally, the proportions are rather correctly estimated (related to the choice of an uniform distribution to model outliers), but the reviewed packages tend to overestimate the true variability of each component, with the worst results obtained with **rebmix** initialisation. **bgmm** sets apart from the others by its reduced bias on the means and standard deviations estimated, a feature left undocumented. However, increasing the number of outliers (Figure 2, panel C) lead also to biased estimations for **bgmm**, while **otrimle**, a dedicated package, is still able to correctly estimate the individual parameters of the components' distributions with a high proportion of outliers. Yet, analysing the code used to implement the **bgmm** reveals that there is no dedicated feature to remove outliers but rather a specific method used to deal with numerical underflow that artificially increases the probability of observing outlying distributions (EM-implementation differences across reviewed packages).

Appendix C: the meta-analysis workflow for the final selection of CRAN and Bioconductor platforms

General workflow

Table 4 lists the terms used in the search, the number of packages returned by the search, the number of packages excluded from review after the search, and the names of the packages ultimately selected for review. Indeed, the CRAN and Bioconductor platforms are the two most popular repositories for R packages, with a constraining review before publication.

Most packages we excluded from review did not focus on the GMM model, but on supplying tools for visualising and asserting the quality of a given clustering. For instance, the search term "cluster" returned many packages implementing other unsupervised clustering methods, such as k-means, KNN or graph clustering, were specifically dedicated to specific data, such as single cell analyses. The search term "mixture" returned either packages dealing only with non-Gaussian components, such as **fitmix** with log-normal distributions or were dedicated to chemical mixture designs.

Table 4: Meta-analysis summary about the selection of packages implementing the estimation of GMMs, on CRAN and Bioconductor.

Platforms	Searched terms	Number of returned packages	Number of packages implementing GMMs	Packages implementing GMMs	Packages kept
Bioconductor	mixture	15	3	epigenomix, fmrs, semisup	∅
Bioconductor	cluster	69	1	Melissa	∅
CRAN	mixture	179	44	AdaptGauss, bgmm bmixture, bpgmm, CAMAN, ClusterR, deepgmm DPP, dppmix, EMCluster, EMMIXgene EMMIXmfa, fabMix, flexmix, fmerPack, GMKMcharlie, IMIX, ManlyMix mclust, MGMM, mixAK, MixAll, mixdist, mixR mixreg, mixsmsn, mixtools, mixture MixtureInf, MMDvariance, nor1mix pcensmix, pgmm, pmclust, polySegratioMM rebmix, Rmixmod, RMixtComp RobMixReg, RPMM, SAGMM, sensory, SMNCensReg	bgmm, EMCluster flexmix, GMKMcharlie mclust, mixtools, Rmixmod
CRAN	cluster	418	16	ClusterR, clustMD, DCEM, EMCluster, HDclassif ManlyMix, mclust, mixAK, MixAll mixture, oclust, otrimle, pmclust, rebmix Rmixmod, tclust	EMCluster mclust, Rmixmod

At this stage, too many packages for a tractable benchmark remained. We hence perform stricter selection of them, based on the following criteria:

- Some of the packages did not implement the unconstrained GMM (no constraint of homoscedascity or equibalanced proportions). Hence, **epigenomix** (Klein and Schaefer 2022) , **EMMIXgene** (Andrew Thomas Jones 2020) , **pcensmix** (Fallah and Hinde 2017) , **mixAK** (Komárek 2022) (homoscedastic components), **mixture** (Pocuca, Browne, and McNicholas 2022) (multi-dimension only), **AdaptGauss** (Thrun, Hansen-Goos, and Ultsch 2020) and **MMDvariance** (X. Li et al. 2018) add constraints on the number of components, on the standard deviation of each component or on mean values of each population, leaving no choice to the user to remove such assumptions. **semisup** (Rauschenberger 2022) restrains on mixtures with two components, for which a part of the observations are labelled. Additionally, it is designed for GWAS or differential analyses. Other packages were designed to deal with high-dimensional data, projecting the data on a smaller subspace using a factor analysis model. Hence, these packages can not learn a GMM for an univariate distribution, as we can not project on a smaller space than the unidimensional space. This led to the exclusion of **HDclassif**, **fabMix** (Papastamoulis 2020) , **EMMIXmfa** and **pgmm** (Paul D. McNicholas et al. 2022) packages. The **sensory** (Franczak, Browne, and McNicholas 2016) package both imputes missing data and performs factor regression on a subspace up to 3 dimensions at most, but requires the user to provide its own initial estimates. Alternatively, **clustvarsel** (Dean, Raftery, and Scrucca 2020) discards the least informative variables, in an attempt to find a locally optimal subset of variables that best discriminate clusters.
- We assume that our original data is continuous. However, some packages are dedicated to deal with discrete data, for instance binned size distributions of medical patients. This led to the exclusion of **mixdist** (Macdonald and Juan Du 2018).
- We restrained our review to packages that use the classic EM algorithm, using maximum likelihood estimation to retrieve the parameters of GMMs. For instance, some packages offer a Bayesian estimation of the parameters of the model using MCMC methods, such as **bmixture** Mohammadi (2021)], **bpgmm** (Lu, Li, and Love 2022), **DPP** (Avila, May, and Ross-Ibarra 2018) , **dppmix** (Xu et al. 2020), **BayesCR** (Garay et al. 2017) and **Melissa** (Kapourani 2022). **polySegratioMM** (Baker 2018) uses the Bayesian framework JAGS's interface in R. Alternatively, other algorithms focusing on maximising the likelihood do exist, but rely on different statistical methods, such as **RPMM** (Houseman et al. 2017) which implements a recursive algorithm, and **SAGMM** (Andrew T. Jones and Nguyen 2019) offering a stochastic approximation.

We then removed the packages in which the MLE estimation of the unconstrained GMM model was an ancillary task:

- We removed the packages that focus on learning mixture of Gaussian regressions such as **fmrs** (Shokoohi 2022) , **mixreg** (Turner 2021) or **fmerPack** (Y. Li and Chen 2021) , an extension of the **flexmix** package with an additional feature selection using the lasso method. **nlsmsn** (Prates, Lachos, and Garay 2021) implements regression of skewed Gaussian mixtures, but in unidimensional space only. **RobMixReg** (S. Cao, Chang, and Zhang 2020) performs robust regression of Gaussian mixtures using five several methods: CTLERob, a component-wise adaptive trimming likelihood estimation; mixbi, bi-square estimation; mixL, Laplacian distribution; mixt, t-distribution; TLE, trimmed likelihood estimation, and flexmix which only performs flexmix regressions with multiple random starts.
- Some packages were built to deal with highly specific tasks. **RMixtComp** (Kubicki, Biernacki, and Grimonprez 2021) and **clustMD** (McParland and Gormley 2017) deal with mixed data (continuous + discrete). The **deepgmm** (Viroli and McLachlan 2020) package learns deep Gaussian mixture models, generalising the classical GMM with multiple layers. **IMIX** (Wang 2022) focuses on clustering multi-omic data that is learnt with the **mclust** package, and **coseq** (Rau 2022) implements RNA-Seq transcriptome clustering using the **Rmixmod** package.
- Some extend the EM algorithm on Gaussian distributions and overcome its main limitations. The **MGMM** (McCaw 2021) package deal with missing data, which is not relevant in unique dimension. The **mixsmsn**

package estimates skewed GMMs. **SMNCensReg** (Garay, Massuia, and Lachos 2022) fit univariate right, left or interval censored data. Some packages offer a robust implementation of the algorithm, automatically trimming possible outliers. **otrimle** models the presence of outliers by an extra component following an improper uniform distribution, while **tclust** and **oclust** automatically removes possible outliers before the estimation step (A small simulation to evaluate the impact of outliers).

- We also removed packages that were limited in their functionalities or complex to install. Indeed, **ClusterR** (Mouselimis 2022) (k-means), **rebmix** (REBMIX), **nor1mix** (univariate dimension only, wrong initialisation process), **MixAll** (Iovleff 2019) (random and small EM) do not allow to perform the EM algorithm with its own initial estimates. The function to provide its own initial estimates for the \pkg{DCEM] package is only internal, and not supposed to be available for the common user. **pmclust** (W.-C. Chen and Ostrouchov 2021) depends on the availability of the OpenMPI framework for its parallelised implementation of the EM algorithm.
- We also removed the **mixR** (Yu 2021) and **CAMAN** (Schlattmann, Hoehne, and Verba 2022) packages which have not been updated in the last two years or are still in beta version.

The popularity of the selected packages varies largely, as illustrated in Figure 3. Among them, **mclust** and **flexmix** are the most popular, followed by **mixtools** and **Rmixmod** packages. We used the **cranlogs** (Csárdi 2019) package to retrieve the daily number of downloads for each of the benchmarked packages, between the 30$^{\text{st}}$ of January, 2023 and the 30$^{\text{th}}$ of April, 2023.

Only the packages dedicated to high-dimensionality, listed in our first bullet point, are relevant to benchmark their performance as a function of the number of dimensions. Indeed, although some packages computing mixtures of regressions do implement features allowing to handle high-dimensional datasets, such as **RobMixReg** and **fmerPack**, they all assume a diagonal covariance structure, and accordingly independent covariates.

The two existing strategies are then limited to projection to a smaller subspace, usually within the theoretical framework of factor analysis or to perform a feature selection strategy. We quickly discarded **fabMix**, since it only retrieves the parameters of GMMs within a Bayesian framework, while we focused on strategies retrieving the MLE via the EM algorithm. The core function `pgmmEM` in the **pgmm** package unfortunately includes a seed for the the algorithm's initialisation that cannot be disabled. Such a feature is generally not recommended for reproducibility, since by defining the seed internally in the function, we were not able to independently generate new reproducible datasets in our benchmark (instead, it is recommended to set the seed value once and for all at the beginning of the virtual simulation). Additionally, while implementing the same AECM variant of the EM algorithm as **EMMIXmfa**, as detailed in Appendix B: Extensions of the EM algorithm to overcome its limitations, its convergence criteria differs from the other benchmarked packages. Indeed, instead of considering a limiting number of iterations along with a prior threshold, either *absolute* or *relative*, it examines only the difference between the current value of the log-likelihood and a corresponding asymptotic estimate, based on the Aitken acceleration (Lindsay 1995). In brief, the asymptotic value of the log-likelihood is the limiting sum of a geometric series, whose common ratio, the so-called Aitken acceleration, is the relative fraction of the log-likelihood gain of the current iteration. Therefore, the use of a different termination criterion precludes any further fair comparison with the other benchmarked packages, as there is no direct equivalence between the two methods.

Finally, **clustvarsel** is not really tailored for datasets with a large number of dimensions, but rather for datasets with a small number of observations. Indeed, by performing a sequential search in the model space in a forward-backward process (namely by adding variables to the null model till we recover the full model, with all features), the algorithm requires intensive computational resources (for instance, there are already $2^{10} = 1024$ models to be tested in dimension 10). In addition, rather than employing a sequential and greedy strategy, an independent and parallelisable feature selection procedure, through the model space, would have sped up by several orders of magnitude the estimation. To that end, (J. Chen and Chen 2008; He and Chen 2016) suggests a stochastic and greedy feature selection strategy, using notably the *eBIC* criteria in order to have an equal chance to draw a model of any dimension[7]. Such a strategy is commonly used in *ensemble learning*.

[7]Indeed, by simply uniformly sampling among the 2^D models available, the probability of getting models with $D/2$ features is

Practical details for the implementation of our benchmark

First, the number of observations ($n = 200$ and $n = 500$ respectively in the univariate and bivariate setting) was chosen enough high to both lower the probability of generating a sample without drawing any observation from one of the components in case of highly-unbalanced clusters and decreases the *margin of error* related to the random sampling error. Specifically, the probability of generating at least one simulation among the N generated fo which less than two observations proceed from component j (the minimal number of elements required to estimate both the mean and the variance of the corresponding cluster), with a two-components mixture of n observations, is given by the following formula (Equation (27)):

$$1 - \Big(1 - (1 - p_j)^n - n \times (1 - p_j)^{n-1} \times p_j\Big)^N \tag{27}$$

Interestingly, the probability of generating a sample among the N repetitions increases exponentially as the level of imbalance increases. For instance, considering $N = 100$ repetitions, $n = 200$ observations per sample and proportion for the minor component $p_j = 0.1$, the probability of generating a degenerate simulation is insignificant: 1.63×10^{-6} while the risk considerably increases, keeping the same general simulation parameters and setting minor proportion to $p_j = 0.05$, with a probability of 0.04. We have focused on one of the impacts of high dimensionality, namely that related to the homogenisation and convergence of any distance norm and the increase in sparsity in relation with the number of features added. We deliberately do not consider the case where the number of dimensions exceeds the number of observations (namely, when $D > n$), since in this configuration, the covariance matrix is no longer of full rank and invertible, implying that the corresponding probability distribution does spans completely over a smaller subspace. However, with so few observations, ($n = 200$ in scenarios identified as a), we reveal the impact in terms of the quality of the estimation when the number of observations is closed to the number of free parameters required to parametrise the full GMM model (with $k = 2$ clusters and $D = 10$ dimensions, $k \times \frac{D(D+1)}{2} + kD + 1 = 131$ are needed.).

Unless stated explicitly, we keep the default hyper-parameters and custom global options provided by each package. For instance, the **flexmix** package has a default option, *minprior*, set by default to 0.05 which removes any component present in the mixture with a ratio below 0.05. Besides, we only implement the fully unconstrained model in both univariate and multivariate settings, as it is the only parametrisation implemented in all the seven packages and the most popular to perform classic GMM clustering, as no restrictive and difficult-to-test assumptions are required.

Additionally, as stated in Parameters estimation in a high-dimensional context, the intrinsic dimension d_j for each cluster j is a hyperparameter, which is generally inferred independently from the GMM estimation itself. While a variety of methods from the field of factor analysis, enumerated in Factor criteria selection, have been developed to estimate the intrinsic dimension, to our knowledge, only two of them have been implemented in CRAN packages: the *Cattell's scree-test* (Cattell 1966) or the dimension selection graph using one of the *penalty metric* discussed in the appendix Model selection (Bergé, Bouveyron, and Girard 2012). However, while **HDclassif** natively implements a performance criterion method for determining the dimension of the spanning space, performed under the hood by function `mixsmsn::hdcc`, none of the other packages evaluated implemented a dimension selection feature. Instead, we infer it for each of the packages dedicated to high-dimensionality with HDclassif, using using the so-called model "AkjBkQkD", for which the intrinsic dimension is common to all components but the characteristics unique for each component Finally, we use among all supplied parametrisations, the least constrained one. Namely, we used the model "AkjBkQkDk" with HDclassif, in which not only the individual features of the covariance matrix but also the spanning dimension are unique for each cluster, and function `mcfa` of the **EMMIXmfa** package, in which the *transition matrix* is common to all components (referred to as the orientation matrix in Appendix Parameters estimation in a high-dimensional context.

If all the seven reviewed packages accept initial estimates provided by the user, both the input and the output format differ between them, requiring an intensive processing to standardise both the initial estimates input, and the output estimates. Notably, a well-known issue with the mixture models is that they identifiable up to a

much higher than drawing models at the boundaries, displaying either few or close to $|D|$ covariates.

permutation of the components (alternatively, changing the index of the labels do not change the likelihood of the model). Assigning one component of the mixture to a specific index is generally immaterial, as the main objective is to return the estimates. However, when it comes to compare the estimated parameters with the true estimates, we must associate unequivocally each component to a specific index. To do so, we set a partial ordering, sorting the components by increasing order of their mean components. Actually, if the ratio or the covariance estimates can be equal for all the components, it is generally not the case for the centroids, as this would result into a degenerate distribution. The consequence and some illustrations of the non-identifiability of the mixture distributions are discussed in section Identifiability of finite mixture models, in Dai and Mukherjea (2001) and in Book Robert and Casella (2010).

We detail below some additional functions we implement to both homogenise input and output of the packages and ease the user's task when comparing the performance of these packages:

- The input observations, mean and covariance matrices have to be transposed compared to the conventional format in packages **bgmm**, **EMCluster**, **GMKMcharlie** and **Rmixmod**, namely $D \times k$ mean matrix and $D^2 \times k$ covariance array (D^2 matrix to store each component variance).
- To save some storage, the **EMCluster** package reshapes the covariance matrix, benefiting from its symmetry. Hence, instead of a three-dimensional array, **EMCluster** expects a compressed $k \times \frac{D(D+1)}{2}$ matrix, each line storing the upper triangular part of the covariance. The memory gain is yet controversial, as decreasing only by a factor two the total space required for the computation. To switch from one format to another, we developed specifically two functions: *trig_mat_to_array()* and *array_to_trig_mat()* in our GitHub package *RGMMBench*, partly inspiring from *vec2sym* function Handy R functions.
- Instead of the covariance matrix, the **mclust** package requires the lower triangular matrix resulting from its Cholesky decomposition. One of the main advantages of this input, in addition to save storage space, is that it ensures that the covariance matrix is indeed positive-definite, as the Cholesky factorisation is only defined if this condition is respected Cholesky decomposition.
- **flexmix** starts by the M-step of the EM algorithm instead of the E-step. Hence, it expects the posterior probabilities assigned to each cluster j for each observation i, $\eta_i(j)$ (Equation (21)), instead of the initial estimates. Both approaches are, however, equivalent.

On the contrary, none of the packages we evaluated that were dedicated to high-dimensional datasets allow the user to provide its own estimates. Thus, when any of the benchmarked initialisation methods listed in Table 1, was internally available in the package, we use it with the same hyperparameters described in main paper, section *Initialisation of the EM algorithm.* If not, we provide instead a vector containing the MAP assignments inferred by the native initialisation method, in a process similar to that used used with hierarchical clustering.

In addition to the plots displaying the bootstrap parameter estimations associated to Scenarios in Tables 5, 10 and 15, we have computed summary statistics to compare the performances of the reviewed packages:

- The *bias* measures the deviation between the sample mean value of the estimate and the true parameter: $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.
- The Mean Squared Error (MSE) summarises both the variability of the estimator and its bias: $\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$, where $\text{var}(\hat{\theta})$ is the empiric variance of each estimator given by the diagonal terms of the empiric covariance matrix.
- We enumerate the number of successes (either the package or the initial method returns an error, or fails in returning a set of parameters enforcing standard constraints of multivariate GMMs, namely the unit simplex constraint over the ratios, positive-definite covariance matrices and in general no missing or infinite value).
- For each scenario, we measured independently the running times taken by the initialisation step and by the estimation of the parameters by the EM algorithm. To do so, the **microbenchmark** package (Mersmann

2021) was used for its higher accuracy and flexibility for the computation of the running times in place of `System.time`.

The main differences across packages as well as performance results obtained across packages in each univariate, bivariate and high-dimensional simulation scenario are thoroughly described in the next section.

Appendix D: comprehensive report from the univariate and multivariate benchmark

Packages used to generate the reports and visualisations

To compute the summary metrics and generate the corresponding boxplots of the bootstrapped parameters, we made extensive use of the facilities provided with the **tidyverse** (Wickham et al. 2019) packages, including:

- **tibble** (Müller and Wickham 2023) to visually and uniformly store the many datasets generated by our benchmark. We then used **readr** (Wickham, Hester, and Bryan 2023) to save and import in a readable format the summary metrics associated with each scenario and the tables listing the main functionalities implemented in the packages studied, **dplyr** (Wickham, François, et al. 2023) to manipulate the data stored in the tibbles and **purrr** (Wickham and Henry 2023) to manipulate the nested tibbles and perform functional programming.
- **ggplot2** (Wickham, Chang, et al. 2023; Wickham 2016) for data visualization, including generating density graphs in the univariate and bivariate framework, and factorial projection for the high-dimensional framework.
- **stringr** (Wickham 2022), for strings, and **forcats** (Wickham 2023), for factors, were particularly useful for customising and ordering the packages in our graphical representations, in order to highlight differences between them.

In addition to the array of packages within the *tidyverse* ecosystem, we utilized the **flextable** (Gohel and Skintzos 2023) and **kableExtra** (Zhu 2021) packages to facilitate the generation of summary reports in HTML and PDF formats.

Furthermore, we would like to express our gratitude for the contributions of **knitr** (Xie 2023, 2015), **rmarkdown** (Xie, Allaire, and Grolemund 2018; Allaire et al. 2023), and the associated wrapper package **rjtools**(O'Hara-Wild et al. 2023), which greatly streamlined the process of creating these HTML and PDF reports.

In a more specialised context, we harnessed the features offered by these packages:

- **ComplexHeatmap** (Gu, Eils, and Schlesner 2016; Gu 2022) to generate the heatmaps of the correlation matrices. This was complemented by **RColorBrewer** (Neuwirth 2022) for effective management of the R colour palette.
- **cowplot** (Wilke 2020), **grid** (R Core Team 2023), and **gridExtra** (Auguie 2017) were used for aggregating and merging multiple plots.
- **ggtext** (Wilke and Wiernik 2022), **glue** (Hester and Bryan 2022), and **scales** (Wickham and Seidel 2022) were employed to enhance the clarity and readability of both x and y ticks and labels.

EM-implementation differences across reviewed packages

Most of the distinct behaviours between the packages result from additional choices external to the EM algorithm itself, aiming at partly overcoming its main limitations (Panel B, Figure 9). We detail below their differences ranked by decreasing order of their leverage effect on the final estimate:

1. Most of the differences between the two classes of packages (Figure (2)) are related to the either relative or absolute choice for the termination criterion of the EM algorithm. Given an user-defined threshold,

the *absolute method* early stops the estimation by comparing the difference between two consecutive log-likelihoods, $|\ell(\hat{\theta}_q|X) - \ell(\hat{\theta}_{q-1}|X)|$, while the *relative method* examines the variation rate, $\left|\frac{\ell(\hat{\theta}_q|X)-\ell(\hat{\theta}_{q-1}|X)}{\ell(\hat{\theta}_{q-1}|X)}\right|$.

2. Several methods can be used to deal with numerical underflow, mostly happening with highly unlikely observations, distant from any centroid.
   - The least elaborate feature is from **Rmixmod**, returning an error when either any of the posterior probabilities or any of the estimated parameters goes below to the precision threshold of the machine (2.22×10^{-16} for most OS).
   - If the maximal value of any posterior probability is null, **bgmm** subtracts the minimal logarithm posterior probability to any log-computed probability. This method avoids numerical underflow by preventing computation of null ratios but the correctness of the estimates is no longer enforced[8].
   - The remaining packages handled numeric underflow in a more convincing manner as they guarantee to return the MLE estimate. The **flexmix**, **GMKMcharlie** and **EMCluster** packages use the same log-rescaling tip detailed in (Application of the EM algorithm to GMMs). The **mixtools** and **mclust** packages use a variant of this trick, taking profit of the factorisation by the greatest element (Equation (28), Equation 3 p.5 Benaglia et al. (2009)), but without exploiting the tip of Taylor's development over $\log(1+x)$:

$$\eta_i(j) = \frac{p_j\,\varphi_{\zeta_j}(x)}{\sum_{j=1}^{k} p_j\,\varphi_{\zeta_j}(x)} = \frac{\frac{p_j\,\varphi_{\zeta_j}(x)}{p_{j_{\min}}\,\varphi_{\zeta_{j_{\min}}}(x)}}{1 + \frac{\sum_{j\neq j_{\min}} p_j\,\varphi_{\zeta_j}(x)}{p_{j_{\min}}\,\varphi_{\zeta_{j_{\min}}}(x)}} \tag{28}$$

In both cases, the computation of the smallest posterior probability, the most proned to be assigned a null value, is avoided, avoiding inconsistent ratios of type 0/0.

- The previous two items deal with specific numeric limitations, but do not directly address one of the main theoretical limitation of the EM algorithm, namely the risk of falling into a suboptimal maximum, plateau or getting trapped on the boundary space (occurs when the proportion of one of the component converges to zero). Some packages specifically handle the case of a vanishing component during the EM optimization: the **mixtools** package performs random re-initialisations in case one of the computed variance goes below a user-defined threshold (default 10^{-8}). **flexmix** and **GMKMcharlie** deal explicitly with the removal of a component, by updating the corresponding MLE parameters. **flexmix** removes any component whose associated weight is by default below 0.05 (such a stringent limitation tends to an underestimation of the true number of components in highly unbalanced mixtures)[9], while GMKMcharlie both implements a lower limit on the proportions of the components and an upper threshold over the ratio of the maximum and minimal eigenvalue resulting from the factorisation of the covariance matrix (Equation (12))[10].

We enumerate below some additional features supplied by the packages:

- In addition to log rescaling, **GMKMcharlie** includes an additional argument, *embedNoise*, to avoid degenerate GMMs by adding a small constant to any diagonal term (by default 10^{-6}). Besides, instead of controlling whether there was a relative change of the log-likelihood, the EM implementation of **GMKMcharlie** controls instead that there was no significant relative difference in the estimated parameters in the ten previous optimisations[11]. Finally, since **GMKMcharlie** has implemented a parallelised version of the algorithm, it ensures using a a time limit that the algorithm indeed terminates (by default, set to one hour).

[8] Additionally, **bgmm** does not update the estimated variances if any newly computed variance is below the criterion stop. A remarkable side-effect of these features, as shown in Figure 2, is that the **bgmm** package is less sensitive to the presence of outliers.

[9] Indeed, at least one of the component was removed in 80% of our estimations in the unbalanced and overlapping case (scenario U9 in 5) and in 20% of the simulations in the unbalanced and well-separated case (scenario U3 in 5).

[10] These options are set respectively to 0 and $+\infty$ by default, thus they did not impact our simulations

[11] In our simulation, the behaviour of the **GMKMcharlie** did not differ significantly from the remaining packages of the second class. However, the use of an Euclidean distance criterion may be problematic when parameters are not on the same order of magnitude, requiring their prior normalisation

- **flexmix** performs an unbiased estimate of the covariance matrix, instead of the corresponding ML covariance estimate (divides by a factor $n-1$ instead of the number of observations n). Such a choice does not affect the results in our simulations, but may have a stronger impact when fitting models to a small number of observations.
- Similarly to flexmix, the **HDclassif** package implements some constraints to preserve numerical stability. The `min.individuals` attribute, like the `minprior` attribute of `flexmix` function, discards any cluster having fewer observations[12]. However, unlike **flexmix**, the algorithm stops instead of re parametrising the mixture problem with a smaller number of components. Coupled with the *Cattell's scree-test*, the `noise.ctrl` attribute is the minimum threshold of a feature's contribution to the overall variance, computed as the corresponding normalised eigenvalue, in order to be included in the mixture of factor analysers. This additional constraint ensures a parsimonious dimension selection process, so that the number of selected intrinsic dimensions cannot be greater than or equal to the order of the discarded eigenvalues.

[12] by default, set to two, i.e. the minimum number of replications to derive an unbiased estimate of the empirical variance of a sample

Supplementary Figures and Tables in the univariate simulation

Table below (5) lists the complete set of parameters used to simulate the univariate Gaussian mixture distribution in our benchmark:

Table 5: The 9 parameter configurations tested to generate the samples of the univariate experiment, with $k = 4$ components.

ID	Entropy	OVL	Proportions	Means	Correlations
U1	1.00	3.3e-05	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U2	1.00	5.7e-03	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U3	1.00	2.0e-02	0.25 / 0.25 / 0.25 / 0.25	0 / 4 / 8 / 12	2 / 2 / 2 / 2
U4	0.96	3.3e-05	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U5	0.96	5.8e-03	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U6	0.96	2.0e-02	0.2 / 0.4 / 0.2 / 0.2	0 / 4 / 8 / 12	2 / 2 / 2 / 2
U7	0.68	2.7e-05	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	0.3 / 0.3 / 0.3 / 0.3
U8	0.68	4.4e-03	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	1 / 1 / 1 / 1
U9	0.68	1.5e-02	0.1 / 0.7 / 0.1 / 0.1	0 / 4 / 8 / 12	2 / 2 / 2 / 2

Figure 4-Figure 8 each summarise the benchmarking results associated with one of the scenarios listed in Table 5.

Summary tables 6- 9 display the average performance for each package of the benchmark with each initialisation method. The best performing pair (lowest bias or MSE) is highlighted in green, and the worst performing in red. The MSE and bias columns were derived by summing respectively the estimated proportions, means and standard deviations associated with the individual components.

Table 6: MSE and Bias associated to scenario U1, in Table 5 (balanced and well-separated components)

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	kmeans	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	quantiles	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	random	0.0061	1.90000	0.19000	0.0290	0.5300	0.1100
	rebmix	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
flexmix	hc	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	kmeans	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	quantiles	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	random	0.0064	1.80000	0.19000	0.0290	0.5300	0.1100
	rebmix	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
mclust / bgmm	hc	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	kmeans	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	quantiles	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	random	0.0062	1.80000	0.19000	0.0290	0.5300	0.1100
	rebmix	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
mixtools	hc	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	kmeans	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	quantiles	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	random	0.0064	1.90000	0.19000	0.0290	0.5300	0.1100
	rebmix	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
Rmixmod / RGMMBench	hc	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	kmeans	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	quantiles	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028
	random	0.0064	1.90000	0.19000	0.0290	0.5300	0.1100
	rebmix	0.0004	0.00077	0.00037	0.0025	0.0068	0.0028

The panels indexed by the B letter, from Figure 4 to Figure 8, display the 0.05, 0.5 and 0.95 quantiles of the distribution of the operating times taken for parameter estimation, for the scenarios listed in Table 5.

First, we note that the execution time grows asymptotically linearly with the number of observations, confirming empirically the expected linear complexity of the EM algorithm. The most important factor playing on the differences observed is related to the complexity of the distribution, and especially the degree of overlap between the components:

- On the one hand hand, when components are well-separated (scenarios 1 and 3 in Table 5), the estimation of the parameters is simple, leading to a reduced number of iterations required to reach the convergence and shorter running times.
- On the other hand, the time taken by the slowest package for the estimation of the parameters increases by a hundred factor with the most complex scenario (see scenario U9, 5, illustrated in Figure 7), compared

Table 7: MSE and Bias associated to scenario U7, in Table 5 (unbalanced and well-separated components)

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.02900	2.8000	0.45000	0.1000	0.840	0.250
	kmeans	0.00730	0.7900	0.13000	0.0260	0.240	0.075
	quantiles	0.16000	19.0000	3.20000	0.6400	6.100	1.800
	random	0.17000	10.5000	1.40000	0.3600	3.100	0.780
	rebmix	0.00027	0.0015	0.00077	0.0025	0.014	0.014
flexmix	hc	0.05500	2.8000	0.45000	0.1100	0.850	0.250
	kmeans	0.00760	0.7800	0.13000	0.0260	0.240	0.075
	quantiles	0.11000	19.0000	3.20000	0.5400	6.000	1.900
	random	0.15000	8.4000	1.00000	0.3000	2.500	0.580
	rebmix	0.00027	0.0015	0.00076	0.0025	0.014	0.011
mclust / bgmm	hc	0.03200	2.8000	0.45000	0.1000	0.850	0.250
	kmeans	0.00740	0.7800	0.13000	0.0260	0.240	0.075
	quantiles	0.14000	19.0000	3.20000	0.6000	6.000	1.900
	random	0.18000	10.4000	1.40000	0.3600	3.100	0.800
	rebmix	0.00027	0.0015	0.00077	0.0025	0.014	0.014
mixtools	hc	0.03200	2.8000	0.45000	0.1000	0.850	0.250
	kmeans	0.00620	0.7600	0.13000	0.0170	0.230	0.079
	quantiles	0.15000	19.0000	3.20000	0.5800	6.000	1.800
	random	0.18000	10.3000	1.40000	0.3600	3.100	0.800
	rebmix	0.00027	0.0015	0.00077	0.0025	0.014	0.014
Rmixmod / RGMMBench	hc	0.02900	2.8000	0.45000	0.1000	0.850	0.250
	kmeans	0.00540	0.7700	0.13000	0.0190	0.230	0.078
	quantiles	0.14000	19.0000	3.20000	0.5900	6.000	1.800
	random	0.17000	10.4000	1.40000	0.3600	3.100	0.800
	rebmix	0.00027	0.0015	0.00077	0.0025	0.014	0.014

to the simplest scenario (see U1, 5, shown in Figure 4). Indeed, the average running time for a complete run of the EM algorithm increases from 0.215 seconds to 10.8 seconds.

To better understand the running times' differences observed between the packages for a given scenario, we perform a three-way anova, taking into account the choice of initialisation method, the programming language and the class of packages[13]:

- With well-separated components (Scenarios U1 and U7 in Table 5), the class of packages (namely the choice of the convergence criterion) has a negligible impact compared to the choice of initialisation algorithm or the programming language. The effect sizes associated to the programming language and the initialisation method are respectively 1.688×10^{-2} (p-value of 3×10^{-60}) and 13×10^{-5} (p-value of 3×10^{-60}), while

[13]To compare whether differences between mean running times or estimation performances differ across packages, we used the between-subjects Anova test `rstatix::anova_test()` to generate the p-values and `rstatix::partial_eta_squared()` to compute the corresponding effect sizes.

Table 8: MSE and Bias associated to scenario U3, in Table 5 (balanced and overlapping components)

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.0170	1.60	0.45	0.1950	1.320	0.93
	kmeans	0.0054	0.81	0.18	0.0125	0.023	0.32
	quantiles	0.0070	0.67	0.30	0.0930	0.590	0.56
	random	0.0440	8.40	1.00	0.0710	0.330	0.63
	rebmix	0.0990	11.00	1.60	0.2000	1.600	0.78
flexmix	hc	0.0260	2.60	0.94	0.1120	1.160	1.22
	kmeans	0.0044	0.67	0.14	0.0036	0.091	0.27
	quantiles	0.0054	0.57	0.27	0.0850	0.670	0.55
	random	0.0420	8.20	1.10	0.0450	0.450	0.68
	rebmix	0.1210	14.30	2.70	0.2700	2.700	1.17
mclust / bgmm	hc	0.0110	2.50	0.84	0.0330	1.160	1.10
	kmeans	0.0068	0.86	0.24	0.0294	0.114	0.36
	quantiles	0.0075	0.70	0.32	0.1110	0.720	0.63
	random	0.0490	9.10	1.20	0.0800	0.320	0.68
	rebmix	0.1410	10.90	2.90	0.2900	1.800	1.47
mixtools	hc	0.0320	2.40	0.80	0.0670	0.360	0.25
	kmeans	0.0415	2.51	1.11	0.1000	0.664	0.74
	quantiles	0.0383	2.40	1.00	0.1170	0.770	0.78
	random	0.0660	9.40	1.80	0.0130	0.340	0.48
	rebmix	0.1090	9.60	2.50	0.2600	1.800	1.33
Rmixmod / RGMMBench	hc	0.0220	2.00	0.67	0.0490	0.370	0.25
	kmeans	0.0318	2.31	0.85	0.0952	0.602	0.67
	quantiles	0.0297	2.19	0.80	0.1210	0.770	0.76
	random	0.0620	9.40	1.70	0.0160	0.310	0.50
	rebmix	0.1140	10.30	2.60	0.2600	1.900	1.31

the choice of the termination criteria did not significantly impact the execution time, with an effect size of 8.119×10^{-4}(p-value of 0.35). Faster running times with packages natively encoded in Fortran or C compared to those encoded in R only were expected, as R is a high-level programming language known to be slower. Indeed, the **flexmix** package is the slowest, preceded by our baseline R implementation. Additionally, **mclust**, followed by **mixtools**, **Rmixmod** and **bgmm** are the fastest.

- On the other hand, with overlapping components (Scenarios U3 and U9 in Table 5), the package class and the programming language have a statistically significant impact on the average running times (the effect sizes associated to the choice of the termination criteria and the programming language are respectively 0.111 (numerical null p-value) and 0.0852 (p-value of 8×10^{-307})) while the initialisation method has no substantial impact (effect size of 2.967×10^{-4} and p-value of 0.32). In the context of highly overlapping mixture, the fastest ones are **mclust** and **GMKMcharlie**, benefiting from both using relative ratios and a fast programming language, while our baseline implementation `emnmix`, preceded by **Rmixmod** and

Table 9: MSE and Bias associated to scenario U9, in Table 5 (unbalanced and overlapping components)

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.230	9.3	0.94	0.78	4.9	1.33
	kmeans	0.094	5.1	0.57	0.50	3.4	0.89
	quantiles	0.230	9.9	1.00	0.80	5.1	1.34
	random	0.270	11.5	0.90	0.63	2.8	0.85
	rebmix	0.330	20.0	2.20	0.53	5.3	0.84
flexmix	hc	0.170	10.5	0.88	0.64	5.2	1.14
	kmeans	0.051	5.6	0.61	0.34	3.6	0.94
	quantiles	0.210	11.3	1.20	0.75	5.6	1.53
	random	0.180	9.5	0.77	0.43	2.7	0.86
	rebmix	0.110	10.0	1.70	0.15	2.3	1.48
mclust / bgmm	hc	0.230	10.2	0.84	0.79	5.1	1.20
	kmeans	0.107	5.5	0.62	0.53	3.6	0.96
	quantiles	0.270	11.4	1.20	0.87	5.6	1.59
	random	0.300	12.2	1.06	0.66	2.9	0.84
	rebmix	0.270	21.0	2.50	0.46	5.2	1.13
mixtools	hc	0.200	9.7	1.19	0.64	3.4	0.69
	kmeans	0.135	7.7	1.16	0.46	2.1	0.48
	quantiles	0.280	11.2	1.60	0.74	4.2	0.72
	random	0.350	15.7	1.62	0.65	2.1	0.64
	rebmix	0.240	22.0	2.70	0.47	5.1	1.18
Rmixmod / RGMMBench	hc	0.210	9.5	1.07	0.69	3.8	0.79
	kmeans	0.113	6.5	0.90	0.46	2.4	0.43
	quantiles	0.240	10.1	1.30	0.74	4.2	0.81
	random	0.320	14.6	1.45	0.61	2.1	0.58
	rebmix	0.250	22.0	2.70	0.49	5.2	1.18

mixtools, are on average a hundred times slower.

Supplementary Figures and Tables in the bivariate simulation

Table below (10) lists the complete set of parameters used to simulate the multivariate Gaussian mixture distribution in our benchmark:

Table 10: The 20 parameter configurations tested to generate the samples of the bivariate experiment.

ID	Entropy	OVL	Proportions	Means	Correlations
B1	1.00	0.15000	0.5 / 0.5	(0,2);(2,0)	-0.8 / -0.8
B2	1.00	0.07300	0.5 / 0.5	(0,2);(2,0)	-0.8 / 0.8
B3	1.00	0.07300	0.5 / 0.5	(0,2);(2,0)	0.8 / -0.8
B4	1.00	0.00078	0.5 / 0.5	(0,2);(2,0)	0.8 / 0.8
B5	1.00	0.07900	0.5 / 0.5	(0,2);(2,0)	0 / 0
B6	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	-0.8 / -0.8
B7	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	-0.8 / 0.8
B8	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0.8 / -0.8
B9	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0.8 / 0.8
B10	1.00	0.00000	0.5 / 0.5	(0,20);(20,0)	0 / 0
B11	0.47	0.06600	0.9 / 0.1	(0,2);(2,0)	-0.8 / -0.8
B12	0.47	0.01600	0.9 / 0.1	(0,2);(2,0)	-0.8 / 0.8
B13	0.47	0.05000	0.9 / 0.1	(0,2);(2,0)	0.8 / -0.8
B14	0.47	0.00045	0.9 / 0.1	(0,2);(2,0)	0.8 / 0.8
B15	0.47	0.03900	0.9 / 0.1	(0,2);(2,0)	0 / 0
B16	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	-0.8 / -0.8
B17	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	-0.8 / 0.8
B18	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0.8 / -0.8
B19	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0.8 / 0.8
B20	0.47	0.00000	0.9 / 0.1	(0,20);(20,0)	0 / 0

Figures 11- 14 are associated to scenarios B11 - B15 of Table 10. Summary tables 11-14 show the average

performance for each combination of a benchmarked package and initialisation method, with the same conventions as discussed in Supplementary Figures and Tables in the univariate simulation.

First, we can directly observe that the OVL increases as the individual variance of each component, the proximity of the centroids of the clusters and the level of imbalance is increased. We demonstrate this statement formally in section An analytic formula of the overlap for univariate Gaussian mixtures. Nonetheless, the influence of the correlation between the x and the y-axis (the off-diagonal term of the covariance matrix) is not immediate, notably the assumption of independent features does not automatically entail a lower OVL or simpler estimation.

From our experiments, we deduce that the highest OVL is obtained when the main axis of the two respective components aligns with the line joining the two centroids. For instance, in our scenario, the lowest OVL is obtained when the correlation term is positive for both clusters (scenario 14, Table 10 and isodensity plot in panel A, Figure 13), whereas the highest OVL is obtained with a negative correlation (scenario 11, Table 10and isodensity plot in panel A, Figure 11). Recall that the slope joining the two centroids of the two components in all our simulated distributions is indeed negative.

Table 11: MSE and Bias associated to scenario B11, in Table 10 (unbalanced, overlapping and negative correlated components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.230	3.90	1.8	0.550	2.30	1.200
	kmeans	0.136	2.80	1.9	0.450	2.30	2.200
	random	0.028	1.27	1.1	0.084	0.12	0.140
	rebmix	0.071	2.20	1.4	0.170	0.66	0.111
flexmix	hc	0.260	3.90	1.9	0.480	2.40	1.300
	kmeans	0.077	2.80	1.9	0.270	2.40	2.300
	random	0.028	0.96	1.0	0.064	0.77	0.720
	rebmix	0.087	1.90	1.0	0.170	1.02	0.468
mclust / bgmm	hc	0.230	3.90	1.8	0.550	2.30	1.200
	kmeans	0.136	2.80	1.9	0.450	2.30	2.200
	random	0.028	1.27	1.1	0.084	0.12	0.140
	rebmix	0.071	2.20	1.4	0.170	0.66	0.111
mixtools	hc	0.210	3.30	1.8	0.470	1.80	1.100
	kmeans	0.131	2.60	1.8	0.380	1.80	1.800
	random	0.051	1.61	1.1	0.129	0.20	0.180
	rebmix	0.093	2.40	1.4	0.210	0.60	0.063
Rmixmod / RGMMBench	hc	0.210	3.30	1.8	0.470	1.80	1.100
	kmeans	0.131	2.60	1.8	0.380	1.80	1.800
	random	0.051	1.61	1.1	0.129	0.20	0.180
	rebmix	0.093	2.40	1.4	0.210	0.60	0.063

Table 12: MSE and Bias associated to scenario B12, in Table 10 (unbalanced, overlapping and opposite correlated components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.00076	0.049	0.16	0.0063	0.056	0.131
	kmeans	0.00076	0.049	0.16	0.0063	0.056	0.131
	random	0.00075	0.049	0.16	0.0057	0.055	0.123
	rebmix	0.00087	0.066	0.17	0.0070	0.063	0.190
flexmix	hc	0.00144	0.049	0.16	0.0101	0.055	0.071
	kmeans	0.00144	0.049	0.16	0.0101	0.055	0.071
	random	0.00144	0.050	0.16	0.0099	0.054	0.067
	rebmix	0.00145	0.048	0.16	0.0142	0.047	0.110
mclust / bgmm	hc	0.00076	0.049	0.16	0.0063	0.056	0.131
	kmeans	0.00076	0.049	0.16	0.0063	0.056	0.131
	random	0.00075	0.049	0.16	0.0057	0.055	0.124
	rebmix	0.00087	0.066	0.17	0.0070	0.063	0.190
mixtools	hc	0.00075	0.050	0.16	0.0049	0.054	0.112
	kmeans	0.00075	0.050	0.16	0.0049	0.054	0.112
	random	0.00075	0.050	0.16	0.0049	0.054	0.112
	rebmix	0.00086	0.066	0.17	0.0061	0.062	0.170
Rmixmod / RGMMBench	hc	0.00075	0.050	0.16	0.0049	0.054	0.112
	kmeans	0.00075	0.050	0.16	0.0049	0.054	0.112
	random	0.00075	0.050	0.16	0.0049	0.054	0.112
	rebmix	0.00086	0.066	0.17	0.0061	0.062	0.170

Table 13: MSE and Bias associated to scenario B14, in Table 10 (unbalanced, overlapping and positive correlated components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.00043	0.044	0.13	0.00081	0.044	0.060
	kmeans	0.00043	0.044	0.13	0.00081	0.044	0.060
	random	0.00043	0.044	0.13	0.00080	0.044	0.060
	rebmix	0.00040	0.044	0.13	0.00120	0.047	0.053
flexmix	hc	0.00043	0.044	0.13	0.00072	0.043	0.035
	kmeans	0.00043	0.044	0.13	0.00072	0.043	0.035
	random	0.00043	0.044	0.13	0.00072	0.044	0.035
	rebmix	0.00040	0.044	0.14	0.00110	0.047	0.044
mclust / bgmm	hc	0.00043	0.044	0.13	0.00081	0.044	0.060
	kmeans	0.00043	0.044	0.13	0.00081	0.044	0.060
	random	0.00043	0.044	0.13	0.00080	0.044	0.060
	rebmix	0.00040	0.044	0.13	0.00120	0.047	0.053
mixtools	hc	0.00043	0.044	0.13	0.00078	0.044	0.060
	kmeans	0.00043	0.044	0.13	0.00078	0.044	0.060
	random	0.00043	0.044	0.13	0.00078	0.044	0.060
	rebmix	0.00040	0.044	0.13	0.00110	0.047	0.053
Rmixmod / RGMMBench	hc	0.00043	0.044	0.13	0.00078	0.044	0.060
	kmeans	0.00043	0.044	0.13	0.00078	0.044	0.060
	random	0.00043	0.044	0.13	0.00078	0.044	0.060
	rebmix	0.00040	0.044	0.13	0.00110	0.047	0.053

Table 14: MSE and Bias associated to scenario B15, in Table 10 (unbalanced, overlapping and uncorrelated components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ
EMCluster / GMKMcharlie	hc	0.1110	2.30	1.30	0.280	1.40	0.90
	kmeans	0.0500	1.50	1.30	0.200	1.05	1.06
	random	0.0290	0.71	0.63	0.070	0.28	0.19
	rebmix	0.0163	0.69	0.78	0.074	0.37	0.44
flexmix	hc	0.1330	2.40	1.40	0.240	1.50	1.05
	kmeans	0.0320	1.60	1.40	0.110	1.21	1.26
	random	0.0370	0.71	0.64	0.048	0.35	0.29
	rebmix	0.0058	0.70	0.84	0.028	0.49	0.62
mclust / bgmm	hc	0.1110	2.30	1.30	0.280	1.40	0.90
	kmeans	0.0500	1.50	1.30	0.200	1.05	1.06
	random	0.0290	0.71	0.63	0.070	0.28	0.19
	rebmix	0.0163	0.69	0.78	0.074	0.37	0.44
mixtools	hc	0.0860	1.90	1.20	0.220	1.10	0.75
	kmeans	0.0470	1.30	1.10	0.170	0.79	0.78
	random	0.0230	0.67	0.66	0.065	0.24	0.19
	rebmix	0.0158	0.69	0.77	0.068	0.30	0.37
Rmixmod / RGMMBench	hc	0.0860	1.90	1.20	0.220	1.10	0.75
	kmeans	0.0470	1.30	1.10	0.170	0.79	0.78
	random	0.0230	0.67	0.66	0.065	0.24	0.19
	rebmix	0.0158	0.69	0.77	0.068	0.30	0.37

In contrast to the univariate setting (Supplementary Figures and Tables in the univariate simulation), the fastest packages are **bgmm**, **EMCluster**, **flexmix**, and **Rmixmod**, and the slowest ones **mclust**, **GMKMcharlie** and **mixtools**, independently from the difficulty of the simulation.

Finally, Figures 15, 16 and 17 represent in a synthetic way less interesting scenarios benchmarked with to the left, the contour maps and to the right the corresponding Hellinger boxplots, with one scenario being illustrated per row.

Supplementary Figures and Tables in the HD simulation

Table below (15) lists the complete set of parameters used to simulate Gaussian distributions in the high dimensional benchmark:

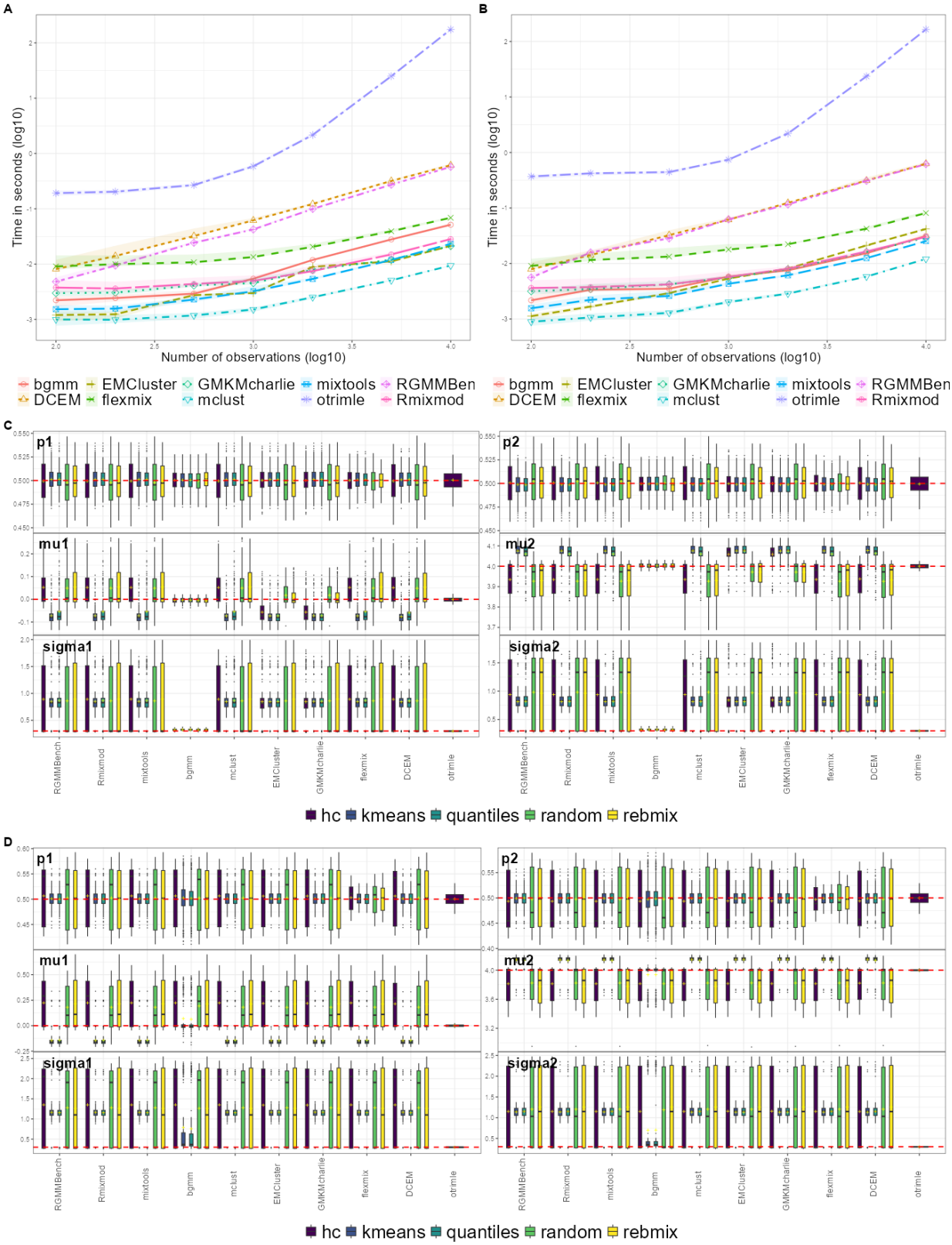


Figure 2: A) Execution times for the nine reviewed packages using hierarchical clustering initialisation, with on the left 2% of outliers in proportion and on the right, 4% of outliers. B) and C) Boxplots of the estimated parameters with $N = 200$ repetitions, $n = 2000$ observations and respectively 2% and 4% of outliers. The red dashed horizontal line corresponds to the true value of the parameters.

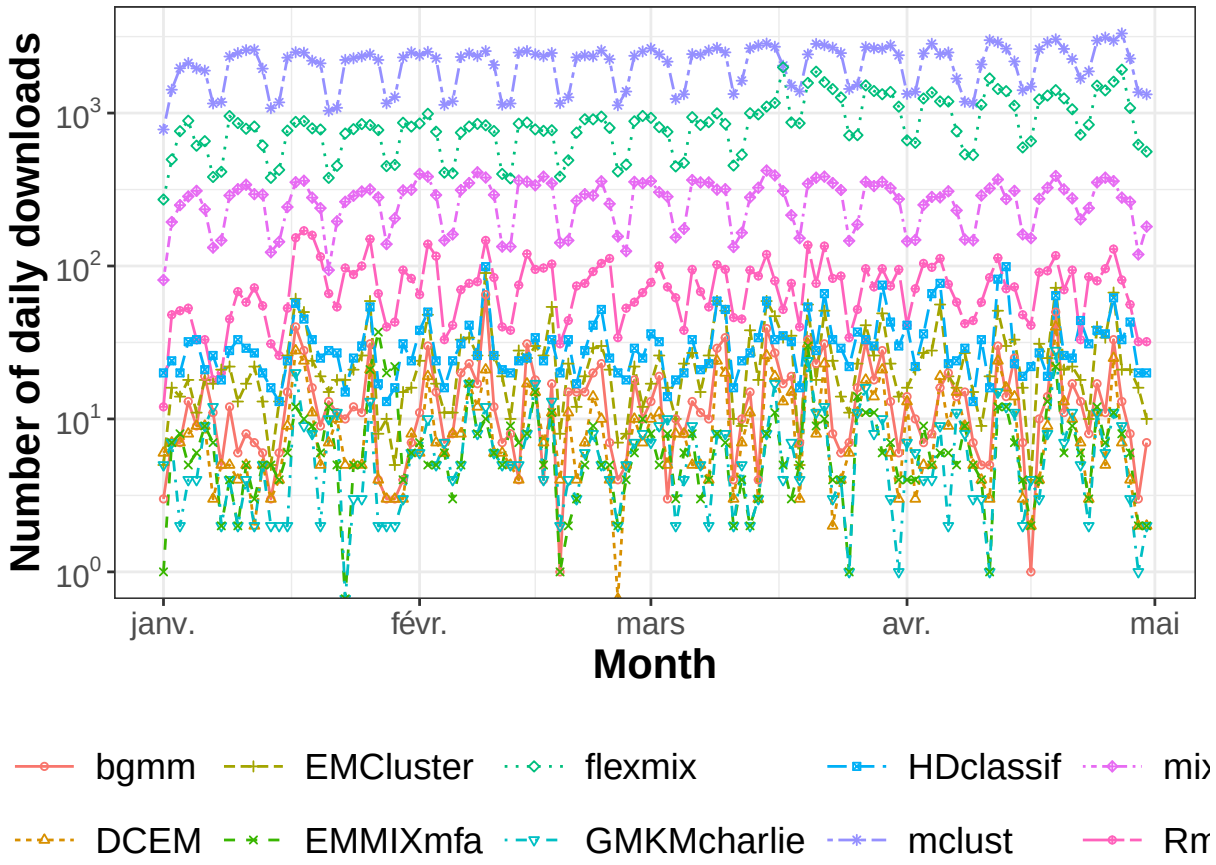


Figure 3: Number of daily downloads (logarithmic scale) from the CRAN mirror from the 1st of January to the 30th April 2022 for the seven R packages reviewed.

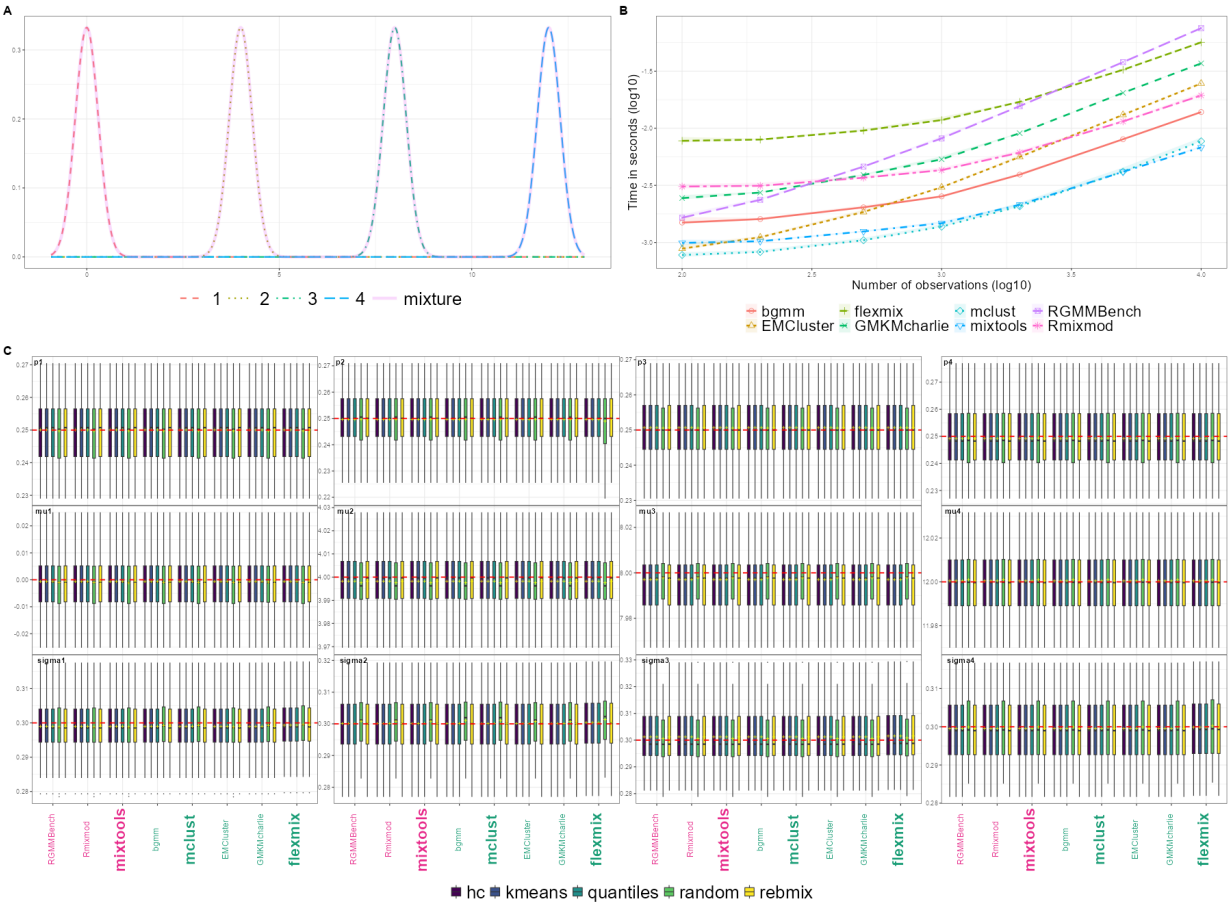


Figure 4: Benchmark summary plots of scenario U1 in Table 5 (balanced and well-separated components), organised as such: The panel A displays the distribution of the global mixture distribution $f_\theta(X)$ (pink solid line) and of each of its constitutive components scaled by their respective proportions (dotted lines). Running times are displayed in Panel B with the k-means initialisation. The number of observations (x-axis) and the running time (y-axis) is in log(10) scale, implying that any linear relationship between the running time and the number of observations is represented by a slope of 1. The points represent median running time. The coloured bands represent the 5th and 95th percentiles of the running time. In panel C are represented the boxplots associated with the distribution of the estimates, with one box per pair of package and initialisation method. The median is displayed with bold black line, the mean with a yellow cross and the 0.25 and 0.95 quantiles match the edges of the rectangular band. Solid black lines extending past the box boundaries represent the 1.5 IQR, estimates above these limits considered as outliers and omitted from the plot. Finally, the true value of the parameter is represented as a dashed red line. The bold black writing in the upper right-hand corner refers to the parameter whose distribution is shown in the corresponding facet. The first, second and third rows are the distributions of the ratios, means and variances of each component, identified by the column index.

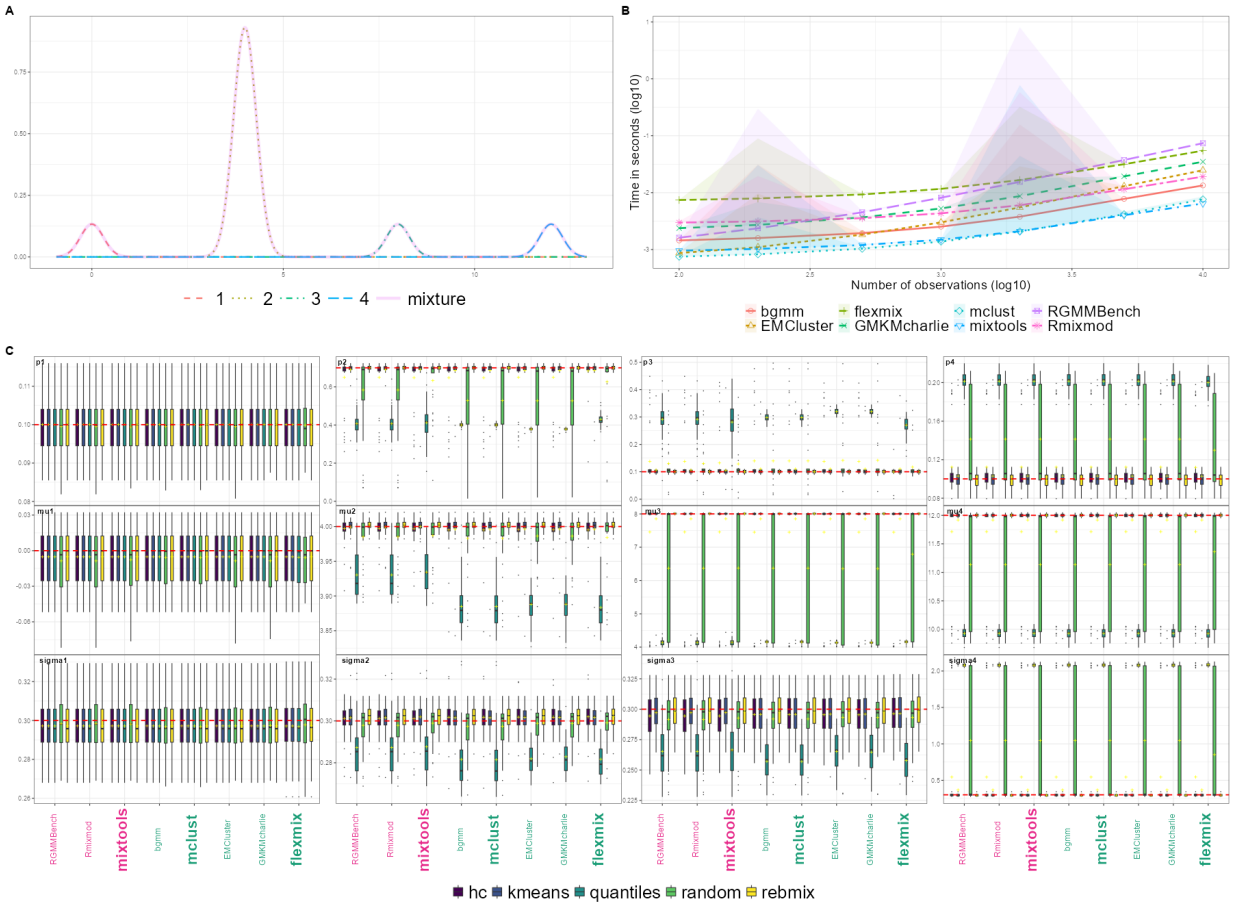


Figure 5: Benchmark summary plots of scenario U7 in Table 5 (unbalanced and well-separated components), with same layout as in Figure 4.

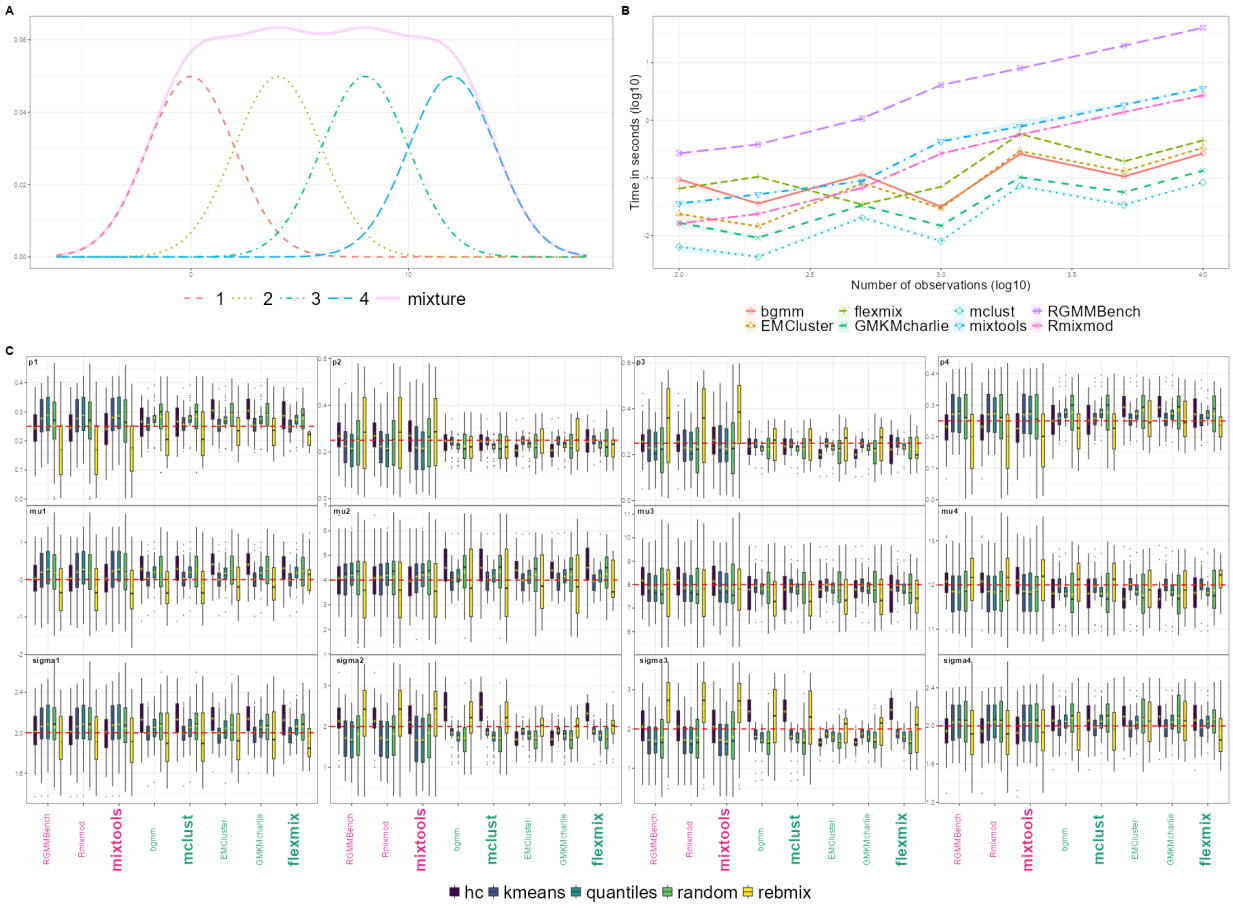


Figure 6: Benchmark summary plots of scenario U3 in Table 5 (balanced and overlapping components), with same layout as in Figure 4.

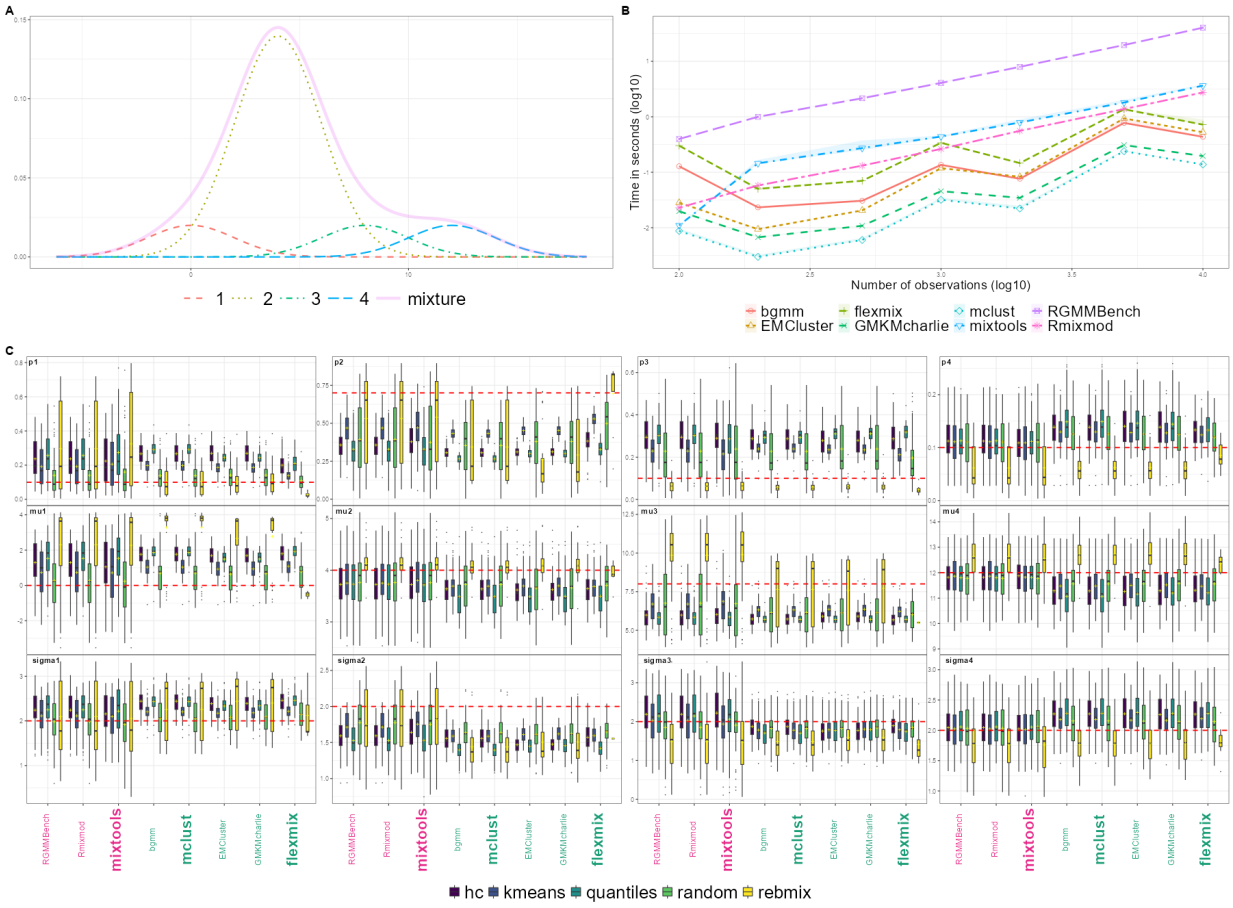


Figure 7: Benchmark summary plots of scenario U9 in Table 5 (unbalanced and overlapping components), with same layout as in Figure 4.

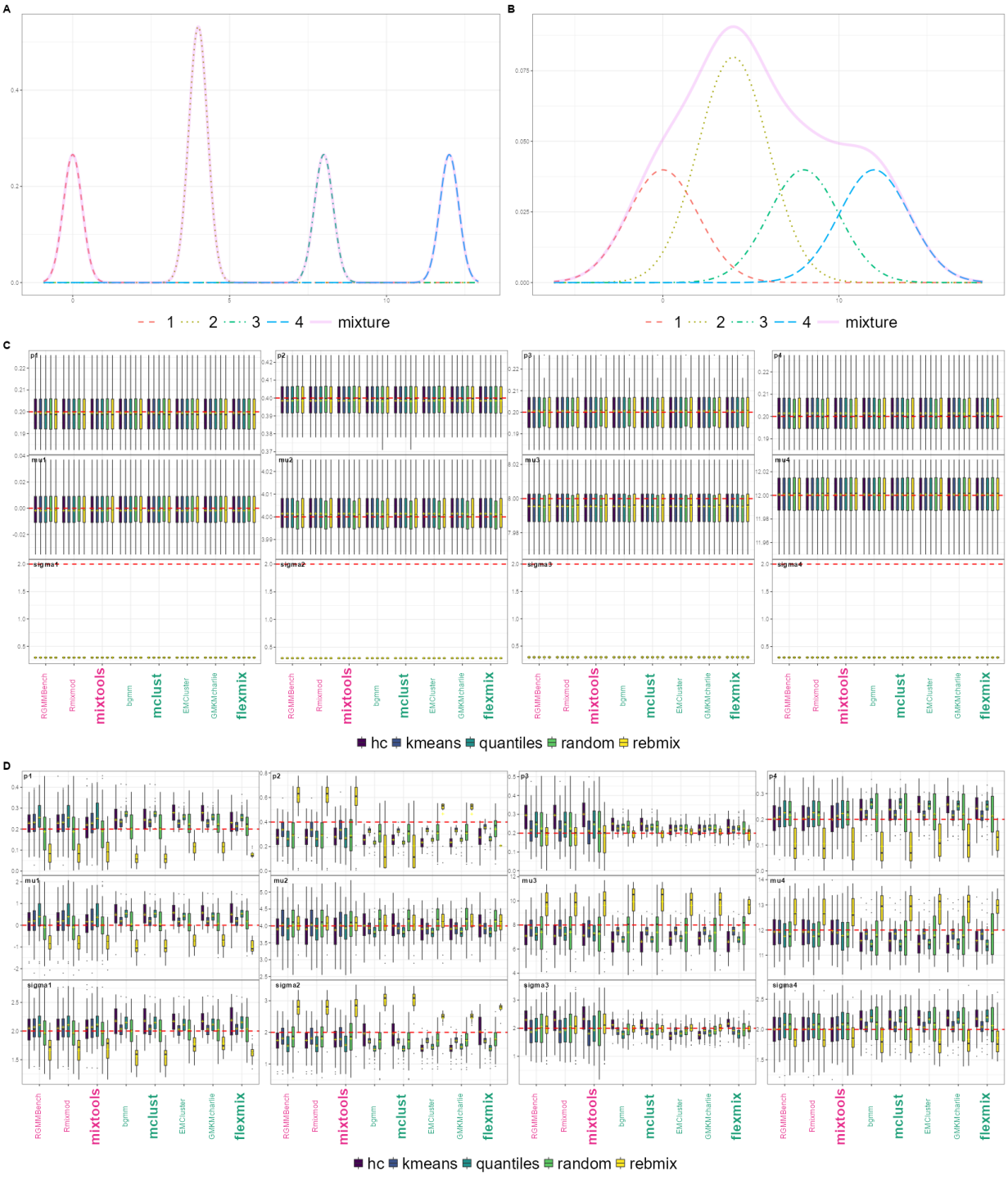


Figure 8: Benchmark summary plots of scenarios U4 and U6 in Table 5 (small unbalance, with additional overlap in scenario U6). Panel A and B display the univariate GMM distributions of respectively scenarios U4 and U6, and Panel C and D the benchmarked distributions of respectively scenarios U4 and U6, built as Panel C of Figure 4.

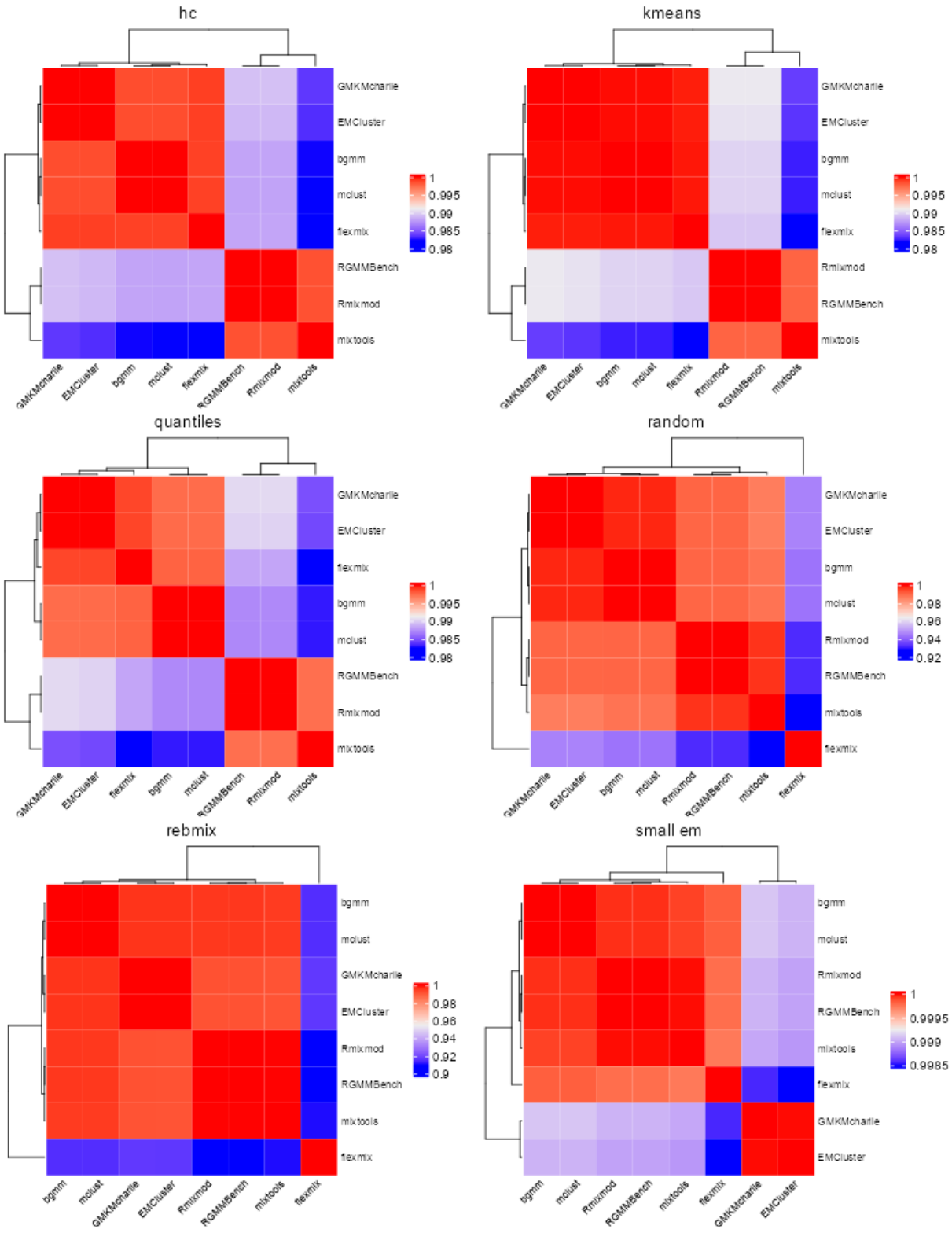


Figure 9: Correlation heatmaps of the estimated parameters extended to the four initialisation methods benchmarked, using the same configuration described in Figure (2), in the bivariate setting.

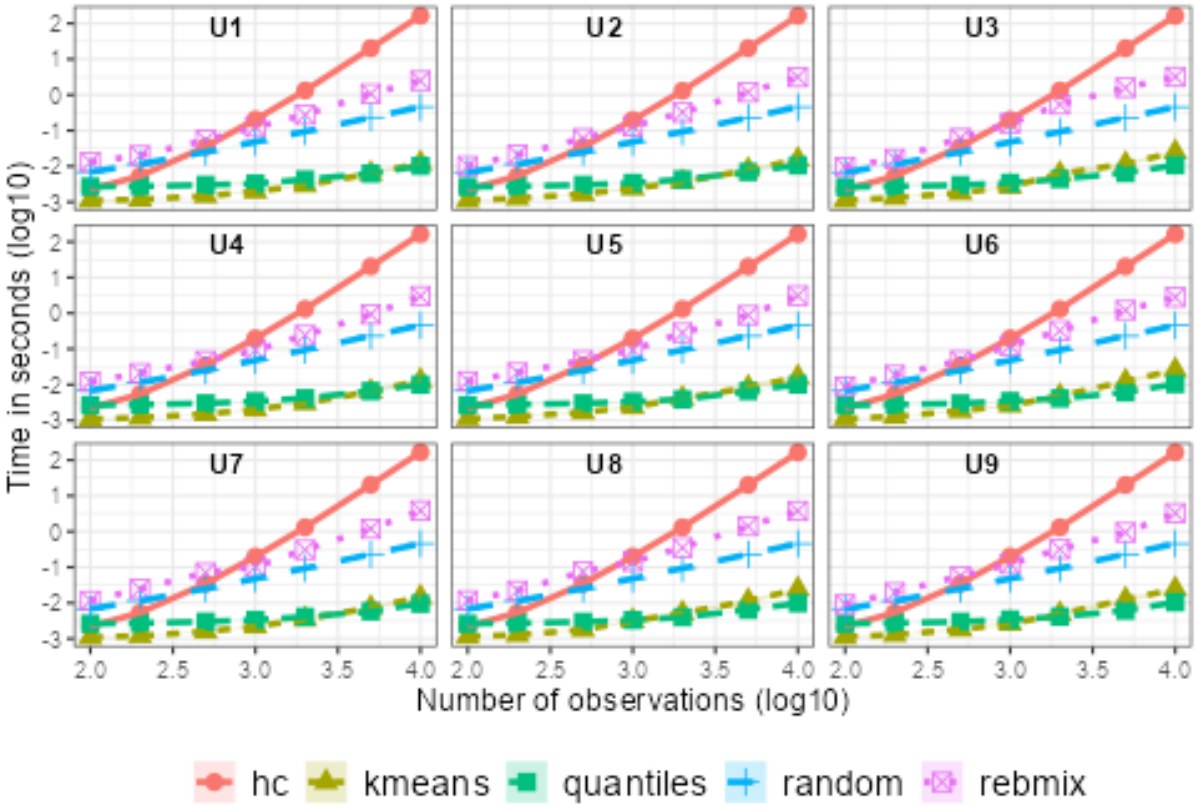


Figure 10: Distribution of the running times taken by each initialisation algorithm enumerated in Table 1, across all scenarios listed in Table 5, sorted by increasing ID number in the lexicographical order.

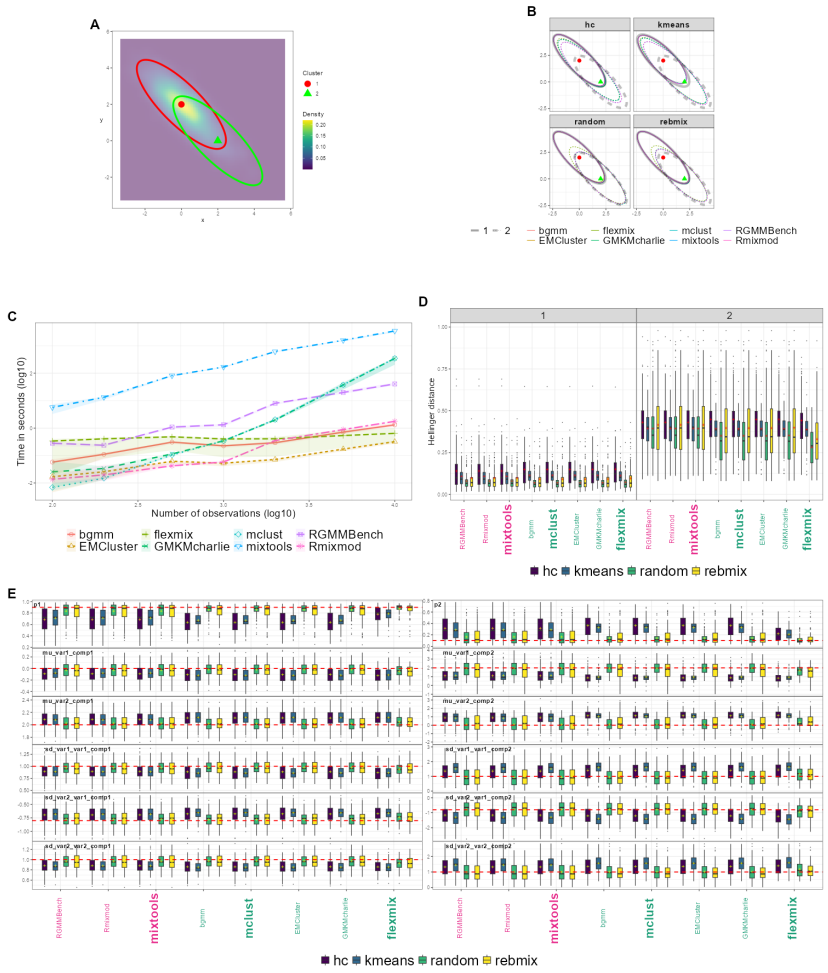


Figure 11: Results of scenario B11 in Table 10 (unbalanced, overlapping and negative correlated components), organised as such: The panel A displays the bivariate contour maps associated to the two-components multivariate Gaussian distribution corresponding to the parametrisation described by the scenario, warmer colours corresponding to regions of higher densities. The two centroids, whose coordinates are given by the mean components' elements, are represented with distinct shaped and coloured point estimates. In both Panels A and B, the ellipsoids correspond to the 95% confidence region associated to each component's distribution. To generate them, we largely inspired from the `mixtools::ellipse()` and website How to draw ellipses. To generate them, we retain for each individual parameter its mean (similar results with the median) over the $N = 100$ sampling experiments, restrained to the random initialisation method. The running times are displayed in Panel C with the k-means initialisation. The number of observations (x-axis) and the running time (y-axis) is in log(10) scale. The coloured bands represent the 5^{th} and 95^{th} percentiles of the running time. The distributions of the Hellinger distances are computed for each component, each initialisation method and each package with respect to the true Gaussian distribution expected for each component. The more dissimilar are the distributions, the higher is the Hellinger distance, knowing it is normalised between 0 and 1. We represent them using boxplot representations in Panel D. In panel E we represent the boxplots associated with the distribution of the estimates, with each column panel associated to the parameters of one component. First row represents the distribution of the estimated ratios, second and third respectively the distributions of the mean vector on the x-axis and on the y-axis, third and fourth the distributions of the individual variances of each feature and finally the fifth row shows the distribution of the correlation between dimension 1 and 2.

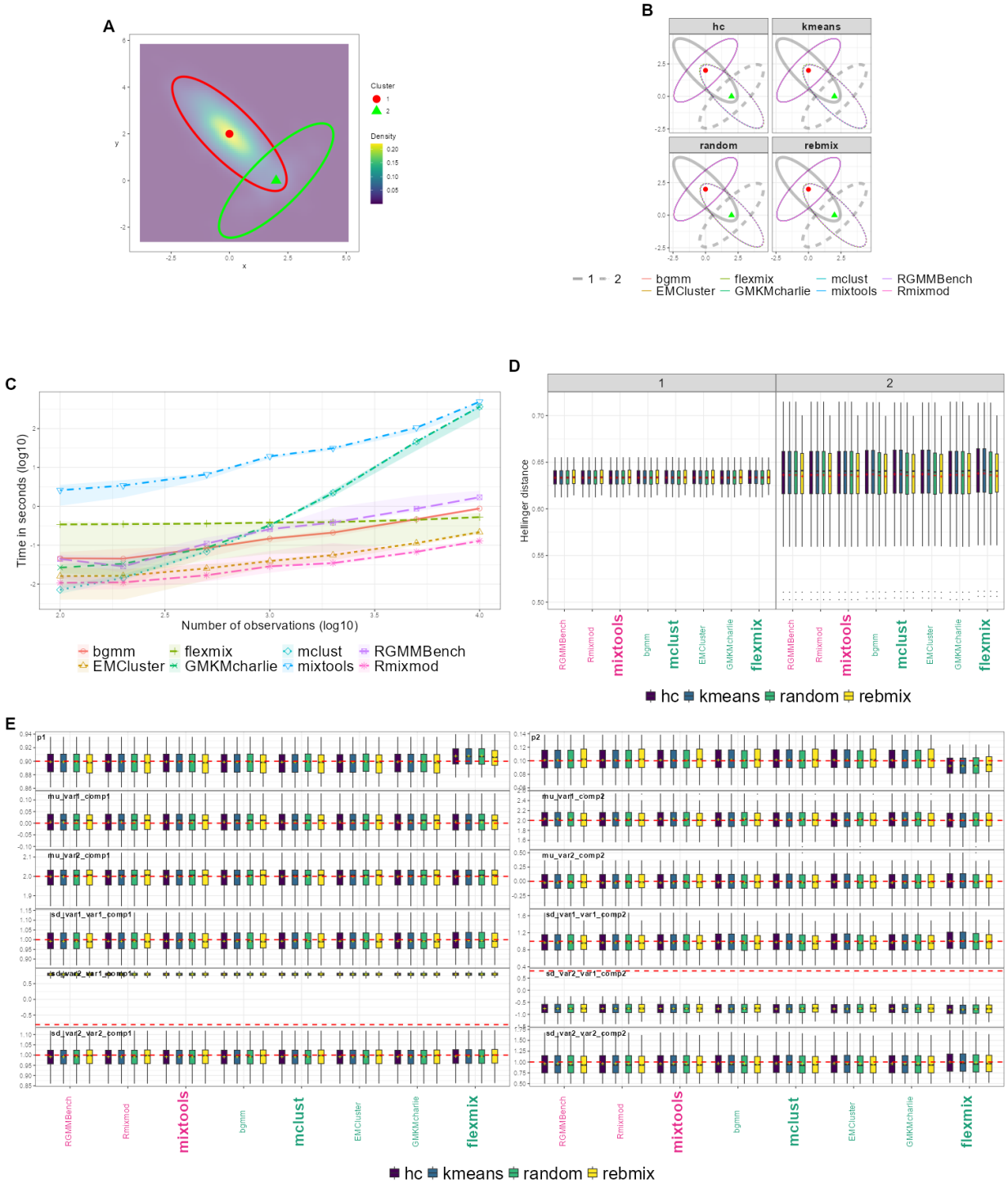


Figure 12: Results of scenario B12 in Table 10 (unbalanced, overlapping and opposite correlated components), with the same layout as Figure 11.

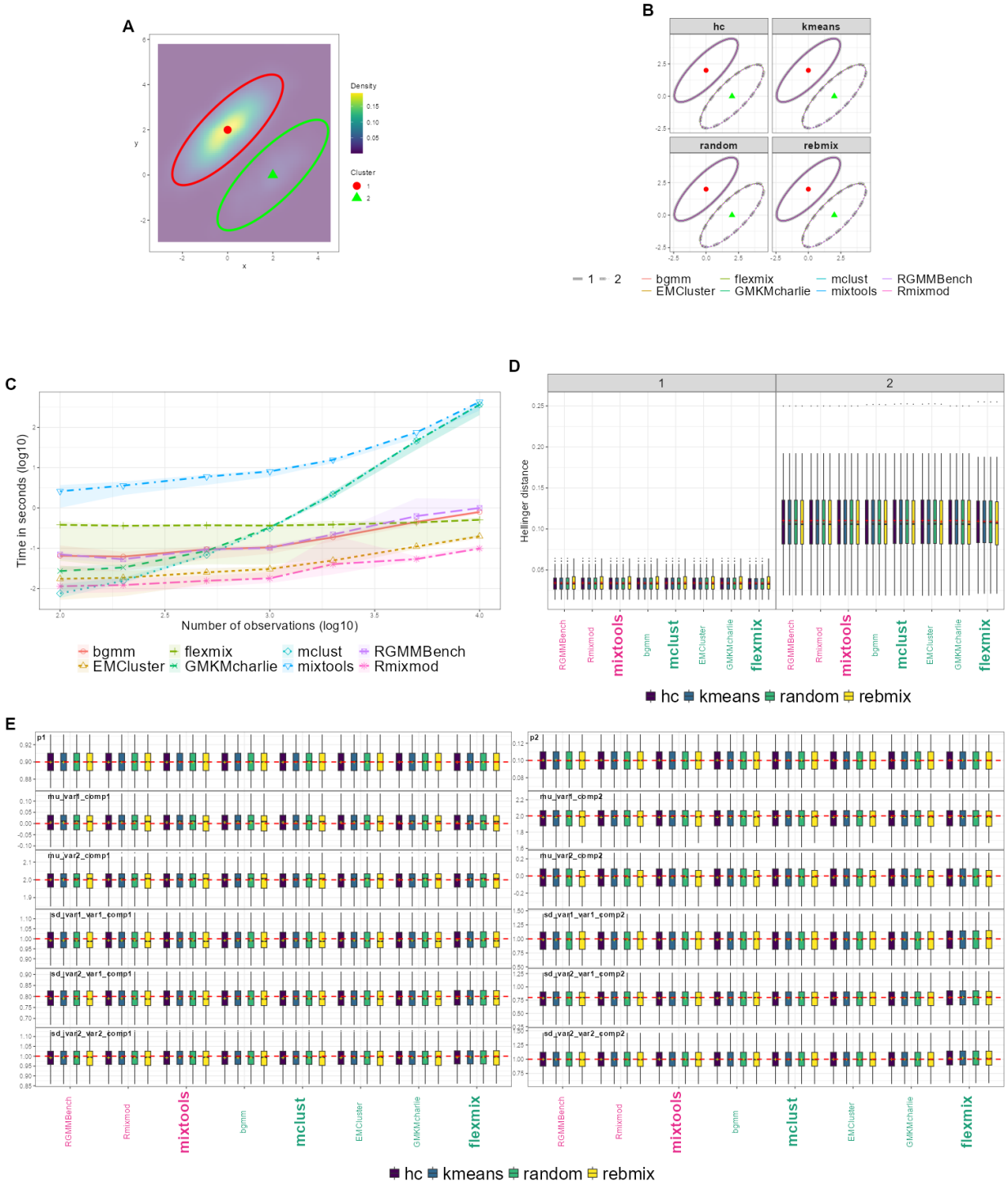


Figure 13: Results of scenario B14 in Table 10 (unbalanced, overlapping and positive correlated components), with the same layout as Figure 11.

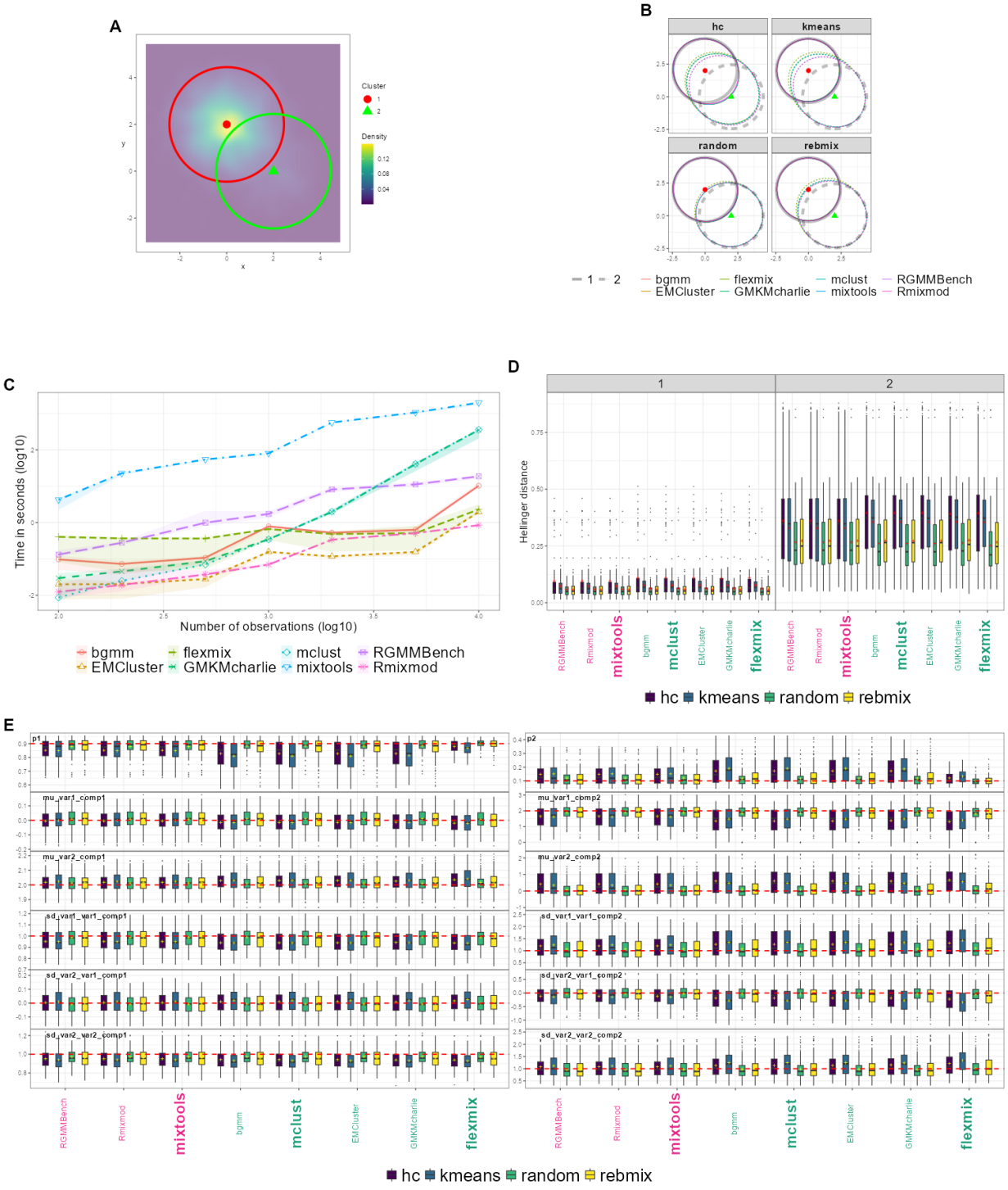


Figure 14: Results of scenario B15 in Table 10 (unbalanced, overlapping and uncorrelated components), with the same layout as Figure 11.

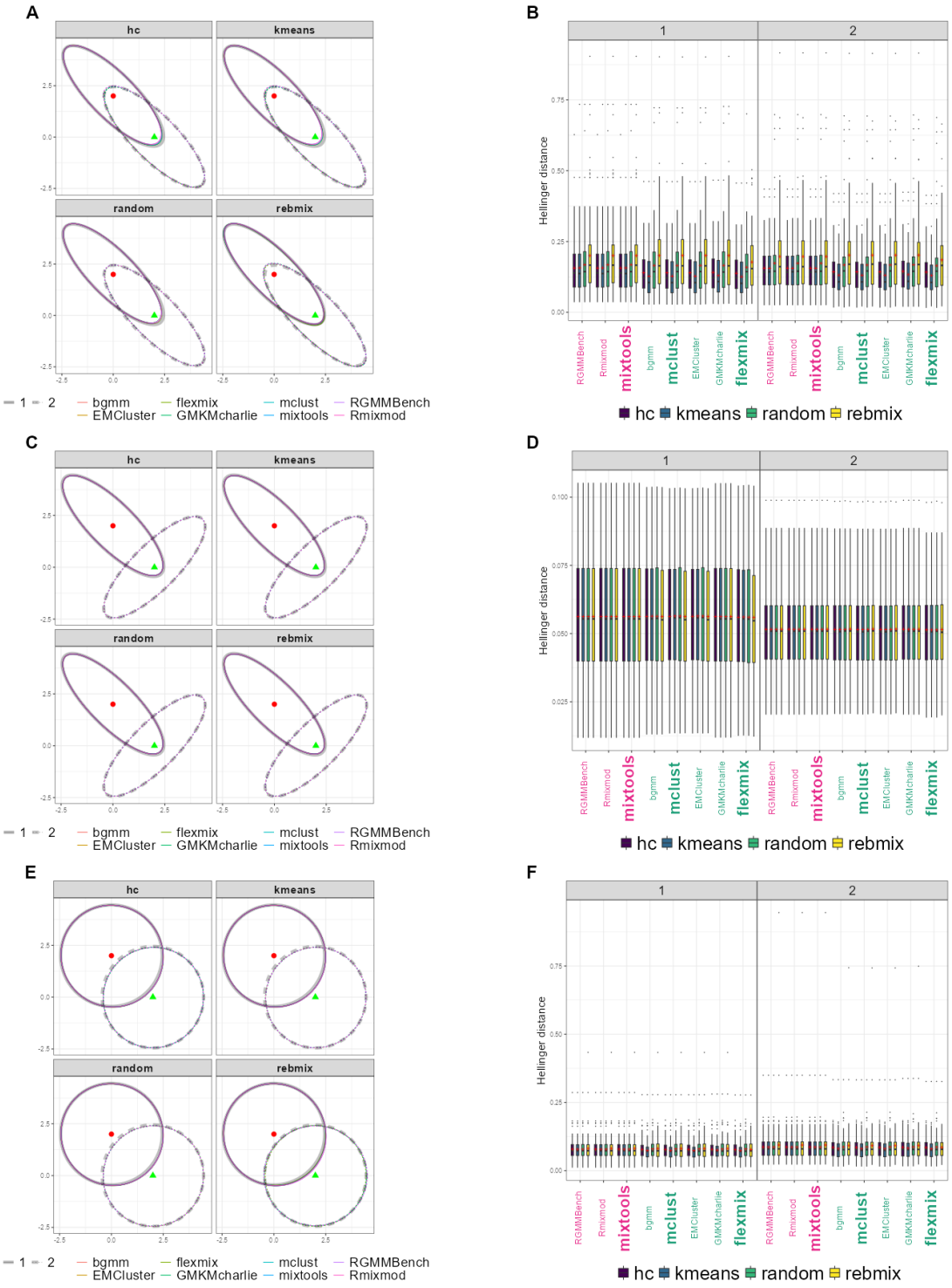


Figure 15: Benchmark summary plots of respectively scenarios B1, B2 and B5 in Table 10 featuring balanced and overlapping clusters. Summary plots of B1, B2 and B5 are represented in this order on each row, with the left column displaying the 95% confidence ellipsoidal regions associated to the mean estimated parameters across each package and the right column the distribution of the Hellinger distances.

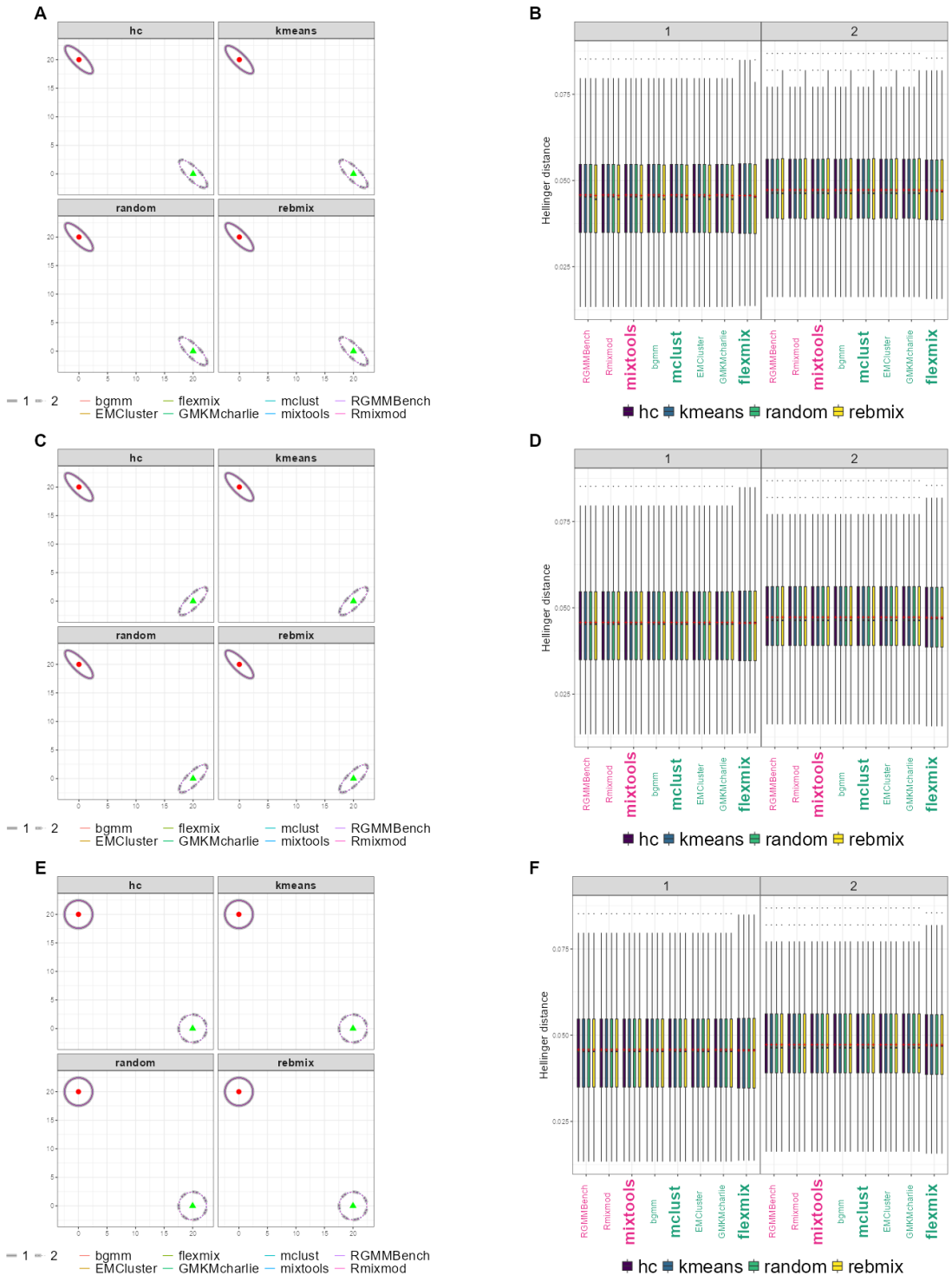


Figure 16: Benchmark summary plots of respectively scenarios B6, B7 and B10 in Table 10 featuring balanced and well-separated clusters, with the same layout as Figure 15.

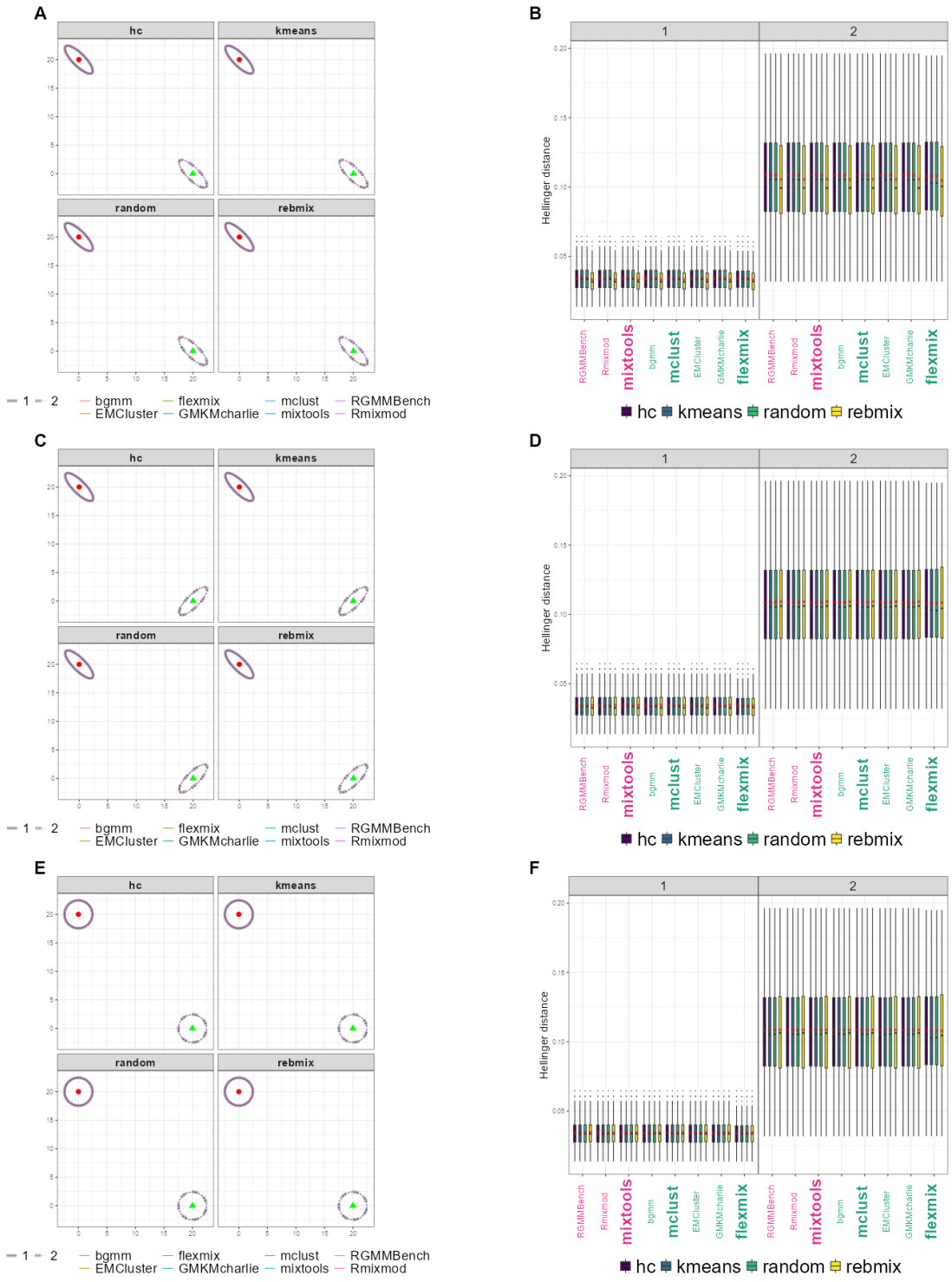


Figure 17: Benchmark summary plots of respectively scenarios B16, B17 and B20 in Table 10 featuring unbalanced and well-separated clusters, with the same layout as Figure 15.

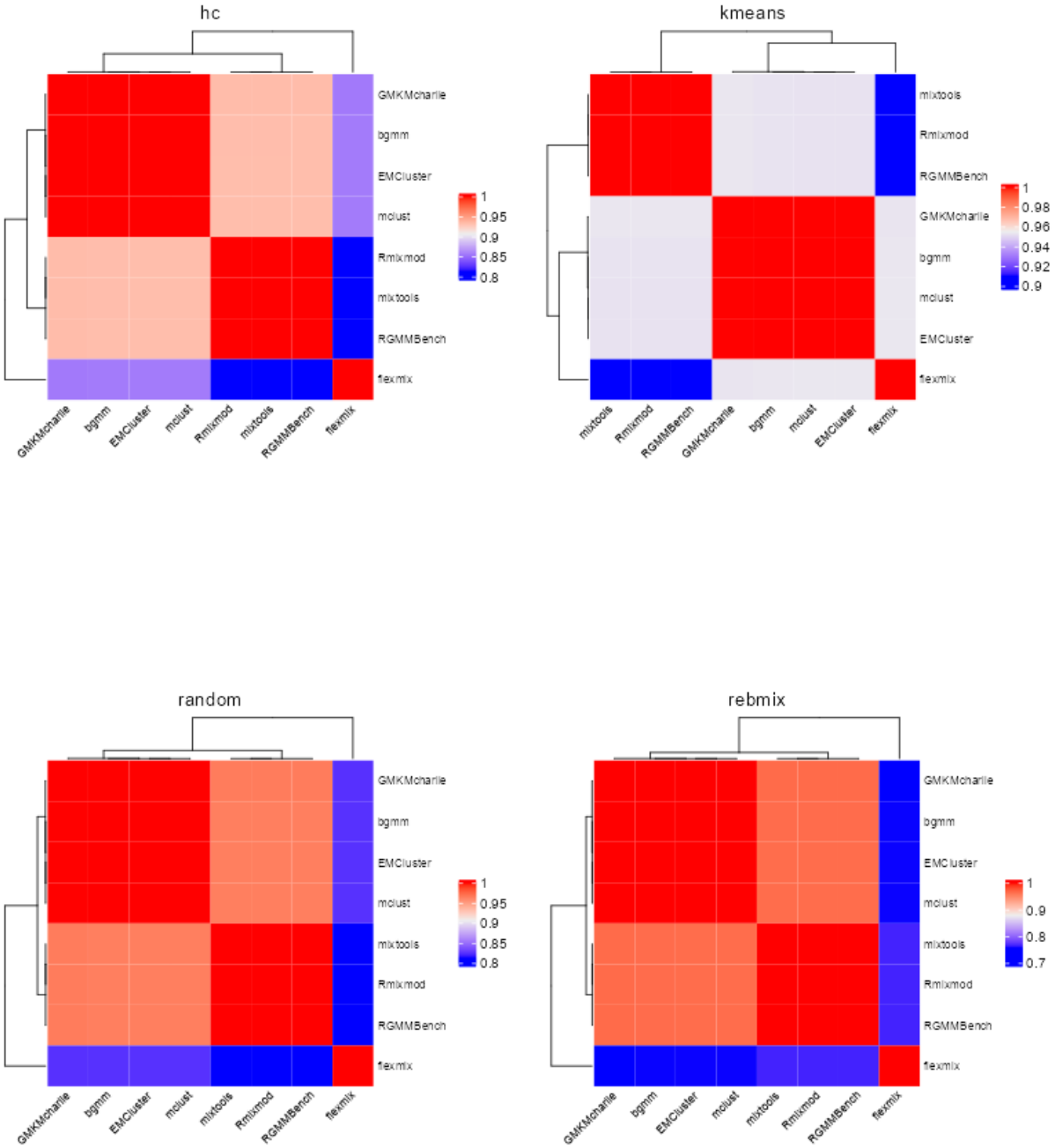


Figure 18: Correlation heatmaps of the estimated parameters in the bivariate setting extended to the four initialisation methods benchmarked, with the most discriminating scenario B11, using the same process described in Figure (2).

Table 15: The 16 parameter configurations tested to generate the samples in a high dimensional context. The first digit of each ID index refers to an unique parameter configuration (identified by its level of overlap, entropy and topological structure, either circular or ellipsoidal, of the covariance matrix, while the lowercase letter depicts the number of observations, a) with $n = 200$ and b) with $n = 2000$.

ID	OVL	Number of observations	Proportions	Spherical
HD1a	1e-04	200	0.5 / 0.5	✅
HD1b	1e-04	2000	0.5 / 0.5	✅
HD2a	1e-04	200	0.19 / 0.81	✅
HD2b	1e-04	2000	0.19 / 0.81	✅
HD3a	1e-04	200	0.5 / 0.5	❌
HD3b	1e-04	2000	0.5 / 0.5	❌
HD4a	1e-04	200	0.21 / 0.79	❌
HD4b	1e-04	2000	0.21 / 0.79	❌
HD5a	2e-01	200	0.5 / 0.5	✅
HD5b	2e-01	2000	0.5 / 0.5	✅
HD6a	2e-01	200	0.15 / 0.85	✅
HD6b	2e-01	2000	0.15 / 0.85	✅
HD7a	2e-01	200	0.5 / 0.5	❌
HD7b	2e-01	2000	0.5 / 0.5	❌
HD8a	2e-01	200	0.69 / 0.31	❌
HD8b	2e-01	2000	0.69 / 0.31	❌

Table 16: MSE and Bias associated to scenario HD4a, in Table 15 (unbalanced, separated and ellipsoidal components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	0.0333	0.0212	0.0106	0.0020	0.056	0.097	100
	kmeans	0.0333	0.0212	0.0106	0.0020	0.056	0.097	100
	rebmix	0.3244	0.1980	0.0845	0.0720	0.395	0.535	98
mclust / flexmix / GMKMcharlie	hc	0.0333	0.0212	0.0106	0.0020	0.056	0.097	100
	kmeans	0.0333	0.0212	0.0106	0.0020	0.056	0.097	100
	rebmix	0.2553	0.1444	0.0924	0.0470	0.364	0.596	85
bgmm	hc	0.0337	0.0214	0.0107	0.0070	0.064	0.096	100
	kmeans	0.0338	0.0216	0.0106	0.0074	0.064	0.096	100
	rebmix	0.4818	0.1152	0.3442	0.0320	0.223	2.329	94
EMCluster	hc	0.0333	0.0212	0.0107	0.0023	0.056	0.096	100
	kmeans	0.0334	0.0213	0.0106	0.0018	0.056	0.096	100
	rebmix	1.5983	1.0992	0.3794	0.3100	2.018	2.575	84
HDclassif	hc	8.4062	8.3936	0.0111	0.0020	10.426	0.149	100
	kmeans	7.9407	7.9282	0.0111	0.0019	10.081	0.149	100
	rebmix	7.9803	7.9514	0.0273	0.0044	10.128	0.262	84
EMMIXmfa	hc	4.0605	3.3317	0.3357	0.6500	5.757	2.772	95
	kmeans	3.8790	3.2175	0.3372	0.5400	5.781	2.777	96
	rebmix	4.0127	3.2715	0.3337	0.5700	5.680	2.757	80

Table 17: MSE and Bias associated to scenario HD7a, in Table 15 (balanced, overlapping and ellipsoidal components).

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	5.8544	2.2153	3.5586	0.0450	2.084	6.704	100
	kmeans	5.4773	1.9490	3.4819	0.0086	2.323	6.861	100
	rebmix	6.9243	2.5185	4.2898	0.0620	2.670	7.626	97
mclust / flexmix / GMKMcharlie	hc	6.0584	2.4737	3.5198	0.0180	2.565	7.624	100
	kmeans	5.6388	2.1597	3.4549	0.0140	2.744	8.266	100
	rebmix	6.5397	2.4738	3.9661	0.0700	2.774	7.764	93
bgmm	hc	9.5015	5.1348	4.1086	0.1000	3.720	10.310	100
	kmeans	8.7930	4.7119	3.8693	0.1500	3.932	10.108	100
	rebmix	10.3630	5.6474	4.4026	0.2700	3.798	10.049	97
EMCluster	hc	6.4022	2.8255	3.5124	0.0120	3.141	9.086	100
	kmeans	6.4333	2.8740	3.5523	0.0110	4.210	11.007	100
	rebmix	6.5527	2.9643	3.4862	0.0580	3.051	9.253	93
HDclassif	hc	15.9010	11.5382	4.2950	0.1400	10.846	10.100	100
	kmeans	15.3377	10.9441	4.3716	0.0087	10.990	10.771	100
	rebmix	16.1231	11.1103	4.9113	0.1600	10.761	10.513	93
EMMIXmfa	hc	4.8606	1.6546	3.1856	0.0160	2.030	7.395	15
	kmeans	4.4039	1.4129	2.9701	0.0260	1.734	6.236	21
	rebmix	5.0984	2.0057	3.0689	0.0470	2.314	7.613	16

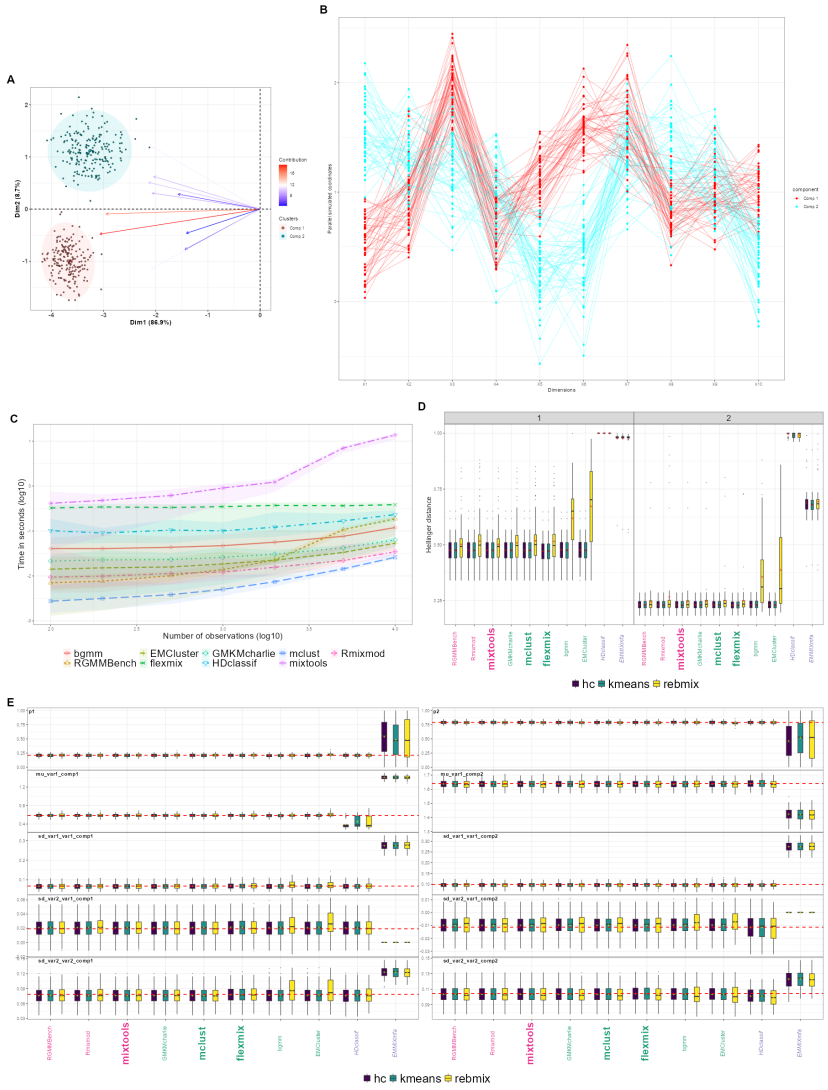


Figure 19: Results of scenario HD4a) in Table 15 (unbalanced, overlapping and negative correlated components), organised as such: The panel A displays the bivariate factorial projection of a random sample drawn from the 10-dimensional multivariate Gaussian distribution parametrised by Table 15. Each component is associated to a specific color, a centroid whose coordinates are given by the mean components' elements in the bivariate projected space and a 95% confidence ellipse. Arrows represent the correlation circle of the dimensional variables. Both panels were displayed respectively using functions `factoextra::fviz_eig` and `factoextra::fviz_pca_biplot` while the underlying computations proceed from the principal component analysis performed by `ade4::dudi.pca` preceded by standard scaling of the sampling dataset. The panel B pictures the *parallel distribution plots* from a random sampling of $n = 100$ observations, generated using `GGally::ggparcoord`, and representing the coordinates of each simulated data point in 10 dimensions. The running times are displayed in Panel C with the k-means initialisation. The number of observations (x-axis) and the running time (y-axis) is in log(10) scale. The distributions of the Hellinger distances are computed for each component in Panel D, each initialisation method and each package with respect to the true Gaussian distribution expected for each component. In panel E we represent the boxplots associated with the distribution of some of the estimates. Since it was impractical to represent all of the $k + kD + k\frac{D\times(D+1)}{2}$ with $k = 2$ and $D = 10$ parameters, we only represent the first component's mean, two first components' variances and their covariance term.

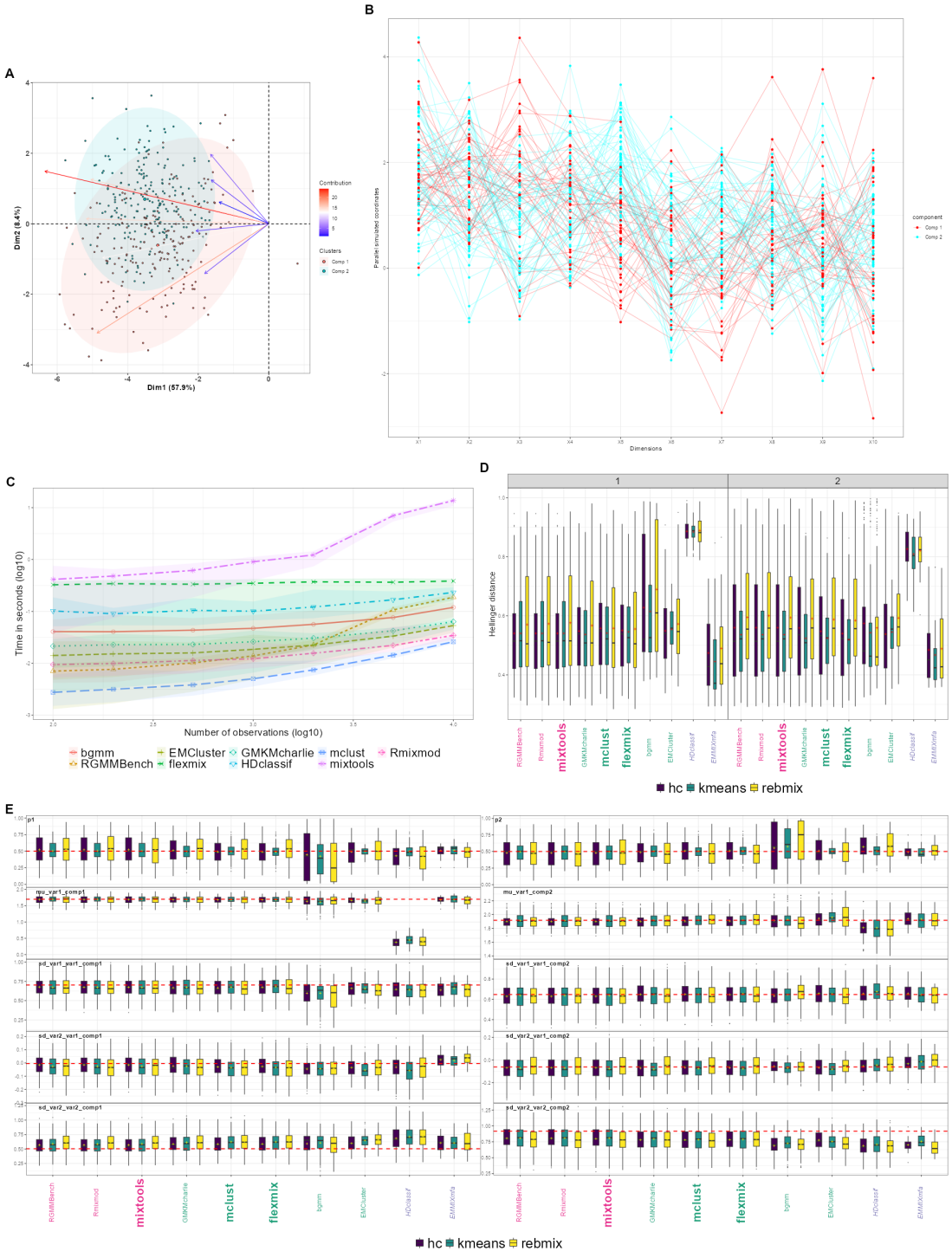


Figure 20: Results of scenario HD7a) in Table 15 (balanced and overlapping components, with full covariance structure), with the same layout as Figure 19.

Table 18: MSE and Bias associated to scenarios HD5a) and HD6a), in Table 15 (overlapping and spherical-distributed components). We delimite each scenario by doubled backslashes with respectively balanced and unbalanced clusters.

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	4.2772 // 19.5172	0.9198 // 1.9835	3.3027 // 17.4943	0.017 // 0.069	0.995 // 0.6	3.571 // 4.381	100 // 100
	kmeans	3.9776 // 17.2212	0.8279 // 1.6336	3.1111 // 15.5684	0.072 // 0.069	0.841 // 0.82	3.023 // 5.034	100 // 100
	rebmix	9.3136 // 25.8028	2.7793 // 4.2893	6.4009 // 21.3519	0.15 // 0.22	3.619 // 2.507	9.061 // 11.826	96 // 80
mclust / flexmix / GMKMcharlie	hc	2.9743 // 18.1175	0.5862 // 1.7729	2.3612 // 16.3168	0.024 // 0.057	0.449 // 0.514	2.127 // 4.412	100 // 100
	kmeans	2.5629 // 15.2959	0.4642 // 1.5608	2.0855 // 13.7206	0.085 // 0.086	0.671 // 1.047	1.67 // 5.801	100 // 100
	rebmix	8.2907 // 23.7588	2.6468 // 4.1629	5.5421 // 19.4579	0.12 // 0.22	3.438 // 2.543	8.792 // 11.94	96 // 69
bgmm	hc	2.4088 // 33.8392	0.7261 // 9.0609	1.6153 // 24.6796	0.12 // 0.038	0.652 // 1.986	1.98 // 10.77	100 // 100
	kmeans	2.0912 // 28.5103	0.5899 // 7.5426	1.4577 // 20.8989	0.091 // 0.025	0.566 // 1.45	1.738 // 9.783	100 // 100
	rebmix	4.6278 // 35.9294	1.9526 // 11.0184	2.5372 // 24.6276	0.048 // 0.22	0.632 // 2.023	2.96 // 12.729	98 // 86
EMCluster	hc	2.5152 // 17.7053	0.5087 // 2.1191	1.9849 // 15.5379	0.024 // 0.12	0.321 // 0.929	1.512 // 5.611	100 // 100
	kmeans	1.793 // 12.8799	0.3527 // 1.6839	1.4344 // 11.155	0.062 // 0.24	0.593 // 2.177	2.547 // 9.595	100 // 100
	rebmix	6.9275 // 23.0817	2.7461 // 5.4713	4.0985 // 17.4511	0.044 // 0.32	3.177 // 3.836	8.535 // 15.437	96 // 70
HDclassif	hc	11.4938 // 49.4328	9.1746 // 12.2155	2.2913 // 36.5886	0.027 // 0.91	8.899 // 9.56	1.98 // 19.55	100 // 100
	kmeans	11.1438 // 40.4749	9.0384 // 11.9946	2.0912 // 28.0385	0.096 // 0.7	9.059 // 9.024	1.682 // 16.35	100 // 100
	rebmix	14.6998 // 47.2364	8.7649 // 12.6715	5.8029 // 33.929	0.22 // 0.92	8.135 // 9.145	8.018 // 21.824	96 // 70
EMMIXmfa	hc	5.6809 // 21.1181	3.7272 // 6.1206	1.7452 // 14.9126	0.41 // 0.019	5.772 // 3.645	4.299 // 12.812	96 // 45
	kmeans	5.7063 // 21.3775	3.6759 // 6.589	1.79 // 14.5681	0.39 // 0.17	5.788 // 4.08	4.357 // 13.352	96 // 40
	rebmix	5.8175 // 19.9703	3.8142 // 6.3202	1.7592 // 13.5389	0.35 // 0.033	5.819 // 4.402	4.349 // 13.812	93 // 34

Table 19: Minimal example setting apart MSE and Bias whether it proceeds from diagonal or offset terms of the covariance matrix, for scenarios HD5a) and HD6a), in Table 15 (overlapping and spherical-distributed components). We delimite each scenario by doubled backslashes with respectively balanced and unbalanced clusters.

Package	Initialisation Method	Global MSE diag(Σ)	Global MSE upper.tri(Σ)	Global Bias diag(Σ)	Global Bias upper.tri(Σ)
mixtools / Rmixmod / RGMMBench	hc	1.1 // 5.9	2.2 // 12	0.9194 // 2.3003	2.651 // 2.081
	kmeans	0.99 // 5.6	2.1 // 10	0.8929 // 2.7422	2.13 // 2.292
mclust / flexmix / GMKMcharlie	hc	0.76 // 5.5	1.6 // 11	0.5698 // 2.418	1.557 // 1.994
	kmeans	0.67 // 5.2	1.4 // 8.5	0.6909 // 3.4316	0.979 // 2.37
bgmm	hc	0.67 // 11	0.94 // 14	0.7755 // 6.9204	1.205 // 3.849
	kmeans	0.58 // 9.1	0.88 // 12	0.6004 // 6.124	1.138 // 3.659
EMCluster	hc	0.62 // 6.1	1.4 // 9.5	0.4985 // 3.4685	1.013 // 2.143
	kmeans	0.48 // 7.5	0.95 // 3.6	0.9269 // 7.6352	1.621 // 1.96
HDclassif	hc	0.72 // 26	1.6 // 11	0.5383 // 17.2225	1.441 // 2.328
	kmeans	0.68 // 20	1.4 // 8.5	0.7156 // 13.7249	0.966 // 2.626
EMMIXmfa	hc	1.6 // 10	0.13 // 4.6	3.7632 // 10.0798	0.536 // 2.733
	kmeans	1.6 // 11	0.17 // 3.9	3.7621 // 10.8057	0.594 // 2.546

Table 20: MSE and Bias associated to scenarios HD1a) and HD1b), in Table 15 (well-separated and spherical-distributed components). We delimite by doubled backslashes for each entry of the summary metrics table respectively the scores with $n = 200$ and $n = 2000$ observations.

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	rebmix	0.5611 // 0.0058	0.2364 // 0.0028	0.3035 // 0.0026	0.019 // 0.00071	0.372 // 0.018	0.915 // 0.04	98 // 100
mclust / flexmix / GMKMcharlie	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0095 // 0.00071	0.053 // 0.018	0.14 // 0.04	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0095 // 0.00071	0.053 // 0.018	0.14 // 0.04	100 // 100
	rebmix	0.3134 // 0.0058	0.1305 // 0.0029	0.1729 // 0.0026	0.0039 // 0.0022	0.2 // 0.02	0.537 // 0.044	88 // 81
bgmm	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0098 // 0.00071	0.053 // 0.018	0.139 // 0.041	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	rebmix	0.7437 // 0.1977	0.3409 // 0.0028	0.3895 // 0.1946	0.02 // 0.00034	0.308 // 0.017	1.602 // 0.926	97 // 99
EMCluster	hc	0.0577 // 0.0058	0.0289 // 0.0028	0.0264 // 0.0026	0.0093 // 0.00061	0.054 // 0.018	0.139 // 0.041	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0092 // 0.00039	0.054 // 0.017	0.139 // 0.04	100 // 100
	rebmix	0.6887 // 0.3391	0.2466 // 0.1276	0.4201 // 0.1969	0.019 // 0.048	0.519 // 0.289	1.721 // 0.875	87 // 81
HDclassif	hc	11.3787 // 11.3322	11.3499 // 11.3293	0.0264 // 0.0026	0.0094 // 0.00071	11.166 // 11.179	0.139 // 0.04	100 // 100
	kmeans	11.3831 // 11.3301	11.3543 // 11.3271	0.0264 // 0.0026	0.0094 // 0.00068	11.162 // 11.181	0.139 // 0.04	100 // 100
	rebmix	11.5949 // 11.6085	11.5167 // 11.6055	0.072 // 0.0026	0.0024 // 0.0022	11.227 // 11.369	0.27 // 0.044	87 // 81
EMMIXmfa	hc	5.9739 // 5.9149	4.0397 // 3.9727	1.8288 // 1.8228	0.32 // 0.47	6.999 // 7.042	4.25 // 4.296	100 // 100
	kmeans	5.972 // 5.8863	4.0431 // 3.9596	1.8283 // 1.8259	0.33 // 0.43	6.997 // 7.051	4.255 // 4.34	100 // 100
	rebmix	5.9835 // 5.9078	4.0477 // 3.9671	1.8257 // 1.8244	0.37 // 0.46	6.994 // 7.045	4.232 // 4.301	87 // 81

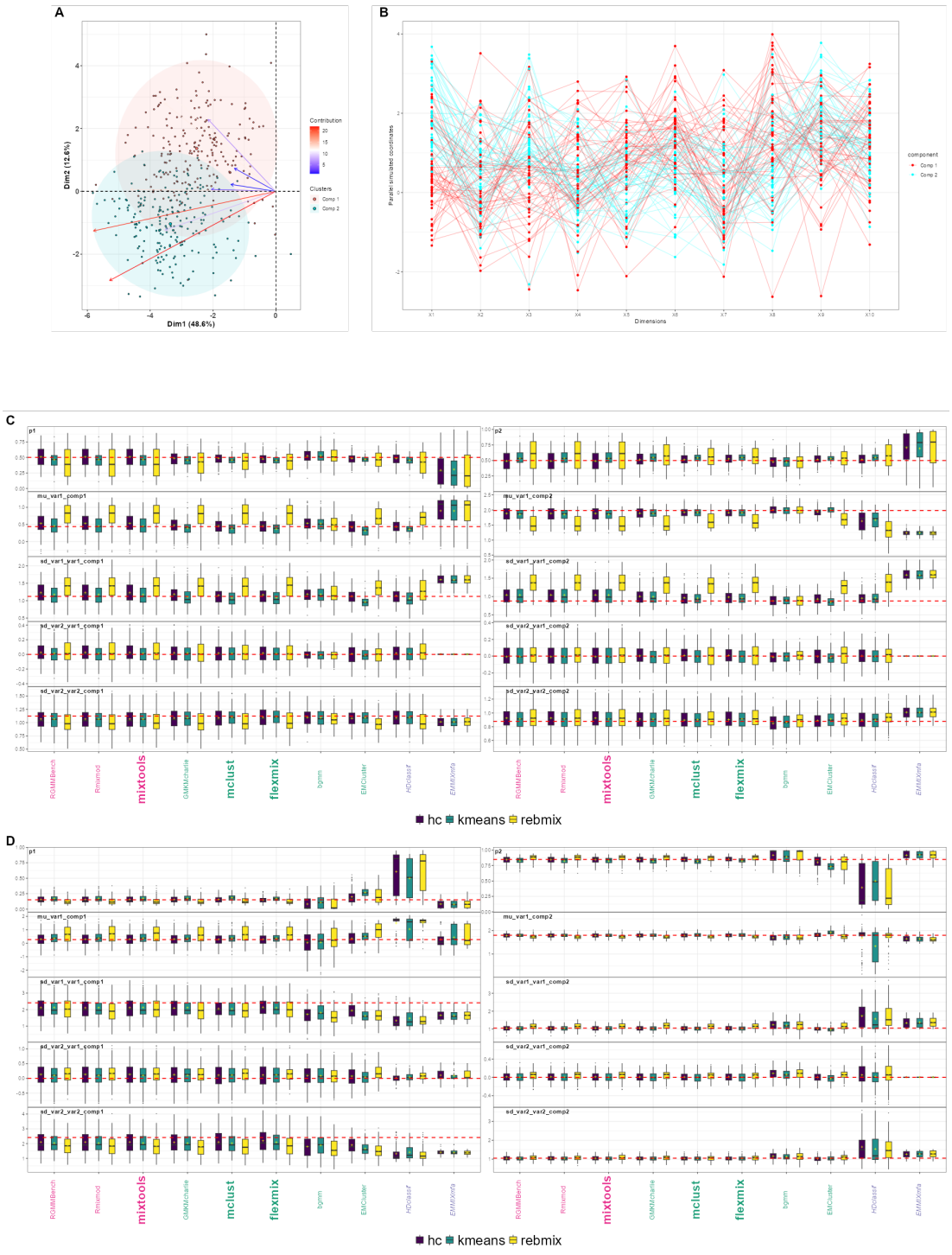


Figure 21: We gathered on the same plot two multivariate benchmark scenarios, in which we consider a strictly spherical structure of the covariance matrix: We represent in Panel A and B, respectively the bivariate projection and parallel distribution plot, associated to scenario HD5a) in Table 15 (balanced and overlapping components, with spherical covariance structure). In Panel C, we display the boxplots associated to scenario HD5a), computing them similarly as in Panel E of Figure 19. In Panel D, we display the boxplots associated to scenario HD6a) (unbalanced and overlapping components, with spherical covariance structure).

Table 21: MSE and Bias associated to scenarios HD8a) and HD8b), in Table 15 (overlapping components with full covariance structure). We delimite by doubled backslashes for each entry of the summary metrics table respectively the scores with $n = 200$ and $n = 2000$ observations.

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	18.6085 // 0.6735	3.566 // 0.0536	14.9495 // 0.6193	0.23 // 0.0017	3.327 // 0.107	14.475 // 0.649	100 // 100
	kmeans	16.7452 // 0.6735	2.9065 // 0.0536	13.7662 // 0.6193	0.2 // 0.0016	2.819 // 0.107	12.552 // 0.649	100 // 100
	rebmix	22.2986 // 0.6738	4.3127 // 0.0536	17.8418 // 0.6196	0.25 // 0.0021	3.768 // 0.108	16.249 // 0.648	95 // 100
mclust / flexmix / GMKMcharlie	hc	20.6916 // 0.728	4.5672 // 0.0696	16.0328 // 0.6557	0.28 // 0.064	4.381 // 0.459	18.656 // 2.07	100 // 100
	kmeans	17.9622 // 0.7169	3.7547 // 0.0671	14.1405 // 0.6474	0.27 // 0.062	3.88 // 0.465	16.802 // 2.049	100 // 100
	rebmix	22.4636 // 0.7553	4.7502 // 0.0678	17.5784 // 0.6853	0.26 // 0.0054	4.165 // 0.158	17.735 // 0.725	94 // 98
bgmm	hc	35.6085 // 13.8411	12.8826 // 3.6502	22.3428 // 10.0718	0.29 // 0.46	6.212 // 5.661	26.812 // 23.753	100 // 100
	kmeans	33.8007 // 12.5545	11.7236 // 3.1419	21.7292 // 9.2934	0.28 // 0.47	6.348 // 5.654	26.546 // 23.141	100 // 100
	rebmix	35.3167 // 13.106	12.2374 // 3.3747	22.6615 // 9.6213	0.37 // 0.42	6.007 // 5.273	26.287 // 23.02	96 // 100
EMCluster	hc	23.4472 // 16.4192	6.2451 // 4.9191	17.0777 // 11.3124	0.35 // 0.51	5.469 // 6.279	23.503 // 22.821	99 // 100
	kmeans	21.0058 // 14.6852	5.9951 // 4.1684	14.9329 // 10.4293	0.38 // 0.42	5.604 // 6.592	24.628 // 24.706	100 // 100
	rebmix	23.0408 // 19.7372	6.4923 // 7	16.3824 // 12.5099	0.36 // 0.35	5.272 // 5.454	23.419 // 23.9	93 // 98
HDclassif	hc	36.5924 // 33.4077	16.108 // 14.4007	20.3809 // 18.8638	0.3 // 0.44	12.706 // 13.085	25.393 // 29.363	100 // 100
	kmeans	34.7935 // 30.1935	15.4329 // 13.78	19.2529 // 16.2816	0.4 // 0.41	12.766 // 12.756	24.988 // 25.151	100 // 100
	rebmix	38.9707 // 24.0327	16.134 // 12.9266	22.6961 // 11.0138	0.25 // 0.21	12.811 // 12.275	25.79 // 15.996	95 // 98

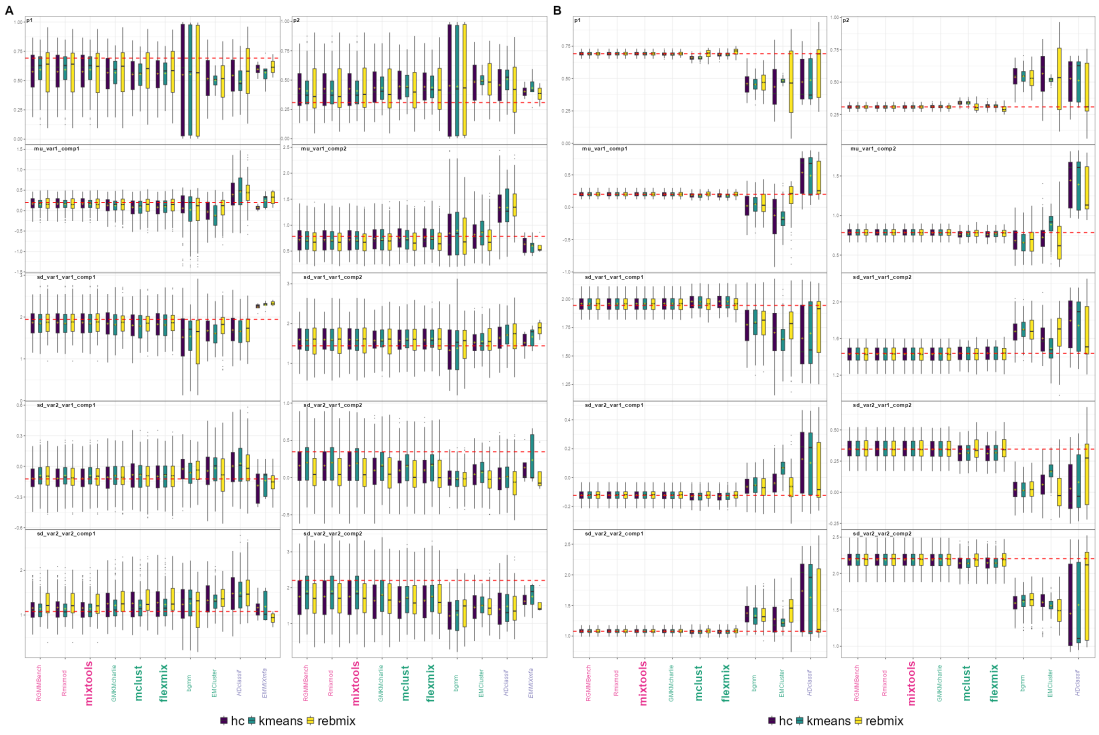


Figure 22: Overview of scenarios HD1 a) and b) and HD8 a) and b) in Table 15comparing the performance of the algorithms in respectively the easiest and most complex scenario. The left-hand column shows box plots of the estimated parameters from simulations with $n = 200$ observations on the left and $n = 2000$ observations on the right.

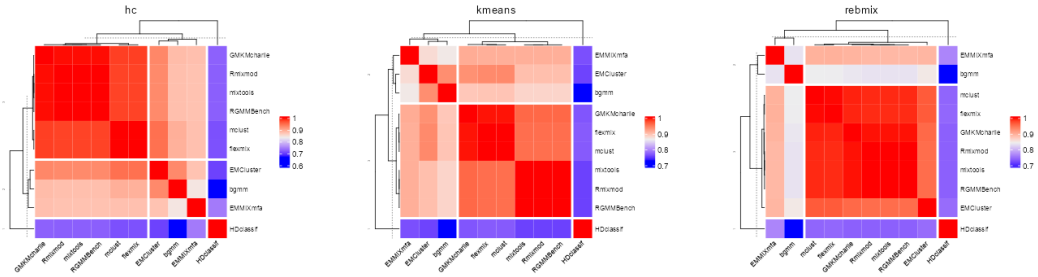


Figure 23: Correlation heatmaps of the estimated parameters in the high dimensional (HD) setting extended to the three initialisation methods benchmarked (respectively hc, k-means and rebmix) in the most discriminating scenario HD8a), using the same process described in Figure (2).

Supplementary bibliography

Alberich-Carramiñana, Maria, Borja Elizalde, and Federico Thomas. 2017. "New Algebraic Conditions for the Identification of the Relative Position of Two Coplanar Ellipses." *Computer Aided Geometric Design*. https://doi.org/10.1016/j.cagd.2017.03.013.

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2023. *rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Avila, Luis M., Michael R. May, and Jeff Ross-Ibarra. 2018. *DPP: Inference of Parameters of Normal Distributions from a Mixture of Normals*.

Bacci, Silvia, Silvia Pandolfi, and Fulvia Pennoni. 2012. "A Comparison of Some Criteria for States Selection in the Latent Markov Model for Longitudinal Data." http://arxiv.org/abs/1212.0352.

Baker, Peter. 2018. *polySegratioMM: Bayesian Mixture Models for Marker Dosage in Autopolyploids*.

Banfield, Jeffrey D., and Adrian E. Raftery. 1993. "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics*. https://doi.org/10.2307/2532201.

Basford, K., D. Greenway, G. Mclachlan, et al. 1997. "Standard Errors of Fitted Component Means of Normal Mixtures." *Computational Statistics*. https://www.researchgate.net/publication/37625647/_Standard/_errors/_of/_fitted/_component/_means/_of/_normal/_mixtures.

Benaglia, Tatiana, Didier Chauveau, David R. Hunter, et al. 2009. "mixtools: An R Package for Analyzing Finite Mixture Models." *Journal of Statistical Software*.

Bergé, Laurent, Charles Bouveyron, and Stéphane Girard. 2012. "HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v046.i06.

Biernacki, Christophe, Gilles Celeux, and Gerard Govaert. 2000. "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. https://doi.org/10.1109/34.865189.

Bobb, Jennifer F., and Ravi Varadhan. 2021. *turboEM: A Suite of Convergence Acceleration Schemes for EM, MM and Other Fixed-Point Algorithms*.

Bouveyron, Charles, Gilles Celeux, and Stéphane Girard. 2011. "Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA." *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2011.07.017.

Bouveyron, Charles, Stéphane Girard, and Cordelia SCHMID. 2007. "High-Dimensional Data Clustering." *Computational Statistics & Data Analysis*. https://doi.org/10.1016/j.csda.2007.02.009.

Browne, Ryan P., and Paul D. McNicholas. 2014. "Estimating Common Principal Components in High Dimensions." *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-013-0139-1.

Cao, Sha, Wennan Chang, and Chi Zhang. 2020. *RobMixReg: Robust Mixture Regression*.

Cao, X., and J. C. Spall. 2012. "Relative Performance of Expected and Observed Fisher Information in Covariance Estimation for Maximum Likelihood Estimates." https://doi.org/10.1109/ACC.2012.6315584.

Cattell, Raymond B. 1966. "The Scree Test For The Number Of Factors." *Multivariate Behavioral Research*. https://doi.org/10.1207/s15327906mbr0102/_10.

Celeux, Gilles, Stéphane Chrétien, and Florence Forbes. 2012. "A Component-wise EM Algorithm for Mixtures."

Celeux, Gilles, Sylvia Fruewirth-Schnatter, and Christian P. Robert. 2018. "Model Selection for Mixture Models - Perspectives and Strategies." arXiv. https://doi.org/10.48550/ARXIV.1812.09885.

Celeux, Gilles, and Gérard Govaert. 1992. "A Classification EM Algorithm for Clustering and Two Stochastic Versions." *Computational Statistics & Data Analysis*. https://doi.org/10.1016/0167-9473(92)90042-E.

Chen, J., and Z. Chen. 2008. "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces." *Biometrika*. https://doi.org/10.1093/biomet/asn034.

Chen, Wei-Chen, and George Ostrouchov. 2021. *Pmclust: Parallel Model-Based Clustering Using Expectation-Gathering-Maximization Algorithm for Finite Mixture Gaussian Model*.

Clark, Katharine M., and Paul D. McNicholas. 2019. *Oclust: Gaussian Model-Based Clustering with Outliers*.

Coretto, Pietro, and Christian Hennig. 2016. "Consistency, Breakdown Robustness, and Algorithms for Robust Improper Maximum Likelihood Clustering." *Journal of Machine Learning Research*.

———. 2021. *Otrimle: Robust Model-Based Clustering.*

Csárdi, Gábor. 2019. *Cranlogs: Download Logs from the RStudio 'CRAN' Mirror.* https://CRAN.R-project.org/package=cranlogs.

Dai, Ming, and Arunava Mukherjea. 2001. "Identification of the Parameters of a Multivariate Normal Vector by the Distribution of the Maximum." *Journal of Theoretical Probability.* https://doi.org/10.1023/A:1007889519309.

Dean, Nema, Adrian E. Raftery, and Luca Scrucca. 2020. *Clustvarsel: Variable Selection for Gaussian Model-Based Clustering.*

Delattre, Maud, and Estelle Kuhn. 2019. "Estimating Fisher Information Matrix in Latent Variable Models Based on the Score Function." https://doi.org/https://doi.org/10.48550/arXiv.1909.06094.

Efron, Bradley, and Robert Tibshirani. 1993. *An Introduction to the Bootstrap.*

Fallah, Lida, and John Hinde. 2017. *Pcensmix: Model Fitting to Progressively Censored Mixture Data.*

Figueiredo, M. A. T., and A. K. Jain. 2002. "Unsupervised Learning of Finite Mixture Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* https://doi.org/10.1109/34.990138.

Fonseca, Jaime. 2008. "The Application of Mixture Modeling and Information Criteria for Discovering Patterns of Coronary Heart Disease." *Journal of Applied Quantitative Methods.* https://doi.org/10.1109/EMS.2008.36.

Fraley, Chris, Adrian E. Raftery, and Luca Scrucca. 2022. *Mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation.*

Franczak, Brian C., Ryan P. Browne, and Paul D. McNicholas. 2016. *Sensory: Simultaneous Model-Based Clustering and Imputation via a Progressive Expectation-Maximization Algorithm.*

Gallegos, María Teresa, and Gunter Ritter. 2005. "A Robust Method for Cluster Analysis." *The Annals of Statistics.* https://doi.org/10.1214/009053604000000940.

Garay, Aldo M., Monique Bettio Massuia, and Victor Lachos. 2022. *SMNCensReg: Fitting Univariate Censored Regression Model Under the Family of Scale Mixture of Normal Distributions.*

Garay, Aldo M., Monique B. Massuia, Victor H. Lachos, et al. 2017. *BayesCR: Bayesian Analysis of Censored Regression Models Under Scale Mixture of Skew Normal Distributions.*

García-Escudero, Luis A., Alfonso Gordaliza, Carlos Matrán, et al. 2008. "A General Trimming Approach to Robust Cluster Analysis." *The Annals of Statistics.* https://doi.org/10.1214/07-AOS515.

Gohel, David, and Panagiotis Skintzos. 2023. *Flextable: Functions for Tabular Reporting.* https://CRAN.R-project.org/package=flextable.

Gu, Zuguang. 2022. "Complex Heatmap Visualization." *iMeta.* https://doi.org/10.1002/imt2.43.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btw313.

He, Yawei, and Zehua Chen. 2016. "The EBIC and a Sequential Procedure for Feature Selection in Interactive Linear Models with High-Dimensional Data." *Annals of the Institute of Statistical Mathematics.* https://doi.org/10.1007/s10463-014-0497-2.

Hester, Jim, and Jennifer Bryan. 2022. *glue: Interpreted String Literals.* https://CRAN.R-project.org/package=glue.

Houseman, E. Andres, Sc.D., Devin C. Koestler, et al. 2017. *RPMM: Recursively Partitioned Mixture Model.*

Iovleff, Serge. 2019. *MixAll: Clustering and Classification Using Model-Based Mixture Models.*

Iscar, Agustin Mayo, Luis Angel Garcia Escudero, and Heinrich Fritz. 2022. *Tclust: Robust Trimmed Clustering.*

Jaki, Thomas, Ting-Li Su, Minjung Kim, et al. 2018. "An Evaluation of the Bootstrap for Model Validation in Mixture Models." *Communications in Statistics: Simulation and Computation.* https://doi.org/10.1080/03610918.2017.1303726.

Jones, Andrew Thomas. 2020. *EMMIXgene: A Mixture Model-Based Approach to the Clustering of Microarray Expression Data.*

Jones, Andrew T., and Hien D. Nguyen. 2019. *SAGMM: Clustering via Stochastic Approximation and Gaussian Mixture Models.*

Kapourani, C. A. 2022. *Melissa: Bayesian Clustering and Imputationa of Single Cell Methylomes.*

Klein, Hans-Ulrich, and Martin Schaefer. 2022. *Epigenomix: Epigenetic and Gene Transcription Data Normalization and Integration with Mixture Models.*

Komárek, Arnošt. 2022. *mixAK: Multivariate Normal Mixture Models and Mixtures of Generalized Linear Mixed*

Models Including Model Based Clustering.

Kubicki, Vincent, Christophe Biernacki, and Quentin Grimonprez. 2021. *RMixtComp: Mixture Models with Heterogeneous and (Partially) Missing Data.*

Kurban, Hasan, Mark Jenne, and Mehmet M. Dalkilic. 2017. "Using Data to Build a Better EM: EM∗for Big Data." *International Journal of Data Science and Analytics.* https://doi.org/10.1007/s41060-017-0062-1.

Langrognet, Florent, Remi Lebret, Christian Poli, et al. 2021. *Rmixmod: Classification with Mixture Modelling.*

Leytham, K. M. 1984. "Maximum Likelihood Estimates for the Parameters of Mixture Distributions." *Water Resources Research.* https://doi.org/10.1029/WR020i007p00896.

Li, Ker-Chau. 1991. "Sliced Inverse Regression for Dimension Reduction." *Journal of the American Statistical Association.* https://doi.org/10.2307/2290563.

Li, Xuan, Yuejiao Fu, Xiaogang Wang, et al. 2018. *MMDvariance: Detecting Differentially Variable Genes Using the Mixture of Marginal Distributions.*

Li, Yan, and Kun Chen. 2021. *fmerPack: Tools of Heterogeneity Pursuit via Finite Mixture Effects Model.*

Lindsay, Bruce G. 1995. "Mixture Models: Theory, Geometry and Applications." *NSF-CBMS Regional Conference Series in Probability and Statistics.* https://www.jstor.org/stable/4153184.

Louis, Thomas A. 1982. "Finding the Observed Information Matrix When Using the EM Algorithm." *Journal of the Royal Statistical Society.* https://doi.org/10.1111/j.2517-6161.1982.tb01203.x.

Lu, Xiang, Yaoxiang Li, and Tanzy Love. 2022. *Bpgmm: Bayesian Model Selection Approach for Parsimonious Gaussian Mixture Models.*

Macdonald, Peter, and with contributions from Juan Du. 2018. *Mixdist: Finite Mixture Distribution Models.*

Maitra, Ranjan, and Volodymyr Melnykov. 2010. "Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms." *Journal of Computational and Graphical Statistics.* https://doi.org/10.1198/jcgs.2009.08054.

Marandon, Ariane, Tabea Rebafka, Etienne Roquain, et al. 2022. "False Clustering Rate Control in Mixture Models." arXiv. https://doi.org/10.48550/ARXIV.2203.02597.

McCaw, Zachary. 2021. *MGMM: Missingness Aware Gaussian Mixture Models.*

McLachlan, G. J., D. Peel, and R. W. Bean. 2003. "Modelling High-Dimensional Data by Mixtures of Factor Analyzers." *Computational Statistics & Data Analysis*, Recent Developments in Mixture Model,. https://doi.org/10.1016/S0167-9473(02)00183-4.

McLachlan, Geoffrey, and David Peel. 2000. *Finite Mixture Models: McLachlan/Finite Mixture Models.* John Wiley & Sons, Inc. https://doi.org/10.1002/0471721182.

Mclachlan, G., and David Peel. 2000. "Mixtures of Factor Analyzers." In. https://doi.org/10.48550/arXiv.1507.02801.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, et al. 2010. "Serial and Parallel Implementations of Model-Based Clustering via Parsimonious Gaussian Mixture Models." *Computational Statistics & Data Analysis.* https://doi.org/10.1016/j.csda.2009.02.011.

McNicholas, Paul David, and Thomas Brendan Murphy. 2008. "Parsimonious Gaussian Mixture Models." *Statistics and Computing.* https://doi.org/10.1007/s11222-008-9056-0.

McNicholas, Paul D., Aisha ElSherbiny, Aaron F. McDaid, et al. 2022. *Pgmm: Parsimonious Gaussian Mixture Models.*

McNicholas, Paul D., and Thomas Brendan Murphy. 2010. "Model-Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Models." *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btq498.

McParland, Damien, and Isobel Claire Gormley. 2017. *clustMD: Model Based Clustering for Mixed Data.*

Meng, Lingyao. 2016. "Method for Computation of the Fisher Information Matrix in the Expectation-Maximization Algorithm." arXiv. https://doi.org/10.48550/ARXIV.1608.01734.

MENG, XIAO-LI, and Donald Rubin. 1993. "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework." *Biometrika.* https://doi.org/10.1093/biomet/80.2.267.

Meng, Xiao-Li, and David Van Dyk. 1997. "The EM Algorithm–an Old Folk-Song Sung to a Fast New Tune." *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* https://doi.org/10.1111/1467-9868.00082.

Mersmann, Olaf. 2021. *Microbenchmark: Accurate Timing Functions.*

Mohammadi, Reza. 2021. *Bmixture: Bayesian Estimation for Finite Mixture of Distributions.*

Mouselimis, Lampros. 2022. *ClusterR: Gaussian Mixture Models, k-Means, Mini-Batch-Kmeans, k-Medoids and Affinity Propagation Clustering*.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Murphy, Kevin. 2012. *Machine Learning A Probabilistic Perspective*. Adaptive Computation; Machine Learning series. https://doi.org/10.1080/09332480.2014.914768.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

Nowakowska, Ewa, Jacek Koronacki, and Stan Lipovetsky. 2014. "Tractable Measure of Component Overlap for Gaussian Mixture Models." https://doi.org/10.48550/arXiv.1407.7172.

O'Hara-Wild, Mitchell, Stephanie Kobakian, H. Sherry Zhang, Di Cook, Simon Urbanek, and Christophe Dervieux. 2023. *rjtools: Preparing, Checking, and Submitting Articles to the "R Journal"*. https://CRAN.R-project.org/package=rjtools.

Oakes, D. 1999. "Direct Calculation of the Information Matrix via the EM." *Journal of the Royal Statistical Society*. https://doi.org/10.1111/1467-9868.00188.

Papastamoulis, Panagiotis. 2020. *fabMix: Overfitting Bayesian Mixtures of Factor Analyzers with Parsimonious Covariance and Unknown Number of Components*.

Pastore, Massimiliano, and Antonio Calcagnì. 2019. "Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index." *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01089.

Petersen, K. B., and M. S. Pedersen. 2008. "The Matrix Cookbook." Technical University of Denmark.

Pocuca, Nik, Ryan P. Browne, and Paul D. McNicholas. 2022. *Mixture: Mixture Models for Clustering and Classification*.

Prates, Marcos, Victor Lachos, and Celso Cabral. 2021. *Mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions*.

Prates, Marcos, Victor Lachos, and Aldo Garay. 2021. *Nlsmsn: Fitting Nonlinear Models with Scale Mixture of Skew-Normal Distributions*.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rau, Andrea. 2022. *Coseq: Co-Expression Analysis of Sequencing Data*.

Rauschenberger, Armin. 2022. *Semisup: Semi-Supervised Mixture Model*.

Redner, Richard A., and Homer F. Walker. 1984. "Mixture Densities, Maximum Likelihood and the Em Algorithm." *SIAM Review*. https://doi.org/239https://doi.org/10.1137/1026034.

Robert, Christian, and George Casella. 2010. *Introducing Monte Carlo Methods with R*. Springer. https://doi.org/10.1007/978-1-4419-1576-4.

Schlattmann, Peter, Johannes Hoehne, and Maryna Verba. 2022. *CAMAN: Finite Mixture Models and Meta-Analysis Tools - Based on c.a.MAN*.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics*. https://doi.org/10.1214/aos/1176344136.

Scrucca, Luca. 2010. "Dimension Reduction for Model-Based Clustering." *Statistics and Computing*. https://doi.org/10.1007/s11222-009-9138-7.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, et al. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal*. https://doi.org/10.32614/RJ-2016-021.

Shokoohi, Farhad. 2022. *Fmrs: Variable Selection in Finite Mixture of AFT Regression and FMR Models*.

Thrun, Michael, Onno Hansen-Goos, and Alfred Ultsch. 2020. *AdaptGauss: Gaussian Mixture Models (GMM)*.

Tipping, Michael E., and Christopher M. Bishop. 1999. "Probabilistic Principal Component Analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1111/1467-9868.00196.

Turner, Rolf. 2021. *Mixreg: Functions to Fit Mixtures of Regressions*.

Viroli, Cinzia, and Geoffrey J. McLachlan. 2020. *Deepgmm: Deep Gaussian Mixture Models*.

Walsh, Gordon Raymond. 1975. *Methods of Optimization*. Wiley.

Wang, Ziqiao. 2022. *IMIX: Gaussian Mixture Model for Multi-Omics Data Integration*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Stringr: Simple, Consistent Wrappers for Common String Operations.* https://CRAN.R-project.org/package=stringr.

———. 2023. *Forcats: Tools for Working with Categorical Variables (Factors).* https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2023. *Purrr: Functional Programming Tools.* https://CRAN.R-project.org/package=purrr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data.* https://CRAN.R-project.org/package=readr.

Wickham, Hadley, and Dana Seidel. 2022. *scales: Scale Functions for Visualization.* https://CRAN.R-project.org/package=scales.

Wilke, Claus O. 2020. *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2".* https://CRAN.R-project.org/package=cowplot.

Wilke, Claus O., and Brenton M. Wiernik. 2022. *ggtext: Improved Text Rendering Support for "ggplot2".* https://CRAN.R-project.org/package=ggtext.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr.* 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xu, Yanxun, Peter Mueller, Donatello Telesca, et al. 2020. *Dppmix: Determinantal Point Process Mixture Models.*

Yang, Yuhong. 2005. "Can the Strengths of AIC and BIC Be Shared? A Conflict Between Model Identification and Regression Estimation." *Biometrika.* https://doi.org/10.1093/biomet/92.4.937.

Yu, Youjiao. 2021. *mixR: Finite Mixture Modeling for Raw and Binned Data.*

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.

Appendix C

Appendix of Article 3

C.1 Appendix of Reference-Based Approaches

Fundamental Assumptions on the Partial Deconvolution Framework

In this section, for the sake of readability, we recall fundamental relations underlying the deconvolution framework.

Traditionally, deconvolution models assume that the total bulk expression is linearly related to the individual cell profiles by the linear equation Equation (C.1):

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X} \times \boldsymbol{p}_i \quad \text{matricial form} \\ y_{gi} &= \sum_{j=1}^{J} x_{gj} \times p_j \quad \text{algebraic form} \end{aligned} \tag{C.1}$$

.

In addition, most deconvolution problems explicitly enforce the *unit-simplex constraint* (Equation (C.2)) on the cellular ratios:

$$\begin{cases} \sum_{j=1}^{J} p_{ji} = 1 \\ \forall j \in \widetilde{J} \quad p_{ji} \geq 0 \end{cases} \tag{C.2}$$

C.1.1 Linear regression and Gauss-Markov theorem

Theorem C.1.1: Normal equations

The ***Normal equations*** *provide the following Ordinary Least Squares (OLS) estimate Equation (C.3):*

$$\hat{p}^{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{C.3}$$

whose existence implies that the design matrix $\boldsymbol{X}$ *is invertible.*

Theorem C.1.2: Gauss-Markov theorems

The Gauss-Markov assumptions encompass:

1. ***Strong exogeneity**: The cell type-specific expression profiles are not random variables but rather fixed and constant observations, underlying implicitly that cell populations do no interact:* $\forall i \in \widetilde{J}, \forall j \in \widetilde{J}, i \neq j, \quad \mathbb{C}\text{ov}[\boldsymbol{x_{.i}}, \boldsymbol{x_{.j}}] = 0.$

2. ***Gaussian-Markov noise:** This hypothesis postulates that the residual error term is described by a white Gaussian noise process, characterised by null mean and variance that is independent on the gene, thus homoscedastic, which yields, in mathematical terms:*

$$y_g = \sum_{j=1}^{J} x_{gj} p_j \ + \epsilon_g, \ \epsilon_g \sim \mathcal{N}\left(0, \sigma_g^2\right)$$

.

By integrating the exogeneity and homoscedasticity assumptions, it is possible to derive the distribution of each transcript, which reveals Gaussian as articulated in Equation (C.4):

$$\boldsymbol{y}_{1:G} | \boldsymbol{X} \sim \mathcal{N}_G(\boldsymbol{X}\boldsymbol{p}, \sigma^2 \mathbf{I}_G) \tag{C.4}$$

The second line highlights that the conditional distribution is identifiable to a spherical multivariate Gaussian distribution.

3. ***Independence:** From the aforementioned Gaussian-Markov and exogeneity assumptions, we readily deduce that the gene expressions of the bulk measures are independent:* $\forall j \in \widetilde{G}, \forall k \in \widetilde{G}, j \neq k, \quad \mathbb{C}\text{ov}[y_j, y_k] = 0.$

4. ***Completeness:** We assume no additional latent variable.*

If they hold, the MLE estimate is then equal to the OLS estimate given by the ***Normal equations*** *(Equation (C.3)). Additionally, the MLE is the unique BLUE (best linear unbiased estimator), i.e. the unbiased estimator with the lowest variance.*

Proof C.1.3: Gauss-Markov proof

Under the Gaussian-Markov assumptions (see Theorem C.1.2 and notably Equation (C.4)) and assumption of independence between samples, then, the global log-likelihood distribution of the response variable $\boldsymbol{y}$ conditioned on $\boldsymbol{X}$ is given by Equation (C.5):

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{MLE}} &= \ell_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{X}) \\
&= \arg\max_{\boldsymbol{\theta}} \left[\sum_{g=1}^{G} \log \left(\mathbb{P}_{\boldsymbol{\theta}}(y_g | \boldsymbol{x}_{g.}) \right] \right. \\
&= \arg\max_{\boldsymbol{\theta}} \left[\sum_{g=1}^{G} \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_g - \sum_{j=1}^{J} x_{gj} p_j\right)^2}{2\sigma^2}} \right) \right] \\
&= K - G\log(\sigma) - \sum_{g=1}^{G} \left(\frac{(y_g - \sum_{j=1}^{J} x_{gj} \times p_j)^2}{2\sigma^2} \right)
\end{aligned} \tag{C.5}$$

with $K = -\frac{G}{2}\log(2\pi)$, the *normalising constant*. Finding the values for which the derivative of the function Equation (C.5) cancels yield the same estimate returned by the OLS method Equation (C.3).
The MLE estimate provides additionally an estimate of the standard deviations:

$$\hat{\sigma}^2 = \frac{1}{G} \sum_{g=1}^{G} y_g - \sum_{j=1}^{J} x_{gj} \times \hat{p_j}$$

Ultimately, to prove that the estimate $\boldsymbol{p}$ is indeed the unique global maxima of the log-likelihood function Equation (C.5), we just have to differentiate the equation once more, and show that the resulting Hessian matrix is indeed *positive definite*.

C.1.2 Robust regression approaches

Definition C.1.4: M-estimates Regression

M-estimates, short for "maximum likelihood estimates" design the class of estimators that maximise a likelihood function. In practice, M-estimates *replaces the equally weighted observations from lls regression with an adaptive function of the residuals:*

$$\hat{\boldsymbol{p}}_M = \arg\min_{\boldsymbol{p}} \sum_{g=1}^{G} \rho\left(y_g | \boldsymbol{x}_{g.}; \boldsymbol{p}\right) \tag{C.6}$$

where ρ is the robust loss function and $\psi = \rho'$ is its derivative called the influence function. *Different loss functions lead to different properties of M-estimators, and the choice of the loss function depends on the distribution of the dataset and the desired properties of the estimator:*

- Tukey's bisquare function *is a softer smoothing function Equation (C.7):*

$$\rho(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{\boldsymbol{x}-\boldsymbol{p}}{c}\right)^2\right)^3\right], & \text{if } |\boldsymbol{x} - \boldsymbol{p}| \leq c \\ \frac{c^2}{6}, & \text{if } |\boldsymbol{x} - \boldsymbol{p}| > c \end{cases} \tag{C.7}$$

  With $c = 4.6885$, its efficiency *is equal to the Huber's estimate (95% of an OLS estimate). Although not implemented independently in any deconvolution paper, the standard rlm (for robust linear modelling) function in the R MASS package, which performs the Tukey's biweight iterative regression, is often used as a gold-standard robust linear regression method in most of the deconvolution benchmark papers ([Stu+19], [Gau13]).*

- *The Least Absolute Deviation (LAD) minimises the absolute differences of the residuals (L1 distance) rather than their squared differences (L2 distance):*

$$\hat{\boldsymbol{p}}_{MAE} = \arg\min_{\hat{p}} \left(|\boldsymbol{x}\boldsymbol{p} - \boldsymbol{y}|\right) \tag{C.8}$$

  where MAE stands for Mean Absolute Deviation. A distribution of these functions is reported in Appendix C.1.4.

Definition C.1.5: Least Trimmed Squared Regression

The LTS method was first proposed in [Rou85], with the idea to select the gene subset that exhibit the smallest residuals altogether. Practically, the estimate is given by Equation (C.9):

$$\hat{\boldsymbol{p}}_{LTS} = \arg\min_{\boldsymbol{p}} \sum_{g=1}^{G} {}^{*} r_g(\boldsymbol{p})^2 \tag{C.9}$$

with $\left|\widetilde{G^*}\right| = G(1-\alpha)+1$ *with* α *the trimming proportion, and* $r_g(\boldsymbol{p})$ *the residuals ordered by increasing order. Taking* $\alpha = \frac{G}{2}$*, LTS asymptotically displays a strong BP of 0.5, implying it is robust to outliers, but a very low efficiency of 0.08. In addition, LTS is an NP-hard problem [Rou85], as any combination of* $\binom{G}{\left|\widetilde{G^*}\right|}$ *observations should be tested, to find the* $\left|\widetilde{G^*}\right|$ *genes with the minimal residual error. [RV06] hence extends the method in high dimension, or with a large number of observations, by proposing a stochastic and faster version of this algorithm. However, its performance is highly dependent on the initial random* $\left|\widetilde{G^*}\right|$*-subset chosen. Last but not least, the trimming ratio is an additional hyper-parameter that plays a key role on the accuracy of the estimate.*

A comprehensive review of robust linear estimates is supplied in [YYB14], with 10 influence functions benchmarked. It notably demonstrates that MM-estimates and RWLSE estimates have overall the best performance in terms of robustness and asymptotic efficiency.

Definition C.1.6: Support Vector Regression

Linear SVR identifies the hyperplane that fits as many data points as possible. However, instead of levering all observations as the standard OLS approach, only a subset of data points, termed as "support vectors" (SVs) impact the prediction. In addition, the optimisation function underlying the SVR framework (Equation (C.10)) aims at finding the sweet spot between minimising the prediction error and maintaining a controlled level of complexity to prevent overfitting:

$$\tau(p,\zeta,\epsilon) = \underbrace{\frac{1}{2}\sum_{j=1}^{J} p_j^2}_{L2\ metric} + \underbrace{C\left(\nu\epsilon + \frac{1}{G}\sum_{g=1}^{G}(\zeta_g + \zeta_g^*)\right)}_{\nu-insensitive\ function} \tag{C.10}$$

where C is a regularisation parameter controlling the trade-off between complexity and error control, $\boldsymbol{p}$ the estimates, referred to as the weights of the model, and ζ_g and ζ_g^* slack variables *to control the number of points outside the ϵ-tube. The* penalty function of the $L2$-norm *in Equation (C.10), which is identical to that employed in ridge regression, penalises the model complexity by putting less weight on the estimated ratios of highly correlated cell types [CM04].*
Finally, each pair of observation and covariates, $y_g, \boldsymbol{x}_{g.}$, are subjected to the following constraints Equation (C.11):

$$\begin{aligned} y_g - \boldsymbol{p}^T\boldsymbol{x}_{g.} - b &\leq \epsilon + \xi_g \\ \boldsymbol{p}^T\boldsymbol{x}_g + b - y_g &\leq \epsilon + \xi_g^* \end{aligned} \tag{C.11}$$

with b is the bias term *(corresponding to the intercept in linear regression models), ϵ the margin of tolerance, and slack variables ξ_i and ξ_i^* the allowed deviations from the margin. The bias term corresponds to the null intercept in standard linear regression framework, and is usually negative in SVM models ([Yan19]).*

To quantify the relative performance of these robust approaches, two metrics are generally used: the *efficiency* of the robust estimate relatively to the OLS estimate when the assumptions of the Gaussian-Markov theorem hold [1], and the *breakdown point*.

The breakdown point is a measure of the robustness of the regression estimator to outliers. Precisely, it represents the proportion of data points that can be perturbed before the estimator's behaviour becomes meaningless and unstable. For instance, the OLS estimate has a small BP of $\frac{1}{G}$, implying that only one single unusual observation can contribute to the mean of the estimated ratios [Rou85].

[1] The OLS estimate is indeed efficient, reaching asymptotically the Cramér-Rao bound

C.1.3 Regularised linear approaches

Definition C.1.7: Regularised linear regression

Historically, the Ridge regression *[HK70] employs a L2-penalty Equation (C.12):*

$$\begin{cases} \hat{\boldsymbol{p}}_{Ridge} = \arg\min_{\boldsymbol{p}} \left[\underbrace{\sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} p_j x_{gj} \right)^2}_{\text{linear regression}} + \underbrace{\lambda \sum_{j=1}^{J} p_j^2}_{\text{penalty function}} \right] \\ \text{subject to } \sum_{j=1}^{J} p_j^2 \leq c \end{cases} \tag{C.12}$$

Ridge regression shrinks the coefficients but not necessarily to zero, implying that there is no hard feature selection. Otherwise, Ridge is particularly useful when multicollinearity is a concern.

Subsequently, the Lasso regression [Tib96] uses a L1-penalty, which allows a hard variable selection:

$$\begin{cases} \hat{\boldsymbol{p}}_{Lasso} = \arg\min_{\boldsymbol{p}} \left[\sum_{g=1}^{G} \left(y - \sum_{j=1}^{J} p_j x_{gj} \right)^2 + \lambda \sum_{j=1}^{J} |p_j| \right] \\ \text{subjected to } \sum_{j=1}^{J} |p_j| \leq c \end{cases} \tag{C.13}$$

Efficiency of this optimisation approach relies strongly on the sparsity of the dataset, inducing that most of the coefficients are truly null (the set of coefficients with non-null values is called the true support*). However, Lasso regression underperforms with highly correlated transcriptomic profiles, especially when the* irrepresentable condition *is violated, namely when the correlation between the explanatory and confusing variables is larger that the within correlation between the explanatory variables. In that case, the Lasso algorithm tends to arbitrarily choose one out of a group of correlated features.*

Elastic net [ZH05] has been developed to keep the middle ground of both worlds Equation (C.14):

$$\hat{\boldsymbol{p}}_{ElasticNet} = \arg\min_{\boldsymbol{p}} \left[\underbrace{\sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} p_j x_{gj} \right)^2}_{\text{regression function}} + \underbrace{\lambda \sum_{j=1}^{J} (1-\alpha) p_j^2 + \alpha |p_j|}_{\text{penalise complexity}} \right] \tag{C.14}$$

in which α is a trade-off parameter between the L2-penalty ($\alpha = 0$) and the L1-penalty ($\alpha = 1$). This formulation enables continuous shrinkage, including hard feature selection, and can even by deployed with highly correlated cell expression profiles.

Interestingly, [Zho+14] demonstrates that the Elastic net problem is identifiable to a linear SVR under specific reparametrisation, allowing to utilise highly-scalable and parallel SVM solvers.

C.1.4 Probabilistic approaches

Definition C.1.8: Latent Dirichlet Allocation: introduction

LDA, as a generative probabilistic model, has first been used in natural language processing and topic modelling, with the goal of inferring the distribution of topics across documents. Precisely, LDA assumes that documents are mixtures of topics, and topics are mixtures of words. Applied to our cellular deconvolution context, the documents, for which only the respective number of words is available, represent each patient or sample bulk transcriptomic profile and the distribution of words represent read counts. Finally, the latent topics describe cell populations that make up each document.
Formally, let's introduce the following couple of independent random variables (T, Z) *in the probabilistic framework, along with* $L = \sum_{g=1}^{G} y_g$*, the total number of counts in the sample (aka the library depth):*

- $Z = Z_{1:L} \in \{1 : J\}^L$*: a discrete latent variable identifying from which reference population each count originates. With that modelling, the cell ratios (document-topic proportions) can be recovered with:*

$$p_j = \frac{\sum_{l=1}^{L} z_l \mathbb{1}_{z_l=j}}{L}$$

  , note that this framework naturally enforces the unit-simplex constraint on the ratios.

- $T = T_{1:L} \in \{1 : G\}^L$*: it its the vectorised transcriptomic expression profile* $\boldsymbol{y}$*. The total expression of a given gene* g *is retrieved by summing all transcripts from* T *associated to this gene:* $y_g = \sum_{l=1}^{L} t_l \, \mathbb{1}_{t_l=g}$*.*

- *Finally, let's introduce the individual purified expression profile for a gene* g *produced by a given cell population* j*:* $x_{gj} = \sum_{l=1}^{L} t_l \, \mathbb{1}_{(t_l=g) \cap (z_l=j)}$*, then the ratio of this specific gene over the total transcriptomic expression for population* j *(topic-word proportions) is given by:* $\beta_{gj} = \frac{x_{gj}}{L_j}$*, with* L_j *the total number of counts in population* j *and* $\boldsymbol{\beta}_j$ *its multidimensional generalisation.*

Definition C.1.9: Latent Dirichlet Allocation: estimation

With this modelling approach, the joint distribution of (T, Z) in the LDA model for a given sample is given by Equation (C.15):

$$\mathbb{P}_{\boldsymbol{\theta}}(Z_{1:L}, T_{1:L}) = \mathbb{P}(\boldsymbol{p}) \times \prod_{l=1}^{L} \sum_{j=1}^{J} \overbrace{\mathbb{P}(Z = j)}^{p_j} \quad \overbrace{\mathbb{P}(T = g|Z = j)}^{\beta_{g,j}} \tag{C.15}$$

which corresponds to a parametric mixture model of multinomial (the generalisation of binomial distributions, with more than two outputs for each generation) distributions, and $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{\beta})$ the minimal set of parameters to estimate (all other quantities of interest can be deduced from them).
Simultaneously optimising both sets of parameters is analytically intractable. Instead, an ECM (Expected Conditional Maximisation) algorithm ([MR93]), a direct extension of the EM framework ([DLR77]), has been used to iteratively optimise the set of parameters until convergence. In brief, the ECM approach consists of replacing the maximisation step of EM with a set of conditional maximisation steps, decomposing here the difficult joint maximisation of the parameters into several easier and conditionally dependent ones:

1. ***Initialisation:*** *LDA requires an initialisation step, initial parameters $\boldsymbol{\theta}_0 = (\boldsymbol{p}_0, \boldsymbol{\beta}_0)$ being drawn from prior distributions that should generate candidates in the support of the solution space.*

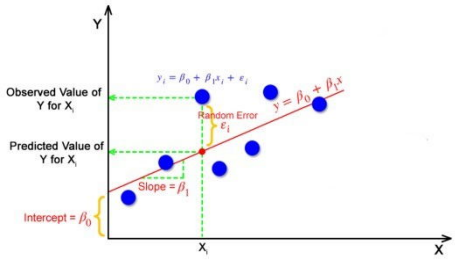
2. ***E-step:*** *At step (q), MAP (maximum a posteriori) is used to assign the production of each gene g for the count $l \in \{1, \dots L\}$ to a given population j Equation (C.16):*

$$\hat{Z}_l^q,\ l \in \{1, \dots, L\} = \arg\max_{j \in \tilde{J}} \left[\mathbb{P}(Z_l = j|T_l = g)\right] \tag{C.16}$$

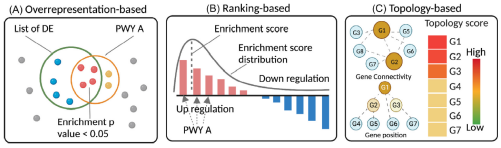
   , using the prior inferred parameters of the mixture of multinomial distributions $\boldsymbol{\theta}^{q-1}$

3. ***M-step:*** *Injecting the latent variables inferred in the previous estimation step, the parametric vector $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{X})$ which maximised the conditional distribution $\mathbb{P}_{\boldsymbol{\theta}}(T_{1:L}|Z_{1:L}) = \prod_{l=1}^{L} \mathbb{P}_{\boldsymbol{\theta}}(T_l|Z_l)$ is returned.*

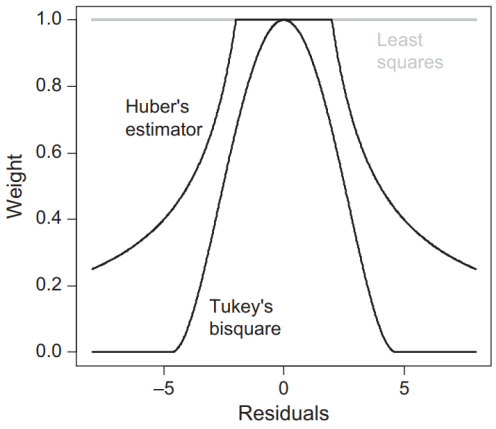
The main advantage of LDA relies on its versatility, since this approach can be applied to various types of data (provided it has been discretised), can be easily interpreted and is close to the biological process. We refer the reader to [Lee+18] and [Xu+23] for a comprehensive report of the main features and limitations provided by this "bag-of-words" approach.
However, determining the number of cell types J can be challenging without proper biological annotation, the method is highly sensitive to preprocessing choices, and struggles with sparse data or short documents (in a biological context, this implies that this method should not be used to characterise rare cell populations, contributing poorly to the final pool of transcripts).



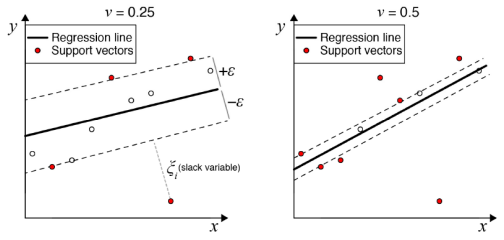
(a) LLS principle. Here, we present briefly the methodology with a simplified univariate regression framework, including an intersection term β_0. Reproduced from [MAL21, Fig. 2].



(c) Overview of three pathway enrichment analysis methods. Over-representation techniques focus on investigating whether a given gene list displays any pathways that are more prevalent than expected by chance when compared to a reference set. (B) In ranking-based methods, the whole gene set is examined to determine whether genes associated with the same pathway exhibit a tendency to cluster at either the top or the bottom of the ordered list of the universe of quantified genes. Such methods return an enrichment score reflecting the amplitude and the sense of the variation induced by the phenotype (C) Topology-based strategies incorporate scores that gauge both gene absolute positions and gene pairwise interactions (up to our knowledge, none of the marker-based methods we reviewed integrate this feature). Reproduced from [ZR23, Fig. 1].



(b) Common influential functions. The weight function distributions for Huber's robust estimator and Tukey's bisquare (or biweight) compared with least squares estimation, in which each observation is assigned the same weight, no matter its contribution to the residuals errors. Reproduced from [Wri09, Fig. 1]



(d) Illustration of support vector regression (SVR).ξ and ξ^* are *slack variables* controlling the upper and lower error margins, respectively. Together, slack variables enable to define boundary decision lines, all points lying outside of the $\epsilon-$ tube making up the set of "support vectors" (red circles). $\nu-$SVR is a recent approach, in which the so-called hyperparameter controls the amount of SVs (for instance, in the right picture, half of the genes lie beyond the confidence boundaries). Interestingly, only the set of SVs is required to predict cellular ratios, avoiding as such overfitting. Reproduced from [New+15, Supplementary Fig. 1].

C.2 Statistical Appendix of Marker-Based Approaches

C.2.1 Gene Set Enrichment Analysis

Definition C.2.1: Principles of GSEA

Gene Set Enrichment Analysis (GSEA) is a bioinformatics method used initially to determine whether a predefined set of genes shows significant differences between two biological states. They hence differ from Differential Gene Expression Analysis (DGEA) analyses, since GSEA operates on groups of genes associated with a biological function or process rather than considering independently one gene after the other. In a second time, GSEA assigns a statistical significance score to each gene set which evaluates the null hypothesis of randomly distributed throughout the ranked gene list against the alternative hypothesis of a clustering pattern at the top or bottom of the ranked list.
The enrichment score (ES) for each gene set returned by GSEA analyses, reflecting the degree to which genes are unequally distributed in the ties of the ranked list, is given by the following running sum statistic, assuming beforehand that $\widetilde{G}$ and $\widetilde{G_j}$ are ranked by decreasing order of fold change (or any relevant metric) Equation (C.17):

$$\begin{aligned} ES\left(\widetilde{G_j} \in \widetilde{G}\right) &= \sup_{g=1}^{|\widetilde{G_j}|} \left| F^*_{g\in\widetilde{G_j}}(g) - F_{g\in\widetilde{G}}(g) \right| \\ \text{with } F^*_{g\in\widetilde{G_j}}(g) = \mathbb{P}^*(\widetilde{G_j} \leq g) &= \frac{R^*(g)}{\left|\widetilde{G_j}\right|}, \quad F_{g\in\widetilde{G}}(g) = \frac{R(g)}{\left|\widetilde{G}\right|} \end{aligned} \tag{C.17}$$

with $\left|\widetilde{G}\right| = G$ the number of genes (I commonly use the second notation for consistency and conciseness reasons, since there is no real risk of confusion), $|\widetilde{G_j}|$ the module, namely the number of genes composing the gene set associated to cell population j, $F^*_{g\in\widetilde{G_j}}(g) = \frac{\text{index of gene } g\text{, alternatively number of genes higher ranked}}{\left|\widetilde{G_j}\right|}$ and $F_{g\in\widetilde{G}}(g)$ are the cumulative distribution functions (CDF) of the gene rankings/positions (ordered by decreasing order of fold change) of gene set G_j ($R^*(g)$ being the index of gene g in gene module $\widetilde{G_j}$), respectively within the module itself and with respect to the total set of genes quantified in the study $\widetilde{G}$ (note the asterisk to set apart both distributions).
Note that this score, without weights, is the standard Kolmogorov-Smirnov running sum statistic, used traditionally to compare empirical distributions and for which the existence for an asymptotic one-sided statistical test of the null hypothesis distribution is known [SL11], and that ES scores can be easily computed in R with the gsva function.

C.2.2 Hypergeometric Distribution

Definition C.2.2: Using Hypergeometric Laws for Gene Pathway Enrichment Analysis:

Hypergeometric distribution is commonly used in gene pathway enrichment analysis, such as in the Gene Ontology (GO) database. The main purpose is to assess whether a particular set of genes, often the set of differentially expressed genes, is statistically over-represented in a predefined gene pathway compared to what would be expected by chance. If the observed overlap is larger than expected, it suggests that the pathway is enriched.
Mathematically, the hypergeometric distribution returns the probability of observing $X = k \equiv \left|\widetilde{G}_j^{\text{diff}}\right|$ *genes (likely the subset of genes differentially expressed) from the set of interest in a pathway of size* $\left|\widetilde{G}_j\right|$*, drawn randomly without replacement from the total set of genes marked as differentially expressed in DGEA* $\widetilde{G}^{\text{diff}} \in \widetilde{G}$*, and is computed by the following probability mass function Equation (C.18):*

$$\mathbb{P}(X = k) = \frac{\binom{\left|\widetilde{G}_j\right|}{k} \cdot \binom{\left|\widetilde{G}\right| - \left|\widetilde{G}_j\right|}{\left|\widetilde{G}^{\text{diff}}\right| - k}}{\binom{\left|\widetilde{G}\right|}{\left|\widetilde{G}^{\text{diff}}\right|}} \tag{C.18}$$

C.2.3 Limitations of Marker-Based Approaches

The GSEA and hypergeometric approaches are constrained by some strong assumptions on the nature of gene pathways. Both methods presuppose that genes are selected independently for inclusion in the cell marker set.

In addition, the statistical assessment of the enrichment score depends on the gene pathway cardinality ([Aba+09]) and the size of the *background set*, denoting the cardinality of the gene universe used in the enrichment analysis. In particular, smaller pathways are consistently prone to being spuriously labelled as enriched. Finally, when multiple pathways are assessed concurrently, multiple testing corrections are required to control the inflation of the false discovery rate (FDR). To mitigate these statistical limitations, it is possible to incorporate additional metrics, such as weighting the rank indexes of fold-change with the p-values derived from Differential Gene Expression Analysis (DGEA). However, hypergeometric tests usually offer less versatility and informativeness compared to GSEA approaches because they treat all genes within a pathway equally, irrespective of the magnitude of gene expression changes.

Both GSEA and hypergeometric methods remain neutral regarding the direction of transcriptomic variation, whether it involves up- or down-regulated genes. In contrast, the algorithm exploiting the Connectivity Map dataset ([Lam07] and [Lam+06]), extends the investigational capacity of Enrichment Scores (Equation (C.17)), with the calculation of two separate metrics for up-regulated and down-regulated genes, possibly followed by their aggregation when biologically relevant.

On a side note, any test for evaluating compositional data, and notably equality of proportions, can be used. It notably encompasses the asymptotic χ^2 statistical test and the *Fisher's exact test*. When quantitative gene expression is available, any test comparing two continuous statistical distributions, including the Pearson correlation score, could alternatively be employed.

C.3 Statistical Appendix of Reference-Free Approaches

Definition C.3.1: Principles of LS-NMF

Least-Square Non-Negative Matrix Factorization (LS-NMF) is originally a dimensionality reduction technique, based on factorising a given non-negative data matrix into a product of two non-negative matrices. LS-NMF enables to reduce the dimensionality of the original data in a meaningful manner, by representing the data as a product of two lower-dimensional matrices while keeping the fundamental linear assumption of linearity in cell deconvolution methods (see Equation (C.1)) and enforcing non-negativity in both the factor matrices $\boldsymbol{P}$ and $\boldsymbol{X}$ (indeed, in both cases, negative values can not be interpreted). More generally, it is a powerful method to extract relevant features or components of the data (here we assume that the subdimensional features match the individual cellular profiles, $\boldsymbol{X}$), while the coefficients in $\boldsymbol{P}$ represent the weights of these features for each data point (in a deconvolution framework, they are assimilated to cellular ratios). The number of hidden components/spanning dimensions J, which is also the rank of $\boldsymbol{X}$ are interpreted as the number of cell populations at the same lineage level.
Given a non-negative data matrix $\boldsymbol{Y} \in \mathbb{R}_+^{G\times N}$, LS-NMF seeks to determine the best two-terms matrix factorisation that approximate $\boldsymbol{Y} \sim \boldsymbol{X}\boldsymbol{P}$, both non-negative matrices, by minimising the Frobenius norm of the difference [a] Equation (C.19):

$$\begin{aligned} &\min_{\boldsymbol{P},\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{X}\|_F^2 \\ &\textit{subject to the non-negativity constraints:} \\ &\boldsymbol{P} \geq 0,\ \boldsymbol{X} \geq 0 \end{aligned} \tag{C.19}$$

This optimisation problem is often intractable, and thus typically solved iteratively using algorithms like multiplicative updates or gradient descent.
However, LS-NMF suffers from two main limitations: it is highly sensitive to the initial set of values provided for $\boldsymbol{P}$ and $\boldsymbol{X}$, and different initialisation can lead to different factorisation and convergence to local optima. The choice of the number of components, which can be interpreted as the rank of $\boldsymbol{X}$ is often arbitrary and critical.

[a] Instead of the Frobenius norm, it is also possible to employ the Kullback-Leibler divergence, as in [Don+20]

C.4 Appendix of Cellular Deconvolution Pipeline

Definition C.4.1: Condition number: general definition

The condition number is defined more precisely to be the maximum ratio of the relative error in the measured value to the relative error made on the input. Consider for an explicit mathematical formula the following variables: $\boldsymbol{p}$ is the input of our problem, $\boldsymbol{y}$ (alternatively $f(\boldsymbol{p})$ the measured value, and $\tilde{f}(\boldsymbol{p})$) (alternatively $\hat{\boldsymbol{y}}$) the predicted value by any algorithm or predictive function. Then, the relative condition number is formally defined by Equation (C.20):

$$\kappa(f, \boldsymbol{p}) = \lim_{\epsilon \to 0^+} \sup_{\|\delta \boldsymbol{p} \leq \epsilon\|} \frac{\|\delta f(\boldsymbol{p})\| \,/\, \|f(\boldsymbol{p})\|}{\|\delta \boldsymbol{p}\| \,/\, \|\boldsymbol{p}\|} \tag{C.20}$$

with $||\quad||$, namely the double vertical bars, the usual typology used to mark any matrix norm [a] and $\|\delta f(\boldsymbol{p})\| = \left\|f(\boldsymbol{p}) - \tilde{f}(\boldsymbol{p})\right\|$ the relative error.

[a] See definitions, properties and popular matrix norm definitions on this Wikipedia page: Matrix Norm.

Theorem C.4.2: Application of the Condition Number as a predictive quality metric for linear-based regression problems

The condition number, associated to the OLS estimate given in Equation (C.3), is Equation (C.21) :

$$\kappa(\boldsymbol{X}) = ||\boldsymbol{X}|| \times ||\boldsymbol{X}^\top|| \tag{C.21}$$

It is then possible to show the following inequality, derived directly from the definition of a matrix norm, holds Equation (C.22):

$$||\boldsymbol{X}|| \times ||\boldsymbol{X}^\top|| \geq ||\boldsymbol{X}\boldsymbol{X}^\top|| \geq \left\|\boldsymbol{X}\boldsymbol{X}^{-1}\right\| = 1 \tag{C.22}$$

, which provides an upper bound on the precision we can achieve with linear regression in the best case scenario. This bound is only reached if, and only if, the condition number of $\boldsymbol{X}$ is equal to 1.
Defining $\|\|$ as the L2 or Euclidean norm, and building the design matrix such that it is normal yields an explicit general formula relating the condition number of the matrix to its eigen values Equation (C.23):

$$\kappa(\boldsymbol{X}) \equiv cond(\boldsymbol{X}) = \frac{\lambda_{max}}{\lambda_{min}} \tag{C.23}$$

with λ_{max} and λ_{min} respectively the largest and smallest eigenvalues resulting from the singular value decomposition of $\boldsymbol{X}$. In R, this condition number can be easily computed with the kappa function.
As such, this metric assesses how small perturbations in the input data can affect the stability and robustness of the regression model, successfully identifying ill-posed or multicollinear regression problems. Indeed, a matrix associated with a high condition number indicates that matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is close to being singular and often exhibits strong Multicollinearity, rendering the task of correlating the variations of the response variable with the dependent challenging.

Theorem C.4.3: Correlating Condition Number with a MLE approach

A probabilistic approach provides meaningful insights to reconsider the LLS regression problem. Remember that we model the error by explicitly adding an error term following a null-centred and homoscedastic Gaussian distribution: $\epsilon \sim \mathcal{N}(0, \sigma^2)$ [a]*. Supposing that the MLE estimate,* $\hat{\boldsymbol{p}}_{mle}$*, is unbiased, its variability is given by Equation (C.24):*

$$\mathbb{V}\mathrm{ar}\left[\hat{\boldsymbol{p}}\right] = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \sigma^2 \tag{C.24}$$

, then, we have the following equality Equation (C.25):

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[\hat{\boldsymbol{p}}\right] &= \sigma^2 \\ &\Leftrightarrow \\ \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} &= 1 \\ &\Leftrightarrow \\ \left\|\boldsymbol{X}^\top \boldsymbol{X}\right\|^{-1} &= \left\|\boldsymbol{X}^\top \boldsymbol{X}\right\| = 1 \\ &\Leftrightarrow \\ \kappa(\boldsymbol{X}) &= 1 \end{aligned} \tag{C.25}$$

In other words, the variability $\mathbb{V}\mathrm{ar}\left[\hat{\boldsymbol{p}}\right]$*, which we can interpret here as the error made by the algorithm oracle, of the estimated ratios, is equal to the measure error made on the response variable* σ^2*, if, and only if, the condition number of the design matrix is equal to 1. In addition, the precision we can achieve on the estimates is bounded by the precision on the response variable.*

However, the condition number as a predictive metric for the robustness of a model suffers from specific limitations. First, hampered by its global encompassing approach of a problem, it can not be used to determine which variables are most influential. In addition, it is often diverted from its original purpose and misused to quantify and predict the impact of numerical stability, while it should not be used to take into account round-off numerical errors nor floating-point accuracy of the computer.

[a] remember from the Gaussian-Markov theorem Proof C.1.3 proves that both approaches are equivalent

C.5 Biological Appendix to the Fate of Deconvolution Algorithms

Mapping, generally employed for highplex RNA imaging assays, consists first to assign each spatially detected cell to its corresponding (scRNA-seq) profile and secondarily, infer a pattern predicting the location of each scRNA-seq cell based on its transcriptome.

Mapping workflow can be subdivided into four main stages, often referred to as the four A's ([Lon+21, Fig. 4]):

- **Adopt** From literature, a subset of the tissues or the populations of interest, with intricate spatial patterns, is selected for further analysis.
- **Assay** Survey the same tissue (to keep the same phenotypical conditions and limit technical variability) with scRNA-sequencing (its higher coverage and unbiased nature makes it a promising candidate for the selection of candidate genes) and spatial barcoding to locate their prevailing location within the tissue. Then, track the spatial and temporal dynamics of this subset of genes with HPRI imaging (recall that this method requires to know in advance the sequence of the genes).
- **Assemble** Using deconvolution and mapping algorithms, generate maps that assigns each coordinate to one cell type. Matching histology images may reveal informative landmarks and help denoising complex areas, such as the tumour leading edge, transition region between cancer and normal tissue.
- **Analyse** The high-dimensionality of ST datasets was use to corroborate ligand-receptor interactions involved in cellular signalling, or to survey evolving dynamics occurring in a disease progressing condition.

Appendix D

Appendix of Article 4

We recall for readability motivations the log-likelihood of DeCovarT's non-constrained generative model, conditioned on the purified and global bulk expression profiles, along with its gradient and its Hessian.

The conditional log-likelihood is readily computed and given by Equation (D.1):

$$\ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p}) = C + \log\left(\operatorname{Det}\left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1}\right) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu})^\top \left(\sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j\right)^{-1} (\boldsymbol{y} - \boldsymbol{p}\boldsymbol{\mu}) \tag{D.1}$$

The Jacobian is given by Equation (D.2):

$$\frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}(\boldsymbol{p})}{\partial p_j} = -2p_j \operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\right) + (\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{.j} + p_j(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\Sigma_j\boldsymbol{\Theta}(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p}) \tag{D.2}$$

The Hessian, $\mathbf{H} \in \mathcal{M}_{J\times J}$, is given by Equation (D.3):

$$\begin{aligned}
\mathbf{H}_{i,i} = \frac{\partial^2 \ell}{\partial^2 p_i} &= -2\operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) + 4p_i^2 \operatorname{Tr}\left(\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right)^2\right) - 2p_i(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\boldsymbol{.i}} - \boldsymbol{\mu}_{.i}^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{\boldsymbol{.i}} - \\
&\quad 2p_i(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - (\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\left(4p_i^2\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_i\right)\boldsymbol{\Theta}(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p}), \quad i \in \widetilde{J} \\
\mathbf{H}_{i,j} = \frac{\partial^2 \ell}{\partial p_i \partial p_j} &= 4p_j p_i \operatorname{Tr}\left(\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\Sigma}_i\right) - 2p_i(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\boldsymbol{.j}} - \boldsymbol{\mu}_{.i}^\top \boldsymbol{\Theta}\boldsymbol{\mu}_{\boldsymbol{.j}} - \\
&\quad 2p_j(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}\boldsymbol{\mu}_{.i} - 4p_i p_j(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p})^\top \boldsymbol{\Theta}\boldsymbol{\Sigma}_i\boldsymbol{\Theta}\boldsymbol{\Sigma}_j\boldsymbol{\Theta}(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{p}), \quad (i,j) \in \widetilde{J}^2, i \neq j
\end{aligned} \tag{D.3}$$

D.1 Optimisation and calculus

D.1.1 Multivariate distributions and basic algebra properties

Definition D.1.1: Multivariate Gaussian distributions

If random vector $\boldsymbol{X}$ *of size* G *follows a random multivariate Gaussian distribution,* $\boldsymbol{X} \sim \mathcal{N}_G(\mu, \boldsymbol{\Sigma})$*, then its distribution is given by:*

$$\mathrm{Det}(2\pi\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)\Sigma^{-1}(\boldsymbol{x}-\mu)^\top\right)$$

in which:

- $\mu = \boldsymbol{X}$ *is the* G*-dimensional mean vector*
- $\boldsymbol{\Sigma}$ *is a* $G \times G$ *positive-definite Definition D.1.2 covariance matrix, whose diagonal terms,* $\mathrm{Diag}(\boldsymbol{\Sigma}) = [(\mathbb{V}\mathrm{ar}[X_{i,j}]), \forall (i,j) \in \widetilde{G}^2, i = j]^\top$ *are the individual variances of each purified gene transcript in population* j *and off-diagonal terms,* $\boldsymbol{\Sigma}_{i,j} = \mathbb{C}ov[X_i, X_j], \forall (i,j) \in \widetilde{G}^2, i \neq j$ *are the covariance between variables. We note* $\Theta = \Sigma^{-1}$*, the inverse of the covariance matrix, called the precision matrix.*

Property D.1.1: Affine invariance property of multivariate GMMs

The two following properties hold for a multivariate Gaussian distribution:

- *if* $\boldsymbol{X} \sim \mathcal{N}_G(\mu, \boldsymbol{\Sigma})$*, then* $p\boldsymbol{X}$*, with* p *a constant, follows itself a multivariate Gaussian distribution, given by:* $p\boldsymbol{X} \sim \mathcal{N}_G(p\mu, p^2\Sigma)$
- *given two independent random vectors* $\boldsymbol{X_1} \sim \mathcal{N}_G(\mu_1, \boldsymbol{\Sigma_1})$ *and* $\boldsymbol{X_2} \sim \mathcal{N}_G(\mu_2, \boldsymbol{\Sigma_2})$ *following a multivariate Gaussian distribution, then the random variable* $\boldsymbol{X_1} + \boldsymbol{X_2}$ *follows itself the multivariate Gaussian distribution:*

$$X + Y \sim \mathcal{N}_G(\mu_1 + \mu_2, \boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_2})$$

By induction, this property generalises to the sum of J *independent random vectors of same dimension* $\mathbb{R}^G$*.*

Deriving the characteristic function of the multivariate GMM yields directly results reported in Property D.1.1.

Definition D.1.2: Definite matrix

A symmetric real matrix $\boldsymbol{A}$ *of rank* G *is positive-definite if:*

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0, \quad \boldsymbol{x} \in \mathbb{R}^G \tag{D.4}$$

To gain a clearer grasp of the positive-definite constraint imposed on the covariance parameter of a multivariate Gaussian distribution, let's delve into the most straightforward scenario, in which we assume that any of the individual features exhibit pairwise independence. This particular setup is parametrised by a covariance matrix containing exclusively diagonal elements.
If the matrix is not strictly positive-definite, then some of the diagonal elements can display negative values, otherwise that the individual variances for some of the covariates are negative. It is not physically possible and leads to improper, degenerate probability distributions.

D.1.2 Matrix calculus

Fundamental algebra calculus formulas used to derive first-order and second-order derivatives of the generative model of `DeCovarT` are reported in Property D.1.2 and Property D.1.3, respectively.

Property D.1.2: First-order matrix calculus

Given two invertible matrices, $A = \boldsymbol{A}(p)$ and $B = \boldsymbol{B}(p)$, functions of a scalar variable p, the following matrix calculus hold:

$$(a)\ \frac{\partial \operatorname{Det}(\boldsymbol{A})}{\partial p} = \operatorname{Det}(\boldsymbol{A}) \operatorname{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p}\right) \qquad (b)\ \frac{\partial \boldsymbol{U}\boldsymbol{A}\boldsymbol{V}}{\partial p} = \boldsymbol{U} \frac{\partial \boldsymbol{A}}{\partial p} \boldsymbol{V} \qquad (c)\ \frac{\partial \boldsymbol{A}^{-1}}{\partial p} = -\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p} \boldsymbol{A}^{-1}$$

From a) and fundamental linear algebra properties, we can readily compute applying the chain rule property on the logarithm:

$$\frac{\partial \log\left(\operatorname{Det}(\boldsymbol{A})\right)}{\partial p} = \operatorname{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p}\right)$$
$$\frac{\partial \log\left(\operatorname{Det}(\boldsymbol{A}^{-1})\right)}{\partial p} = -\operatorname{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p}\right)$$

Finally, injecting these first-order matrix derivatives, we obtain:

$$\begin{aligned} \frac{\partial (\boldsymbol{y} - \boldsymbol{x}p)^\top \Theta (\boldsymbol{y} - \boldsymbol{x}p)}{\partial p} &= -2(\boldsymbol{y} - \boldsymbol{x}p)^\top \Theta \boldsymbol{x} \\ &= -2\boldsymbol{x}^\top \Theta (\boldsymbol{y} - \boldsymbol{x}p) \end{aligned}$$
$$\text{with } \boldsymbol{A} = \boldsymbol{D} = -\boldsymbol{x} \in \mathbb{R}^G, \quad \boldsymbol{b} = \boldsymbol{e} = \boldsymbol{y}, \quad \boldsymbol{C} = \boldsymbol{\Theta} \text{ symmetric}$$

Property D.1.3: Second-order matrix calculus

Given an invertible matrix $\boldsymbol{A}$ depending on a variable p, the following calculus formulas hold:

$$(a)\ \frac{\partial^2 \boldsymbol{A}^{-1}}{\partial p_i \partial p_j} = \boldsymbol{A}^{-1} \left(\frac{\partial \boldsymbol{A}}{\partial p_i} \boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p_j} - \frac{\partial^2 \boldsymbol{A}}{\partial p_i \partial p_j} + \frac{\partial \boldsymbol{A}}{\partial p_j} \boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p_i} \right) \boldsymbol{A}^{-1} \qquad (b)\ \begin{matrix} \frac{\partial \operatorname{Tr}(\boldsymbol{A})}{\partial p_i} \\ \operatorname{Tr}\left(\frac{\partial \boldsymbol{A}}{\partial p_i}\right) \end{matrix} =$$

Combining Property D.1.2 with the linear property of the trace operator yields:

$$\frac{\partial^2 \log\left(\operatorname{Det}(\boldsymbol{A}^{-1})\right)}{\partial^2 p} = -\operatorname{Tr}\left[\boldsymbol{A}^{-1} \frac{\partial^2 \boldsymbol{A}}{\partial^2 p_i}\right] + \operatorname{Tr}\left[\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p_i}\right)^2\right]$$

D.1.3 First and second-order derivation of constrained DeCovarT

To reparametrise the log-likelihood function (Equation (D.1)) in order to explicitly handling the unit simplex constraint (Equation (C.2)), we consider the following mapping function: $\boldsymbol{\psi} : \boldsymbol{\theta} \to \boldsymbol{p} \mid \quad \boldsymbol{\theta} \in \mathbb{R}^{J-1}, \boldsymbol{p} \in]0,1[^J$ (Equation (D.5)):

1.
$$\boldsymbol{p} = \boldsymbol{\psi}(\boldsymbol{\theta}) = \begin{cases} p_j = \frac{e^{\theta_j}}{\sum_{k<J} e^{\theta_k} + 1}, \; j < J \\ p_J = \frac{1}{\sum_{k<J} e^{\theta_j} + 1} \end{cases} \tag{D.5}$$

2. $\boldsymbol{\theta} = \boldsymbol{\psi}^{-1}(\boldsymbol{p}) = \left(\ln\left(\frac{p_j}{p_J}\right)\right)_{j \in \{1,\dots,J-1\}}$

that is a C^2-diffeomorphism, since $\boldsymbol{\psi}$ is a bijection between $\boldsymbol{p}$ and $\boldsymbol{\theta}$ twice differentiable. Its Jacobian, $\mathbf{J}_{\boldsymbol{\psi}} \in \mathcal{M}_{J\times(J-1)}$ is given by Equation (D.6):

$$\mathbf{J}_{i,j} = \frac{\partial p_i}{\partial \theta_j} = \begin{cases} \frac{e^{\theta_i} B_i}{A^2}, & i = j, \; i < J \\ \frac{-e^{\theta_j} e^{\theta_i}}{A^2}, & i \neq j, \; i < J \\ \frac{-e^{\theta_j}}{A^2}, & i = J \end{cases} \tag{D.6}$$

with i indexing vector-valued $\boldsymbol{p}$ and j indexing the first-order order partial derivatives of the mapping function, $A = \sum_{j'<J} e^{\theta_{j'}} + 1$ the sum over exponential (denominator of the mapping function) and $B = A - e^{\theta_i}$ the sum over ratios minus the exponential indexed with the currently considered index i.

The Hessian of the multi-dimensional mapping function $\boldsymbol{\psi}(\boldsymbol{\theta})$ exhibits symmetry for each cell ratio component j, as anticipated in accordance with Schwarz's theorem. It is is a third-order tensor of rank $(J-1)(J-1)J$, given by Equation (D.7):

$$\frac{\partial^2 p_i}{\partial k \partial j} = \begin{cases} \frac{e^{\theta_i} e^{\theta_l}\left(-B_i + e^{\theta_i}\right)}{A^3}, \; (i < J) \wedge ((i \neq j) \oplus (i \neq k)) \quad (a) \\ \frac{2 e^{\theta_i} e^{\theta_j} e^{\theta_k}}{A^3}, \; (i < J) \wedge (i \neq j \neq k) \quad (b) \\ \frac{e^{\theta_i} e^{\theta_j}\left(-A + 2e^{\theta_j}\right)}{A^3}, \; (i < J) \wedge (j = k \neq i) \quad (c) \\ \frac{B_i e^{\theta_i}\left(B_i - e^{\theta_i}\right)}{A^3}, \; (i < J) \wedge (j = k = i) \quad (d) \\ \frac{e^{\theta_j}\left(-A + 2e^{\theta_j}\right)}{A^3}, \; (i = J) \wedge (j = k) \quad (e) \\ \frac{2 e^{\theta_j} e^{\theta_k}}{A^3}, \; (i = J) \wedge (j \neq k) \quad (f) \end{cases} \tag{D.7}$$

with i indexing $\boldsymbol{p}$, j and k respectively indexing the first-order and second-order partial derivatives of the mapping function with respect to $\boldsymbol{\theta}$. In line (a), $\oplus$ refers to the Boolean XOR operator, $\wedge$ to the AND operator and $l = \{j, k\} \setminus i$.

To derive the log-likelihood function in Equation (D.2), we reparametrise $\boldsymbol{p}$ to $\boldsymbol{\theta}$, using a standard *chain rule formula*. Considering the original log-likelihood function, Equation (D.1), and the mapping function, Equation (D.5), the differential at the first order and at the second order is given by Equation (D.8) and Equation (D.9), respectively defined in $\mathbb{R}^{J-1}$ and $\mathcal{M}_{(J-1)\times(J-1)}$:

$$\left[\frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}}{\partial \theta_j}\right]_{j<J} = \sum_{i=1}^{J} \frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}}{\partial p_i} \frac{\partial p_i}{\partial \theta_j} \tag{D.8}$$

$$\left[\frac{\partial \ell^2_{\boldsymbol{y}|\boldsymbol{\zeta}}}{\partial \theta_k \theta_j}\right]_{j<J,\, k<J} = \sum_{i=1}^{J}\sum_{l=1}^{J}\left(\frac{\partial p_i}{\partial \theta_j} \frac{\partial^2 \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}}{\partial p_i \partial p_l} \frac{\partial p_l}{\partial \theta_k}\right) + \sum_{i=1}^{J}\left(\frac{\partial \ell_{\boldsymbol{y}|\boldsymbol{\zeta}}}{\partial p_i} \frac{\partial^2 p_i}{\partial \theta_k \theta_j}\right) \quad (d) \tag{D.9}$$

D.2 A MCMC Algorithm for the Joint Distribution of Purified Profiles and Ratios

We introduce two variations of the MCMC algorithm, namely the *Metropolis-Hasting* and the *Gibbs sampling* algorithms. They are respectively tailored to approximate distributions for which no explicit form is known (Definition D.2.1) or streamline the optimisation of strongly dependent parameters (Definition D.2.2).

D.2.1 An Introduction to Gibbs and Metropolis Hasting Samplers

Definition D.2.1: Metropolis-Hasting algorithm

First of all, we introduce some key notations:

- *Function $f(\theta|.)$ is the target distribution, usually the posterior distribution that encompasses both prior knowledge and new data.*
- *The distribution $q(\theta|\theta^{(q-1)})$ is the proposal distribution, alternatively known as the transition kernel, and the transient value sampled from it, $\theta^{(*)}$, is the proposal estimate.*
- *The probability of accepting the proposal is naturally called the acceptance probability.*

Each iteration, indexed by (q), of the MH algorithm includes the following steps:

1. *Draw a proposal, $\theta^{(*)}$, from conditional distribution $q(\theta|\theta^{(q-1)})$.*
2. *Compute the acceptance probability parameter, noted α:*

$$K(\theta^{(*)}) = \min\left(\frac{f(\theta^{(*)}|.)}{f(\theta^{(q-1)}|.)}\frac{q(\theta|\theta^{(q-1)})}{q(\theta^{(q-1)}|\theta)}, 1\right)$$

   Typically, choices of the acceptance probability and the kernel distribution are tailored to satisfy the balance condition of the MCMC chains and ensure that the chain behaviour reproduces the sampling pattern of the desired distribution.

3. *The decision of whether to accept or reject the new state proposal is determined by the previously computed acceptance probability parameter, denoted as α. This parameter sets the threshold for accepting a value drawn from a standard Uniform distribution, $u \sim \mathbb{U}[0,1]$. This mathematical protocol is further described in Equation (D.10):*

$$\theta^{(q)} = \begin{cases} \theta^{(*)}, & u \leq K(\theta^{(*)}) \\ \theta^{(q-1)}, & u > K(\theta^{(*)}) \end{cases} \tag{D.10}$$

   In the first case, we say that the proposal is accepted*, while in the second case, it is* rejected*.*

One of the major advantages of the Metropolis-Hastings (MH) algorithm lies in the design of its acceptance probability function, denoted as $K(\theta^{(*)})$. Indeed, by involving the computation of a ratio between two density functions, the normalisation constant, which represents the value of the marginal likelihood and is usually intractable to compute is naturally cancelled out.

In our modelling framework, the acceptance function further simplifies with the choice of proposing a new proposal by adding an error term following a null-centred, multivariate and symmetric distribution (Property D.2.1):

Property D.2.1: Random walk Metropolis-Hastings

Indeed, from the analytical properties of the kernel distribution (Equation (D.11) and Equation (D.12)):

$$\begin{aligned} q(\theta^{(*)}|\theta^{(q-1)}) &= q(\epsilon) \\ q(\theta^{(q-1)}|\theta^{*}) &= q(-\epsilon) \end{aligned} \tag{D.11}$$

$$q(\epsilon) = q(-\epsilon) \tag{D.12}$$

,
the acceptance probability function simplifies to Equation (D.13) ([Tab21]):

$$K(\theta^{(*)}) = \min\left(\frac{f(\theta^{(*)}|.)}{f(\theta^{(q-1)}|.)}, 1\right) \tag{D.13}$$

Definition D.2.2: Gibbs sampling

The fundamental concept behind Gibbs sampling is to break down the joint posterior distribution of the parameters, into a product of conditional distributions of the parameters. To that end, it is generally assumed that there exists a natural partition of the hidden parameters allowing them to factorise in a meaningful way.

It is usually implemented when the joint conditional posterior distribution is intractable to compute, whereas the conditional distribution for a subset of the parameters, conditioned on all others, is rather straightforward to derive. This is especially the case when the set of hidden parameters is linked to each other, and that numerous numerical constraints linking them must be endorsed.

Using our notations, the following joint posterior distribution $f(\boldsymbol{p}, \boldsymbol{X}|\mathcal{D})$, with $\mathcal{D}$ denoting the observed data, here $\boldsymbol{y}$ is analytically complex to derive, while $f(\boldsymbol{p}|\boldsymbol{X}, \mathcal{D})$ and $f(\boldsymbol{X}|\boldsymbol{p}, \mathcal{D})$, the posterior cellular ratios and purified individual cell expression profiles, respectively, can be simply computed. In practice, like any MCMC framework, you start to initiate the values for all the parameters. Then, the iterated Gibbs process samples each parameter (or subset of parameters), one at a time, and updates its value conditioned on the other parameters at their current values. It can be proven that after a sufficient number of iterations, the corresponding Markov chain of parameters converges and approximates well the desired joint distribution.

D.2.2 Pseudo-code Gibbs sampler

We detail in Algorithm 0 a potential pseudo-code to generate MCMC chains of the joint distribution of the parameters of interest, in which variable q denotes the running index, B the number of *burn-in* iterations to be discarded after sampling, and Q the actual length of the resulting Markov chain.

Algorithme 0 : Pseudo-code of the iterated optimisation method, to retrieve the parameters of DeCovarT's generative model.

Input :

- Prior estimates of the mean, $\left[\boldsymbol{\mu}_{ji} \in \mathbb{R}^G\right]$, $j \in \{1, \ldots, J\}$ and covariance, $\left[\boldsymbol{\Sigma}_{ji} \in \mathcal{M}_{\mathbb{R}^{G\times G}}\right]$, $j \in \{1, \ldots, J\}$ of each cell population.
- Initial estimates of cellular ratios, $\boldsymbol{p}_{0i}$, and purified cell expression profiles, $\boldsymbol{X}_{0i}$, for each individual. They should align with both the fundamental linear deconvolution assumption (Equation (C.1), and the unit-simplex constraint (Equation (C.2)).
- Standard deviation, $\sigma_{0(\boldsymbol{\rho})}$ and $\sigma_{0(\boldsymbol{X})}$ of the additional residual term added to each cellular ratio and each individual cell profile of the *transition kernel*, respectively. [SK19], [And+18], [VK21] and [Mar+20] suggests tuning these hyper-parameters such that the acceptance rates in the long term are bounded between 0.234 and 0.574.

for $q \leftarrow 1$ **to** $(B + Q)$ **do**
  for $i = 1 : N$ **do**
    $\epsilon_{i(\boldsymbol{\rho})} \sim \mathcal{N}_{J-1}(0, \sigma^2_{0(\boldsymbol{\rho})}\, \mathbb{I}_{J-1})$
    $\boldsymbol{\rho}_i^{(*)} = \boldsymbol{\rho}_i^{(q-1)} + \epsilon_{i(\boldsymbol{\rho})}$
    for $j = 1 : (J - 1)$ **do**
      $u \sim \mathbb{U}(0, 1)$
      if $u < \min\left(1, K_\rho\left(\rho_j^{(q-1)} \rightarrow \rho_j^{(*)}\right)\right)$ **then**
        $\rho_{ji}^{(q)} \leftarrow \rho_{ji}^{(*)}$
      else
        $\rho_{ji}^{(q)} \leftarrow \rho_{ji}^{(q-1)}$
      end
      $\boldsymbol{p}_i^{(q)} = \boldsymbol{\psi}(\boldsymbol{\rho}_i^{(q)})$ $\quad (i)$
      MCMC.$\boldsymbol{p_i} \leftarrow \boldsymbol{p}_i^{(q)}$
    end
    for $j = 1 : (J - 1)$ **do**
      $\varepsilon_{ji(\boldsymbol{X})} \sim \mathcal{N}_G(0, \sigma^2_{0(\boldsymbol{X})}\, \mathbb{I}_G)$
      $\boldsymbol{x}_{ji}^{(*)} = \boldsymbol{x}_{ji}^{(q-1)} + \varepsilon_{ji(\boldsymbol{X})}$
      $u \sim \mathbb{U}(0, 1)$
      if $u < \min\left(1, K_{\boldsymbol{x}}\left(\boldsymbol{x}_{ji}^{(q-1)} \rightarrow \boldsymbol{x}_{ji}^{(*)}\right)\right)$ **then**
        $\boldsymbol{x}_{ji}^{(q)} \leftarrow \boldsymbol{x}_{ji}^{(*)}$
      else
        $\boldsymbol{x}_{ji}^{(q)} \leftarrow \boldsymbol{x}_{ji}^{(q-1)}$
      end
    end
    $\boldsymbol{x}_{Ji}^{(q)} \leftarrow \frac{\boldsymbol{y}_i - \sum_{j=1}^{J-1} \boldsymbol{x}_{ji}^{(q)} p_{ji}^{(q)}}{p_{Ji}^{(q)}}$ $\quad (ii)$
    MCMC.$\boldsymbol{X}_i \leftarrow \boldsymbol{X}_i^{(q)}$
  end
end

We implemented two reparametrisations at each global iteration (outer loop) to ensure that the kernel distribution generates proposals that fall within the "support" of the target distribution, and notably the adhesion to the fundamental linear deconvolution relation (Equation (C.1)) and the unit-simplex constraint over the cellular ratios (Equation (C.2)):

1. First, the mapping function, described in (i) and in eq. (D.5), enforces that the estimated cellular ratios adhere to the unit-simplex constraint (Equation (C.2)).

2. Second, the fundamental linearity of deconvolution, Equation (C.1), is endorsed by the update formula (ii). Formula Equation (C.1) implies that the last J cellular expression profile, $\boldsymbol{x}_{.j}$ is not a free parameter, and can be rewritten as a combination of the others, given by Equation (D.14):

$$\begin{aligned} \boldsymbol{y}_i = \sum_{j=1}^{J-1} p_{ji}^{(q)} \boldsymbol{x}_{ji}^{(q)} + p_{Ji}^{(q)} \boldsymbol{x}_{Ji}^{(q)} = \sum_{j=1}^{J} p_{ji}^{(q-1)} \boldsymbol{x}_{ji}^{(q-1)} \\ \iff \\ \boldsymbol{x}_{Ji}^{(q)} = \frac{\boldsymbol{y}_i - \sum_{j=1}^{J-1} \boldsymbol{x}_{ji}^{(q)} p_{ji}^{(q)}}{p_{Ji}^{(q)}} \end{aligned} \tag{D.14}$$

D.2.3 Derivation of the Acceptance Probability Function

By utilising a Random Walk MH approach, additionally cancelling out the normalisation constant (Equation (D.13)), the acceptance probability function to compute is simply the product of the prior distributions and the likelihood of the observed data, $f(\boldsymbol{\theta}) \times f(\mathcal{D}|\boldsymbol{\theta})$.

To simplify further this product of distributions, we preliminary suppose that the density distribution characterising the priors is improper, in other words, that $f(\boldsymbol{\theta})$ is always equal to 1 on $\boldsymbol{\Theta}$. As we provide a closed form of the log-likelihood of our generative model (eq. (D.1)), and not the likelihood, we need to apply an exponential transformation to recover the desired acceptance probability (Equation (D.15)):

$$K(\theta^{(*)}) = \min\left(1, \exp\left(\ell(\theta^{(*)}|.) - \ell(\theta^{(q-1)}|.)\right)\right) \tag{D.15}$$

with:

- $\boldsymbol{\theta}^{(*)}$, the current proposal (either $\boldsymbol{p}^{(*)}$ or $\boldsymbol{X}^{(*)}$)
- $\ell(\boldsymbol{\theta}^{(*)}|.) = \log\left(f(\boldsymbol{Y}|\boldsymbol{\theta}^{(*)}, \boldsymbol{\zeta})\right)$, the log-likelihood of the currently observed data, conditioned on the current values of the latent parameters to estimate and the user-defined parameters $\boldsymbol{\zeta}$, given in our generative model by Equation (D.1).
- $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\quad \boldsymbol{\mu} = (\boldsymbol{\mu}_{.j})_{j \in \widetilde{J}} \in \mathcal{M}_{G \times J}$, $\quad \boldsymbol{\Sigma} = \sum_{j=1}^{J} p_j^2 \boldsymbol{\Sigma}_j \in \mathcal{M}_{G \times G}$ denote the parameters provided by the user before conducting the study.

Appendix E

Article 5: Gene clustering applied to primary Sjögren's disease

In this appendix, we present a comprehensive transcriptome clustering method aimed at expanding upon and elucidating the patient clustering outlined in Chapter 4 (see also [Sor+21]) by adopting a holistic paradigm. Actually, the Gaussian mixture-based clustering approach we employed in Chapter 4 assumes that each covariate input, such as gene expression, is independent to each other. In addition, the high dimensionality of the dataset used for clustering Sjögren's patients prevents from uncovering the key drivers contributing to the biological heterogeneity observed among patients.

Hence, the primary rationale driving the unsupervised inference of gene modules is to understand the variety of molecular profiles observed across patients afflicted by the Sjögren's syndrome, by explicitly integrating gene co-expression structures. Indeed, studying the biological mechanisms at the gene pathway, or module, level, reduces by several factors the dimensionality and the complexity of the data space studied. It thus enables better understanding of highly-connected biological mechanisms, and the exploratory capability of downstream analyses.

To summarise the major output of the paper, we identified 13 Consensus gene Modules (CMs) that contribute the most to the variability of the transcriptome across pSD patients. We retrieved them using unsupervised clustering methods on four different transcriptomic datasets, all processes from blood samples. Then, gene set enrichment analyses were used to annotate each module based on its connection with cell populations or biological function. Finally, flow cytometry data and cytokine measurements were used to validate the biological annotations inferred from pathway enrichment analysis. Figure E.1 details the major steps of the clustering pipeline, as well as the major biological results.

In details, the four datasets originate from both private sources, all composing the NECESSITY consortium meta-analysis (ASSESS [Got+13], PRECISEADS [Bar+18], and UKPSSR [Ng+11]) and publicly available repositories (GSE84844 [Tas+17]).To reduce partly the intrinsic high dimensionality of transcriptomic data [1], often prone to high confusing technical noise, we tailored a dedicated analysis workflow, summarised in Figure E.1(a).

[1] With 20 000 identified genes coding for proteins, the number of pairwise interactions reaches the staggering number of 4×10^8 correlation coefficients.

Initially, each cohort's gene expression matrix was transformed into an affinity matrix representing the gene co-expression network. Each pairwise Pearson correlation coefficient was mapped to a non-linear but monotonic custom affinity function. This transformation provides similar benefits to the sigmoid function by effectively reducing low and insignificant correlation coefficients towards zero, as reported in [Wan+14]. Subsequently, an additional filtration step allows for further refinement of the constructed network, retaining only the most significant pairwise interactions between co-expressed genes, thereby generating a sparse weighted graph. Secondly, we used the Similarity Network Fusion (SNF) algorithm ([Wan+14] and Figure E.1(b)) to integrate multiple affinity matrices, in order to optimise shared correlation patterns across the four distinct cohorts of pSD patients.

Finally, we applied Louvain clustering [Blo+08] to this sparse, *consensus graph* and identified 13 CMs. In brief, Louvain's method is one of the graph clustering algorithms that focuses on maximising the *modularity* of the network. This metric, bounded between -1 and 1, aims at computing the averaged density of edges within clusters with respect to the interaction density between communities, a graph composed only of *cliques* unconnected to each other displaying a score of 1. However, Louvain's paper implements two innovative features: an approximate heuristic algorithm, briefly described in Figure E.1(c) to increase the scalability of the method to larger datasets (it has even been applied to social network, with tens of millions of user nodes) and an internal hierarchical approach, enabling to adjust the level of granularity to user requirements.

We found that these CMs were highly consistent and correlated across cohorts. We found out that CM1 was correlated with type 1 interferon signalling, CM7 corroborates cell cycle-related genes and out of the 11 remaining, 9 modules were significantly enriched in pathways involved in lymphoid and myeloid development and signalling.

We also looked into the therapeutic effect of a drug combination of hydroxychloroquine and leflunomide on the blood transcriptome of pSD patients and revealed that the expression of some modules was significantly linked to treatment outcome, suggesting these modules could be leveraged as predictive biomarkers.

While the development of the primary components of the pipeline and the biological interpretation of the results were primarily conducted by another member of my research team, my contributions to this paper encompass the following tasks:

- I largely contributed with my industrial research team to the release of a consensual pipeline, dedicated to the pre-processing, normalisation and downstream analysis of transcriptomic datasets (see Appendix A).
- While various clustering methods were assessed for identifying closely related transcript networks, the Gaussian mixture approach exhibited significantly lower performance compared to the Louvain's method. This could be attributed to certain assumptions associated with standard GMM not being met, such as the requirement for the number of observations (genes) to significantly exceed the number of variables (biological samples).

  Parsimonious parametrisations or projections to sub-dimensional space, as detailed in Appendix B combined with careful optimisation of the hyper-parameters (number of components, initial estimates, ...), may improve the performance of the clustering.

Consensus gene modules strategy identifies candidate blood-based biomarkers for primary Sjögren's disease

Cheïma Boudjeniba[1,2,3], Perrine Soret[1], Diana Trutschel[3], Antoine Hamon[4], Valentin Baloche[9], Bastien Chassagnol[1], Emiko Desvaux[1], Antoine Bichat[1], Audrey Aussy[1], Philippe Moingeon[1], Céline Lefebvre[1], Sandra Hubert[1], Marta Alarcón-Riquelme[5], Wan-Fai Ng[6], Jacques-Eric Gottenberg[7], Benno Schwikowski[3], Michele Bombardieri[8], Joel A.G. van Roon[9], Xavier Mariette[10], Mickaël Guedj[1], Etienne Birmele[11], Laurence Laigle[1] and Etienne Becht[1]

[1] Translational Medicine, Servier, Research and Development, Gif-Sur-Yvette, France
[2] Laboratoire MAP5 UMR 8145, Université Paris Cité, Paris, France
[3] Computational Systems Biomedicine Lab, Institut Pasteur, Université Paris Cité, F-75015 Paris, France
[4] Lincoln, Research and development, Paris, France
[5] GENYO, Centre for Genomics and Oncological Research. Pfizer, University of Granada, Spain
[6] Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK
[7] Rhumatologie, hôpitaux universitaires Strasbourg, CHU de Strasbourg, Strasbourg, France
[8] Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Barts and the London, School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK
[9] Department of Rheumatology and Clinical Immunology and Center for Translational Immunology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[10] Department of Rheumatology, Université Paris-Saclay, INSERM UMR1184, AP-HP, Hôpital Bicêtre, Le Kremlin Bicêtre, France
[11] Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS, Strasbourg, France

Corresponding author: etienne.becht@servier.com

Abstract *Primary Sjögren disease (pSD) is an autoimmune disease characterized by lymphoid infiltration of exocrine glands leading to dryness of the mucosal surfaces and by the production of autoantibodies. The pathophysiology of pSD remains elusive and no treatment with demonstrated efficacy is available yet. To better understand the biology underlying pSD heterogeneity, we aimed at identifying Consensus gene Modules (CMs) that summarize the high-dimensional transcriptomic data of whole blood samples in pSD patients. We performed unsupervised gene classification on four data sets and identified thirteen CMs. We annotated and interpreted each of these CMs as corresponding to cell type abundances or biological functions by using gene set enrichment analyses and transcriptomic profiles of sorted blood cell subsets. Correlation with independently measured cell type abundances by flow cytometry confirmed these annotations. We used these CMs to reconcile previously proposed patient stratifications of pSD. Importantly, we showed that the expression of modules representing lymphocytes and erythrocytes before treatment initiation is associated with response to hydroxychloroquine and leflunomide combination therapy in a clinical trial. These consensus modules will help the identification and translation of blood-based predictive biomarkers for the treatment of pSD.*

Keywords Precision Medicine, Sjögren Disease, Unsupervised learning, Integrated analysis.

Introduction

Primary Sjögren Disease (pSD) is a chronic, disabling inflammatory autoimmune disease characterized by lymphoid infiltration of exocrine glands leading to dryness of the mucosal surfaces, such as the mouth and eyes and by the production of specific auto-antibodies[1–3]. Long-term complications include ocular and dental diseases, systemic involvement, organ damages and increased risk of lymphoma with excess mortality[4, 5]. This pathology is affecting between 0.05% and 0.4% of the adult population[6–9] and is the second most common systemic autoimmune disease[10]. It affects women more often than men (9:1) and the peak frequency of the disease is around fifty years of age[11].

The advent of new technologies has provided a path towards the development of classification criteria for autoimmune diseases that are based on molecular patterns representing disease mechanisms and molecular pathways[12, 13]. By applying computational methodologies to clinical and multi-omic datasets, several pSD disease taxonomies have recently been proposed. Indeed, Tarn et al. proposed a symptom-based stratification of patients with pSD[14], while Soret et al.[15] and Trutschel et al.[16] proposed a molecular classification of pSD based on whole blood transcriptomic profiles of pSD patients. These classifications may provide useful clinical insights on disease subtypes of pSD patients but remain limited in the characterization of the biology underlying the disease in each patient subgroup. Indeed, pathogenesis of autoimmunity involves dysfunction of the entire immune system, and many cellular or functional components, including neutrophils, dendritic cells (DCs), macrophages, T and B cells, cytokine signaling pathways or autoantibodies[17, 18].

The clinical manifestations and biological disturbances associated with pSD are indeed highly heterogeneous among individuals which complicates its diagnosis. Mechanistically, the pathophysiology of pSD remains elusive[19]. No targeted therapy is therefore currently approved and only symptomatic treatments are offered[20, 21]. Precision Medicine approaches designed to better address the needs of patients based on the specific biological mechanisms underlying their symptoms would greatly improve the management of patients suffering from pSD.

The IMI 2 NECESSITY European consortium was launched in 2019 to identify a new composite clinical endpoint, biomarkers for stratifying patients and predictive biomarkers of treatment response for pSD, and test them in a prospective clinical trial. To achieve these goals, members of the NECESSITY consortium share clinically-annotated datasets, including whole blood transcriptomic datasets of pSD patients. These transcriptomes allow the identification of biological heterogeneity across pSD patients and its potential link with response to treatments, but were produced using diverse transcriptomic technologies, making their combined analysis challenging.

In order to jointly analyze independent whole blood transcriptomic datasets of pSD patients, we used a graph theoretical approach to unify four correlation networks into a consensus graph linking positively correlated genes. By clustering this unified representation of multiple cohorts, we identified 13 consensus transcriptomic gene modules that summarize the pathophysiology of pSD at the blood level. We annotated each of these modules for correspondence with cell types or molecular pathways, and validated these biological interpretation with matching flow cytometry data or cytokine measurements whenever available. We used these modules to better characterize and reconcile previously-published pSD patient stratifications[15, 16]. Importantly, we investigated clinical trial data to decipher the impacts of treatments on the peripheral blood of patients and propose a model predictive of the response to leflunomide-hydroxychloroquine combination therapy.

Results

Identification of thirteen consensus gene modules (CMs) from whole blood transcriptomes of pSD patients

We analyzed four whole blood transcriptomic datasets from pSD patients. Three were provided by the NECESSITY consortium: ASSESS[22] ($n = 371$), PreciseSADS[12] ($n = 341$) and UKPSSR[23] ($n = 144$). We also included the publicly-available GSE84844[24] dataset ($n = 30$). Our goal was to identify consistent signals across these four sources, and in particular consensus gene modules (CMs) of coexpressed genes. Transcriptomic data sets are however high dimensional which can hamper the correct identification of gene modules. Indeed, spurious correlations may appear due to the size and noisiness of the data: 20,000 protein coding genes indeed correspond to 400×10^6 correlation coefficients. To ensure that the CMs we identify were reproducible across a large range of blood transcriptomic data sets (from distinct pSD cohorts), we used a dedicated analysis workflow summarized in **Figure 1A**. We first converted each cohort's gene expression matrix to an affinity matrix (gene co-expression network). This affinity is non-linearly and monotonically linked to the observed correlation between two genes and shrinks low correlation coefficients towards 0 (See **Methods** and Wang *et al.*[25]). We applied Similarity Network Fusion (SNF)[25], a computational method designed for the merging of multiple affinity matrices, generating a consensual representation of genes' pairwise

similarities in the blood of pSD patients across these four independent cohorts (**Figure 1B**). We pruned the consensual affinity matrix to obtain a sparse weighted graph with edges corresponding to highly co-expressed genes (**Supplementary Figure 1**). Finally, Louvain clustering[26] of the sparse graph (see **Methods**) identified 13 CMs (**Supplementary Table 1**). We confirmed a posteriori that these CMs are reproducible groups of highly co-expressed genes that are reproducible across the four datasets (**Figure 1C**).

Biological interpretation of the CMs

The 13 CMs represent the main axes of heterogeneity of the blood transcriptome across pSD patients and can therefore facilitate the interpretation of high dimensional transcriptomic data by summarizing it using 13 dimensions. In order to biologically interpret these 13 axes of variation, we annotated each of them as corresponding to cell types or biological functions by using gene set enrichment analyses using gene sets from the Gene Ontology[27] and Altman *et al.*[28] databases (**Figure 2A, 2B**), as well as their average expression in transcriptomic profiles of sorted blood cell subsets[29] (**Figure 2C**).

CM1 was enriched in Interferon related as well as response to viruses pathways, and we interpreted it as representing type 1 IFN signaling. CM7 was enriched in cell cycle-related genes, and we interpreted it as a transcriptomic signature of mitosis within blood cells.

Out of the 11 other modules, 9 represent different cell types. We found four modules corresponding to lymphoid cells: CM4, CM5 and CM11 were respectively enriched in pathways associated with T cells, NK cells and B cells functions **(Figure 2A, 2B)** and that were overexpressed in the transcriptome of the corresponding purified cell types **(Figure 2C)**. CM8 was enriched in genes associated with gene transcription and overexpressed across the transcriptomes of purified lymphocytes (T, B and NK cells) and therefore represents a shared gene transcription signature across all lymphocytes **(Figure 2C)**. We found six modules (CM2, CM6, CM9, CM10, CM12, CM13) representing myeloid cell subsets. CM2 was enriched in erythrocytes-annotated gene sets and CM10 in platelets-annotated gene sets. Module CM6 was overexpressed in the transcriptome of eosinophils. CM9 and CM13 were enriched in inflammation and neutrophil-related gene sets and overexpressed in the transcriptome of purified granulocytes and neutrophils. CM13 was in addition enriched in genes from the I-κB kinase/NF-κB signaling pathway, an inflammatory transcription factor expressed by neutrophils[30]. Finally, CM12 was enriched in gene sets related to monocytes and overexpressed in the transcriptome of cells derived from monocytes.

Among the 13 CMs, CM3, which contains the highest number of genes (n=1247), was the least co-expressed, had the lowest absolute expression levels **(Supplementary Figure 2)** module and showed inconsistent characterization results **(Figure 2A, 2B)**. We therefore did not take it into consideration for further analysis. In summary, we interpreted CM1 as type 1 interferon (IFN) activation, CM2 as representing the frequency of erythrocytes within the blood, CM3 as residual variance, CM4, CM5, CM6 as the frequencies of respectively T cells, NK cells and Eosinophils, CM7 as a signature of cell proliferation, CM8, CM10, CM11 and CM12 as the frequencies of respectively lymphocytes, platelets, B cells and monocytes, and CM9 and CM13 as representing neutrophils.

Validation of the biological interpretations of the CMs

To confirm the biological interpretations of the CMs representing cell types, we compared their average expressions (**Material and Methods**) to the corresponding cellular frequencies measured by flow cytometry in matching samples whenever available (**Figure 3A**). For functional modules, we compared them to previously-published gene signatures (**Figure 3B**) or cytokines concentrations (**Figure 3C**).

For all the cellular modules for which we had matching cytometry data, we observed a high and significant correlation of the average module expression with the frequency among live single cells measured by flow cytometry (**Figure 3A**). More precisely, we observed correlation coefficients of 0.71 between the CM4 module and the frequency of T cells, of 0.51 between the CM5 modules and NK cells, of 0.39 between CM6 and eosinophils, 0.75 (respectively 0.64) between CM9 (respectively CM13) and

neutrophils, 0.84 between CM11 and B cells, 0.67 between CM12 and monocytes, and 0.62 between CM8 and lymphocytes (all p-values $< 2 \times 10^{-12}$).

For functional modules, we observed a strong correlation (Pearson's $r > 0.94$) of the CM7 with genes signatures corresponding to phases of the mitotic cycle identified with single cell RNA-sequencing data[31]. The other functional module CM1 was highly correlated with the concentration of type 1 IFN (measured by SIMOA) in the blood ($r = 0.65, p = 3.3 \times 10^{-11}$) (**Figure 3C**). Collectively, these analyses confirm the interpretation of the CMs derived from gene set enrichment analyses.

The consensus gene modules identify consistency and heterogeneity across pSD patient stratifications

Three studies have proposed pSD patients stratifications according to molecular and clinical features of the disease[14–16]. Two methods were based on blood transcriptomic profiles of pSD patients on two distinct cohorts[15, 16]. Both studies identified four clusters of patients hereafter referred to as S1, S2, S3 and S4 (respectively T1, T2, T3 and T4) for the Soret (respectively Trutschel) classification. These stratifications were established using unsupervised clustering methods. Algorithmic classifiers to stratify new pSD cohorts according to these classification systems are however currently lacking, and no direct comparison has been performed so far.

Briefly, from Soret *et al.*, cluster S1 exhibited high levels of interferon (IFN) activity and an increased frequency of B lymphocytes in the blood. Cluster S2 showed a similar expression profile to that of healthy volunteers. Cluster S3 displayed a high IFN signature, along with a more prominent involvement of B cell components compared to other clusters, including an increased frequency of B cells in the blood. Lastly, cluster C4 was characterized by an inflammatory signature driven by monocytes and neutrophils. Confirming the findings of *et al.*[15], our analysis confirmed the defining characteristics of these patient clusters. We consistently observed an upregulation of the Interferon module CM1 in S1 patients, the Neutrophils module CM9 in S4 patients, and the B cell module CM11 in S3 patients (Figure 4A). Our analysis further revealed that S3 is defined by a high abundance of lymphocytes (B, T, and NK cells represented by the CM11, CM4, and CM5 modules, respectively) associated with cell proliferation (CM7). Cluster S4 is characterized by a high abundance of platelets (CM10), erythrocytes (CM2), and neutrophils (CM9 and CM13). S1 is distinguished by high activation of type 1 IFN (CM1), while S2, described as normal-like by Soret *et al.*, has fewer monocytes (CM12) and more T cells (CM4) compared to the cohort's averages.

In a separate study by Trutschel *et al.*, four patient clusters were also identified. These clusters were based on two modules: IFN-stimulated genes (ISGs) and the erythroid module (ERM). Cluster T1 showed high expression of both these modules, while cluster T2 had low ISG expression but high ERM expression. Cluster T3 had high ISG expression and low ERM expression, and cluster T4 had low expression in both ISGs and ERM. We observed a high interferon signature (CM1) in clusters T1 and T3, with cluster T1 exhibiting a higher platelet presence compared to cluster T3 (Figure 4B). Cluster T2 had a lower abundance of monocytes (CM12), while cluster T4 had a high neutrophil signature (CM13). Cluster T1 had a high presence of erythrocytes, cluster T3 had fewer eosinophils (CM6), and clusters T3 and T4 had a higher abundance of lymphocytes (CM8).

To formally study the correspondence between the Soret and Trutschel classification systems, we computed Pearson correlation coefficients across centroids computed on mean-centered and unit variance-scaled module expression scores. This comparison highlighted a very high concordance between cluster S2 and T2 (r = 0.9), good concordance between clusters S1 and T1 (r = 0.6), moderate across clusters S3 and T3 (r = 0.4), and poor concordance across clusters S4 and T4 (r = 0) (**Figure 4E**). This analysis shows that there is a substantial overlap between the two classification systems, especially in the identification of T2 patients.

It therefore appears that cluster S1 of the Soret classification corresponds to cluster T1 of the Trutschel classification, marked by high type 1 IFN signaling (CM1) **(Figure 4C, 4D)**. Cluster S3 matches cluster T3, as identified by high type 1 IFN signaling (CM1) in the context of a lower abundance of platelets (CM10) and erythrocytes (CM2). Cluster S2 matches cluster T2, with the lowest type 1 IFN signature (CM1). Cluster S4 in resembles cluster T4, as both have the highest

expression of the Neutrophil activation module (CM13), although other modules such as platelets (CM10) and erythrocytes (CM2) had discordant expression levels across the two patient classification systems. In general, there were no differences in the lymphoid modules (CM4, CM5 and CM11) across Trutschel clusters.

Tarn et al. propose a stratification model based on patient-reported symptoms and identified four clusters of patients: Low symptom burden (LSB), high symptom burden (HSB), dryness dominant with fatigue (DDF), and pain dominant with fatigue (PDF). We were unable to see any significant difference in the level of expression of any CM across the four subgroups of patients **(Supplementary Figure 3)**. Consistently, we observed -in the PreciseSADS and ASSESS cohorts- weak correlations of the CMs expression scores with the ESSDAI[32] and ESSPRI[33] disease activity scores **(Supplementary Figure 4)**. We however noted that unlike other components of the ESSDAI and ESSPRI disease activity scores, the presence of autoantibodies (anti-SSA, anti-SSB, PFLC, IgG) was positively-associated with the CM1 module representing type 1 IFN **(Supplementary Figure 5)**. These observations suggest that among pSD clinical manifestations, the presence of autoantibodies is the most associated with a specific blood transcriptomic profile.

CM8 and CM2 are associated with response to hydroxychloroquine and leflunomide combination

Many clinical trials for Sjögren's patients have shown poor results especially for response to treatment[34–37] but, negative clinical trials can still provide valuable information about the efficacy of a particular treatment and can help guide future research. However, positive trials provide a unique opportunity to compare responder and non-responder patients' characteristics. Within the IMI2 NECESSITY, data from both positive and negative clinical trials are available for exploratory retrospective analyses. RepurpSS-1[38] is a placebo-controlled, double-blinded, phase 2A randomized clinical trial that evaluated the combination therapy of hydroxychloroquine and leflunomide and is one of the first positive clinical trials in pSD.

Firstly, we validated the co-expression of the genes within each CM on this cohort independent of those used for the identification of the modules, highlighting the reproducibility and generalizability of the CMs to independent pSD blood transcriptomic datasets **(Supplementary Figure 7)**.

Secondly, we looked at the evolution of the expression of each module between treatment initiation and completion. We observed that lefluonomide-hydroychloroquine combination led to a decrease in the expression of CMs representing T cells, platelets and B cells, and an increase expression of the CMs representing monocytes and neutrophils, thus suggesting that this treatment combination favored the number of myeloid immune cells over lymphoid immune cells in the blood (**Figure 5A**). While treatments received by patients before blood transcriptomic profiling were more heterogeneous in the PreciseSADS cohort, we consistently observed an influence of the type of treatment received on the expression level of the CMs (**Supplementary Figure 6**).

Finally, we examined whether the heterogeneity of the patients encompassed in the modules could help identify responders in the RepurpSS-1 trial before treatment initiation. To do so, we focused on the recently developed STAR clinical endpoint[39]. The CM8 Lymphoid Lineage module was significantly overexpressed in responders before treatment initiation ($q = 0.013$) (**Figure 5B, 5C, Supplementary Figure 8)**. Conversely, a trend for higher expression in non-responders of the CM2 module representing erythrocytes was also found ($q = 0.055$). By combining CM2 and CM8, we were able to perfectly separate responders and non-responders in this clinical trial (**Figure 5D**). These analyses suggest that these cell populations could represent biomarkers predictive of therapeutic efficacy of this treatment combination.

Discussion

Primary Sjögren's disease (pSD) is a debilitating and clinically heterogeneous disease with no well-established causal mechanism, nor approved targeted therapy. There is therefore an urgent need to identify biomarkers able to inform treatment selection as well as to stratify patients in clinical trials in the context of personalized medicine. High throughput transcriptomic profiling is an appealing

technology for biomarker discovery as it allows the interrogation of tens of thousands of genes for differential expression across groups of patients, such as responders and non-responders to a drug in a clinical trial. The interpretation of transcriptomic profiles is however difficult, as groups of differentially expressed genes may represent dysregulation of functional pathways or changes in the cellular composition of samples, or both. In addition, the very high dimensionality of whole transcriptome assays makes difficult distinguishing true and replicable biological signal from noise.

To overcome these difficulties in the interpretation of the transcriptome in the context of pSD, we jointly analyzed four independent transcriptomic datasets profiling whole blood samples from pSD patients. We used clustering methods to identify the main axes of variation across these four datasets. As clustering algorithms are sensitive to noise, we implemented a method to perform a gene clustering analysis on a joint representation of the pairwise gene correlations matrix across the four datasets, rather than on each dataset separately. To do so, we recast the four observed matrices of pairwise gene correlations as graphs and used the SNF[25] algorithm to obtain a consensus graph representation of the gene correlation network across the four cohorts, on which we applied the Louvain graph clustering algorithm. We importantly showed that the gene modules we identified are reproducible across the four cohorts on which they were discovered **(Figure 1C)** as well as on an independent cohort (**Supplementary Figure 7**). These modules therefore represent the main biological features contained in the transcriptomic profile of the whole blood in pSD patients, therefore facilitating its interpretation for translational research.

In order to make the CMs more biological meaningful, we interpreted them using distinct public databases of pathways and blood cells transcriptomes[29]. This allowed us to identify both functional modules (interferon signaling or cell proliferation) or modules reflecting the cellular composition of the patients' blood. Importantly, we observed highly significant correlations between the expression of the gene modules and corresponding cellular frequencies or cytokine levels, thus validating these computationally derived biological interpretations. In the recent years, so called transcriptomic deconvolution methods have been proposed in order to infer cellular proportions from transcriptomic measurements[40]. Most of these methods rely on a reference averaged transcriptomic profiles of cell types, usually derived from purified cells from the blood of healthy donors and use genes that are discriminative across cell populations in a given context, such as cancer[29]. In contrast, our approach is driven by the observed variations in the blood of pSD patients across multiple cohorts, ensuring that the gene signatures of the identified cell types are valid in this context. In addition, this data driven approach allowed us to define gene modules indicative of rare cell populations such as eosinophils or signatures of non-immune cell types such as erythrocytes or platelets which are not typically quantified by deconvolution algorithms[41]. Moreover, we found functional modules (CM1 type 1 IFN and CM7 Cell Cycle) that do not correspond to variations in the frequencies of blood cell types. The consensus gene modules described herein therefore could help understanding the complex pathophysiology of pSD as they represent biologically meaningful, reproducible, and sensitive sources of heterogeneity in the blood transcriptome of pSD patients.

The gene modules that we identified can serve as a building block for translational research in pSD, by providing a concise list of potential biomarkers provided by whole blood transcriptomic profiling. Multiple independent studies have recently focused on the stratification of the disease into discrete patient subgroups, based on whole blood transcriptomic profiles[15, 16] or clinical characteristics[14]. These classifications systems may become relevant in future clinical trials, as new treatments may benefit only to a restricted subset of patients. Our approach complements these classifications by highlighting the functional and cellular composition differences across patient subgroups, as well as highlighting the consensus and differences across classification systems. Our analyses notably suggest that the patient subgroups in published transcriptomic-based patient stratification systems can be distinguished based on the measurement of three variables: the frequency of neutrophils in the peripheral blood, the concentration of type 1 IFN, as well as the frequency of either erythrocytes or platelets within the blood (**Figure 4C, 4D**). These observed differences across patient subgroups may provide clinically actionable biomarkers for disease stratification in settings where whole blood transcriptomic

profiling is impractical. Indeed, these key features of pSD drive disease heterogeneity and altogether may be useful predictors of response.

Some medications are designed to target specific genes or proteins, altering their activities and ultimately leading to changes in cellular behavior. Understanding the complex relationship between medications and gene expression is an important area of research that includes Drug Repurposing computational activities and may eventually lead to the definition of more effective treatment strategies for a wide range of diseases and conditions. Our analyses showed that the CMs can be used to understand the effect of drugs on the composition and functional orientation of the peripheral blood **(Figure 5, Supplementary Figure 7)**. We also confirmed, in two independent cohorts, the correlation between the presence of anti-SSA and anti-SSB autoantibodies and the level of type 1 IFN in the peripheral blood. The pathogenic role of the IFN pathway has been extensively described: type I IFN signature is correlated with the development of systemic extra-glandular manifestations, and a substantial production of autoantibodies and inflammatory cytokines[42]. Moreover, in the context of systemic autoimmune manifestations, pSD patients may present with hematologic abnormalities including anaemia, leukopenia (mainly neutropenia or lymphopenia), and thrombocytopenia[43, 44]. These three components are indeed evaluated in the haematological domain of the ESSDAI scale. As these patient characteristics are recapitulated by our CMs, whole blood transcriptomic profiling thus appears informative in the context of pSD translational research.

The CMs we identified indeed provide a succinct list of candidate blood-based biomarkers that recapitulate whole transcriptome profiles in a biologically interpretable manner. These modules can therefore be examined in exploratory and clinical research for their potential association with the response to a treatment or to study drug mechanism of action. We exemplified this idea by retrospectively analyzing data from the RepurpSS-1 phase IIa clinical trial[38] which evaluated a combination of leflunomide and hydroxychloroquine for the treatment of pSD. Longitudinal whole blood transcriptomic profiling allowed us to show that this combination led to a decreased expression of CMs corresponding to T cells, platelets and B cells, and an increase in modules representing monocytes and neutrophils. Our results therefore show that this combination of treatments influence the cellular composition of the peripheral blood in pSD patients.

Importantly, we investigated the relationship between each CM expression levels before treatment initiation and the observed clinical response upon completion of the clinical trial. Our results show that responders to this treatment combination featured higher expression of the module representing lymphocytes and a trend for lower expression of the module representing erythrocytes. These observations are consistent with the mechanism of action of leflunomide, an immunomodulatory drug known to inhibit de novo synthesis of pyrimidine, preventing lymphocytes from expanding in inflammatory context[45]. While the mechanism of hydroxychloroquine is less clear considering its initial use as an antimalarial drug, this molecule has widely been used in rheumatic autoimmune diseases such as systemic lupus erythematosus[46]. Studies have shown that hydroxychloroquine can contribute to regulate inflammation by blocking Toll-like receptors (TLR) leading to type I IFN pathway inhibition[47]. Hydroxychloroquine has also demonstrated inhibitory effect on platelet activation[48], in accordance with modulations seen on CM relating to platelets in the RepurpSS-1 clinical trial. Our results suggest that clinical efficacy for this treatment combination may be restricted to patients with high lymphoid frequency and low erythrocytes frequency, thus providing new hypotheses guiding the treatment strategy of pSD patients and the design of future clinical trials.

Our work is therefore expected to facilitate translational and clinical research on primary Sjögren's disease by presenting a set of reproducible and annotated gene modules that capture the major variations in the blood transcriptome of patients, which will open up the path for identifying biomarkers in clinical trials for this disease that is still poorly managed.

Acknowledgements

Funding: This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement number 806975. JU receives support from the European Union's

Horizon 2020 research and innovation program and EFPIA. The present article reflects only the authors' view and JU is not responsible for any use that may be made of the information it contains.

The UKPSSR is established with the funding provided by the Medical Research Council (G0800629), with additional infrastructural support from the British Sjogren's syndrome association, NIHR Newcastle Clinical Research Facility and the NIHR Newcastle Biomedical Research Centre.

Author contributions

Conceptualization: C.B., M.G., E.Be, E.Bi, L.L.

Methodology: C.B., M.G., E.Be, E.Bi, A.H., B.C.

Validation: D.T, P.S, E.D.

Formal analysis: C.B, A.H, B.C., A.B, E.Be

Writing – Original Draft: C.B., E.Be, E.Bi, L.L.

Writing – Review and Editing: P.S., D.T., A.H., B.C, C.L., A.A., S.H., P.M., E.D, A.B, M.A.R, W.F.N, J.R, J-E.G., B.S., X.M, M.G.

Resources: J-E.G, M.E.A.R, W.F.N, J.R.

Supervision: M.G., E.Be, E.Bi, L.L.

We, the authors of this manuscript, confirm that we have collectively agreed to submit this work for publication. We have read and approved the final draft and take full responsibility for its content, including the accuracy of the data presented. We have also ensured that the statistical analysis, where applicable, was conducted appropriately and accurately. As authors, we are committed to upholding the highest standards of scientific integrity and ethical conduct, and we affirm that this work represents our best efforts to contribute to the advancement of knowledge in our field.

Declaration of Interests

While engaged in the research project, C.B., B.C. and E.D. were phD students financed by Institut de Recherches Internationales Servier when they contributed to the research project. P.S., A.B., A.A., P.M., M.G., L.L., C.L., S.H., and E.Be were employees Institut de Recherches Internationales Servier when they contributed to the research project. W.F.N. has provided consultation for Novartis, GlaxoSmithKline, Abbvie, BMS, Sanofi, MedImmune, Argenx, Janssen, Resolves Therapeutics, Astella and UCB.

Figures

Figure 1 A) Schematic summary of the work. pSD = primary Sjögren Disease **B)** Heatmap of the consensus pairwise gene affinity computed by Similarity Network Fusion (SNF). Side annotations represent gene modules.**C)** Heatmaps of Pearson's correlation matrices of the four input datasets, with genes grouped by their consensus gene modules.

Figure 2 A) For each module, the two most significantly-enriched pathways in the Chaussabel database[28]. B) Most significantly-enriched pathways in the GO database[27] C) Average expression of modules in transcriptomes of purified cells

Figure 3 A) Significant Pearson's correlations between the average expression of the CMs and cell types abundances measured by flow cytometry. Scatter plots of average CMs expression and matching cellular frequencies. B) Scatter plots illustrating the average expression of CM7 versus averages of cell cycle signatures C) Scatter plot of the average expression of CM1 type 1 IFN and dosage of type 1 IFN

Figure 4 CMs scores across patient subgroups of A) the Soret classification B) the Trutschel classification. Average expression of the CM1 type 1 IFN, CM2 Erythrocytes, CM10 Platelets and CM13 Neutrophils.2 CMs in the C) Soret classification and D) Trutschel classification. E) Correlation across cluster centroids of the two stratification systems.

Figure 5 A) Boxplots illustrating the evolution of the modules significantly differentially-expressed at baseline (BL) versus Week 24 for treated patients B) Heatmap of baseline average gene expression of the CMs. Patients are split by their responder status according to the STAR clinical endpoint. Right side annotations indicate FDR corrected p-value (qvalue) C) Avegage expression of CM8 and CM2 at baseline in responders versus non-responders D) Dotplot of average expression of the CM8 and CM2 modules, colored by response statuses.

Material and Methods

Data collection

Gene expression and associated clinical and biological data was obtained through tranSMART, the NECESSITY consortium data sharing platform for the ASSESS (Assessment of Systemic complications and Evolution in Sjögren's Syndrome) cohort[22], PRECISESADS[12] and UKPSSR[23] cohort. Data from the fourth cohort was downloaded from the Gene Expression Omnibus repository, under the accession number GSE84844[24].

Transcriptomic data pre-processing

The UKPSSR RNA-seq count data was transformed as in[14]. RNA-seq data from the PreciseSADS cohort was normalized as in Soret *et al.*[15]. The ASSESS Affymetrix Clariom S microarray data were normalized as in[16].

The GSE84844 Affymetrix Human Genome U133 Plus 2.0 Array data was pre-treated by filtering out probesets indistinguishable from background noise. For that purpose, we modeled probesets expression after applying a $log2(x+1)$ transformation by a two component Gaussian mixture model[**dempster˙maximum˙1977**] with the first peak corresponding to unexpressed genes, and the second peak to expressed genes. We retrieved the parameters of the mixture distribution using the function *normalmixEM* from the *mixtools* R package. The 0.95^th quantile of the first component of the distribution was used as a threshold. Probesets whose expression were below that threshold in more than 95% of the samples were removed. Finally, the fRMA function from the fRMA R Package[**McCall˙Bolstad˙Irizarry˙1970**] was used to normalize probesets intensities across samples.

Finally, to have comparable data sets, the intersection of the 80% most varying common genes across all the data sets was selected (5443 genes).

Integrated affinity network

The construction of the integrated network involves two steps: First, gene affinity ($affi$) is computed independently on each data set as follow : for each pair of genes (x, y), we consider the affinity between x and y as $affi_{(x,y)} = exp((1 - cor(x, y))/\sigma)$ where cor is the Pearson correlation coefficient and $\sigma = 3$, as suggested by Wang *et al.*[25]. The four networks are then merged into an integrated affinity network by using the Similarity Network Fusion (SNF) method[25], with 30 neighbours per gene and 20 iterations. The SNF algorithm produces a weighted fully connected graph with $5000^2 = 2.5 \times 10^6$ edges. Visual inspection of the distribution of the weights showed that their distribution was bimodal, with a largely preponderant low weight peak **[Supplementary Figure 1]**. To convert the fully connetected output of the SNF algorithm to a sparse graph, we removed edges below the $0.9775th$ quantile of the weights distribution **(Supplementary Figure 1)**.

Consensus modules identification

Consensus gene modules were identified by applying the Louvain clustering algorithm[26] on the fused and truncated graph of pairwise gene affinities. This method is based on a modularity optimization algorithm that aims to partition genes into communities with high within-group affinity and low between-group affinity. The modularity score of a community structure is calculated as the difference between the weighted proportion of intra-community edges and the expected weighted proportion of such edges if the edges were randomly distributed.

Gene modules summarization

We used the mean expression the genes contained in a module to represent that module's expression as performed in Becht et al[29].

Gene set enrichment analysis

Enrichment analysis is performed by applying a Fisher-exact tests on the human blood-derived transcriptomic modules of Altman *et al.*[28] as well as the Gene Ontology database[27]. P-values were corrected using the Benjamini-Hochberg procedure to select pathways by controlling the false discovery rate at a 0.05 level.

Mapping with purified and sorted immune cells

To identify modules representing the abundances of blood cell types, we used the GSE86362 dataset[29], which consists of 1936 gene expression profiles from immune cell populations, non-immune non-malignant cell populations and non-hematopoietic cancer cell lines. For consistency with our sample types, we only retained samples corresponding to blood cell populations ($n = 1095$).

Correlation between CMs and cell type abundances measured by Flow Cytometry

On the PreciseSADS cohort, proportions of relevant cell types using flow cytometry custom marker panels were analyzed for samples where matched transcriptomic profiles and cytometry data were available. Correlations were performed between summarized CM expression levels and log-frequencies of the corresponding cell populations among live single cells, as previously described[29]. We corrected the p-values by Benjamini-Hochberg (BH) procedure by controlling the False Discovery Rate (FDR) at a 0.05 level.

Correlation between CMs and cytokines

On the PreciseSADS cohort, relevant cytokines were measured as in[12]. A log transformation was applied on the concentrations. Finally, we computed correlations tests between the average expression of the CMs and the cytokines levels we corrected the p-value by controlling the FDR at a 0.05 level (BH procedure).

Application to clinical trial

RepurpSS-1 (registered under trial number EudraCT, 2014–003140–12) was a phase II a placebo-controlled clinical trial testing a combination of Leflunomide and Hydroxychloroquine[38]. Gene expression and associated biological and clinical data for the RepurpSS-1 trial was obtained through the NECESSITY consortium. Transcriptomes of samples with a RIN < 6 or DV200 > 70 were excluded, resulting in the analysis of 16 patients. Pre-treatment and post-treatment (at week 24) CM expression levels were compared using paired t-tests with Benjamini-Hochberg correction. Responder status was determined based on the STAR clinical composite endpoint[39]. Patients with a STAR score of 5 or above were classified as responders. Difference in CM expression levels between responders and non-responders were assessed using univariate t-tests with BH FDR correction.

Supplementary materials

Table 1. List of genes (SYMBOL) in each Concensus Modules (CMs)

Supplementary Fig1. Histogram showing the distribution of weights in the SNF matrix. The x-axis denotes the weight range (logged) and the y-axis represents the frequency of weights. A vertical red line indicates the discretization threshold corresponding to the 0.975^{th} quantile (for better visualization).

Supplementary Fig2. A) Average correlation of the 4 input datasets **B)** Average of average correlation matrices **C)** Average gene expression levels for each CM in cohorts profiled by RNA-sequencing

Supplementary Fig3. CMs scores across patient subgroups of the Tarn classification in UKPSSR cohort

Supplementary Fig4. Pearson's correlation between average CMs expression and ESSDAI and ESSPRI scores in A) PRECISESADS and B) ASSESS cohorts

Supplementary Fig5. Pearson's correlation between average CMs expression and autoantibodies levels in A) PRECISESADS and B) ASSESS cohorts

Supplementary Fig6. A)T-test between average CMs expression and treatment. q = corrected p-value **B)**CMs expression scores across patients stratified by treatments received. AM = Antimalarials, STD = Steroids, IS = Immunosupressors **C)**Significant diffrences observed in treated versus untreated patients.

Supplementary Fig7. Correlation matrix in REPURPSS-1 cohort, sorted by CMs.

Supplementary Fig9. Boxplots of average expression of the CMs at baseline versus after treatment splitting patients by treatment and placebo.

Supplementary Fig8. Boxplots of average expression of the CMs versus response status.

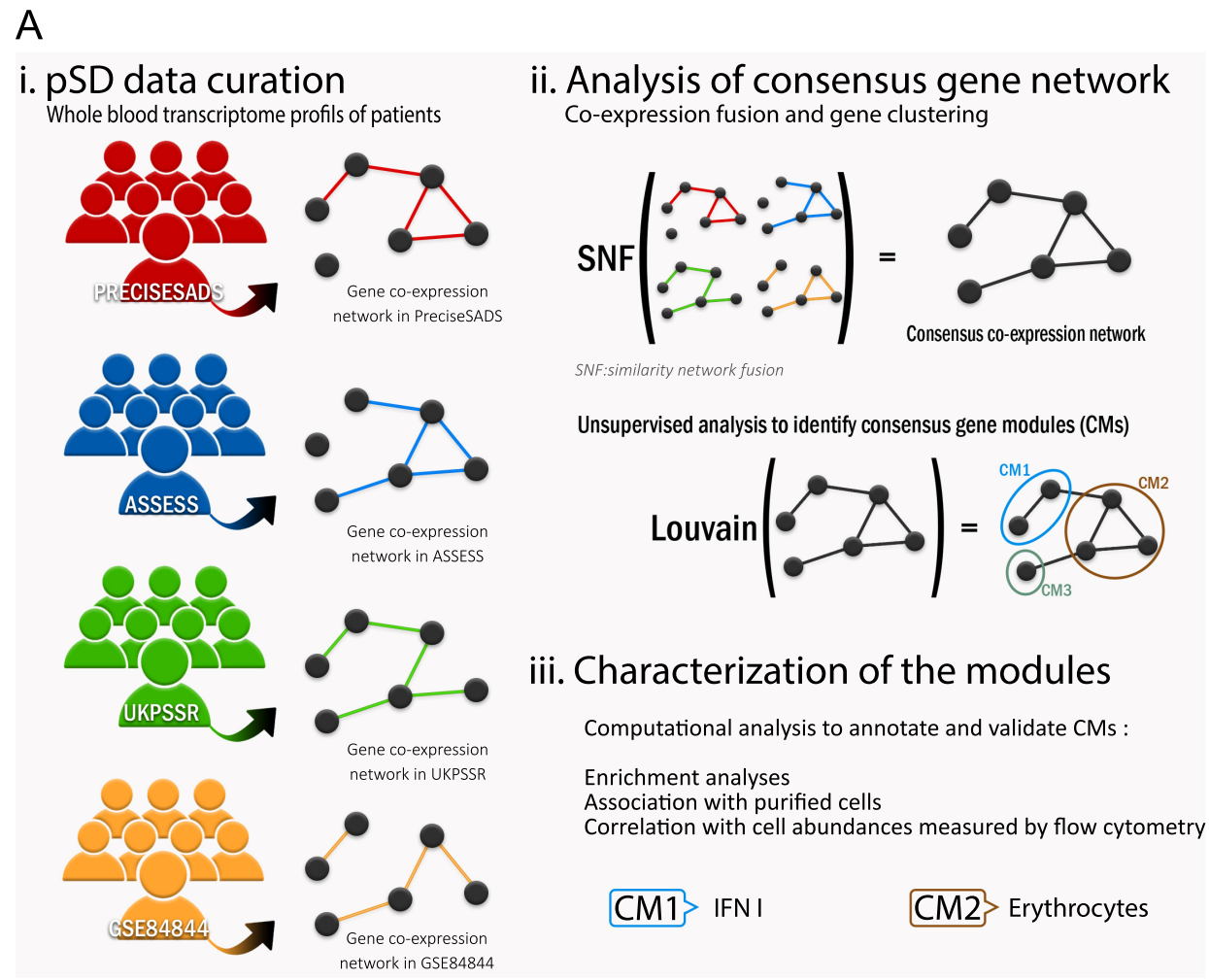
References

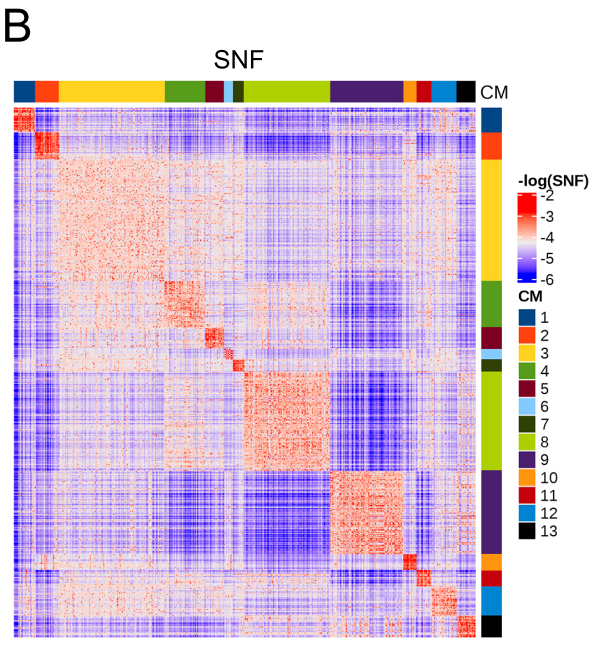
1. Mariette, X. & Criswell, L. A. Primary Sjögren's Syndrome. *New England Journal of Medicine* **378** (ed Solomon, C. G.) 931–939 (Mar. 2018).
2. Brito-Zerón, P. *et al.* Sjögren syndrome. *Nature Reviews Disease Primers* **2** (July 2016).
3. Parisis, D., Chivasso, C., Perret, J., Soyfoo, M. S. & Delporte, C. Current State of Knowledge on Primary Sjögren's Syndrome, an Autoimmune Exocrinopathy. *Journal of Clinical Medicine* **9,** 2299 (July 2020).
4. Solans-Laqué, R. *et al.* Risk, Predictors, and Clinical Characteristics of Lymphoma Development in Primary Sjögren's Syndrome. *Seminars in Arthritis and Rheumatism* **41,** 415–423 (Dec. 2011).
5. Nocturne, G., Pontarini, E., Bombardieri, M. & Mariette, X. Lymphomas complicating primary Sjögren's syndrome: from autoimmunity to lymphoma. *Rheumatology* (Mar. 2019).
6. Narváez, J., Sánchez-Fernández, S. Á., Seoane-Mato, D., Dıaz-González, F. & Bustabad, S. Prevalence of Sjögren's syndrome in the general adult population in Spain: estimating the proportion of undiagnosed cases. *Scientific Reports* **10** (June 2020).
7. Mavragani, C. P. & Moutsopoulos, H. M. The geoepidemiology of Sjögren's syndrome. *Autoimmunity Reviews* **9,** A305–A310 (Mar. 2010).
8. Anagnostopoulos, I. *et al.* The prevalence of rheumatic diseases in central Greece: a population survey. *BMC Musculoskeletal Disorders* **11** (May 2010).
9. Maldini, C. *et al.* Epidemiology of Primary Sjögren's Syndrome in a French Multiracial/Multiethnic Area. *Arthritis Care &amp Research* **66,** 454–463 (Feb. 2014).
10. Vivino, F. B. Sjogren's syndrome: Clinical aspects. *Clinical Immunology* **182,** 48–54 (Sept. 2017).
11. Qin, B. *et al.* Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Annals of the Rheumatic Diseases* **74,** 1983–1989 (June 2014).
12. Barturen, G., Beretta, L., Cervera, R., Vollenhoven, R. V. & Alarcón-Riquelme, M. E. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nature Reviews Rheumatology* **14,** 75–93 (Jan. 2018).
13. Barturen, G. *et al.* Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis &amp Rheumatology* **73,** 1073–1085 (Apr. 2021).
14. Tarn, J. R. *et al.* Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials. *The Lancet Rheumatology* **1,** e85–e94 (Oct. 2019).
15. Soret, P. *et al.* A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nature Communications* **12** (June 2021).
16. Trutschel, D. *et al.* Variability of Primary Sjögren's Syndrome Is Driven by Interferon-α and Interferon-α Blood Levels Are Associated With the Class II HLA–DQ Locus. *Arthritis &amp Rheumatology* **74,** 1991–2002 (Nov. 2022).
17. Fu, X., Liu, H., Huang, G. & Dai, S.-S. The emerging role of neutrophils in autoimmune-associated disorders: effector, predictor, and therapeutic targets. *MedComm* **2,** 402–413 (July 2021).
18. Negrini, S. *et al.* Sjögren's syndrome: a systemic autoimmune disease. *Clinical and Experimental Medicine* **22,** 9–25 (June 2021).

19. Bombardieri, M. *et al.* One year in review 2020: pathogenesis of primary Sjögren's syndrome. *Clinical and experimental rheumatology* **38 Suppl 126,** 3–9. ISSN: 0392-856X (4 **Jul-Aug** 2020). ppublish.
20. Saraux, A., Pers, J.-O. & Devauchelle-Pensec, V. Treatment of primary Sjögren syndrome. *Nature Reviews Rheumatology* **12,** 456–471 (July 2016).
21. Ritter, J., Chen, Y., Stefanski, A.-L. & Dörner, T. Current and future treatment in primary Sjögren's syndrome – A still challenging development. *Joint Bone Spine* **89,** 105406 (Nov. 2022).
22. Gottenberg, J.-E. *et al.* Serum Levels of Beta2-Microglobulin and Free Light Chains of Immunoglobulins Are Associated with Systemic Disease Activity in Primary Sjögren's Syndrome. Data at Enrollment in the Prospective ASSESS Cohort. *PLoS ONE* **8** (ed Re, V. D.) e59868 (May 2013).
23. Ng, W.-F., Bowman, S. J. & and, B. G. United Kingdom Primary Sjogren's Syndrome Registry–a united effort to tackle an orphan rheumatic disease. *Rheumatology* **50,** 32–39 (Aug. 2010).
24. Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **76,** 1458–1466 (May 2017).
25. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11,** 333–337 (Jan. 2014).
26. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (Oct. 2008).
27. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29 (May 2000).
28. Altman, M. C. *et al.* Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications* **12** (July 2021).
29. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17** (Oct. 2016).
30. Castro-Alcaraz, S., Miskolci, V., Kalasapudi, B., Davidson, D. & Vancurova, I. NF-κB Regulation in Human Neutrophils by Nuclear IκBα: Correlation to Apoptosis. *The Journal of Immunology* **169,** 3947–3953 (Oct. 2002).
31. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (Apr. 2016).
32. Seror, R. *et al.* EULAR Sjögren's syndrome disease activity index: development of a consensus systemic disease activity index for primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **69,** 1103–1109 (June 2009).
33. Seror, R. *et al.* EULAR Sjögren's Syndrome Patient Reported Index (ESSPRI): development of a consensus patient index for primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **70,** 968–972 (Feb. 2011).
34. Devauchelle-Pensec, V. *et al.* Treatment of Primary Sjögren Syndrome With Rituximab. *Annals of Internal Medicine* **160,** 233–242 (Feb. 2014).
35. Bowman, S. J. *et al.* Randomized Controlled Trial of Rituximab and Cost-Effectiveness Analysis in Treating Fatigue and Oral Dryness in Primary Sjögren's Syndrome. *Arthritis &amp Rheumatology* **69,** 1440–1450 (June 2017).
36. Ship, J. A. *et al.* Treatment of Primary Sjogren's Syndrome with Low-Dose Natural Human Interferon-alpha Administered by the Oral Mucosal Route: A Phase II Clinical Trial. *Journal of Interferon &amp Cytokine Research* **19,** 943–951 (Aug. 1999).
37. Zandbelt, M. M. *et al.* Etanercept in the treatment of patients with primary Sjögren's syndrome: a pilot study. *The Journal of rheumatology* **31,** 96–101. ISSN: 0315-162X (1 Jan. 2004). ppublish.
38. Van der Heijden, E. H. M. *et al.* Leflunomide–hydroxychloroquine combination therapy in patients with primary Sjögren's syndrome (RepurpSS-I): a placebo-controlled, double-blinded, randomised clinical trial. *The Lancet Rheumatology* **2,** e260–e269 (May 2020).
39. Seror, R. *et al.* Development and preliminary validation of the Sjögren's Tool for Assessing Response (STAR): a consensual composite score for assessing treatment effect in primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **81,** 979–989 (Apr. 2022).

40. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy* **67,** 1031–1040 (Mar. 2018).
41. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35,** i436–i445 (July 2019).
42. Papa, N. D. *et al.* The Role of Interferons in the Pathogenesis of Sjögren's Syndrome and Future Therapeutic Perspectives. *Biomolecules* **11,** 251 (Feb. 2021).
43. Stergiou, I. E., Kapsogeorgou, E. E., Tzioufas, A. G., Voulgarelis, M. & Goules, A. V. Clinical Phenotype and Mechanisms of Leukopenia/Neutropenia in Patients with Primary Sjögren's Syndrome. *Mediterranean Journal of Rheumatology* **33,** 99 (2022).
44. Wen, W. *et al.* Clinical and serologic features of primary Sjögren's syndrome concomitant with autoimmune hemolytic anemia: a large-scale cross-sectional study. *Clinical Rheumatology* **34,** 1877–1884 (Oct. 2015).
45. Breedveld, F. C. Leflunomide: mode of action in the treatment of rheumatoid arthritis. *Annals of the Rheumatic Diseases* **59,** 841–849 (Nov. 2000).
46. Shippey, E. A., Wagler, V. D. & Collamer, A. N. Hydroxychloroquine: An old drug with new relevance. *Cleveland Clinic Journal of Medicine* **85,** 459–467 (June 2018).
47. Kužnik, A. *et al.* Mechanism of Endosomal TLR Inhibition by Antimalarial Drugs and Imidazoquinolines. *The Journal of Immunology* **186,** 4794–4804 (Apr. 2011).
48. Erkan, D. *et al.* 14th International Congress on Antiphospholipid Antibodies Task Force Report on Antiphospholipid Syndrome Treatment Trends. *Autoimmunity Reviews* **13,** 685–696 (June 2014).

Figures





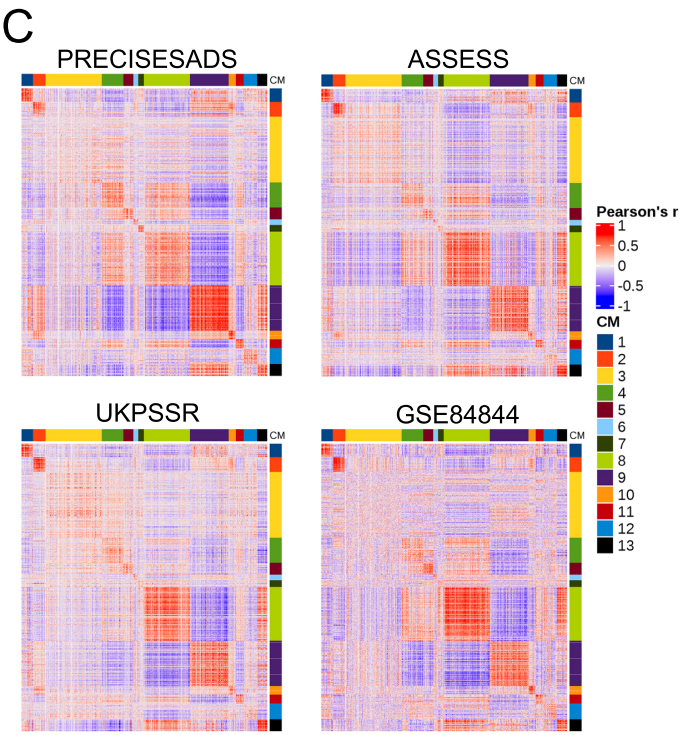


Fig. 1. A) Schematic summary of the work. pSD = primary Sjögren Disease B) Heatmap of the consensus pairwise gene affinity computed by Similarity Network Fusion (SNF). Side annotations represent gene modules. C) Heatmaps of Pearson's correlation matrices of the four input datasets, with genes grouped by their consensus gene modules.

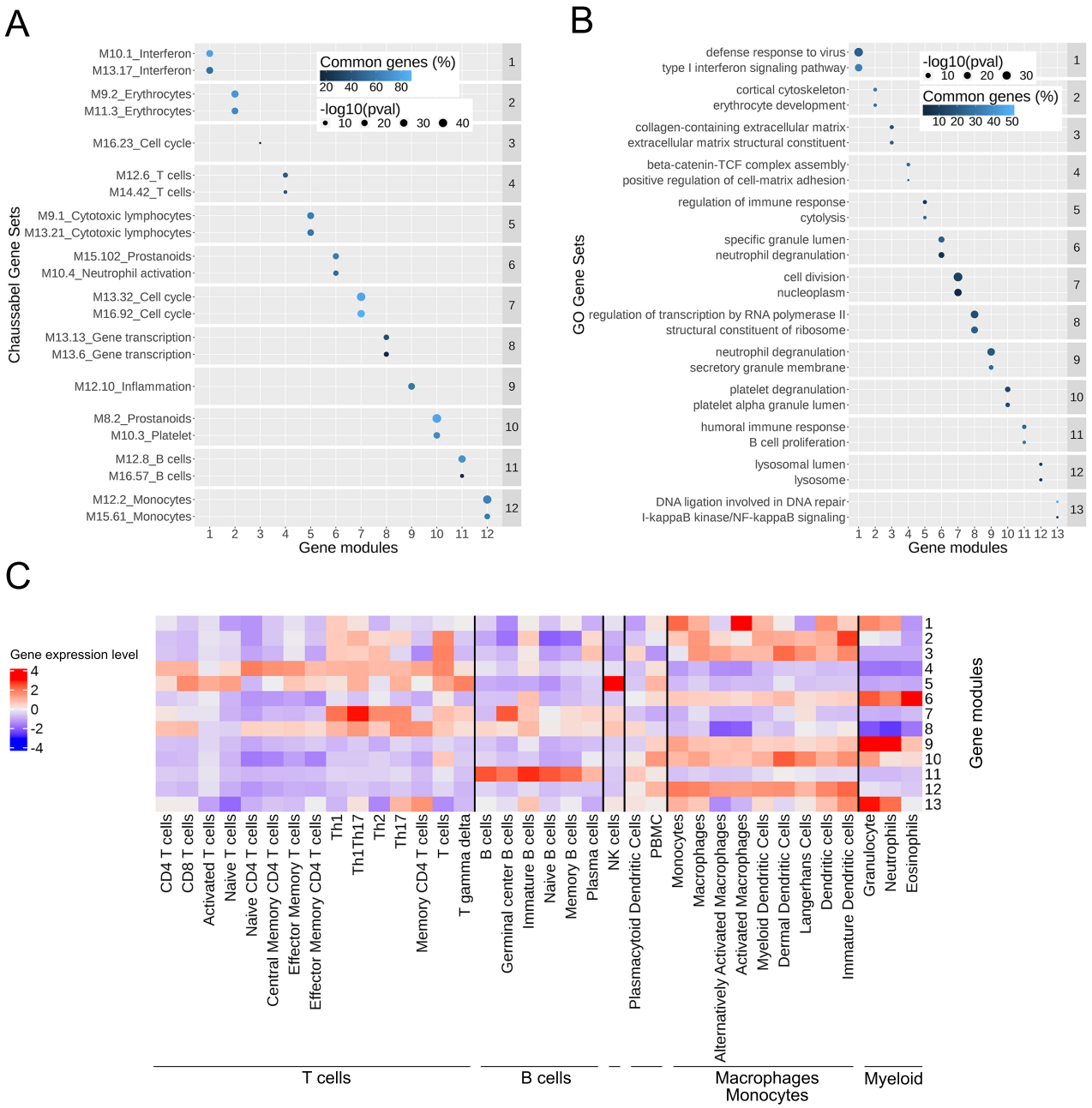


Fig. 2. A) For each module, the two most significantly-enriched pathways in the Chaussabel database[**?**]. B) Most significantly-enriched pathways in the GO database[**?**] C) Average expression of modules in transcriptomes of purified cells

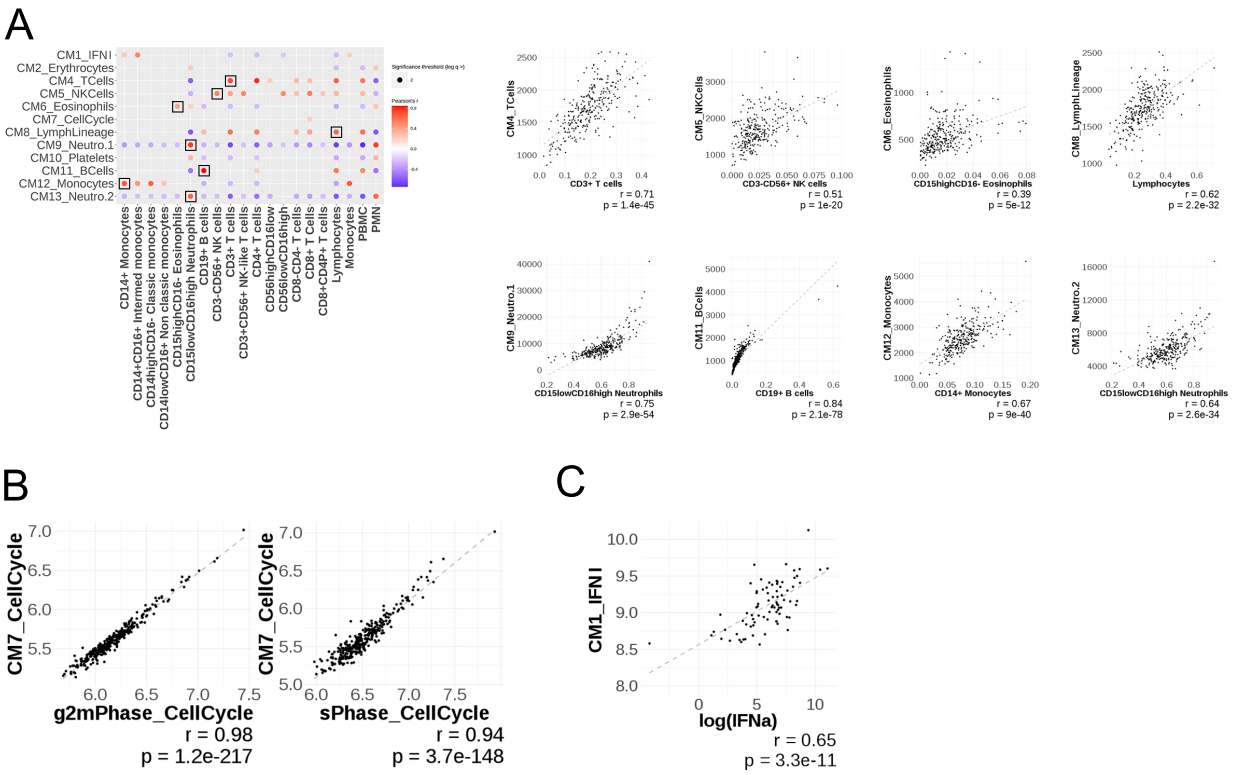


Fig. 3. A) Significant Pearson's correlations between the average expression of the CMs and cell types abundances measured by flow cytometry. Scatter plots of average CMs expression and matching cellular frequencies. B) Scatter plots illustrating the average expression of CM7 versus averages of cell cycle signatures C) Scatter plot of the average expression of CM1 IFN-α and dosage of IFN-α

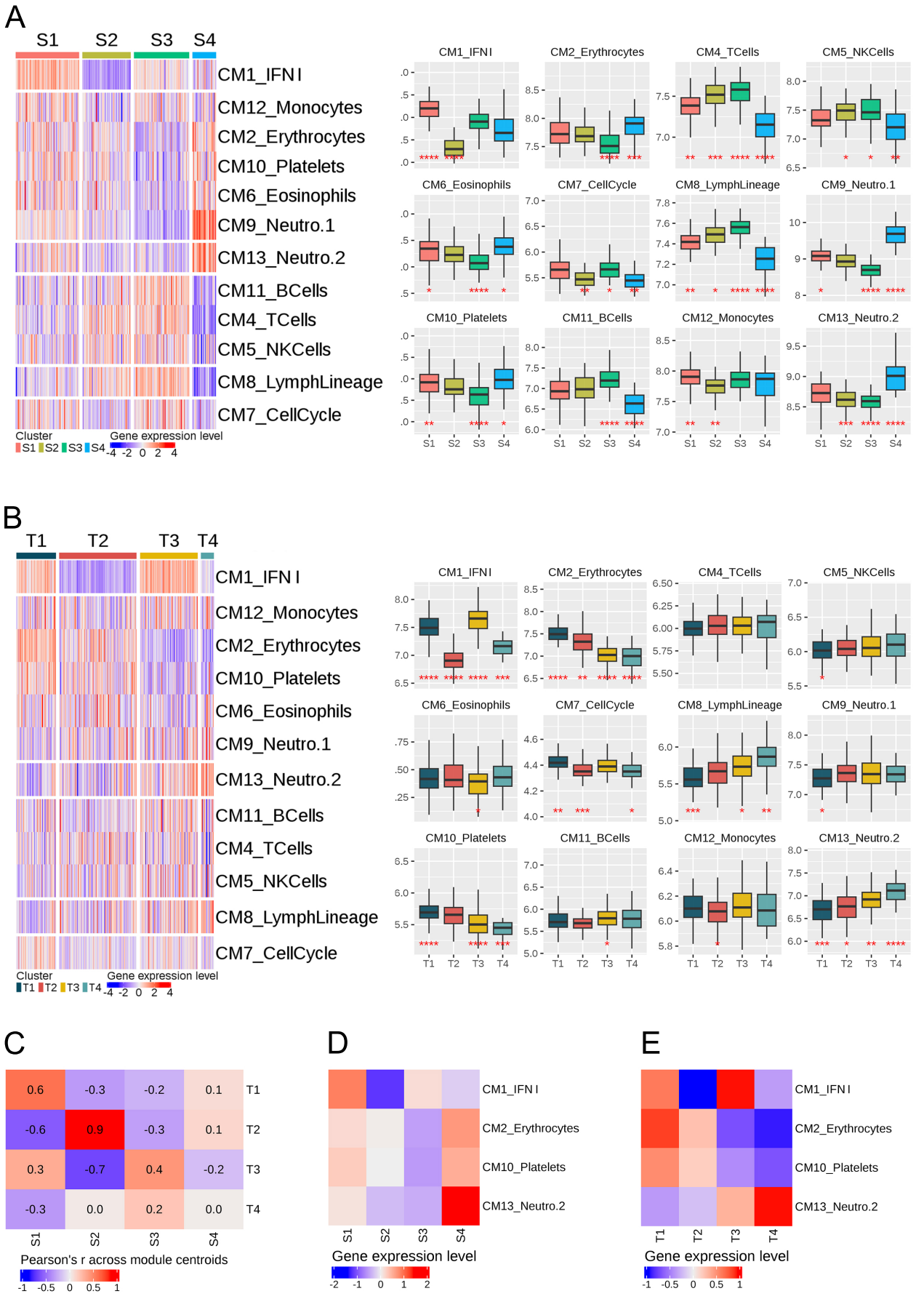


Fig. 4. CMs scores across patient subgroups of A) the Soret classification B) the Trutschel classification and ANOVA tests for each clusters. Average expression of the CM1 IFN-α, CM2 Erythrocytes, CM10 Platelets and CM13 Neutrophils.2 CMs in the C) Soret classification and D) Trutschel classification. E) Correlation across cluster centroids of the two stratification systems.

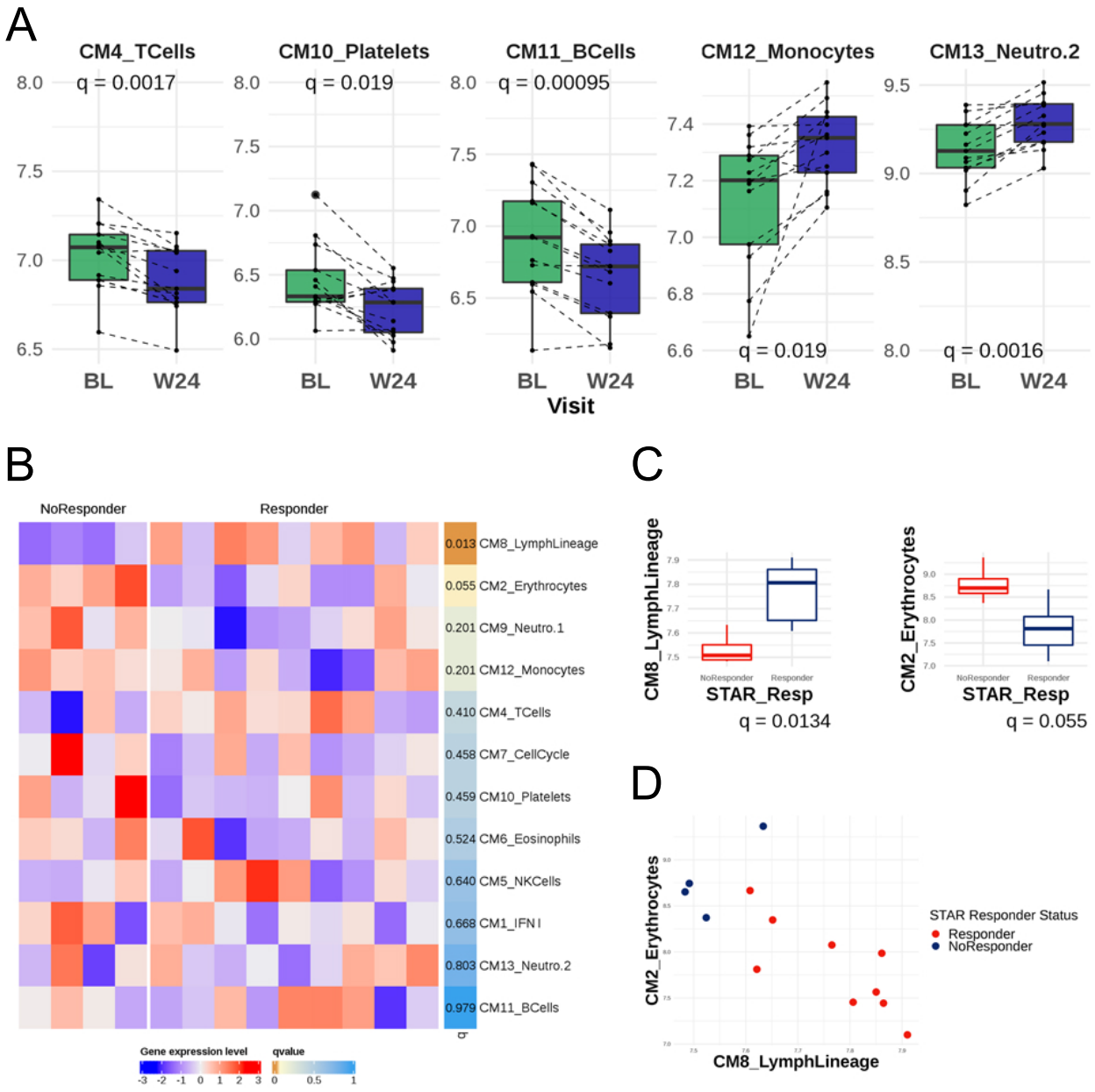
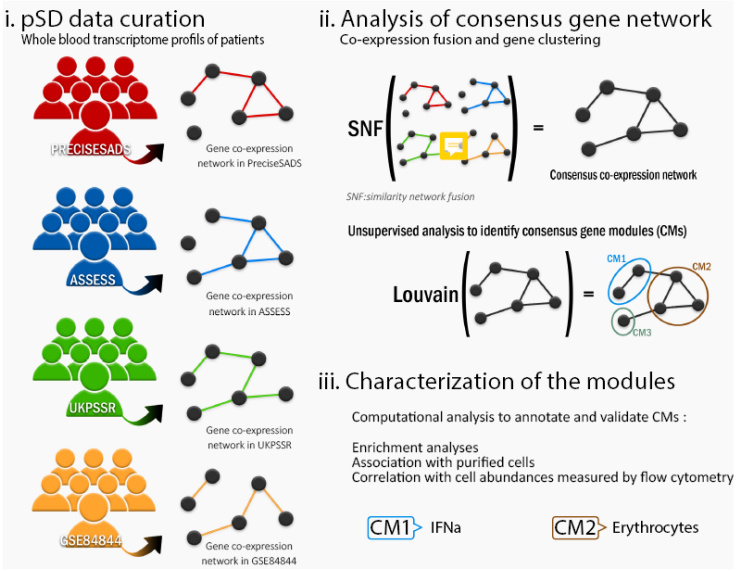
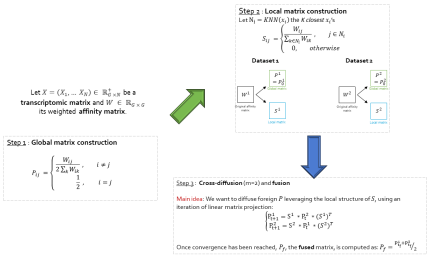


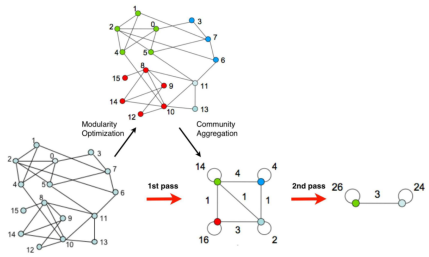
Fig. 5. A) Boxplots illustrating the evolution of the modules significantly differentially-expressed at baseline (BL) versus Week 24 for treated patients B) Heatmap of average gene expression of the CMs. Patients are split by their responder status according to the STAR clinical endpoint C) Avegage expression of CM8 and CM2 at baseline in responders versus non-responders D) Dotplot of average expression of the CM8 and CM2 modules, colored by response statuses.



(a) Schematic representation of the pipeline used in this paper, in which step i) encompasses all the construction steps to retrieve a fused transcriptomic network, step ii) covers the network and clustering operations to return consensual modules across all studied datasets and iii) includes all the post statistical and biological experiences to assert the soundness and relevance of the inferred gene clusters.



(b) SNF ([Wan+14]) is a *cross-diffusion* process that outputs enhanced metrics by averaging in an integrated manner multiple similarity measures. Briefly, it consists first of computing similarity matrices for each dataset, discarding remote nodes. Then, iterative steps of matrix projection on the same graph space and a final averaged operation results in a fused global network that concatenates relevant neighbourhood information across all datasets.



(c) A schematic visualisation of Louvain's algorithm [Blo+08] each pass (alternatively *epoch*) is made of two phases: a local optimisation phase, where modularity is increased only by local changes of community assignment followed by an aggregation phase, where sub-communities are merged in order to build a global network of clusters. The passes are repeated iteratively until the total modularity score no longer increases.

Figure E.1: Infographic of Chapter 5, about a practical use case of high dimensional clustering

E.1 Conclusion

In summary, this study revealed that 13 gene modules, inferred in an unsupervised framework from a collection of 4196 genes, were enough to apprehend most of the heterogeneity characterising the blood transcriptome across pSD patients. We were additionally able to annotate these modules with either a cell type signature, or metabolic function.

Overall, we demonstrated with this analysis that adopting a holistic paradigm for the study of the transcriptome, enables enhanced understanding of the pathophysiology of pSD, and notably the intertwined mechanisms involved in the evolution of this complex disease. Furthermore, this approach, by projecting noisy and high-dimensional data into meaningful functional gene groups, not only increases the robustness and reproducibility of our predictive models, but also improves their interpretability by medical experts. We indeed demonstrate with real-world experiences that these modules could be used as accurate biomarkers to predict the response to a treatment.

E.2 Perspectives

One of the major challenges we encountered when applying unsupervised and interconnected clustering methods in this study was determining the optimal threshold for removing spurious edges from the graph representing pairwise transcript interactions. In this work, we selected this threshold arbitrarily by visually inspecting the modes of the associated distribution of correlation weights. However, it's important to note that this threshold can vary between datasets, as it is heavily influenced by the sequencing technology and normalization procedures employed.

Truncated parametric mixture models, which naturally adhere to the inherent constraints of correlation coefficients (bounded between -1 and 1) or non-parametric, agnostic methods (see details in Section 3.3) could alleviate this issue in a robust and reproducible setting. Indeed, such approaches are able to automatically determine the threshold and adjust to the nature and shape of transcriptomic datasets, reducing the need for manual intervention.

We also found out that one of the modules, with the largest number of genes ($n = 1247$) and the lowest average expression value, exhibited inconsistent biological characterization. Accordingly, we did not consider it for downstream analyses, hypothesising that it was a residual, spurious nuisance cluster. This phenomenon is common in studies involving the clustering of gene expression profiles, as documented in several previous works ([Slo+13], [LdCL09],[ZA18], [AH05], [IBB04]).

The existence of this residual cluster can be attributed to a combination of factors, both intrinsic and extrinsic. Intrinsic factors stem from biological variability, arising from the stochastic nature of gene expression regulation ([IBB04]). Extrinsic factors, on the other hand, result from technical issues, such as errors during data collection and laboratory contaminations ([LdCL09]). Additionally, the high dimensionality of the datasets under analysis and the potential overlap with other clusters further mitigate the signal of interest ([GKT05]).

Hence, it may be advantageous to explore alternative approaches with noisy datasets exhibiting strong technical batch effect. Low-dimensional representations of transcriptomic data, such as Independent Component Analysis (ICA) ([ZA18]), or dedicated clustering methods to account for noise ([AH05]), could potentially yield more robust transcriptomic modules, by uncovering latent and irrelevant factors contributing to transcriptomic variability.

Ultimately, we underscore that the validation of the biological utility of the identified modules, as markers of the prognosis of pSD evolution, necessitates further clinical validation and in vitro experimentation.

Appendix F

Article 6: Network-based repurposing applied to COVID-19

The main part of this appendix chapter focuses on paper [Des+21], in which we developed a new industrial, computational repurposing approach named *Patrimony* and applied it to identify candidate therapeutic drugs that could help control the progression of severe inflammation during COVID-19. I summarised the key concepts related in the following graphical infographic, in Figure F.1.

Indeed, the COVID-19 pandemic, caused by SARS-CoV-2 strain virus, leads to millions of hospitalization in intensive care, with an estimated requirement of ICU transfer ranging from 5% to 32% with respect to the country, 768 560 000 confirmed cases worldwide and 6 951 000 deaths, according to the most updated statistics provided by the world health organisation by July 2023 [Aba+20]. And at the time we published the paper, no current approved FDA medication nor vaccine were widely available, while there were strong concerns about the effectiveness of vaccines against emerging new variants of the virus.

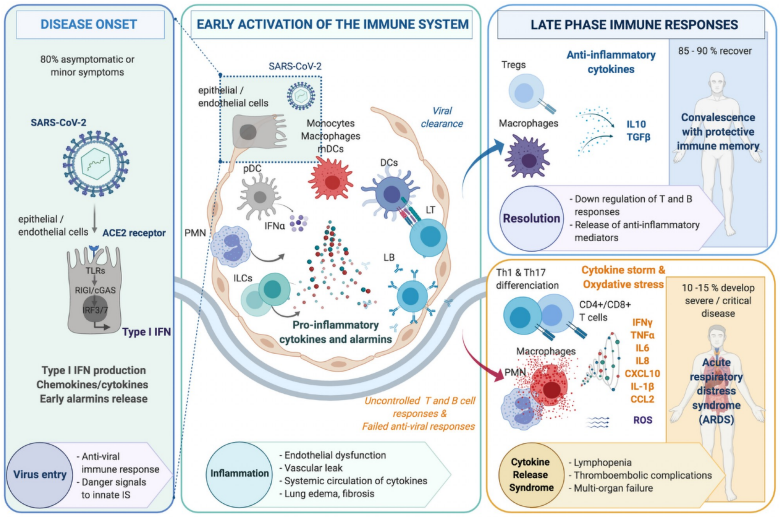
Practically, we focused on repurposing therapeutic drugs addressing severe lung inflammation caused by COVID19 pathophysiology.

To identify driver proteins and transcriptomic pathways involved in the severe cases of COVID-19, we have collected in a first time various public gene expression data from both SARS-CoV-2 infected and control pulmonary cells [1] and then identified proteins significantly related with early lung inflammation and the severity of cytokine storm events [MM20]. After a comprehensive pre-processing step, which included removing outliers and discarding genes with low expression, we ran standard Differential Gene Expression Analysis (DGEA), comparing the COVID-19 patients.

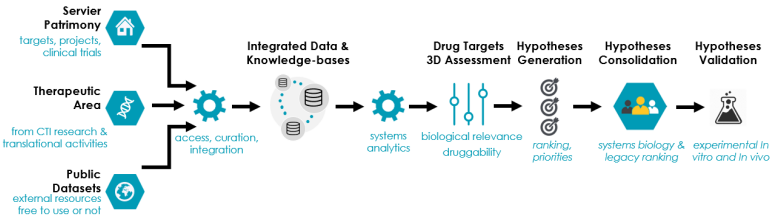
The identification of new drug targets itself was performed through two complementary drug repurposing strategies:

- In the network-based approach to drug repurposing, we capitalise on the existing integrated network of PPI of [CKB19], compiling 15 894 distinct proteins with up to 213 861 significant

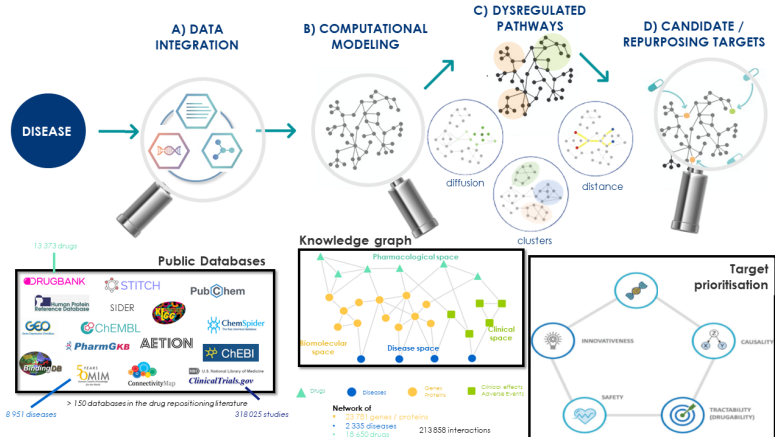
[1] We leveraged two distinct cell lines: bronchial epithelial (NHBE, [Bla+20]) and human lung epithelial cancer (Calu-3, [Ack+20]) cells



(a) The tree step progression describing the pathophysiology of COVID-19 in the airways, inspired from [Sum+21, Fig. 1].



(b) Patrimony general framework.First, data sources are integrated into a *knowledge graph*, then refined with omics sources specific to each disease. Finally, mining algorithms are used to extract putative, biologically-relevant targets. Reproduced from [Gue+22, Fig.1].



(c) Main steps of the computational Patrimony platform. Once data sources are curated, they are merged together to build the Patrimony "knowledge graph", encompassing biomolecular, pharmacological, disease, and clinical datasets. Once key driver genes or clusters of highly-interconnected entities have been identified, they are ranked according to a global score aggregating five meaningful criteria, including druggability and safety. Reproduced from [Gue+22, Fig.2 and Fig.3].

Figure F.1: Infographic of Appendix C, describing our internal repositioning platform "Patrimony".

pairwise interactions, all derived from a compendium of 15 distinct protein databases. We then fused this integrated network with two drug target databases: the Therapeutic Target Database (TTD, [CJC02]) and Drugbank [Wis+08], gathering 3092 drugs. The *network proximity* between disease-related proteins and drug targets within the PPI network was used as proxy of the relevance of drugs to the disease. Measure similarity was estimated through two complementary metrics: a *topological distance* and a *diffusion-based* distance, the first one returning the averaged shortest path length between the disease-related proteins and the drug targets and the second quantifying the similarity of the perturbations induced on disease-related proteins by biological mechanisms dysregulated by the COVID-19 on one side and drug targets on the other. The second metric is largely inspired from the Diffusion State Distance (DSD, [Cao+13]), a proven *metric* [Cao+13, Lemma 1] that leverages asymptotic random walk transition probabilities to estimate the distance between two nodes [2]. Then, for each metric used, bootstrap distributions for nodes having the same degree in the graph were computed to derive p-values, then combined using the Fisher's probability test and finally the drugs were ranked by decreasing order of aggregated p-value.

- To strengthen the connectivity findings, we then take profit of the Connectivity Map (CMap, [Lam+06]), perturbagen database, which aggregates more than 3000 compounds known as distributing the transcriptomic expression. We then compute CMap scores (compiling different comparison methods, we determine that Pearson correlation was the most robust and reproducible one for our experiment) reflecting the potential therapeutic effect of each of the listed perturbagen. Indeed, the core idea underlying the use of connectivity maps is to extract a set of perturbagens displaying a reversed gene expression profile, a negative correlation metric being an indicator of a reversed profile and accordingly a suggested therapeutic indication of the perturbagen.

F.1 Drug repurposing: a brief historical overview

Overview: Drug Development Process Drug development is an essential process for the discovery and availability of life-saving and life-improving medications. However, from the identification of potential biological markers to post-marketing surveillance, it is also a complex and challenging journey with numerous obstacles and failures. Throughout this process, pharmaceutical companies face various setback, with many drug candidates failing at different stages due to issues with efficacy, safety, or lack of sufficient clinical evidence. In addition, the experimental, human and computational burden deter numerous small companies into developing their own molecules, instead relying on the resiliency of bigger pharmaceutical groups.

While the ancestry of each successfully developed drug is quite unique, most of them follow generally the following storyline development (Figure F.2(a)):

1. The first phase is *Discovery and Target Identification*, where researchers identify potential drug targets (e.g., proteins, enzymes, receptors) that are involved in a disease process, leveraging high-throughput screening, omics combined with prior knowledge or literature review techniques to identify ligands interacting with these targets. The next stage consists of selecting the most promising drug candidates, termed "lead compounds" that interact with the target, displaying both strong molecular activity and specificity towards the target. Numerous steps make up the *early drug discovery*, that for obvious reasons of brevity, we

[2] This metric has some interesting asymptotic properties, notably proof of asymptotic convergence [Cao+13, Lemma 2] and even derived explicit form in the limit [Cao+13, Claims 2 and 3].

can not detail in this section. We hence recommend the interest reader the reading of [Hug+11], which encompasses the description of the key preclinical stages, ranging from initial target identification and validation, through assay development, high throughput screening, hit identification, lead optimisation up to the final selection of a reduced subset of candidate molecule for clinical development.

2. Before testing on humans, the selected lead compounds undergo extensive *preclinical testing*, which involves in-vivo testing on animals, in-vitro testing on cell lines and recently in-silico simulations, with the development of "twin models" expected to reproduce the complexity of the highly connected biological mechanisms, with extensive safety assessments to evaluate the drug's efficacy, toxicity, and pharmacokinetics.

3. Following the Investigational New Drug Application to the regulatory authorities (e.g., FDA in the US), ascertaining the validity of the preclinical testing and outlining the human clinical trials, there is the most critical phase of drug development, since it involves human patients and enables the final commercialisation of the drug. This step generally decomposes in three stages:

   - Phase 1: Small-scale trials on healthy volunteers to assess safety, dosage, and side effects.
   - Phase 2: Trials on a larger group of patients with the target disease to evaluate efficacy and safety further, in some cases further splinted into a step a) to evaluate the pharmacokinetics and a phase b) to evaluate the pharmacodynamics.
   - Phase 3: Large-scale trials involve a much bigger human cohort, to confirm efficacy and safety and monitor on the other hand any rare side effects.

4. If the results from Phase 3 trials are favourable, a New Drug Application is submitted to the regulatory agency for approval to market the drug, which, after conclusively stating the drug's benefits outweigh the risks for its intended use, approve the drug for marketing and use by the public.

5. The process does not stop once the drug enters the market, indeed, its safety and efficacy are continuously monitored through post-marketing surveillance, notably to detect any adverse reactions or long-term effects. This phase has gained recently increasing interest, related to horrendous clinical scandals, as well as the enumeration of side effects may help towards the development of repositioned drugs, with new indications, or suggest more efficient combination of therapies as detailed in next Appendix F.1. The commercialisation, the study of the side effects and the potential additional therapeutic use cases describe the *Life Cycle Management*.

Even after successful completion of clinical trials and regulatory approval, drugs previously approved might be ultimately withdrawn from the market, resulting from post-marketing surveillance that detect unveiled drug's toxicity or introduction of a new compound with undeniable efficiency, one of the most controversial case hitting recently the headlines certainly being the Mediator health scandal [21]. However, even failures help researchers and industrials into refining their approaches, while potentially suggesting new therapeutic uses, within a drug repurposing strategy (see Appendix F.1).

Motivation for drug repurposing The main purpose of drug repurposing is to identify existing drugs, generally clinically approved for a given medical use case, but displaying additional therapeutic value for other diseases or conditions. Determining new biological applications of existing drugs reveal paramount, since, despite increasing investments, the number of newly released molecules keep on decreasing.

Thereby, by determining new drug use case, repositioning is expected to speed up the drug development process, avoiding to reproduce the pharmacokinetics studies, and to reduce the sky-rocketing costs of drug development, compared to design molecules from scratch. Finally, and of utmost significance, the attrition rates are expected to significantly decrease, notably when anterior early-stage trials were conducted and positively evaluated the risk of safety failure. [Nos16] has estimated that the average cost of introducing a repurposed drug to the market is approximately 300 million dollars, against on average around $2-3$ billion dollars for a new molecule. In addition, the development process to release a new drug on the market has been estimated to 15 years while promising candidates undermined within the exploratory phase undergo a strikingly drop-out rate of 90% reaching Phase II clinical trials, mostly due to safety worries or inadequate effectiveness. This pattern of decreased R&D efficiency, gauged by the count of new drugs delivered to patients for each dollar expended and which decreased by half every 10 years since 1950, is commonly known as Eroom's Law, or the "valley of death", illustrated by the attrition funnel diagram in Figure F.2(a). This trend contrasts with the familiar Moore's Law that describes the exponential growth in computational power per unit of surface on a microchip over time.

Notably, such approaches are particularly relevant to provide novel treatment options for *orphan diseases*, rare affections with still unmet medical needs, or in case of emergency, such as outbreaks like the COVID-19, for which the standard drug development process storyline is not adjusted Appendix F.1.

Historically, many drug discoveries were accidental. For example, sildenafil citrate was originally developed as an antihypertensive drug, but then successfully repurposed to treat erectile dysfunction [YCh], an opportunistic process commonly known as serendipity. Yet, the recent fast accumulation of organised biomedical and omics datasets, coupled with innovative computational methods, have promoted the development of "data-driven" approaches. By analysing and integrating various types of data (e.g., chemical structure, omics, electronic health records, ...), such approaches have identified numerous candidate drugs and targets in an agnostic manner, uncovering unforeseen and unexpected connections between diseases and therapeutic compounds (see [Pus+19, Table 1 and Box 1], for a comprehensive review of repurposed drug success stories).

Data-driven repurposing strategies Notably, data-driven approaches are classically separated between experimental and purely computational, data-driven strategies, the latter being subdivided into the following categories:

- *Signature matching* is the process of comparing the unique characteristics, such as transcriptomic, structural, or adverse effect profiles, all combined composing the so-called drug or disease "signature", with those of another drug or disease phenotype. Transcriptomic signature can be used to unravel novel drug-disease and drug-drug associations, aiming to identify shared mechanisms of action between dissimilar drugs, alternative drug targets and/or potential off-target side effects. The underlying approach of this computational method relies on the *signature reversion principle*, where it is assumed that a drug displaying a reverse transcriptomic profile can counterbalance the dysregulated expression patterns

making up the hallmark for a given disease phenotype. In other terms, we hypothesise that an opposite drug signature should shift back the abnormal profile towards a healthy state (see Figure F.2(b), blue section). Chemical signature matching involves comparing the chemical composition of drugs with molecular patterns or sequences known for their biological activity. These approaches rely heavily on publicly accessible gene expression data, the largest and most famous one being certainly the Connectivity Map (CMap) project [Lam+06], [Lam07], and are largely inspired from previously developed "pathway enrichment analysis" methods [ZR23] [Sub+05] and [Lam+03]. We give a brief overview of the actual method used to compute enrichment scores (ES) in Definition C.2.1, and discuss briefly of its limitations in Appendix C.2.3.

- *Biomolecular or pathway network-based approaches* involve studying the impact of drug or disease on omics networks, based on gene expression patterns, protein interactions or Genome wide association study (GWAS). The main purpose is to identify key gene drivers, through genetic variant information combined with tissue-specific interaction networks, that could either mitigate downstream the dysregulation pattern induced by disease-related mechanistic disorders, or used upstream as a biomarker indicator of the efficiency of a treatment. For instance, the perturbation analysis of gene expression data induced by respiratory viruses led in study [Smi+12] to unravelling 67 common biological pathways associated with respiratory viral infections, which were ultimately checked against the DrugBank database to identify drugs targeting host-viral interactions (see details in Figure F.2(b), red section).

- *Clinical and side-effects databases*, either retrospective or undergoing, such as electronic health records (EHRs), post-marketing surveillance data and publicly available clinical trial data, has led to numerous drug repurposing successes, since the phenotypic effect observed is straightforward compared to the previously described methods. However, a systematic approach for analysing clinical data may provide additional drug repurposing opportunities, by not only considering explicit repurposing strategies, but also eliciting intricate similarity drug patterns by considering as well adverse effects [3]. Among them, EHRs offer the most comprehensive source of information to identify drug-disease or drug-drug associations, however, leveraging these resources reveal a highly challenging task, including ethical and legal obstacles related to personal health records and painstaking extraction of highly unstructured information. New natural language processing (NLP) combined with efficient computational power and alleviated open access from companies and governments to these health records could nonetheless keep pace of drug repurposing based on clinical trial and pharmacovigilance studies ([Pai+15] and Figure F.2(b), green section).

- *Molecular docking* predicts binding compatibility between a ligand (e.g., a drug) and a target receptor (e.g., a protein, see Section 2.1.3 for ligand and receptor definition). This technique can thus be used to identify potential drug-receptor interactions but further development of the tool is hindered by missing 3D structures for certain proteins [Hua+18], inconsistent target databases, and limited prediction ability, notably in detecting emergent *entropic forces* occurring at the total molecular system [PST17].

- The number of *Genome-wide association studies* highly increased in the recent years, with a common goal of identifying genetic variants linked to specific disease traits. Interestingly, it turns out that the genes highlighted by GWAS studies, are more likely to code for "druggable" proteins, and thereby, supply potential targets for drug development. Furthermore, the

[3] This principle of matching drugs based on indirectly shared side-effects, or diseases by similar drugs, is known as "guilt-by" association, see [Pai+15] and Figure F.2(b), green section, for practical examples

continuous discovery of new gene variants and functions contributes to a enhanced and dynamic understanding of the biological mechanisms, concurring altogether to maintain the homoeostasis of the human genome [San+12]. However, setting apart gene variants that have a causal effect on disease phenotypes from spurious relations is a highly intractable task, especially in gene-rich loci displaying strong linkage disequilibrium [WZ13].

- Finally, large-scale in vitro drug screens paired with genomic data (*pharmacogenomic interactions*), EHR-linked large biobanks and self-reported patient data, are promising avenues and databases sources for drug repurposing. For instance, human cancer cell lines (CCLs, [HV16]) have been used to test the impact of hundreds of compounds on cell viability and thereby identify molecular characteristics of the cells related to drug response. Recent studies showed thus recently that CCL datasets were able to clinically replicate pharmacogenomic interactions of primary tumours, and among newly discovered interactions relating drugs to tumoral cell lines survival, many involved drugs already used against distinct types of cancer. Specifically, this higher resolution up to the genomic and cell type level can contribute to personalised cancer therapy, by targeting groups of patients displaying specifically identified genomic variants associated with stronger drug response [Ior+16].

In addition to these data-driven approaches, machine learning methods promote the development of high-throughput experimental techniques, including *phenotypic screening*, in which the idea is to rapidly test a vast amount of putative compounds and subset those with a specific phenotype [Ilj+09], and *binding assays*, to identify novel target interactions from known drugs, using proteomic methods such as affinity chromatography or mass spectrometry [Als+16]. It is important to acknowledge that all these methods, compiled in Figure F.2(c), are now increasingly utilised in synergy as they are complementary.

However, a bunch of barriers related specifically with drug repurposing, including patent licenses, regulatory considerations and pharmaceutical organisational hurdles, hinder further development of computational or experimental repositioning. Legally approved drug patents are required to obtain the market exclusivity for drugs and thus ensure the economic sustainability of its commercialisation. However, repurposed drugs pose specific challenges: indeed, to register a new drug therapeutic indication (termed new method-of-use, MOU), you must enforce that the new repurposed use is innovative enough, leading thereby to early removal of promising candidates, since their extensive use was already described in the scientific literature.

Repurposed drugs with an orphan indication display a stronger potency of patent, since the EU approves 10 years of market exclusivity. However, in other disease cases, data exclusivity does not generally hold for other indications relying only on variations to existing marketing authorisations, while in the US, a new use of a previously marketed drug receives only 3 years of data exclusivity, a period usually not sufficient to recover the investment costs. Finally, off-label use of generic repurposed drugs does not always promote their commercial value, consider for instance a drug originally developed to cure for cancers, repurposed to cure common diseases [Mur+14].

The heterogeneity and the scattering of the available data across several companies and countries, on par with the requirement of innovative machine learning and network-based methods, promote stronger collaboration between small Biotech firms, academic communities and big pharmaceutical companies. For instance, initiatives like the AstraZeneca Open Innovation Platform or the Pfizer's Centers for Therapeutic Innovation [AT04b] endorse external collaborations, yet, challenging data management arise when the repurposed indication falls outside the company's

disease area, the development of the compound has been long stopped or no cured repository of the known side effects or parmacokinetics of a drug exist.

F.2 Introduction to the Patrimony initiative

To handle the ingestion of the ever-increasing data generated from genetics, multi-omics, molecular interactions or real-life evidence sources, pharmaceutical industries are increasingly developing dedicated computational environments, with the final purpose to identify both disease targets and promising drugs, in other words, to facilitate decision-making in drug development based on agnostic and data-driven approaches. Computing platform encompasses hardware, software and user interface components, while integrating diverse biomedical databases. While existing public or public-private initiatives, like Open Targets [Kos+17] (reach an extensive description of consortium projects in [Gue+22, Supplementary. Table 2]), were already implemented, and after an extensive benchmark on external solutions and examination of projects from numerous start-ups, my former manager decided to implement a corporate computational solution, offering reactivity, flexibility, better integration of internal data sources and relapsing legacy concerns. To that end, the Computational biology team, whom I am an active member, contributed to the implementation of the high-throughput computing platform "Patrimony", so called since we capitalise on both proprietary and public data to foster innovation and decrease the strong attrition rates in drug development [4].

The development of this computational platform involves first identifying and curating relevant data sources, both in-house and public, to integrate them into an uniformed knowledge graph. Then, machine-learning algorithms combined with graph theory were developed to mine the network. Finally, we adjust the platform by supplementing it with application-specific add-ins, tailored to the disease of interest, including immuno-inflammatory, oncological or even neurological disorders. This whole process is summarised in Figure F.1(c), ranging from data acquisition to experimental validation, through hypothesis generation and target prioritisation.

The proprietary knowledge graph underlies the core of Patrimony, by displaying in a compact and interpretable structure a wide set of entities and their relationships. To explore the resulting complex network, we applied different techniques from the graph theory and statistical field, which enable to identify *hubs* (key driver genes displaying a high degree) and *clusters* (highly interconnected cliques of vertices) within the graph or measure using an integrative and connected approach the impact of a drug or a disease on the biological pathways, benefiting from various diffusion and propagation algorithms [5]. Furthermore, a distance value alone is not meaningful, as highly dependent on the level of graph sparsity, hence, we derive an ad-hoc p-value to evaluate statistically the dissimilarity observed between two conditions, generating empirical distribution from bootstrapping nodes defined by the same degree in the graph. In addition, we enrich each node or interaction with attributes derived from ontology databases, while we make the tool versatile to each disease case by complementing each node with statistics resulting from multi-omics analyses (e.g. differential expression) and gene–gene co-expression values to weight

[4] Servier has not released any new drug on the market, nor receive additional NDA (New Drug Application) for 15 years, and its sale margins are comparable to those of generic drug company, which, by definition, only rely on existing drugs [16].

[5] One of the main challenge in developing a relevant similarity metric is underscored by the *hub protein bias* [Fis+21], occurring when key driver proteins display a high degree = high level of pairwise connections in the network, while strong sparsity is generally required to generate discriminatory metrics. Contrary to common beliefs, it was discovered that the distribution of the degree of vertices, i.e., the number of connections per gene or protein, did not follow a *scale-free* hypothesis, in other words, that the number of highly-connected genes was not significantly smaller than the number of weakly-connected regions of the graph [BC19].

pairwise interactions. Once a disease-specific knowledge graph has been established, therapeutic targets to a given condition of interest are prioritised through a summarised criteria. This global metric integrates five scores: *Biological Relevance* that quantifies the level of statistical evidence supporting a gene or protein's involvement in the disease pathophysiology, through for instance similarity metrics; *Causality*, which is used to heuristically counterbalance the absence of direction in our summarised graph, aiming notably at discriminating the cause and effect of a biological response; *Tractability*, which evaluates the "druggability"; *Safety*, which considers the adverse events associated with drugs known to bind to the target of interest and *Innovativeness*, which assesses the novelty of the application and the marketing opportunity associated. These criteria are then combined together in a global scoring system to rank the most promising therapeutic targets, and returned through interactive, appealing and user-friendly target "ID cards" that summarise the attributes assigned to it. After sub-setting the most promising biological targets, the biologists within our team conducted a comprehensive literature review to confirm the target's involvement in disease pathways and its druggability. Wet-lab experimental validation is then potentially performed to demonstrate the target's pharmacological activity when still not described in the literature.

The industrialised implementation of the Patrimony computing platform at the scale of Servier involved three stages (Figure F.1(b)): a *proof-of-concept*, a *structuration* and ultimately an *industrialisation step* for scalability. The Agile project management approach [SHZ15] was used for a rapid implementation of the methods, combining Python and R languages to implement the machine learning algorithms, Google Cloud Platform to host the network remotely, BigQuery for scalability and Neo4J [LC15] to support graph visualisation. Still, generalisation and adjustment of the method was particularly challenging, with the integration of large-scale, scattered and multidimensional data, subjected to access restrictions and regulations, especially regarding patient-level clinical information, keeping in mind the FAIR (findability, accessibility, interoperability and reusability) principles [Wil+16] to ensure the robustness and consistency of the integrated datasets.

It rapidly turned out that implementing the Patrimony initiative in an "old-fashioned" and highly compartmentalised pharmaceutical industry was a tedious task: the process required transversality between multidisciplinary and numerous teams with no universal terminology, the consolidation step, to validate the scientific rationale of predicted targets, demands extensive human resources, with extensive literature search and experimental validation, finally, biological interpretation of the model outputs was made challenging by the complexity and multidimensionality of the integrated biomedical data sets. The performance of the method was partly hampered as well by the lack of consistency across databases, and the existing gaps in medical knowledge, for instance, the current Human Interactome [Men+15] covers only around 25% of all molecular interactions described in the specialised literature. Experimentally, it appears that protein expression was a more consistent and straightforward approach to understanding intricate disease pathological mechanisms than gene expression alone. Another challenge in reconstructing valuable knowledge graphs was raised by the lack of overlap between omics, and we experimentally observed in our proof-of-concepts that the protein expression, measured through proteomics or flow cytometry, provided more meaningful and faithful insights into disease pathological mechanisms compared to transcriptomic expression alone, an empirical statement confirmed in [Mei+13].

Of note, I contribute to evaluate and complement the first two proofs of concept, respectively on an immuno-inflammatory condition (Sjögren's disease, see Chapter 4), and a viral pandemic COVID-19-repurposing).

F.3 Repurposing applied to severe COVID-19 cases

PLOS ONE

RESEARCH ARTICLE

Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19

Emiko Desvaux[1☯], Antoine Hamon[2☯], Sandra Hubert[1☯], Cheïma Boudjeniba[1], Bastien Chassagnol[1], Jack Swindle[2], Audrey Aussy[1], Laurence Laigle[1], Jessica Laplume[1], Perrine Soret[1], Pierre Jean-François[1], Isabelle Dupin-Roger[1], Mickaël Guedj[1☯], Philippe Moingeon[1]*

1 Servier, Research and Development, Suresnes Cedex, France, **2** Lincoln, Research and Development, Boulogne-Billancourt Cedex, France

☯ These authors contributed equally to this work.
* philippe.moingeon@servier.com

Citation: Desvaux E, Hamon A, Hubert S, Boudjeniba C, Chassagnol B, Swindle J, et al. (2021) Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19. PLoS ONE 16(7): e0254374. https://doi.org/10.1371/journal.pone.0254374

Editor: Svetlana P. Chapoval, University of Maryland School of Medicine, UNITED STATES

Received: March 5, 2021

Accepted: June 24, 2021

Published: July 22, 2021

Data Availability Statement: The RNA Seq data on genes differentially expressed in SARS-CoV-2 infected NHBE or Calu-3 human lung epithelial cells are publicly available from repository Gene Expression Omnibus (GEO, accession number GSE147507). Drug-target links can be retrieved from the Therapeutic Target Database (version 7.1.01) and from Drugbank. All other sources of data used in the present study related to various aspects of COVID-19 pathophysiology were obtained from the scientific literature. A

Abstract

While establishing worldwide collective immunity with anti SARS-CoV-2 vaccines, COVID-19 remains a major health issue with dramatic ensuing economic consequences. In the transition, repurposing existing drugs remains the fastest cost-effective approach to alleviate the burden on health services, most particularly by reducing the incidence of the acute respiratory distress syndrome associated with severe COVID-19. We undertook a computational repurposing approach to identify candidate therapeutic drugs to control progression towards severe airways inflammation during COVID-19. Molecular profiling data were obtained from public sources regarding SARS-CoV-2 infected epithelial or endothelial cells, immune dysregulations associated with severe COVID-19 and lung inflammation induced by other respiratory viruses. From these data, we generated a protein-protein interactome modeling the evolution of lung inflammation during COVID-19 from inception to an established cytokine release syndrome. This predictive model assembling severe COVID-19-related proteins supports a role for known contributors to the cytokine storm such as IL1β, IL6, TNFα, JAK2, but also less prominent actors such as IL17, IL23 and C5a. Importantly our analysis points out to alarmins such as TSLP, IL33, members of the S100 family and their receptors (ST2, RAGE) as targets of major therapeutic interest. By evaluating the network-based distances between severe COVID-19-related proteins and known drug targets, network computing identified drugs which could be repurposed to prevent or slow down progression towards severe airways inflammation. This analysis confirmed the interest of dexamethasone, JAK2 inhibitors, estrogens and further identified various drugs either available or in development interacting with the aforementioned targets. We most particularly recommend considering various inhibitors of alarmins or their receptors, currently receiving little attention in this indication, as candidate treatments for severe COVID-19.

comprehensive list of those publications and detailed references are provided in Supporting information file S1 Table.

Funding: We confirm that Servier and Lincoln only provided financial support in the form of authors' salaries. These companies did not play a role in the study design, data collection and analysis, decision to publish, nor in the preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: We confirm that our commercial affiliations do not alter our adherence to PLOS ONE policies on sharing data and materials.

Introduction

Since the emergence of the new strain of Coronavirus SARS-CoV-2 in December 2019, the ongoing crisis associated with the COVID-19 disease has affected more than 170 million individuals worldwide, causing over 3.5 million deaths (World Health Organization Dashboard, June 1st, 2021), mainly as the consequence of an Acute Respiratory Distress Syndrome (ARDS). The pandemic is still progressing actively despite lockdown measures throughout the world, with the recent emergence of highly transmissible viral strains [1]. To date, the only proven medications for reducing either viral loads, hospitalization rates, invasive mechanical ventilation or patient mortality include corticosteroids such as dexamethasone, the antiviral remdesivir, the anti-IL6R tocilizumab as well as neutralizing monoclonal antibodies directed to the spike protein of the virus [2–5]. Many additional drugs have been tested, including the lopinavir antiviral, the anti-malarial hydroxychloroquine or IFNβ with as of today disappointing efficacy results [6].

Recently, several vaccines have been approved by regulatory authorities based on remarkable efficacy results, with evidence that they can protect against infection by eliciting high titers of neutralizing antibodies against the Spike protein of the SARS-CoV-2 virus [7]. Whereas such vaccines will very positively transform the course and gravity of the COVID-19 pandemic, a recent concern is whether they will be fully effective against emerging new variants of the virus bearing point mutations in the Spike protein [1]. Furthermore, the challenge of manufacturing and administering billions of vaccine doses in order to establish a protective herd immunity at a worldwide population level will not be met in a short time frame.

During the time needed to deploy preventive vaccines at such a scale, the repurposing of existing drugs is a valid solution to better address severe forms of COVID-19 and alleviate the burden on health services in a time and cost-effective manner. Previous repurposing strategies have been undertaken in the context of a limited understanding of COVID-19 pathogenesis, prompting to use related viruses such as SARS-CoV and MERS-CoV as proxies to model SARS-CoV-2 infection [8–13]. Several network computing studies have been successful to predict drug disease associations for repurposing in COVID-19. Many of those initial approaches were aiming to identify existing compounds to prevent viral infection by either targeting mechanisms involving the viral receptor ACE2 (angiotensin converting enzyme 2), the TMPRSS2 transmembrane protease serine 2, or clathrin-mediated endocytosis [14–16]. In the present repurposing study, we rather focused on drugs predicted to interfere with pro-inflammatory mediators identified by modelling immune dysregulations caused in the airways by SARS-CoV-2 infection.

Since a vast majority of patients infected with SARS-CoV-2 develop no or only mild symptoms, we reasoned that ideal candidate drugs to repurpose should rather inhibit severe airways inflammation in the course of the disease. Lung inflammation is the main cause requiring hospitalization in up to 20% of COVID-19 cases, with life threatening ARDS affecting 75% of COVID-19 patients transferred to intensive care units [17]. In this subset of patients with severe lung inflammation, persisting proinflammatory immune responses result in a cytokine release syndrome (CRS) linked to the activation of myeloid cells secreting cytokines such as IL1β, IL6 and TNFα [18–20].

Capitalizing on the most recent scientific insights on the pathophysiology of COVID-19, we undertook computational network analyses to integrate a wide variety of data sources encompassing extensive molecular profiling of SARS-CoV-2 infected epithelial or endothelial cells, genetic susceptibilities and immune dysregulations linked to severe COVID-19 as well as molecular mechanisms elicited during lung infection by other respiratory viruses. From this approach, a short list of COVID-19 disease-related proteins considered as potential

therapeutic targets was established and used to computationally assess a topological proximity with drug targets within the comprehensive human protein-protein interactome [21, 22]. Herein, we report on the identification of candidate therapeutic targets, as well as drugs predicted to interact with some of those targets which could be repurposed to prevent or slow down severe lung inflammation during COVID-19.

Materials and methods

Sources of data on COVID-19 pathophysiology

To identify proteins related to lung inflammation in COVID-19, we selected relevant categories of data from the scientific literature (detailed in S1 Table), such as genes differentially expressed following SARS-CoV-2 infection of (i) primary normal human bronchial epithelial cells (*NHBE*) or of the ACE2-expressing lung-epithelial *Calu-3* cell line, (ii) endothelial cells or cells recovered from bronchoalveolar lavages or lung biopsies of patients with severe COVID-19 [23–25]. We also mined public data regarding immunological signatures obtained in the blood or in tissues of patients, distinguishing those with mild COVID-19 from others rather affected by severe forms of the disease [26–34]. We included as well information from previous studies on lung inflammation caused by other respiratory viruses (including asthma exacerbation), in light of an involvement of monocytes, macrophages, myeloid dendritic cells, innate lymphoid cells in those conditions similarly to COVID-19 [18, 35–38].

Identification of disease-related proteins

COVID-19 disease-related proteins predicted to be involved in early lung inflammation and in the transition to the cytokine storm were identified following data mining from scientific publications listed in S1 Table. To establish molecular pathways dysregulated during lung inflammation due to COVID-19, we first used RNAseq data from *NHBE* (normal human bronchial epithelial) and *Calu-3* (human lung epithelial cancer) cells infected or not with SARS-CoV-2. These data were pre-treated by removing outlier samples whose total sum of counts was below 5 000 000. In order to filter out genes undistinguishable from background noise, we modelled gene expression after applying a log2(x + 1) transformation by a two component Gaussian mixture model, with a first peak corresponding to unexpressed genes, and the second peak to truly expressed genes. Numbers of genes pre and post-filtering were 17557 and 21797, respectively. We retrieved the parameters of the mixture distribution using function normalmixEM from mixtools package and determined that the 0.95 quantile for the noise distribution was 1.6. We subsequently removed all genes whose expression was below that threshold in more than 95% of samples. We performed a differential analysis (COVID versus mock) in each cell line using the *limma R* package and *eBayes* function (with mock group corresponding to healthy & no treatment patients). Disease signatures were then extracted by considering differentially expressed genes (DEG) as those with adjusted *p*-value below 0.05 with an absolute fold change superior to 1.3 (commonly used as a threshold for biological significance). Canonical pathway enrichment analyses were subsequently performed by using the Ingenuity Pathway Analysis (IPA) software.

Network-based drug repurposing

Network-based drug repurposing relies on the hypothesis that the closer a target is to a group of disease related genes in the PPI network, the higher the chance of having a significant impact on the disease. Many approaches focus on the shortest path to determine proximity, with some variations in order to avoid hub protein bias [15, 39]. The latter bias occurs from

certain proteins that have an extremely high degree in the network and thereby cause a highly dense graph. Other approaches take advantage of the diffusion process to define proximity [40] while considering all the topological features of the graph. Diffusion based metrics have a comparable advantage over shortest path distances when in highly dense graphs such as PPI graphs [41]. Other metrics distinct from shortest path and diffusion can be used such as such as largest connected component -based methods [42].

Our computational repurposing approach (Fig 1A) takes advantage of the proximity between disease-related proteins and drug targets through an established network of protein-protein interactions (PPIs, referred to as an *interactome*). Drug-target links were gathered from the Therapeutic Target Database (TTD, version 7.1.01) and Drugbank [43, 44]. The PPIs network was derived from previous work by Cheng et al [45]. It was built from 15 different databases such as BioGRID and HPRD by compiling binary PPIs tested by high-throughput yeast-two-hybrid (Y2H) systems, kinase-substrate interactions from literature-derived low-throughput and high-throughput experiments, high-quality PPIs from three-dimensional (3D) protein structures, and signaling networks from literature-derived low-throughput experiments.

Relevance of drugs to the disease was assessed based on proximity of their targets to disease-related proteins according to two complementary metrics, namely a simple *topological distance* and a more advanced *diffusion-based distance*.

The *topological distance* (d_{topo}) corresponds to the shortest path length in the PPIs network between the disease-related proteins and the drug targets, computed according to the following formula:

$$d_{\text{topo}}(P,T) = \frac{1}{\|T\|}\sum_{t\in T}\min_{p\in P} SP(p,t)$$

With P the set of nodes corresponding to the disease-related proteins, T the set of nodes corresponding to the drug targets, and $SP(p,t)$ the shortest path length between a node p of P and another node t of T. When calculating a topological distance, we generate a distribution from bootstrapping similar nodes defined by same degree in the graph. From the given distribution, we calculate a z-score (and p-value).

The *diffusion-based distance* (d_{diff}) is computed based on the similarity of the impact on the network of perturbations starting from disease-related proteins on one side and drug targets on the other. The impact of a perturbation starting from a given node n_i on the network is assessed by use of a *diffusion* algorithm. Let (n_i,n_j) being a pair of nodes, then $\mathbb{P}\left(n_i, n_j\right)$ represents the random walk-based probability that a perturbation starting from n_i reaches n_j. It allows us to define a numerical vector $V(n_i)$ representing the impact perturbation of n_i on the whole interactome:

$$V(n_i) = [\mathbb{P}(n_i,\ n_1),\ \mathbb{P}(n_i,\ n_2),\ \ldots,\ \mathbb{P}(n_i, n_n)]$$

The similarity between two perturbations starting from n_i and n_j is then assessed by computing the Manhattan distance between $V(n_i)$ and $V(n_j)$. In order to extend this principle to the distance between sets of nodes, we derived the following formula:

$$d_{\text{diff}}(P,T) = \frac{1}{\|T\|}\sum_{t\in T}\min_{p\in P} MD(p,t)$$

With P the set of nodes corresponding to the disease-related proteins, T the set of nodes corresponding to the drug targets, p one given node of P, t one given node of T, and MD (p,t)

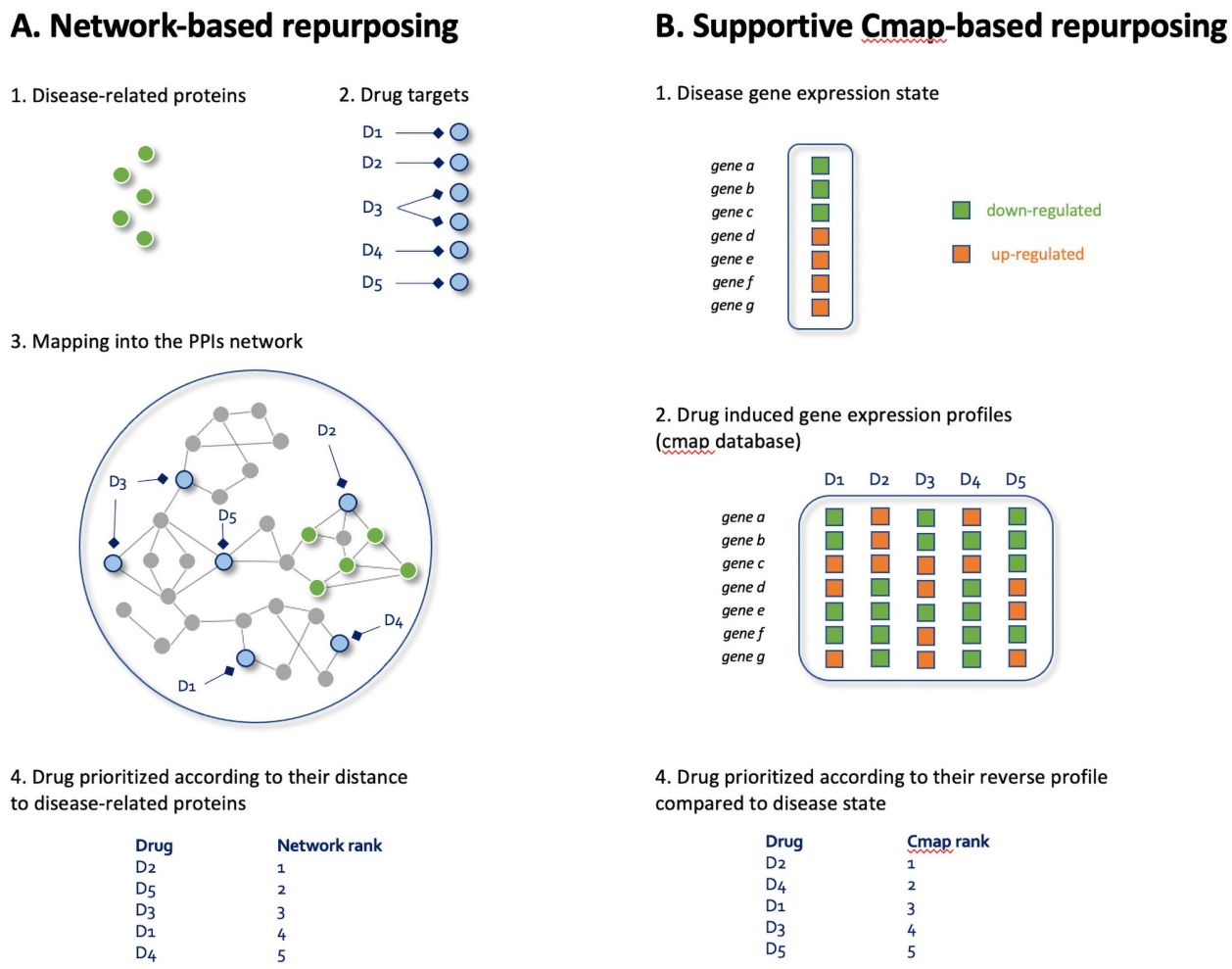


Fig 1. General principles of network and Cmap-based repurposing approaches. A) Network-based repurposing. Disease-related proteins and drug targets are mapped into a network of protein-protein-interactions (PPI). Drugs are prioritized according to their distance to disease-related proteins. B) Supportive Cmap-based repurposing. In those supportive analyses, disease-related as well as drug induced gene expression states are compared in order to identify drugs eliciting reverse profiles compared to those found in the disease.

https://doi.org/10.1371/journal.pone.0254374.g001

the Manhattan distance between $V(p)$ and $V(t)$. This diffusion-based distance was implemented via the DSD algorithm [46]. For each diffusion-based distance, we also calculate associated z-scores (and p-values). Note that DSD is by construction normally distributed. In order to prioritize drugs from this network-based repurposing approach, we defined a network rank resulting from the mean rank aggregation of d_{topo} and d_{diff}. Given that we have p-values for both of our distance measures, we perform a Fisher's combined probability test to obtain a unique combined p-value per drug. Using the DSD algorithm, we generated a computed distance matrix of 15 894 X 15 894 encompassing all proteins in our interactome.

Cmap-based drug repurposing

We complemented the network-based approach by using Cmap as a supportive method (Fig 1B). Cmap identifies drugs inducing a reverse gene expression profile compared to the disease state using a method of similarity [47]. The Cmap database comprises human cancer cell lines either treated or not with chemical drugs, referred to as perturbagens. We used the *R* package

ccdata which encompasses expression profiles for 1309 perturbagens over 13832 genes. Disease state was obtained from gene expression profiles induced in *NHBE* and *Calu-3* cells following infection by SARS-CoV-2. We compare expression profiles induced by disease state with those induced by perturbagens, using mainly the Pearson correlation between transcriptome values of the query signature and the perturbagen signature. A negative correlation score provides a potential therapeutic indication of the perturbagen. Cmap scores (the smaller the better) were first computed on both *NHBE* and *Calu-3* data and then averaged.

Results and discussion

Identification of COVID-19 disease-related proteins

Based on recent scientific advances, the pathophysiology of COVID-19 can be summarized as three sequential steps (Fig 2). We reasoned that treatments suitable to control severe COVID-19 should interfere with molecular pathways involved in the evolution from mild to severe

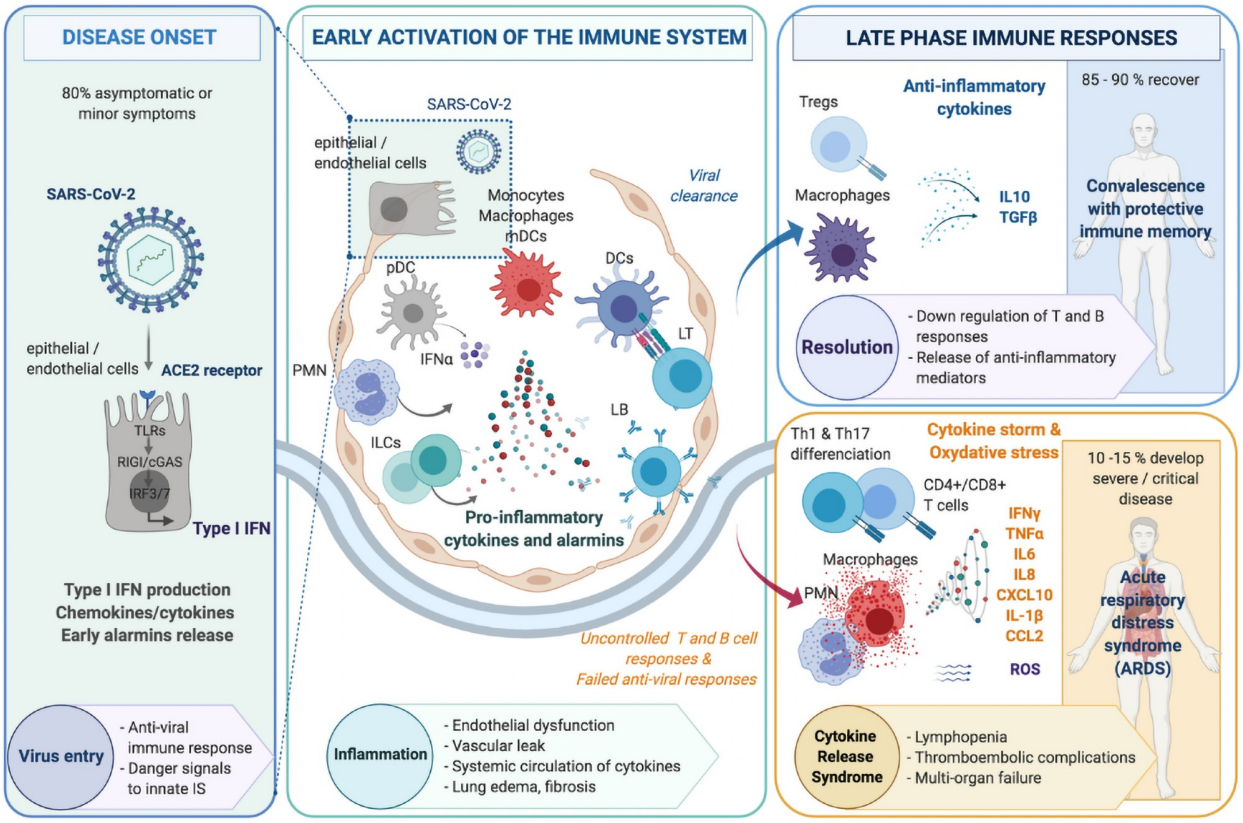


Fig 2. Three step progression towards severe COVID-19. The pathophysiology of COVID-19 in the airways encompasses schematically three successive steps, including **(i)** Disease onset following viral infection of alveolar epithelial or endothelial cells expressing the ACE2 receptor (left panel) leading to the activation of the innate immune system, with IFNα production by plasmacytoid dendritic cells (pDC). **(ii)** An early inflammatory phase within lung tissues where a cross-talk between infected epithelial/endothelial cells and innate immune cells such as monocytes, macrophages, myeloid dendritic cells (mDC) and innate lymphoid cells (ILCs) leads to a release of pro-inflammatory alarmins, cytokines and chemokines (center panel). This results in the activation of adaptive immunity, involving both CD4+ T cell help, CD8+ T cells cytotoxic for virally-infected cells as well as production of neutralizing antibodies against surface viral antigens. **(iii)** A late inflammatory phase with two potential outcomes: 85 to 90% of cases evolve towards resolution of inflammation with downregulation of T and B cell responses concomitant with the release of anti-inflammatory mediators (right upper panel); whereas 10 to 15% patients rather exhibit major tissue damage and severe acute respiratory distress syndrome (ARDS) caused by a deleterious uncontrolled inflammation linked with persisting T cell activation, excessive myeloid cell activation associated with a cytokine storm as well as oxidative stress (right lower panel).

https://doi.org/10.1371/journal.pone.0254374.g002

lung inflammation (Fig 2, central panel), while preserving anti-viral protective immune mechanisms. We thus compiled a comprehensive list of genes differentially upregulated in *NHBE* and *Calu-3* human epithelial cells following SARS-CoV-2 infection, providing important quantitative information [23]. We cross-validated this list in comparison with molecular signatures reported at the level of endothelial cells, bronchoalveolar lavage cells or lung biopsies in other studies to be associated with severe COVID-19 or exposure to other respiratory viruses (S1 Table). The latter was further completed with deep immunophenotyping, RNA seq and cytokine profiling data related to dysregulated innate or adaptive immune responses in the blood or the lungs of patients with severe COVID-19. A compilation of the most relevant COVID-19 disease related-proteins thus obtained, together with data sources supporting their relevance to lung inflammation in COVID-19 are presented in S1 Table.

Ingenuity pathway analyses were then performed on this list, allowing to confirm that genes/proteins upregulated following SARS-CoV-2 infection in the airways belong to multiple well-known pro-inflammatory pathways (Fig 3, S2 Table). Further data interpretation led us to classify disease-related proteins in two distinct sets of highly represented proinflammatory mediators and cytokines termed *Alarmins* and *Cytokine storm*, respectively (S1 Table). Alarmins represent a family of immunomodulatory proteins acting as damage-associated molecular patterns provided by injured stromal cells to recruit and activate various innate immune cells such as monocytes, macrophages, innate lymphoid cells as well as myeloid dendritic cells. Multiple proteins belonging to this family (*i.e.* defensins, HMGB1, IL1α, IL25, IL33, TSLP, S100A4, S100A7, S100A8, S100A9, S100A12, S100B, S100P) as well as their receptors such as IL1R1, RAGE, ST2 were predicted by our model to be involved in the evolution towards severe lung inflammation in COVID-19.

Our study also draws attention on disease-related proteins linked to the cytokine storm occurring in severe forms of COVID-19. The latter includes proinflammatory cytokines produced by activated myeloid cells such as IL1β, IL6 and TNFα directly involved as a cause of the CRS observed in COVID-19 [18, 35, 36]. Other potential targets associated with the cytokine storm include various cytokines (*e.g.* IL1β, IFNγ, IL2, IL12, IL15, IL17, IL23, IL32), chemokines (*e.g.* CCL5, CCL20, CXCL5, CXCL10, CXCL11), as well as selected proinflammatory factors (*e.g.* JAK1, JAK2, C5a) (S1 Table) [19, 20, 26–28, 36, 48–50].

Mapping into the interactome and identification of drug candidates for repurposing

COVID-19 disease-related proteins were mapped in parallel with known drug targets into the human complete interactome made of 15894 proteins (including 951 known drug targets) and 213861 interactions (Fig 4). From this, 3092 drugs were ranked according to computational proximity of their targets to each of the alarmins and cytokine storm sets by using a network-based method (S3 Table). Both COVID-19-related proteins as well as some functionally-related proteins in the interactome (such as the NR3C1 glucocorticoid receptor or receptors for reproductive steroids) were identified as candidate therapeutic targets.

Table 1 provides a list of selected targets as well as drugs interacting with those targets predicted to be of interest in severe COVID-19. Specifically, several high-ranking drugs were identified to treat severe COVID-19, such as anti-IL1β, anti-IL6 and IL6R or anti-TNFα antibodies. Our model supports as well the interest of corticosteroids such as dexamethasone, broadly used currently to treat severe COVID-19 [2]. Other high-ranking candidates for repurposing identified in our study are JAK2 inhibitors, with drugs not yet approved such as momelotinib or gandotinib previously shown by structure-based virtual screening to interact with ACE2 and the SARS-CoV-2 main protease, but also baricitinib, as well as other JAK1/

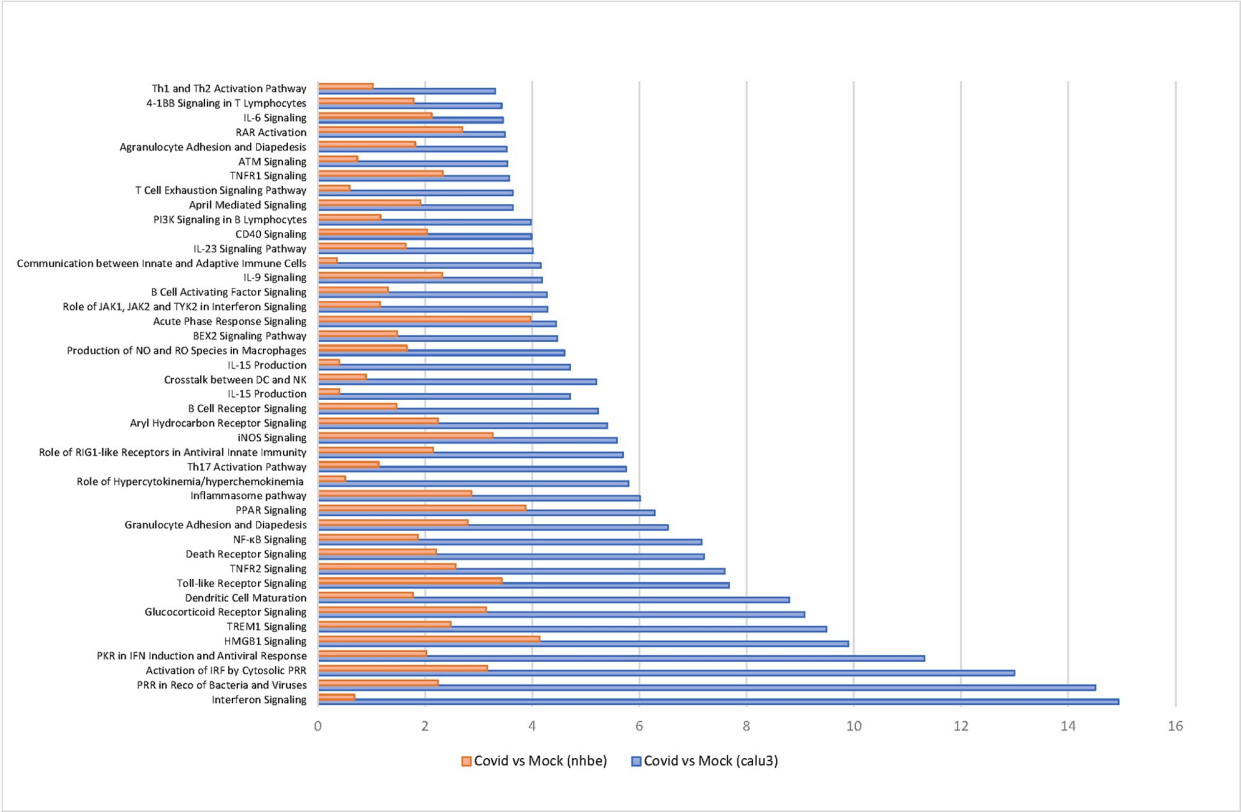


Fig 3. Pathway enrichment analysis from disease signatures (COVID-19 versus mock) in epithelial cell lines infected by SARS-CoV-2. The top 40 most significantly dysregulated immunological canonical pathways in either the Calu-3 (yellow) and NHBE (brown) infected cell lines are represented in a radar plot according to -log (*p*-value). Pathway enrichment *z*-scores, based on fold change direction, represent predicted up-regulation (green dots) or down-regulation (blue dots) for positive or negative values, respectively.

https://doi.org/10.1371/journal.pone.0254374.g003

JAK2 inhibitors currently being evaluated in COVID-19 patients (Table 1). Interestingly, some network computing approaches aiming to repurpose drugs inhibiting cell infection by SARS--CoV-2 also concluded to the interest of blocking antibodies against IL1β, IL6 and TNFα as well as JAK inhibitors in treating COVID-19 patients, in agreement with the present study [15, 16]. In addition, we also identify several reproductive steroids (estrogens and progesterone) as interesting candidates for treating COVID-19 patients.

Whereas the previous targets and some of the drugs directed to them could be expected from the current state of knowledge, our modeling study provided as well interesting hypotheses regarding other therapeutic options receiving less attention as of today. For example, drugs interacting with alarmins were also strongly suggested to be useful in COVID-19. To our knowledge, only three clinical studies have been initiated in COVID-19 with anti-alarmins, despite the availability of multiple additional drug candidates in this class (Table 1). Noteworthy, since Alarmins of the S100 family activate Toll-like receptors such as TLR2 and TLR4, a therapeutic option might be to target specific TLRs downstream of alarmins. Indeed, several TLR-antagonists are currently undergoing clinical evaluation in order to restore immune-homeostasis in patients with COVID-19 [51].

Similarly, anti-IL17 antibodies rank very high in our repurposing analysis, suggesting that inhibitory drugs directed to this well-known pro-inflammatory cytokine as well as the

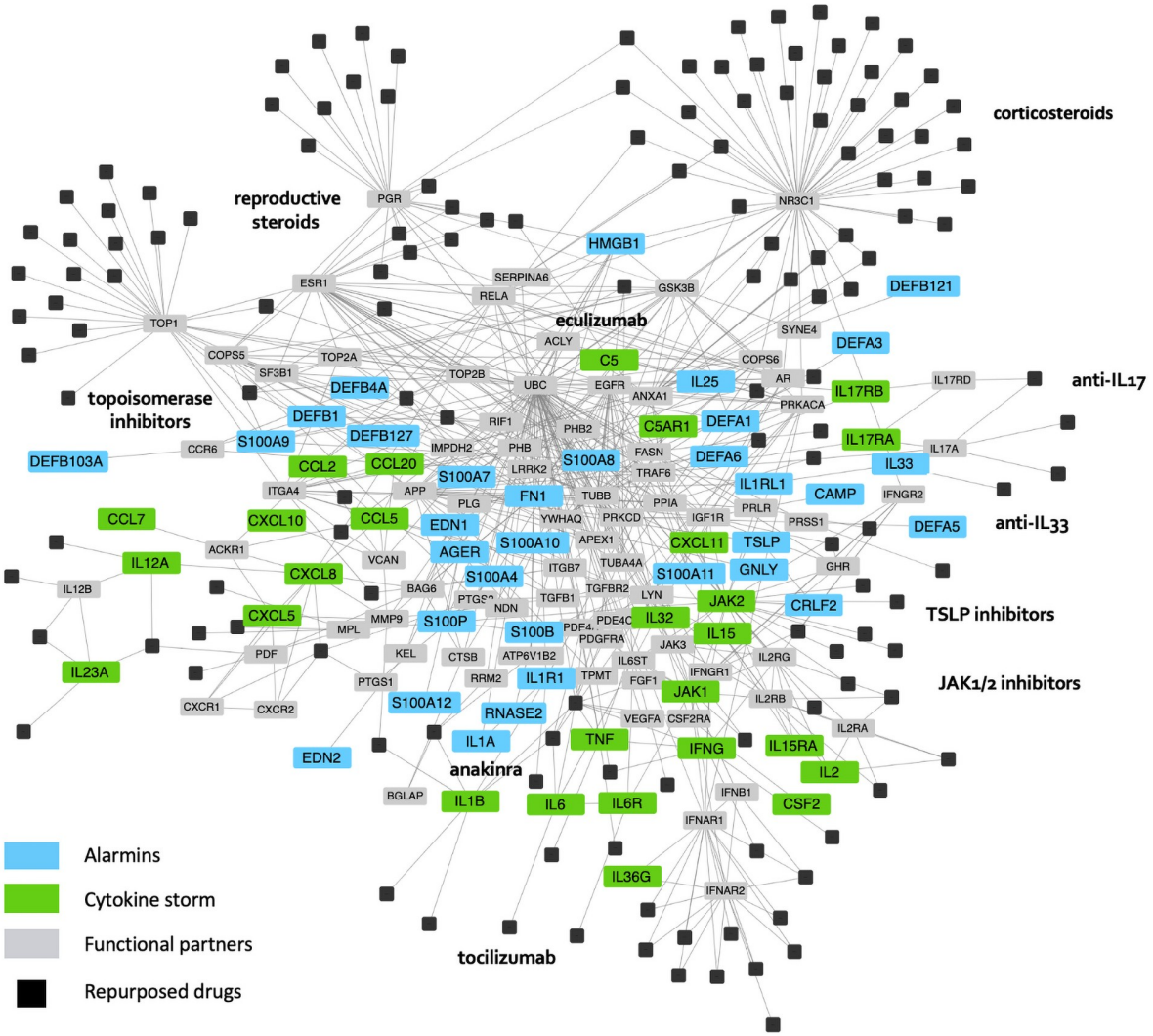


Fig 4. Druggable interactome of proteins contributing to lung inflammation in COVID-19. Extraction of the interactome encompassing proteins predicted to contribute to COVID-19 evolution towards a cytokine storm. Following SARS-CoV-2 infection of lung tissues and ensuing activation of innate and adaptive immune cells, different categories of proteins represent potential therapeutic targets to prevent or slow down lung inflammation associated with severe COVID-19. The latter include *Alarmins*, as well as cytokines, chemokines and selected proinflammatory factors associated with the *Cytokine storm*. For clarity, this figure only displays the disease related proteins (Alarmins & Cytokine storm) identified in our model, our top ranking repurposed drugs as well as some functional partners. The latter represent additional proteins needed in order to form a minimal principal component graph.

https://doi.org/10.1371/journal.pone.0254374.g004

functionally related IL23 cytokine or their receptors should be further investigated in COVID-19, with only one ongoing clinical trial in COVID-19 as of today [52]. In addition, the C5 complement inhibitor eculizumab is also predicted to represent an interesting treatment option, in agreement with recent evidence that the C5a-C5aR axis contributes to severe lung inflammation in COVID-19 patients [53]. As a strong chemoattractant, C5a provides in parallel to alarmins a link between innate and adaptive immune responses during severe COVID-19.

The thrombopoietin receptor appears as well to be a valid therapeutic target for agonists in light of the high incidence of thrombocytopenia associated with COVID-19 infection [54]. Rather unexpectedly, Topoisomerase 1 inhibitors, currently used as cytotoxic drugs in

Table 1. Overview of main therapeutic targets and clinical-stage candidate drugs for repurposing in COVID-19- related lung inflammation.

Therapeutic targets [Disease-related genes]	Candidate drugs for repurposing [Company name]	Modalities	Marketed drugs: Yes/ No	Clinical status in COVID-19 [Clinical trial ref]	Ref.
Cytokine Release Syndrome: IL1β, IL6, TNFα and their receptors	**Anti-IL1 β** Canakinumab [Novartis]	Antibody	Yes	Completed phase 2 in COVID-19 severe pneumonia [NCT04476706]. No impact on survival without the use of an invasive artificial respirator.	[4, 5, 18, 35, 56–58]
	Anti-IL1 β GLS1027 [GeneOne Life Science]	Small molecule	No	Recruitment planned for phase 2 in severe COVID-19 pneumonia [NCT04590547].	
	Anti-IL6 Clazakizumab [CSL Limited]	Antibody	No	Ongoing phase 2 in life-threatening COVID-19 infection [NCT04343989].	
	Anti-IL6 Olokizumab [R-Pharm]	Antibody	Yes	Completed phase 3 in acute respiratory distress syndrome [NCT04380519]. Results not yet available.	
	Anti-IL6 Siltuximab [EUSA Pharma]	Antibody	Yes	Ongoing phase 3 in acute respiratory Distress Syndrome [NCT04616586].	
	Anti-IL6 Sirukumab [Johnson & Johnson]	Antibody	No	Ongoing phase 2 in severe COVID-19 infection [NCT04380961].	
	Anti-IL6R Sarilumab [Sanofi]	Antibody	Yes	Completed phase 3 in severe or critical COVID-19 infection [NCT04327388], which did not meet its primary endpoint. Some improvement in survival when treating critically ill COVID-19 patients in association with dexamethasone.	
	Anti-IL6R Tocilizumab [Roche]	Antibody	Yes	Several trials completed in severe COVID-19 showing only limited efficacy [NCT04381936]. Some improvement in survival when treating critically ill COVID-19 patients in association with dexamethasone.	
	Anti-TNFα Infliximab [Johnson & Johnson]	Antibody	Yes	Ongoing phase 3 in COVID-19 [NCT04593940].	
	Anti-TNFα Adalimumab, [AbbVie]	Antibody	Yes	Ongoing phase 3 in mild to moderate COVID-19 [NCT04705844].	
	TNF-α inhibitor XPro-1595 [INmune Bio]	Peptide	No	Ongoing phase 2 in pulmonary complications of COVID-19 [NCT04370236].	
	Anti-TNFα Etanercept [Amgen]	Fusion protein	Yes	No evaluation yet in COVID-19.	
Glucocorticoid receptor NR3C1	**Corticosteroids** Dexamethasone [Mylan], Hydrocortisone [Sanofi-Aventis], Prednisolone [Mylan]	Small agonist molecules	Yes	Positive results obtained in the RECOVERY phase 3 study [NCT04381936], confirmed by a WHO-sponsored meta-analysis of 7 randomized clinical trials, collectively providing evidence for a reduced mortality of critically ill patients.	[2, 59]
				Dexamethasone is broadly used as a treatment for severe COVID-19.	
JAK1, JAK2	**JAK1/JAK2 inhibitor** Baricitinib [Eli Lily]	Small molecule	Yes	Ongoing phase 2 in moderate pneumonia [NCT04358614]. Recent evidence that Baricitinib can inhibit viral entry by clathrin-mediated endocytosis.	[60, 61]
	JAK/JAK2 inhibitor Ruxolitinib [Novartis]	Small molecule	Yes	Ongoing phase 2 in severe COVID-19 pneumonia [NCT04359290].	
	JAK2 inhibitor Jaktinib [Suzhou Zelgen Biopharmaceutical]	Small molecule	No	Completed phase 2 in severe and acute exacerbation of COVID-19 pneumonia [ChiCTR2000030170].*	
	JAK2 inhibitor Pacritinib [CTI BioPharma]	Small molecule	No	Ongoing phase 3 in severe COVID-19 [NCT04404361].	
	JAK2 inhibitor TD-0903 [Theravance Biopharma]	Small molecule	No	Ongoing phase 2 in symptomatic acute lung injury associated with COVID-19 [NCT04402866].	
Reproductive steroids: Estrogens, progesterone and their receptors	**Receptor agonists** Ethinylestradiol + Norelgestromin [Johnson & Johnson]	Small molecules	Yes	Planned phase 2 in non-severe COVID-19 patients [NCT04539626].	[63]

(*Continued*)

Table 1. (Continued)

Therapeutic targets [Disease-related genes]	Candidate drugs for repurposing [Company name]	Modalities	Marketed drugs: Yes/ No	Clinical status in COVID-19 [Clinical trial ref]	Ref.
Cytokines: IL2, IL15, IL17	**IL2Rβ superagonist** Bempegaldesleukin [Nektar]	Recomb protein	No	Ongoing phase 1b in mild COVID-19 [NCT04646044].	[52]
	IL15 super agonist ALT803 [Altor Biosciences]	Recomb protein	No	Planned phase 1 study in mild to moderate COVID-19.	
	Anti-IL17 Secukinumab [Novartis]	Antibody	Yes	Ongoing phase 2 in mild and severe COVID 19 [NCT04403243].	
	Anti-IL17, -IL17R, -IL23	Antibodies	Yes	No evaluation yet in COVID-19. Anti IL17 [Ixekizumab, Eli Lilly], anti IL17R [Brodalumab, Astra Zeneca/ Amgen], anti IL23 [Ustekinumab, Johnson & Johnson; Tildrakizumab, Merck] antibodies are commercialized as treatments for inflammatory diseases.	
C5, C5aR	**Anti C5** Eculizumab [Alexion]	Antibody	Yes	Proof-of-concept evidence suggesting that eculizumab provides some benefit in severe COVID-19. Ongoing phase 2 in moderate, severe or critical COVID-19 pneumonia [NCT04346797].	[53, 64–66]
	Anti C5aR Avdoralimab [Innate Pharma]	Antibody	No	Ongoing phase 2 in severe COVID-19 pneumonia [NCT04371367].	
Alarmins and their receptors: IL1α, TSLP, IL33	**IL1R1 antagonist** Anakinra [Sobi]	Peptide	Yes	Completed phase 2 in severe COVID-19 [NCT04366232]. Results not yet available.	[71, 75, 77]
	Anti-IL33R [ST2] AMG282-Astegolimab [Genentech]	Antibody	No	Ongoing phase 2 in severe COVID-19 Pneumonia [NCT04386616].	
	TSLP inhibitor HY-209- NuSepin [Shaperon] agonist for G protein-coupled TGR5 receptor	Small molecule	No	Ongoing phase 2 in COVID-19 pneumonia [NCT04565379].	
	Anti IL25, -IL33, -TSLP	Antibodies	No	No evaluation yet in COVID-19. Anti IL25 [ABM-125, Abeome], Anti-IL33 [REGN3500, Regeneron] and anti TSLP [Teepelumab, Amgen] are in clinical evaluation as treatments for asthma or atopic dermatitis.	
	Anti S100A4, -S100A7,—S100P	Antibodies	No	No evaluation yet in COVID-19. Antibodies in preclinical development in cancer or autoimmune diseases by Cancer Res Technol and Lykera Biomed.	
Thrombopoietin receptor	**Receptor agonist** Romiplostim [Amgen]	Peptibody [peptide agonist fused to Fc IgG1]	Yes	Case study documenting platelet recovery following treatment by Romiplostim of a pediatric patient with thrombocytopenia due to COVID-19.	[54]

All clinical trial information are available in Clinical trials gov: https://www.clinicaltrials.gov/ or* in Chinese clinical trial Registry: http://www.chictr.org.cn/.

https://doi.org/10.1371/journal.pone.0254374.t001

oncology, were also identified as of potential interest in COVID-19, with as of today only pre-clinical evidence that they can inhibit SARS-CoV-2 inflammation and death in animal models [55].

Supportive Cmap-based for drug repurposing

Given the rather limited set of transcriptomics data available and the small Cmap coverage for repurposable drugs (*i.e.* only 17% of molecules in our drug database, with none of the biologics), results were taken as supportive in the present study. Among the top network-based drugs proposed for repurposing, only 2 corticosteroids (betamethasone and hydrocortisone) were confirmed to elicit a reversed gene expression profile (Cmap score < -0.3) when compared to the disease gene expression state.

Conclusion

This study was designed to identify existing drugs which could be repurposed in a short time frame as a treatment for severe forms of COVID-19. We reasoned that such drugs should target those molecular pathways involved in the transition from mild lung inflammation caused by viral infection up to the cytokine storm associated with advanced stages of the disease (Fig 2, central and right lower panels). To this aim, using multiple sources of molecular profiling data from the literature relevant to distinguish mild from severe forms of the disease at the level of tissues and immune cells, we established a model of lung inflammation associated with COVID-19 in the form of an interactome of disease-related proteins. Combined with pharmacological knowledge of drug targets, this interactome allowed us to identify existing compounds which could be made available to patients in a short time frame.

Our network computational analyses identified several candidate therapeutic targets and corresponding drugs to repurpose which were confirmatory of existing knowledge (Table 1). This includes for example therapeutic antibodies interfering with either IL1β, IL6, TNFα or their receptors directly contributing to the CRS associated with severe COVID-19. Various inhibitory antibodies directed to these targets have already been evaluated in COVID-19 patients, such as anti-IL1® (canakinumab), anti-IL6R (tocilizumab, sarilumab) or anti-TNFα (infliximab, adalimumab) antibodies [4, 56]. Overall, these drugs yielded conflicting efficacy results, likely explained by evidence that such anti-cytokine treatments are rather effective if administered to patients before they develop advanced COVID-19 [57]. Nonetheless, a recent study evaluating the anti-IL6R antibodies tocilizumab and sarilumab demonstrated some improvement in survival when treating critically ill COVID-19 patients, even more so when these drugs were associated with dexamethasone [4, 5, 58]. Corticosteroids, are also predicted by the present study to be useful in severe COVID-19, in agreement with positive results previously obtained in multiple randomized clinical trials, eventually leading to a broad use of dexamethasone as a treatment for severe COVID-19 [2, 59]. JAK1 and JAK2 inhibitors came out also as interesting candidates for repurposing, with several inhibitors being actively tested in COVID-19 patients [60]. In this therapeutic class, the JAK1/JAK2 inhibitor baricitinib is currently raising most of the interest in light of recent evidence that it interferes with virus entry mediated by clathrin-associated endocytosis (Table 1) [61]. We also identified drugs interfering with reproductive steroids or their receptors as valid candidates for repurposing. This observation makes sense in light of the strong bias towards males among patients with severe COVID-19, perhaps explained in part by the upregulation by androgens of the expression of the SARS CoV-2 receptor [62]. In contrast estrogens and progesterone are rather considered to be protective in light of their anti-inflammatory properties as well as their capacity to promote proliferation and repair of respiratory epithelial cells [63]. On this basis, treatment with estrogens are being considered in patients with mild COVID-19 (Table 1).

Perhaps more interestingly, our repurposing study sheds light on other therapeutic classes which as of today receive insufficient attention as potential treatments for severe COVID-19. We predict that inhibitors of the well-known IL17 and IL23 proinflammatory cytokines (or their receptors) could be useful in COVID-19, with to our knowledge a single clinical trial evaluating as of today the anti-IL17 antibody secukinumab in COVID-19 [52]. Multiple monoclonal antibodies blocking those cytokines have been registered as treatments for other inflammatory diseases, which thus could be promptly repurposed in COVID-19 (Table 1). Similarly, the C5 complement inhibitor eculizumab was also identified to represent a valid therapeutic option, in agreement with recent evidence that the C5a-C5aR axis promotes severe lung inflammation in COVID-19 patients by mediating recruitment and activation of pro-inflammatory myeloid cells [53, 64]. Only proof of concept studies have been conducted so far

in human with eculizumab, suggesting that this antibody may provide some benefit in severe COVID-19 [65, 66], with a confirmatory trial ongoing in a larger cohort of patients. Noteworthy, another clinical study has been recently initiated to evaluate as well in this indication the anti C5a receptor antibody avdoralimab (Table 1). Also, approaches combining JAK1/2 inhibitors with blockade of C5a with eculizumab are being considered as a treatment of severe pulmonary damage in COVID-19 patients [67]. Moreover, drugs such as romiplostim acting as an agonist for the thrombopoietin receptor are also predicted to be useful to treat COVID-19-associated thrombocytopenia, in agreement with a recent case study documenting platelet recovery following treatment with this drug of a COVID-19 pediatric patient [54].

The most significant outcome of our repurposing study is the prediction that several members of the alarmin family such as defensins, HMBG1, IL1α, IL25, IL33, TSLP, S100A4, S100A7, S100A8, S100A9, S100A12, S100B, S100P likely contribute to lung inflammation during COVID-19 (Fig 4) [68–70]. The role of each individual alarmin in this regard remains to be investigated, with presumably some of them (*e.g.* IL25, TSLP) rather contributing to the initial recruitment of myeloid cells and innate lymphoid cells following epithelial or endothelial cell infection, whereas others (IL33, S100 members) are likely being involved in later stages of lung inflammation culminating in the cytokine storm. The later assumption is consistent with recent observations that some alarmins can stimulate the production of both IL1β, IL6 and TNFα as well as multiple other proinflammatory cytokines and chemokines [71]. Furthermore, blood levels of IL1α, calprotectin (a heterodimer made of S100A8 and S100A9), S100A12, S100B and HGBM1 appear to correlate with COVID-19 severity [72–76] (S1 Table). Also, IL33 has been recently proposed to play a broad role in the pathophysiology of COVID-19 pneumonia by dampening both the antiviral interferon response as well as regulatory T cells, while promoting thrombosis and activating pro-inflammatory type 2 innate lymphoid cells and γδ T cells [77]. To our knowledge, only few clinical studies are being conducted as of today in COVID-19 with a TSLP inhibitor or with blocking antibodies directed to receptors for IL1α or IL33 (i.e. ST2), whereas multiple additional blocking monoclonal antibodies directed to IL25, IL33 or TSLP are well under clinical evaluation to treat severe forms of asthma or atopic dermatitis [62, 69]. Furthermore, various inhibitors of the S100 family of proteins currently in preclinical development may represent promising drug candidates for the future (Table 1). We thus recommend considering existing anti-alarmins therapies to treat severe COVID-19, most particularly in the context of the converging rationale from this computational study as well as recent wet-lab evidence that this important class of proteins conveying proinflammatory signals plays a critical role in the pathophysiology of severe COVID-19. Lastly, this first model of severe lung inflammation in COVID-19 should be updated as new data are generated to better distinguish at an early stage patients with a high risk of evolving towards severe lung inflammation from those who will only develop mild forms of the disease.

Supporting information

S1 Table. Candidate COVID-19 related disease genes.
(PDF)

S2 Table. Pathways enrichment analysis.
(PDF)

S3 Table. Drug repurposing.
(XLSX)

Acknowledgments

The authors are thankful to Dorothée Piva for providing excellent secretarial assistance.

Author Contributions

Conceptualization: Emiko Desvaux, Sandra Hubert, Audrey Aussy, Laurence Laigle, Mickaël Guedj, Philippe Moingeon.

Data curation: Emiko Desvaux, Sandra Hubert.

Formal analysis: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Audrey Aussy, Laurence Laigle, Jessica Laplume, Perrine Soret, Pierre Jean-François, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Investigation: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Jessica Laplume, Perrine Soret, Pierre Jean-François, Isabelle Dupin-Roger.

Methodology: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Jessica Laplume, Perrine Soret, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Software: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Perrine Soret.

Supervision: Mickaël Guedj, Philippe Moingeon.

Validation: Audrey Aussy, Laurence Laigle, Jessica Laplume, Pierre Jean-François, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Visualization: Emiko Desvaux, Sandra Hubert.

Writing – original draft: Mickaël Guedj, Philippe Moingeon.

Writing – review & editing: Emiko Desvaux, Sandra Hubert, Audrey Aussy, Laurence Laigle, Isabelle Dupin-Roger.

References

1. Callaway E. Could new COVID variants undermine vaccines? Labs scramble to find out. Nature. 2021; 589: 177–178. https://doi.org/10.1038/d41586-021-00031-0 PMID: 33432212
2. Group TRC. Dexamethasone in Hospitalized Patients with Covid-19—Preliminary Report. New England Journal of Medicine. 2020 [cited 18 Jan 2021].
3. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the Treatment of Covid-19—Final Report. N Engl J Med. 2020; 383: 1813–1826. https://doi.org/10.1056/NEJMoa2007764 PMID: 32445440
4. Salvarani C, Dolci G, Massari M, Merlo DF, Cavuto S, Savoldi L, et al. Effect of Tocilizumab vs Standard Care on Clinical Worsening in Patients Hospitalized With COVID-19 Pneumonia: A Randomized Clinical Trial. JAMA Intern Med. 2021; 181: 24–31. https://doi.org/10.1001/jamainternmed.2020.6615 PMID: 33080005
5. Hermine O, Mariette X, Tharaux P-L, Resche-Rigon M, Porcher R, Ravaud P, et al. Effect of Tocilizumab vs Usual Care in Adults Hospitalized With COVID-19 and Moderate or Severe Pneumonia: A Randomized Clinical Trial. JAMA Intern Med. 2021; 181: 32–40. https://doi.org/10.1001/jamainternmed.2020.6820 PMID: 33080017
6. Repurposed Antiviral Drugs for Covid-19—Interim WHO Solidarity Trial Results. New England Journal of Medicine. 2021; 384: 497–511. https://doi.org/10.1056/NEJMoa2023184 PMID: 33264556
7. Krammer F. SARS-CoV-2 vaccines in development. Nature. 2020; 586: 516–527. https://doi.org/10.1038/s41586-020-2798-3 PMID: 32967006
8. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020; 6: 14. https://doi.org/10.1038/s41421-020-0153-3 PMID: 32194980

9. Nabirotchkin S, Peluffo AE, Bouaziz J, Cohen D. Focusing on the Unfolded Protein Response and Autophagy Related Pathways to Reposition Common Approved Drugs against COVID-19. 2020 [cited 18 Jan 2021]. https://doi.org/10.20944/preprints202003.0302.v1

10. Li X, Yu J, Zhang Z, Ren J, Peluffo AE, Zhang W, et al. Network Bioinformatics Analysis Provides Insight into Drug Repurposing for COVID-2019. 2020 [cited 18 Jan 2021]. https://doi.org/10.20944/preprints202003.0286.v1

11. Ciliberto G, Cardone L. Boosting the arsenal against COVID-19 through computational drug repurposing. Drug Discovery Today. 2020; 25. https://doi.org/10.1016/j.drudis.2020.04.005 PMID: 32304645

12. Chowdhury KH, Chowdhury MR, Mahmud S, Tareq AM, Hanif NB, Banu N, et al. Drug Repurposing Approach against Novel Coronavirus Disease (COVID-19) through Virtual Screening Targeting SARS-CoV-2 Main Protease. Biology. 2021; 10: 2. https://doi.org/10.3390/biology10010002 PMID: 33374717

13. Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, et al. COVID-19: combining antiviral and anti-inflammatory treatments. Lancet Infect Dis. 2020; 20: 400–402. https://doi.org/10.1016/S1473-3099(20)30132-8 PMID: 32113509

14. Gysi DM, Valle ÍD, Zitnik M, Ameli A, Gan X, Varol O, et al. Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19. arXiv:200407229 [cs, q-bio, stat]. 2020 [cited 1 Jun 2021]. Available: http://arxiv.org/abs/2004.07229 PMID: 32550253

15. Fiscon G, Conte F, Farina L, Paci P. SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19. PLOS Computational Biology. 2021; 17: e1008686. https://doi.org/10.1371/journal.pcbi.1008686 PMID: 33544720

16. Fiscon G, Paci P. SAveRUNNER: An R-based tool for drug repurposing. BMC Bioinformatics. 2021; 22: 150. https://doi.org/10.1186/s12859-021-04076-w PMID: 33757425

17. Tzotzos SJ, Fischer B, Fischer H, Zeitlinger M. Incidence of ARDS and outcomes in hospitalized patients with COVID-19: a global literature survey. Critical Care. 2020; 24: 516. https://doi.org/10.1186/s13054-020-03240-7 PMID: 32825837

18. Vardhana SA, Wolchok JD. The many faces of the anti-COVID immune response. J Exp Med. 2020; 217. https://doi.org/10.1084/jem.20200678 PMID: 32353870

19. Moore JB, June CH. Cytokine release syndrome in severe COVID-19. Science. 2020; 368: 473–474. https://doi.org/10.1126/science.abb8925 PMID: 32303591

20. de la Rica R, Borges M, Gonzalez-Freire M. COVID-19: In the Eye of the Cytokine Storm. Front Immunol. 2020; 11. https://doi.org/10.3389/fimmu.2020.558898 PMID: 33072097

21. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nature Communications. 2016; 7: 10331. https://doi.org/10.1038/ncomms10331 PMID: 26831545

22. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018; 9: 2691. https://doi.org/10.1038/s41467-018-05116-5 PMID: 30002366

23. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, Møller R, Panis M, Sachs D, et al. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. bioRxiv. 2020; 2020.03.24.004655. https://doi.org/10.1101/2020.03.24.004655

24. Ackermann M, Verleden SE, Kuehnel M, Haverich A, Welte T, Laenger F, et al. Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. New England Journal of Medicine. 2020; 383: 120–128. https://doi.org/10.1056/NEJMoa2015432 PMID: 32437596

25. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nature Medicine. 2020; 26: 842–844. https://doi.org/10.1038/s41591-020-0901-9 PMID: 32398875

26. Laing AG, Lorenc A, del Molino del Barrio I, Das A, Fish M, Monin L, et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. Nature Medicine. 2020; 26: 1623–1635. https://doi.org/10.1038/s41591-020-1038-6 PMID: 32807934

27. Hadjadj J, Yatim N, Barnabei L, Corneau A, Boussier J, Péré H, et al. Impaired type I interferon activity and exacerbated inflammatory responses in severe Covid-19 patients. medRxiv. 2020; 2020.04.19.20068015. https://doi.org/10.1126/science.abc6027 PMID: 32661059

28. Ng LFP, Hibberd ML, Ooi E-E, Tang K-F, Neo S-Y, Tan J, et al. A human in vitro model system for investigating genome-wide host responses to SARS coronavirus infection. BMC Infect Dis. 2004; 4: 34. https://doi.org/10.1186/1471-2334-4-34 PMID: 15357874

29. Brodin P. Immune determinants of COVID-19 disease presentation and severity. Nature Medicine. 2021; 27: 28–33. https://doi.org/10.1038/s41591-020-01202-8 PMID: 33442016

30. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discovery. 2020; 6: 1–18. https://doi.org/10.1038/s41421-020-0168-9 PMID: 32377375

31. Burke H, Freeman A, Cellura DC, Stuart BL, Brendish NJ, Poole S, et al. Inflammatory phenotyping predicts clinical outcome in COVID-19. Respiratory Research. 2020; 21: 245. https://doi.org/10.1186/s12931-020-01511-z PMID: 32962703

32. Wu M, Chen Y, Xia H, Wang C, Tan CY, Cai X, et al. Transcriptional and proteomic insights into the host response in fatal COVID-19 cases. PNAS. 2020; 117: 28336–28343. https://doi.org/10.1073/pnas.2018030117 PMID: 33082228

33. Combes AJ, Courau T, Kuhn NF, Hu KH, Ray A, Chen WS, et al. Global absence and targeting of protective immune states in severe COVID-19. Nature. 2021; 1–10. https://doi.org/10.1038/s41586-021-03234-7 PMID: 33494096

34. Arunachalam PS, Wimmers F, Mok CKP, Perera RAPM, Scott M, Hagan T, et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science. 2020; 369: 1210–1220. https://doi.org/10.1126/science.abc6261 PMID: 32788292

35. Merad M, Martin JC. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nature Reviews Immunology. 2020; 20: 355–362. https://doi.org/10.1038/s41577-020-0331-4 PMID: 32376901

36. Vabret N, Britton GJ, Gruber C, Hegde S, Kim J, Kuksin M, et al. Immunology of COVID-19: Current State of the Science. Immunity. 2020; 52: 910–941. https://doi.org/10.1016/j.immuni.2020.05.002 PMID: 32505227

37. Choreño-Parra JA, Jiménez-Álvarez LA, Cruz-Lagunas A, Rodríguez-Reyna TS, Ramírez-Martínez G, Sandoval-Vega M, et al. Clinical and immunological factors that distinguish COVID-19 from pandemic influenza A(H1N1). medRxiv. 2020; 2020.08.10.20170761. https://doi.org/10.1101/2020.08.10.20170761

38. Atamas SP, Chapoval SP, Keegan AD. Cytokines in chronic respiratory diseases. F1000 Biol Rep. 2013; 5. https://doi.org/10.3410/B5-3 PMID: 23413371

39. Wang M, Withers JB, Ricchiuto P, Voitalov I, McAnally M, Sanchez HN, et al. A systems-based method to repurpose marketed therapeutics for antiviral use: a SARS-CoV-2 case study. Life Science Alliance. 2021; 4. https://doi.org/10.26508/lsa.202000904 PMID: 33593923

40. Stolfi P, Manni L, Soligo M, Vergni D, Tieri P. Designing a Network Proximity-Based Drug Repurposing Strategy for COVID-19. Front Cell Dev Biol. 2020; 8. https://doi.org/10.3389/fcell.2020.545089 PMID: 33123533

41. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017; 18: 551–562. https://doi.org/10.1038/nrg.2017.38 PMID: 28607512

42. Song J-S, Wang R-S, Leopold JA, Loscalzo J. Network determinants of cardiovascular calcification and repositioned drug treatments. FASEB J. 2020; 34: 11087–11100. https://doi.org/10.1096/fj.202001062R PMID: 32638415

43. Chen X, Ji Z-L, Chen Y. TTD: Therapeutic Target Database. Nucleic acids research. 2002; 30: 412–5. https://doi.org/10.1093/nar/30.1.412 PMID: 11752352

44. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36: D901–906. https://doi.org/10.1093/nar/gkm958 PMID: 18048412

45. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nature Communications. 2019; 10: 1197. https://doi.org/10.1038/s41467-019-09186-x PMID: 30867426

46. Cao M, Zhang H, Park J, Daniels NM, Crovella ME, Cowen LJ, et al. Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. PLOS ONE. 2013; 8: e76339. https://doi.org/10.1371/journal.pone.0076339 PMID: 24194834

47. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313: 1929–1935. https://doi.org/10.1126/science.1132939 PMID: 17008526

48. Sokulsky LA, Garcia-Netto K, Nguyen TH, Girkin JLN, Collison A, Mattes J, et al. A Critical Role for the CXCL3/CXCL5/CXCR2 Neutrophilic Chemotactic Axis in the Regulation of Type 2 Responses in a Model of Rhinoviral-Induced Asthma Exacerbation. The Journal of Immunology. 2020; 205: 2468–2478. https://doi.org/10.4049/jimmunol.1901350 PMID: 32948685

49. Ye Q, Wang B, Mao J. The pathogenesis and treatment of the 'Cytokine Storm' in COVID-19. J Infect. 2020; 80: 607–613. https://doi.org/10.1016/j.jinf.2020.03.037 PMID: 32283152

50. Leisman DE, Ronner L, Pinotti R, Taylor MD, Sinha P, Calfee CS, et al. Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. The Lancet Respiratory Medicine. 2020; 8: 1233–1244. https://doi.org/10.1016/S2213-2600(20)30404-5 PMID: 33075298

51. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. National Center for Biotechnology Information. [cited 19 May 2021]. https://www.ncbi.nlm.nih.gov/

52. Pacha O, Sallman MA, Evans SE. COVID-19: a case for inhibiting IL-17? Nature reviews Immunology. 2020; 20: 345–346. https://doi.org/10.1038/s41577-020-0328-z PMID: 32358580

53. Carvelli J, Demaria O, Vély F, Batista L, Chouaki Benmansour N, Fares J, et al. Association of COVID-19 inflammation with activation of the C5a–C5aR1 axis. Nature. 2020; 588: 146–150. https://doi.org/10.1038/s41586-020-2600-6 PMID: 32726800

54. Schneider CW, Penney SW, Helfrich AM, Hartman KR, Lieuw K. A Novel Use of Romiplostim for SARS-CoV-2–induced Thrombocytopenia. Journal of Pediatric Hematology/Oncology. 2021;Publish Ahead of Print. https://doi.org/10.1097/MPH.0000000000001961 PMID: 33003146

55. Ho JSY, Mok BW-Y, Campisi L, Jordan T, Yildiz S, Parameswaran S, et al. Topoisomerase 1 inhibition therapy protects against SARS-CoV-2-induced inflammation and death in animal models. bioRxiv. 2020; 2020.12.01.404483. https://doi.org/10.1101/2020.12.01.404483 PMID: 33299999

56. Robinson PC, Liew DFL, Liew JW, Monaco C, Richards D, Shivakumar S, et al. The Potential for Repurposing Anti-TNF as a Therapy for the Treatment of COVID-19. Med (N Y). 2020; 1: 90–102. https://doi.org/10.1016/j.medj.2020.11.005 PMID: 33294881

57. De Stefano L, Bobbio-Pallavicini F, Manzo A, Montecucco C, Bugatti S. A "Window of Therapeutic Opportunity" for Anti-Cytokine Therapy in Patients With Coronavirus Disease 2019. Front Immunol. 2020; 11. https://doi.org/10.3389/fimmu.2020.572635 PMID: 33123149

58. Della-Torre E, Campochiaro C, Cavalli G, De Luca G, Napolitano A, La Marca S, et al. Interleukin-6 blockade with sarilumab in severe COVID-19 pneumonia with systemic hyperinflammation: an open-label cohort study. Ann Rheum Dis. 2020; 79: 1277–1285. https://doi.org/10.1136/annrheumdis-2020-218122 PMID: 32620597

59. WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group, Sterne JAC, Murthy S, Diaz JV, Slutsky AS, Villar J, et al. Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19: A Meta-analysis. JAMA. 2020; 324: 1330–1341. https://doi.org/10.1001/jama.2020.17023 PMID: 32876694

60. Luo W, Li Y-X, Jiang L-J, Chen Q, Wang T, Ye D-W. Targeting JAK-STAT Signaling to Control Cytokine Release Syndrome in COVID-19. Trends in Pharmacological Sciences. 2020; 41: 531–543. https://doi.org/10.1016/j.tips.2020.06.007 PMID: 32580895

61. Seif F, Aazami H, Khoshmirsafa M, Kamali M, Mohsenzadegan M, Pornour M, et al. JAK Inhibition as a New Treatment Strategy for Patients with COVID-19. Int Arch Allergy Immunol. 2020; 181: 467–475. https://doi.org/10.1159/000508247 PMID: 32392562

62. Fagone P, Ciurleo R, Lombardo SD, Iacobello C, Palermo CI, Shoenfeld Y, et al. Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. Autoimmunity Reviews. 2020; 19: 102571. https://doi.org/10.1016/j.autrev.2020.102571 PMID: 32376402

63. Pinna G. Sex and COVID-19: A Protective Role for Reproductive Steroids. Trends in Endocrinology & Metabolism. 2021; 32: 3–6. https://doi.org/10.1016/j.tem.2020.11.004 PMID: 33229187

64. Peffault de Latour R, Bergeron A, Lengline E, Dupont T, Marchal A, Galicier L, et al. Complement C5 inhibition in patients with COVID-19—a promising target? Haematologica. 2020; 105: 2847–2850. https://doi.org/10.3324/haematol.2020.260117 PMID: 33256385

65. Annane D, Heming N, Grimaldi-Bensouda L, Frémeaux-Bacchi V, Vigan M, Roux A-L, et al. Eculizumab as an emergency treatment for adult patients with severe COVID-19 in the intensive care unit: A proof-of-concept study. EClinicalMedicine. 2020; 28. https://doi.org/10.1016/j.eclinm.2020.100590 PMID: 33173853

66. Diurno F, Numis FG, Porta G, Cirillo F, Maddaluno S, Ragozzino A, et al. Eculizumab treatment in patients with COVID-19: preliminary results from real life ASL Napoli 2 Nord experience. Eur Rev Med Pharmacol Sci. 2020; 24: 4040–4047. https://doi.org/10.26355/eurrev_202004_20875 PMID: 32329881

67. Giudice V, Pagliano P, Vatrella A, Masullo A, Poto S, Polverino BM, et al. Combination of Ruxolitinib and Eculizumab for Treatment of Severe SARS-CoV-2-Related Acute Respiratory Distress Syndrome: A Controlled Study. Front Pharmacol. 2020; 11: 857. https://doi.org/10.3389/fphar.2020.00857 PMID: 32581810

68. Yalcin Kehribar D, Cihangiroglu M, Sehmen E, Avci B, Capraz A, Yildirim Bilgin A, et al. The receptor for advanced glycation end product (RAGE) pathway in COVID-19. Biomarkers. 2021; 1–5. https://doi.org/10.1080/1354750X.2020.1861099 PMID: 33284049

69. Roth A, Lütke S, Meinberger D, Hermes G, Sengle G, Koch M, et al. LL-37 fights SARS-CoV-2: The Vitamin D-Inducible Peptide LL-37 Inhibits Binding of SARS-CoV-2 Spike Protein to its Cellular

Receptor Angiotensin Converting Enzyme 2 In Vitro. bioRxiv. 2020; 2020.12.02.408153. https://doi.org/10.1101/2020.12.02.408153

70. Idris MM, Banu S, Siva AB, Nagaraj R. Downregulation of Defensin genes in SARS-CoV-2 infection. medRxiv. 2020; 2020.09.21.20195537. https://doi.org/10.1101/2020.09.21.20195537

71. Yang D, Han Z, Oppenheim JJ. ALARMINS AND IMMUNITY. Immunol Rev. 2017; 280: 41–56. https://doi.org/10.1111/imr.12577 PMID: 29027222

72. Chen L, Long X, Xu Q, Tan J, Wang G, Cao Y, et al. Elevated serum levels of S100A8/A9 and HMGB1 at hospital admission are correlated with inferior clinical outcomes in COVID-19 patients. Cellular & Molecular Immunology. 2020; 17: 992–994. https://doi.org/10.1038/s41423-020-0492-x PMID: 32620787

73. Silvin A, Chapuis N, Dunsmore G, Goubet A-G, Dubuisson A, Derosa L, et al. Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. Cell. 2020; 182: 1401–1418.e18. https://doi.org/10.1016/j.cell.2020.08.002 PMID: 32810439

74. Zuniga M, Gomes C, Carsons SE, Bender MT, Cotzia P, Miao QR, et al. Autoimmunity to the Lung Protective Phospholipid-Binding Protein Annexin A2 Predicts Mortality Among Hospitalized COVID-19 Patients. medRxiv. 2021; 2020.12.28.20248807. https://doi.org/10.1101/2020.12.28.20248807

75. Aceti A, Margarucci LM, Scaramucci E, Orsini M, Salerno G, Di Sante G, et al. Serum S100B protein as a marker of severity in Covid-19 patients. Scientific Reports. 2020; 10: 18665. https://doi.org/10.1038/s41598-020-75618-0 PMID: 33122776

76. Zeng Z, Hong X-Y, Li Y, Chen W, Ye G, Li Y, et al. Serum-soluble ST2 as a novel biomarker reflecting inflammatory status and illness severity in patients with COVID-19. Biomarkers in Medicine. 2020; 14: 1619–1629. https://doi.org/10.2217/bmm-2020-0410 PMID: 33336592

77. Zizzo G, Cohen PL. Imperfect storm: is interleukin-33 the Achilles heel of COVID-19? The Lancet Rheumatology. 2020; 2: e779–e790. https://doi.org/10.1016/S2665-9913(20)30340-4 PMID: 33073244

F.4 Conclusion

This repurposing study identified putative drug targets to reduce severe cases of inflammation occurring in patients suffering from COVID-19 (see [Yal+21], [Rot+20] and [Idr+20]).

Interestingly, we also highlighted the potential therapeutic effect of drugs that received insufficient attention till now for COVID-19 treatment, such as IL17 and IL23 inhibitors [PSE20], the C5 complement inhibitor eculizumab [Car+20], and agonists for the thrombopoietin receptor intervening in mitigating thrombocytopenia (when the platelet population, key for tissue clogging, is too low, [Sch+21]).

In parallel, we highlighted the role of alarmin family proteins (e.g. defensins, HMGB1, IL1*alpha*, IL25, IL33, and S100 proteins, a comprehensive review is proposed in [YHO17]) in COVID-19 cases. Last but not least, Topoisomerase 1 inhibitors [Ho+20], currently used as cytotoxic drugs in oncology, were also identified as potential candidates in inhibiting SARS-CoV-2 inflammation and death in animal models.

Overall, our comprehensive data-driven and agnostic approach enabled to identify disease-related proteins and potential drugs for repurposing in the treatment of severe lung inflammation during COVID-19 scourge, as demonstrated in [Yad+], [Bel+21] and [Tai+22].

In this paper, I mostly contribute by retrieving a transcriptomic signature of COVID-19-infected lung cells, leveraging the RNASeq expression extracted from Calu-3 and NHBE. To perform these standard differential analyses, we notably capitalise with my PhD partner on our newly industrial and homogenised RNASeq pipeline, see Appendix A for details. Then, I profit from the same pipeline to homogenise the perturbagen transcriptomic profiles of the CMap database, ensuring notably to reduce batch and normalisation artefacts between the infected samples and the ones from the molecular profile collection. Finally, I compared several enrichment analyses or distant-metrics scores to assert statistically how similar two gene expression profiles are, providing complementary material to section `Supportive CMap-based for drug repurposing` of [Des+21]. In our study, it appears that a simple Pearson correlation metric, while making strong naive assumption of linearity between two compared profiles, showcases the most consistent results with respect to the drug-disease links returned by the network-based methodology.

F.5 Perspectives

Given the rather limited set of transcriptomics data available and the small CMap coverage for repurposable drugs (i.e. only 17% of molecules in our drug database), results were only used to support the present study. Indeed, among the top network-based drugs, only 2 corticosteroids (betamethasone and hydrocortisone) exhibited a significantly inverted gene expression profile (CMap score $<$ -0.3) when compared to the COVID-19 gene expression profile. To that end, the L1000 project is a promising avenue by extending by several orders of magnitude the CMap expression profiles database. Through the new L1000 technology [6], an experimental high-throughput profiling platform, the large-scale L1000 database embraces a much larger diversity of perturbagen signature profiles, ranging from drugs to genetic manipulations (mostly elicited from Knock out or Knock down experiences) through mutagen factors. While the general principle of treating human cells with different perturbagens is comparable to the CMap underlying principle, the generation of the L1000 database involves an additional step, since, instead of measuring the expression level of the entire genome, only a subset of *landmark genes* are directly acknowledged,

[6] hence named, as this method represents both a 1,000-fold scale-up of the CMap profiling database and only accounts for the expression of 1000 key genes

carefully selected such that they are representative of the entire transcriptome and can be used as proxies to infer the expression of remaining genes. Interestingly, [Sub+17] demonstrate that this protocol compared favourably to RNA sequencing in terms of robustness and accuracy to infer the expression levels of approximately 81% of transcripts coding for proteins, that were not yet directly measured. Ultimately, by taking up to a much larger scale the CMap compendium, the measured aggregation of dozens of thousands of perturbagen profiles is expected to provide a priceless resource for understanding cellular responses to different stimuli and hereby elucidating novel biological pathways and insights into disease mechanisms, while providing an integrative way to explore shared transcriptomic mechanisms induced by various treatments, facilitating drug discovery or repositioning (see also their web, user-friendly API, available here clue.io).

A significant limitation of the methodology devised with Patrimony is the potential loss of meaningful and interpretable biological information, since we fuse diverse datasets, from multiple omics to drug interaction databases, into a single global knowledge graph. Furthermore, searching for drug similarities and interactions reveal a highly memory- and time-consuming task, primarily due to the high-dimensionality and number of datasets that were integrated in the Patrimony project. By reviewing the recent repurposing literature, I was stunned by the iCell method [Mal+19] to tackle these scalability and integration issues, since iCell was precisely set up to determine clusters of highly co-expressed genes across dissimilar interaction databases.

Briefly, the iCell framework merges three molecular interaction networks (Protein-protein interaction: PPI, Genetic interaction: GI and gene Coexpressions: COEX), all represented as indicator adjacency matrices (symmetric matrices in which an off-diagonal 1 encodes for an existing interaction, 0 otherwise), by projecting them into a smaller subspace of shared cluster of genes. Precisely, these three matrices were simultaneously decomposed into a common matrix $\boldsymbol{G}$, describing gene cluster annotations shared across all networks, and a compressed, omic-specific $\boldsymbol{S}_i$ matrix, showing how gene clusters are related to each other. This decomposition was achieved by minimizing a Multiple Symmetric Non-negative Matrix Tri-Factorization (MSNMTF) objective function.

Since then, this methodology was extended to integrate completely distinct sources of information, firstly in a naive manner, by taking profit of the cluster-indicator matrix returned by iCell as compact and informative input for other biological applications, such as patient clustering and stratification [Xen+23] or inferring new drug-pathway interactions [Mal+23]. [Mal+23] generated new drug repurposing hypotheses to cure severe cases of COVID-19 cases, computed with the previously described iCell methodology [7], with drug-target interactions (DTIs) and Drug Chemical Similarity (DCS) networks from DrugBank, the latter being represented by its *Laplacian* version (the diagonal degree matrix subtracted to the drug-drug adjacency matrix, in other words a matrix whose diagonal stores the degree vertices, and the off-terms pairwise interactions, with a -1 positively coding for a direct connection). Precisely, the main idea was to retrieve the low-dimensional gene clusters-drug interaction matrix, resulting from both a matrix factorisation with an additional regularisation term to account for the known structure of the DCS network, which predicts the best the global drug-target interaction matrix while projected into a much lower dimensional space (see [Mal+23, Eq 1.] for the explicit optimisation problem). Ultimately, the new drug-target interaction matrix, supposedly with higher completion as inferred from disease-specific omics, was reconstructed to predict individual drug-target interactions. Indeed, each entry in the reconstructed matrix stands for an *association score*, supposed to be significant if the predicted drug-target interaction scores higher than the mean of existing interactions. [Xen+23] implements a novel Non-negative Matrix Tri-Factorization

[7] Interestingly, the same two cell lines that we used to retrieve enriched pathways in [Des+21], namely Calu3 and NHBE, were used to that purpose.

(NMTF) strategy to overcome the lack of genetic samples (five samples make up the cohort) to identify novel disease-related genes for antithrombin resistance, since GWAS, while relevant to find disease-associated variants for common diseases, are not suited to rare diseases due to the sparsity of genetic data. And precisely, matrix factorisation techniques can deal with low sample size by integrating prior knowledge and combining different types of biological and medical data.Precisely, the NMTF method was used to integrate four different data types: germline mutations, protein-protein interactions (PPIs), co-expressions (COEXs) and genetic interactions (GIs), the last three being summarised into an indicator matrix $\boldsymbol{G}$ in which hard assignment to a gene cluster is indicated by a 1, 0 otherwise. Then, the genetic patient information matrix $\boldsymbol{M}$ is simultaneously decomposed into the product of three lower dimensional factors, $\boldsymbol{P}$, $\boldsymbol{S}$ and $\boldsymbol{G}$; the latter, storing the k_1 gene clusters, being the standard output of the iCell methodology and $\boldsymbol{P}$ storing the k_2 patient clusters (see [Xen+23, Eq.1] for details).

This heuristic method to fuse heterogeneous data entities was further extended to capture in an unified framework all systematic interactions while providing gene and patient clusters and supplying drug target candidates, with respectively applications to COVID-19 repurposing again [Zam+21], cancer stratification to generate personalised treatment [GMP15] or an ongoing work to determine Parkinson's trajectories through multiple single-cell RNA-seq time points [, Poster ID: 8328, session P236-M]. These matrix factorisation methods embody a new powerful way-of-thinking to add in the computational toolbox.

To conclude, by explicitly incorporating interactions across multiple biological entities and systems in an integrated way, thus keeping the mechanistic biological interpretation and facilitating prior knowledge integration, these approaches have the potential to bridge the gap between powerful yet less biologically interpretable and robust AI-based models and more standard statistical methods that struggle with high-dimensional datasets, particularly in the presence of a limited number of replicates.

However, the methods described earlier are not well-suited for identifying causal genes and reconstructing the chronological sequence of genes involved in pathway signalling. In other words, they may not effectively distinguish between downstream resulting phenotypes and upstream activation events in biological processes, since the inferred graphs are not directed. In this regard, Bayesian networks and their natural extension to Causal networks appear to be particularly well-suited for capturing complex and mechanistic interactions of biological phenomena, and such in their entirety.

For instance, the study conducted by [Li+19], Bayesian networks were used to infer pathways specifically involved in Systemic Lupus Erythematosus (SLE), from a cohort of 1760 SLE patients. Of note, the analysis identified and confirmed the significance of both the JAKSTAT and the Interferon signature, while capitalising on both prior knowledge and novel inferred interactions. Precisely, the pipeline to infer the structure of Bayesian networks comprised three main steps: first, co-expression networks were constructed using the R WGCNA package, returning 3 gene modules highly correlated to the phenotype, encompassing 431 differentially expressed genes. Second, text mining was employed to uncover literature-based gene pairwise connections. Third, both observational data (transcriptomic expression) and prior edges were combined to generate the Bayesian networks:

1. Random graphs were initially created, by integrating the edges inferred through text mining.

2. Subsequently, the igraph package was utilized to remove *bidirectional edges* and *cycles* in the graph. Indeed, one of the primary limitations of Bayesian networks is that they are restricted to the space of Directed Acyclic Graphs (DAGs), which makes them less valuable for describing positive or negative feedback regulation loops.

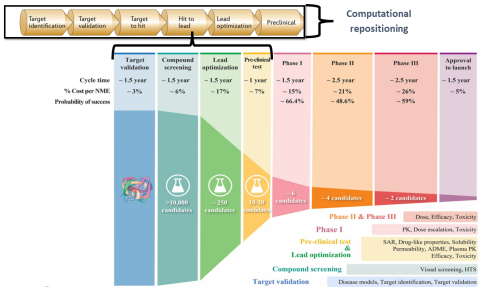
3. A score-based method from the bnlearn package, employing a hill-climbing algorithm to maximise the likelihood of the generated network structure, was then used to generate an *ensemble model* composed of 100 graphs, in which edges appearing with low frequency were discarded.

A similar process was used in [Mos+18], to build a molecular network of the aging human frontal cortex, and identifies specifically the pathways leading to Alzheimer's degeneracy and cognitive decline. [Mos+18] includes an additional validation step to highlight key driver genes, through direct interventions, in the form of knock-outs to directly silence the expression of genes one by one, and evaluated by conducting standard ANOVA analysis.

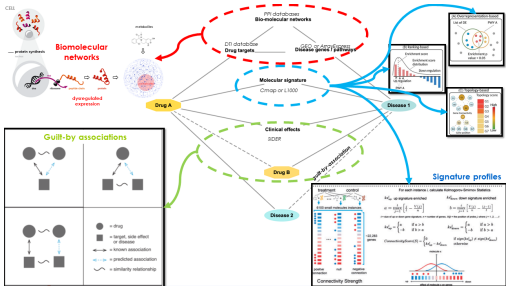
However, instead of integrating causal information posterior to the Bayesian network construction, it is also possible to directly combine interventional and observational information when learning the structure of Bayesian networks. We already discuss a possible framework to do so in Section 6.3, in relation with paper [RJN13]. An alternative approach, in the absence of interventional datasets, is to integrate biological causal prior information. For example, by incorporating the knowledge that DNA nodes should precede RNA nodes, which in turn should be parents of protein nodes. This strategy was demonstrated in [Gru+16] using the proprietary REFS^{TM} causal inference engine, where they combined transcriptomic data, clinical features and treatment annotations in an ensemble model of 256 Bayesian networks with 30 084 variables. The goal was to identify severe pathways involved in myeloma severity and identify patient subpopulations that would benefit most from stem cell transplantation treatment. The RIMBANET algorithm was also used to distinguish simple correlations from causal relations by leveraging DNA-based variations (expression quantitative trait loci, eQTL, or SNP, for single nucleotide polymorphisms), respectively revealing unexplored interconnections between lipid metabolism and glucose regulation pathways in type 2 diabetes [Coh+21] and identifying key regulators of Alzheimer's disease [Bec+20].

However, it is important to note that the performance of Bayesian networks is partly hindered by their lack of scalability to high-dimensional datasets (the sampling space of possible causal node orderings grows exponentially with the number of variables), the strong assumptions they make on the distribution of the datasets (most Bayesian algorithms were developed to accommodate either discretised or Gaussian-shaped distributions), and their high sensitivity to even minor alterations or noise present in the samples.

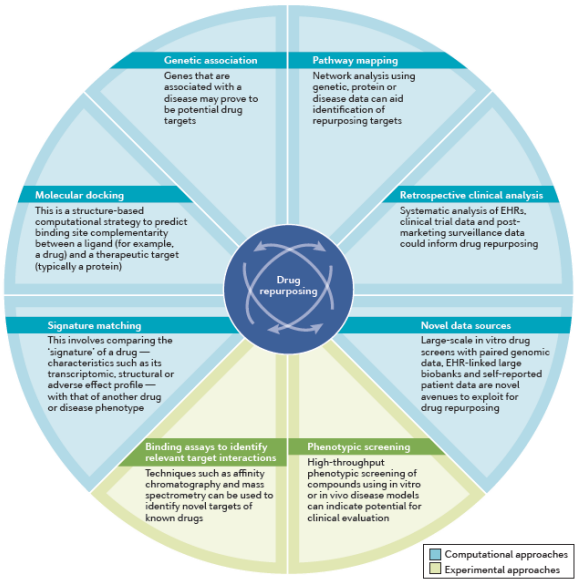
[FC05], [Goë+07], [GE13], [SD03], [PF17], [Lou23] [Cal+22], [De +17] [MJ18], [Che+11] [EL13], [LA18], [Pla05], [Le +01] [ZH14], [Mur+02] [Pat+06], [Kuo+06], [Sev+06], [Sîr+12] [Daw23], [SGH19] [DrS17], [DML23], [Cle22], [Sîr+12], [Cal11], [Amb20], [GM13], [Ngu21] [MA23], [LHM21] [LGN16], [Moh+22], [Nak20], [Gro+16], [RA15], [hae11], [Shr22], [DMW19], [Gal23], [dBru+21]



(a) The "valley-of-death" in drug-development. We detail the main stages of the drug-development cycle, the chances of success being modelled by an "attrition funnel"([phi22, Fig.1]).



(b) Main computational repositioning strategies implemented in Patrimony. a) First approach consists of constructing knowledge graphs, combining molecular profiles them with drug targets. The second approach matches transcriptomic profiles between perturbagens and disease phenotype, with [Mus+17, Fig.1], detailing the principle underlying the CMap enrichment score. Various *guilt-by-association* strategies, reproduced from [Hod+16, Fig.3].



(c) The most popular **Drug repurposing strategies** are enumerated in that diagram. Reproduced from [Pus+19, Fig. 1].

Figure F.2: Infographic 2) of Appendix C

Contents

Remerciements v

Publications ix

Oral communications x
Peer-reviewed communications . . . x
Invited seminars . . . xi
Poster sessions . . . xi

Abstract xii

Résumé long de la thèse xiv

Contents xxiii

I Biological introduction 1

1 Study of the transcriptome 2
1.1 Regulation of the Transcriptome . . . 2
1.1.1 Overview: Importance of meticulous Regulation of Gene Expression . . . 2
Regulation of the chromatin structure . . . 3
Co-expression networks of transcription factors . . . 3
Post-transcriptional regulation . . . 4
Post-translational regulation . . . 5
1.1.2 Epigenetics: Implications in Molecular Profile Diversity . . . 6
1.2 Tools for exploring the transcriptome . . . 8
1.2.1 Microarray technology to quantify gene expression . . . 8
1.2.2 RNASeq technology . . . 9
Historical development of RNA-Seq . . . 9
Outline of RNA-Seq analysis . . . 9
Microarray vs RNA-Seq . . . 13
1.2.3 Perspectives: single cell and spatial transcriptomics . . . 15
1.2.4 Conclusion: The Significance of Transcriptomic Data in Computational Medicine . . . 18

2 Introduction to the Immune System **21**
2.1 Key actors of the Immune System . . . 21
2.1.1 The innate system . . . 21
2.1.2 The adaptive system . . . 23
2.1.3 Exchange of goodwill between the two immune systems . . . 24
2.1.4 Immune dysregulation . . . 26
2.2 Physical methods for studying changes of cellular Composition . . . 28
2.2.1 Cytometry analyses . . . 28
2.2.2 Imaging methods . . . 29

II Transcriptome and mixture models **34**

3 Article 1: Gaussian Mixture Models in R **35**
3.1 Article 1 . . . 35
3.2 Main results . . . 58
3.3 Perspectives . . . 58

4 Article 2: A new molecular classification in primary Sjögren's syndrome **60**
4.1 Article 2 . . . 63
4.2 Main results . . . 82
4.3 Limitations and perspectives . . . 84

III Cell populations and deconvolution algorithms **86**

5 Article 3: review of cellular deconvolution methods **87**
5.1 Article 3 . . . 88
5.2 Conclusion: major Limitations of existing Deconvolution Algorithms Solutions 122

6 Article 4: DeCovarT, a deconvolution algorithm leveraging co-expression networks **123**
6.1 Article 4 . . . 123
6.2 Conclusion . . . 133
6.3 Publication Outline . . . 133

Thesis overview **135**

A Appendix of Chapter 1 **139**
A.1 Data import and cleaning . . . 144
A.1.1 Import relevant files . . . 144
A.1.2 Data wrangling with ExpressionSet . . . 145
Sample annotation . . . 147
Gene annotation . . . 150
Cell type annotation . . . 154
A.2 Pre-processing of raw counts, using a Nextflow pipeline . . . 159
A.3 From raw counts to normalised gene expression and downstream analyses . . . 160
A.3.1 Sample filtering . . . 160
A.3.2 Gene filtering . . . 160
A.3.3 Normalization and transformation . . . 164

A.3.4 Quality control and data exploration . . . 169
Control the distribution of read counts . . . 169
Multidimensional projections to identify patterns of expression . . . 171
A.3.5 Batch effect correction . . . 173
Combat . . . 176
Surrogate Variable Analysis (SVA) . . . 176
A.4 Downstream analyses . . . 178
A.4.1 Differential expression analysis . . . 178
From scratch . . . 179
DESeq2 framework . . . 180
Limma framework . . . 181
Visualisations . . . 183
A.4.2 Multi-level classification of cell populations . . . 190
A.5 Conclusions and perspectives . . . 193
A.6 Appendix A: Gene notations . . . 195
A.6.1 Gene terminologies . . . 195
A.6.2 Automated methods for gene annotation . . . 196

B Appendix of Article 1 **200**

C Appendix of Article 3 **265**
C.1 Appendix of Reference-Based Approaches . . . 265
Fundamental Assumptions on the Partial Deconvolution Framework . . . 265
C.1.1 Linear regression and Gauss-Markov theorem . . . 265
C.1.2 Robust regression approaches . . . 268
C.1.3 Regularised linear approaches . . . 271
C.1.4 Probabilistic approaches . . . 272
C.2 Statistical Appendix of Marker-Based Approaches . . . 275
C.2.1 Gene Set Enrichment Analysis . . . 275
C.2.2 Hypergeometric Distribution . . . 276
C.2.3 Limitations of Marker-Based Approaches . . . 276
C.3 Statistical Appendix of Reference-Free Approaches . . . 277
C.4 Appendix of Cellular Deconvolution Pipeline . . . 278
C.5 Biological Appendix to the Fate of Deconvolution Algorithms . . . 279

D Appendix of Article 4 **281**
D.1 Optimisation and calculus . . . 282
D.1.1 Multivariate distributions and basic algebra properties . . . 282
D.1.2 Matrix calculus . . . 284
D.1.3 First and second-order derivation of constrained DeCovarT . . . 285
D.2 A MCMC Algorithm for the Joint Distribution of Purified Profiles and Ratios . 286
D.2.1 An Introduction to Gibbs and Metropolis Hasting Samplers . . . 287
D.2.2 Pseudo-code Gibbs sampler . . . 288
D.2.3 Derivation of the Acceptance Probability Function . . . 290

E Article 5: Gene clustering applied to primary Sjögren's disease **291**
E.1 Conclusion . . . 313
E.2 Perspectives . . . 313

F Article 6: Network-based repurposing applied to COVID-19 **314**
F.1 Drug repurposing: a brief historical overview 316
F.2 Introduction to the Patrimony initiative 321
F.3 Repurposing applied to severe COVID-19 cases 323
F.4 Conclusion 342
F.5 Perspectives 342

Contents **348**

Glossary **352**

Glossary **352**

Bibliography **364**

Glossary

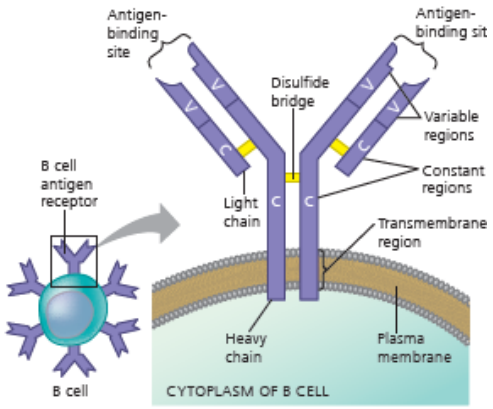
Anti-SSA and anti-SSB Anti-SSA autoantibodies (anti-Sjögren's syndrome A-related autoantibodies) are antinuclear autoantibodies associated with many autoimmune diseases, such as systemic lupus erythematosus (SLE), Sjögren's syndrome (SS) [FC05] [Goë+07], or rheumatoid arthritis (RA). Anti-Ro autoantibodies are often associated with autoantibody anti-La/SSB, displaying similar molecular structure [GE13]. 82

antibodoy Antibodies, also known as *immunoglobulins* (Ig), are Y-shaped proteins composed of four polypeptide chains: two identical *heavy chains* and *light chains*. Each chain is additionally split into (1) a variable region at each end of the y-arm, responsible for binding to the antigen and (2) a *constant region* at the stem of the antibody, which determines the effector functions of the antibody, such as activating complement, recruiting other immune cells or neutralising toxins (left Panel).

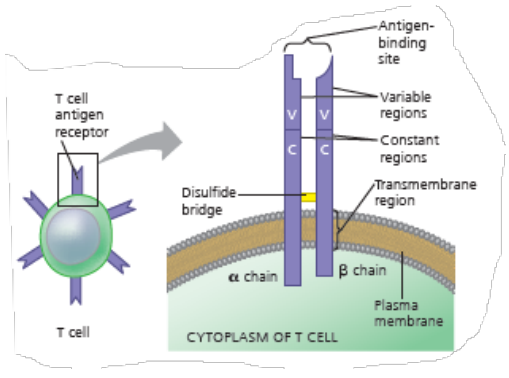
On the other hand, the structure of the antigen receptor of T cells slightly differs, with only one antigen-binding site and only two different polypeptide chains (α and β). However, akin to the antibody structure, its base is a constant region that anchors the molecule in the cell's plasma membrane, while the outer tip of the molecule is a variable region that gives the specificity of the epitope-receptor bound (right Panel).

B cells can express five different *classes* of immunoglobulin (IgA, IgD, IgE, IgG, and IgM):

- All B cells display the same antigen receptor, known as IgD, at its surface.
- **IgM antibodies** are the first deployed on the battlefield and act as an early activator of the innate immune system. They are composed of five antibodies merged together at their hips, speeding up additionally the activation of the complement system.



(a) The structure of a B cell antigen receptor



(b) The structure of a T cell antigen receptor

- **IgG antibodies** are the most specialised: some are involved in *passive immunity*, passing trough mother activated antibodies to the fetus, while some down regulate the activity of the innate system, preventing chronic inflammation.
- **IgA antibodies** It mostly intervenes in the multiple openings of the human body, acting as a guardian of the mucous tissues: respiratory and digestive tract, sexual organs .... IgA antibodies generally work by pairs, preventing access to their constant regions, and attack multiple targets by clumping them together. Additionally, IgA antibodies are transmitted when mothers are breastfeeding their babies, providing them simultaneously with the breast milk. These antibodies then saturate the gut of the newborn and enforce a well-balanced microbiota.
- **IgE antibodies** are involved in allergic reactions, overreacting to innocuous foreign substances, from the pollen of plants to peanuts or seafood. Originally, they used to target multicellular enemies, such as parasitic worms. But the strong development of prophylactic measures in developed countries since early 1900's deprived them from their original cumbersome enemies, and they just in turn found new irrelevant targets, a theory detailed in [Det21, Chapter 39] The distinctive structure of an antibody is what imparts upon it its singular capacity for identifying with an high specificity foreign molecules, acting as a lock-and-key system.

The antibody does not bind to the entire foreign particle, but rather to small peptide fragments commonly known as *epitopes*. B cells and circulating antibodies can directly bind to epitopes present in the extracellular medium, such as in the blood or lymphoid system or to antigens protruding from the surface of pathogens.
Conversely, T cells exclusively identify antigens that are exhibited on the surface of host cells via the major histocompatibility complex (MHC). 349

B-cell B cells, also known as B lymphocytes, are usually classified in circulating whole blood samples into the three major cell subtypes:

1. **Naïve B cells** are fully differentiated, but have not yet encountered any antigen. They circulate in the bloodstream system, constantly scanning for foreign invaders.

   Plasma Cells, aka Effector B Cells, are the terminally differentiated form of B cells. They become activated (and so not anymore naive) when they encounter an antigen matching their B cell receptor. They are the massive antibody producing plants of the human organism (see details in Section 2.1.2). Accordingly, their primal function is to neutralise pathogens by tagging them for destruction and activating the complement system.

2. The final stage corresponds to **memory B cells**, long-lived cells that "remember" previous encounters with pathogens and are thus crucial for the enhanced secondary immune response (see Section 2.1.2). Notably, if the same pathogen infects the body again, memory B cells can quickly differentiate into plasma cells and provide faster and more effective protection.

3. Other residual B cells subsets have been identified in early differentiation stage, but they are not usually found in whole blood, such as B-1 cells and T-Independent B Cells, found primarily in body cavities and involved in the early defence against bacterial infections (the latter is even able to stimulate B cell activation without the help of T cells and primarily produce IgM antibodies) and are critical for early responses to certain pathogens.

Follicular B Cells are typically found in the the spleen and involved in the production of high-affinity antibodoy. Regulatory B Cells play an immuno-suppressive role by regulating the immune responses through tolerance mechanisms.

We should note however that B cell subtypes and functions can overlap, since immunologists and pathologists tend to use a distinct terminology approach, the former focusing on clearly identified protein markers while the latter classified them based on their localisation in the micro-anatomical compartments and differentiation status ([SD03] and [PF17]). 159, 353

cell markers Cells markers, expressed either on the cell surface or intracellular, (proteins, lipids, glycosylation, ...) can be used to set apart unique cell types 16

cell-mediated In the *cell-mediated immune response*, another subset of T cells, the TCD 8, proliferate and differentiate into cytotoxic T cells. Initiation of the cell-mediated response first requires activation signals from Th Helper, while the identification of the cells to eliminate relies on the detection of an antigen fragment displayed by the class I MHC of infected cells. In a process similar to NK cells (2.1.1), destruction of host cells involves the release of perforines that puncture the membrane structure (generate "pores"), leading to cell lysis, and granzymes that cleave essential proteins, leading to cell apoptosis. Cytotoxic T cells are also equipped with an accessory protein, the CD8 marker, that guaranties strong binding all along the destructive process [Cam+20, Figure 22, Chapter 43]. Even though not directly targeting virus, depriving the pathogens from potential hosts reduces practically the magnitude and the propagation of the infection. 24

cellular communication Cellular communication enumerates the process by which cells coordinate their activities and functions. This communication can be either direct (e.g. through gap junctions) or indirect (e.g. through the release of signalling molecules).
Generally, the intermediate signalling molecules are special proteins named *hormones* (from the Greek "horman", to excite) that circulate throughout the body. The communication between cells is further classified with respect to two criteria: the type of secreting cell and the distance between the signal and its target:

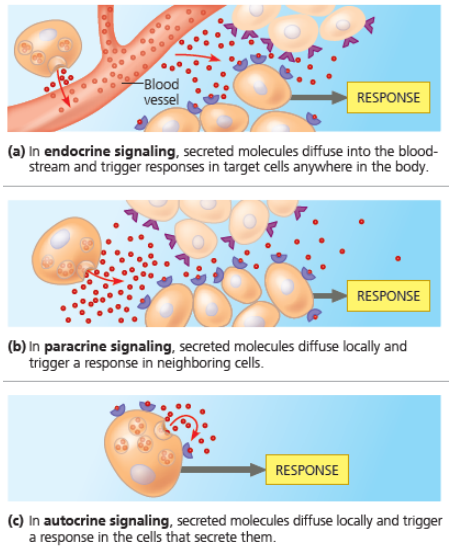


Figure F.4: Categories of intercellular communication. This figure is reproduced from [Cam+20, Fig. 45.2, p. 1000].

5
Contigs Contig are contiguous DNA sequences, assembled from overlapping DNA reads through computational algorithms, and issued from the same origally fragmented genome. One of the main advantages of contigs is that they allow biologists to study specific regions of the genome, even when the complete genome sequence is not available, see also glossary entry Genomic scaffolding. 11, 12, 14, 351

endotype An endotype is an intermediate compartment between the genotype and phenotype scale, describing a disease condition. It is characterised by genetic mutations, activation of specific pathways and the chronological phases of the disease. Endotypes have notably been identified in chronic inflammatory conditions and Alzheimer's disease ([Cal+22] and [De +17]). Understanding endotypes is crucial for the development of personalised medicine, as it enables to tailor targeted treatments to subgroup of patients within the same disease that share the same underlying pathophysiological mechanisms of the disease. 60
epithelium The epithelium is one of the four types of tissues making up the organs of the body, along with the connective (support function), muscular and nervous components. Similarly to the frontiers of a country, they have a protective role against potential intruders (mucous tissues, skin) and an exchange role (transfer of nutrients to the blood in the digestive tract and transfer of oxygen while flushing away carbon dioxide in the respiratory tract). Finally, as main component of glands, they play a key role in maintaining the homoeostasis, releasing hormones in the blood system. 22
eukaryotic Eukaryotic organisms exibit more complex cells (compartmented nucleus that houses their genetic material, highly-specialised organoids, ...), and encompass the animals, plants, fungi, and protists realms. They are defined with respect to prokaryotic organisms, such as bacteria and archea, which lack membrane-bound true nucleus and organelles. 12

FASTA FASTA files gather a collection of sequences (a string of characters, such as A, U, G, C for RNA, that can span multiple lines and is not case-sensitive), each uniquely identified by a *header line* (starting with symbol >, this line provides the name, source or any relevant context related to the sequence). FASTA files are easily readable by both humans and computational tools. 10

Genomic scaffolding Genomic scaffolding is a crucial intermediate step in genome assembly, designed to connect individual Contigss into larger and more complete genome sequences. To do so, the relative order, distance (with potential uncovered gaps) orientation of contigs must be addressed.
Genomic scaffolding is typically achieved by providing to automated aligner algorithms either "mate-pair" (come from non-adjacent regions) or "paired-end" (sequences from opposite ends of a DNA fragment) sequences. The resulting alignment is a *scaffold*, namely a sequence of connected contigs with estimated gap sizes in-between.

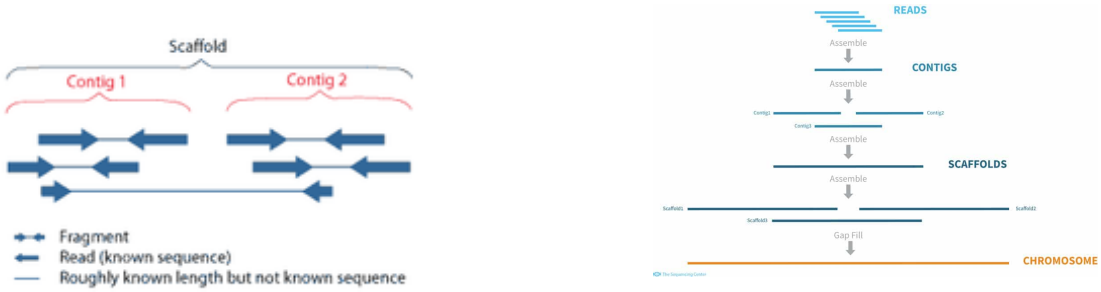


Figure F.5: De novo Assembly Process:. Overlapping reads from paired-end sequencing form *contigs*; overlapping contigs combined with gaps of known length form *scaffolds*; sets of scaffolds are joined into a single *chromosome*. Subfigures a) and b) are reproduced from [Cal11].

11, 14, 350

hematopoiesis The general process describing the generation and the renewal of the blood cells. 23

Heteroskedasticity Heteroskedasticity in RNA-Seq data refers to the phenomenon where the variance of gene expression varies within and between samples. We now detail the reasons, ordered by categories, of the common observation that genes with low read counts exhibit, on average, a higher Signal to Noise Ratio (SNR) ratio, compared to highly expressed genes (report to [MJ18] and [Che+11] for details):

- The major reason proceeds from pure statistical considerations. Indeed,, RNA-Seq data is inherently counts, which are usually modelled by Poisson or negative binomial distributions. For both of them, the variance is linearly correlated to the mean, and even equal for Poisson:

$$y_g \sim \text{Poisson}(\mu_g) \longrightarrow \mathbb{V}\text{ar}\left[y_g\right] = \mu_g \tag{F.1}$$

  . Consequently, the standard deviation of the gene expression, modelled by such distributions, tends to evolve relatively faster than the mean for smaller values.

  This related consequence is further highlighted by SNR formula:

$$\text{SNR}(y_g) = \frac{\mu_g}{\sqrt{\mathbb{V}\text{ar}\left[y_g\right]}} \tag{F.2}$$

  with $\sigma_g = \sqrt{\mathbb{V}\text{ar}\left[y_g\right]}$ the standard deviation of gene expression. , and is used to measure how strong the true signal is relative to the background noise. Coupling the SNR formula with the variance of a Poisson's model mathematically implies that lowly expressed genes exhibit a lower SNR, in other words display a signal closer in magnitude to the noise.

- **Technical noise** also contribute to the higher variability of lowly expressed genes, indeed, since they are close to the detection limit of the sequencing technology, even minor fluctuations or noise in the measurement process can strongly impact the quantification of observed counts.

- Ultimately, mechanistic biological factors that intervene on the regulation of the gene expression tend to reduce the variability observed for highly-expressed genes. Indeed, they are often involved in key cellular processes and are thus more tightly regulated through

robust feedback, leading to more stable expression patterns across samples. Among them, Housekeeping genes genes are continuously expressed in all cells of an organism since they play fundamental roles in maintaining the basic cellular functions necessary for the survival of all cells.

To summarise, genes with high expression levels tend to exhibit lower SNR in RNA-Seq compared to lowly expressed genes. 87, 166

Housekeeping genes Housekeeping genes, also known as constitutive genes, are continuously expressed in all cells, regardless of their type, developmental stage, or environmental conditions. We qualify such expression as Ubiquitous, since present in all tissues and cell types. This broad expression pattern distinguishes them from tissue-specific or condition-specific genes, which are only active in specific situations or cell types.
They are characterised by a Constant and Stable Expression, they are usually involved in fundamental cellular functions, such as energy production (e.g., genes involved in glycolysis and oxidative phosphorylation), structural maintenance (e.g., genes encoding cytoskeletal proteins), and cellular metabolism (e.g., genes involved in protein synthesis and degradation). Some housekeeping genes encode proteins that are involved in regulating key cellular processes. Their consistent expression makes them relevant in gene expression studies, allowing researchers to normalise data and compare the expression of other genes ([EL13]). 169, 352

humoral The humoral immune response involves the release of antibodies by plasma cells, that in turn promote bacteria elimination by facilitating phagocytosis (act as markers of foreign cells) and promoting the complement response. The released antibodies do not directly kill pathogens, but by targeting circulating toxins or epitopes cradled by the class I MHC of infected cells, prevent viral infection or bacterial activity through *neutralisation* or *opsonisation* mechanisms. The activation of B cells requires a two-factor authentication process: the first contact with a circulating antigen antigen in lymph nodes turns a naive B-cell into a mature B cell and starts clonal amplification. In the mean time, the activated lymphocyte engulfs some foreign molecules by endocytosis and displays them through its MHC class II molecules. Ultimately, it is only when the B cell meets an helper T cell that was activated in parallel by a macrophage or a dendritic cell that it can turn into an efficient Plasma Cell, a true antibody-secreting factory. This two-step activation is described in [Cam+20, Figure 19, Chapter 43] and in [Det21, Figures 1 and 2, Chapter 21]. 24

Illumina While the chemical strategy underlying the core Illumina sequencing protocol has long been known, relying on "sequencing-by-synthesis", a process similar to modern Sanger sequencing, the Illumina protocol differs by its enhanced parallel sample-throughput, with the ability to sequence thousands of reads simultaneously thanks to its proprietary clustering and clonal amplification (see Section 1.2.2 and [Amb20]).
This achievement is realised through a combination of physically isolated sequencing lines coupled with the utilization of *multiplexing*. Besides reducing the risk of introducing technical biases, early-stage multiplexing is a cost-saver by curtailing both reagent consumption and labour demands.
Specifically, the Illumina flow cell incorporates eight physically segregated *lanes*, enabling to process up to eight distinct samples, further extended by multiplexing technique which permits the simultaneous sequencing of multiple libraries within the same lane. This is accomplished by uniquely identifying each read to a sample using a specific barcode added during the library preparation, the final operation consisting of assigning unequivocally each read to a sample is so-called the *demultiplexing* procedure.
Due to the nature of the reads generated by Illumina and its prevailing market position, it is often referred to short-read sequencing (SRS) protocol, or Second Next Generation Sequencing,

extending the historical Sanger method. Some key features about Illumina company: they released the first sequencing tool enable to reconstruct from scratch an entire human genome for less than $1000 and foresees an expected $100 commercial target in a near future ([ill23, Chapter 3: Illumina DNA-to-Data NGS Solutions]). 355, 356

inferferon signature There are two main classes of interferons: type I IFNs and type II IFNs [LA18] [Pla05]. While there are many distinct type I IFNs (including IFN-α and IFN-β) binding to the same cell surface receptor, there is only one type II IFN, IFN-γ, which binds to a distinct cell surface receptor. Type I and type II IFNs activate common and distinct STAT (signal transducer and activator of transcription) pathways, hereby playing an important role in regulating the expression of transcriptome and the intensity of the immune response.
Traditionally, type I IFNs are linked to the humoral immune response [Le +01], namely the activation of B cells and the release of antibodies directly targeting foreign invaders, such as viruses or bacteria [ZH14].
On the other hand, type II IFN conducts the cell-mediated response ([Mur+02]), by activating the production of TCD4 and TCD8 cells, which in turn target self cells displaying aberrant phenotypic activity. 82

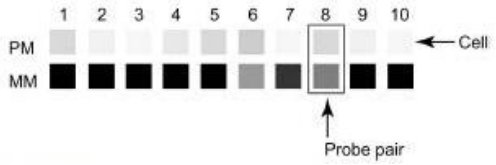
Ion Torrent Proton In this method, first, DNA templates are attached and amplified on beads, which are then loaded into wells on a semiconductor chip. Then, sequential addition of nucleotide to the DNA strand is detected by the induced changes in pH: precisely, when a complementary nucleotide is included, it releases a proton that causes a specific pH change which is detected by ion-sensitive sensors (see [**kchouk_etal17**]). This technology allows for fast sequencing of relatively short read lengths (see [Ngu21], [**kchouk_etal17**] and [GM13]). 356

k-mers K-mers are contiguous sequences of a given size, extracted from longer DNA sequences, and commonly found and shared across many individuals of the same species. Their application in bioinformatics is manifold, including *sequence alignment*, *genome assembly* and *Homology Search* studies. 12

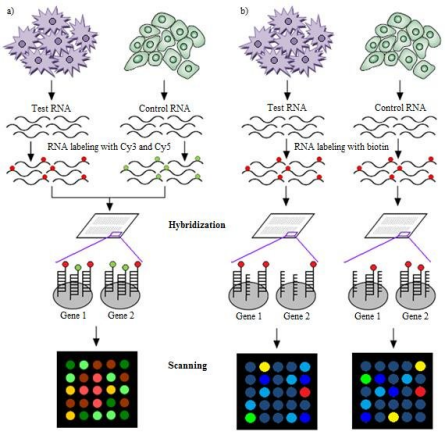
linkage disequilibrium Linkage disequilibrium refers to the genetic observation that alleles at different loci on a chromosome tend to be inherited together more often than expected by chance, implying that the genetic recombination during inheritance was not performed independently. This phenomena is usually triggered by the physical proximity of the linked variants on the chromosome.
Notably, strong linkage disequilibrium may negatively impact bioinformatic analysis. When conducting genome-wide association studies (GWAS), identifying Causal Variants responsible for a particular trait is much more challenging, since identify the one variant driver truly responsible for the observed effect is much harder in a set of highly related genetic variants due to strong LD, leading to numerous spurious associations. Similarly, it negatively impacts the sensibility of Genetic Risk Prediction, since they assume that genetic variants are independently contributing to the risk of a disease. Strong LD can violate this assumption, leading to overestimation or underestimation of the true risk. 320

microarray The quantification protocol is common to most of microarray-based technologies. First, the mRNA samples, called the “targets”, are extracted from the investigated biological sample. Then, the targets hybridise to their complementary probe on the microarray and are labelled using fluorescence or radioactive labels.
Two competitive methods are used to quantify gene expression between two biological conditions, as presented in right panel below. Either the two mRNA preparations are hybridised at the same time, the RNA strands competing for the access to the probes, or each sample is assigned its own array, but with the same dying label (an extensive comparison of the pros and perks of both

approaches is reviewed in [Pat+06], [Kuo+06], [Sev+06] and [Sîr+12]). In both cases, each probe is generally duplicated thousands of time in the same "probe cell" (see right panel, below), called a *spot*, to trap the target mRNA fragment.
High-spatial resolution pictures of the plate are subsequently taken, on which a spot usually matches more than one single pixel. The *segmentation* stage sets apart pixels proceeding from a spot, marked as relevant signal, from background noise. Ultimately, the *quantification* stage sums the individual signals from a spot, and the resulting intensity is used as proxy of the abundance of mRNA. The output is filed in a CEL document.



(a) Probe set structure. A micro-array chip is composed of a large collection of probe oligonucleotides. To enhance detection power and assess variability, the probes are further organised into *probe sets*: a collection of perfect match probes (PM) that precisely match the target sequence paired with mismatch probes (MM) that only differ by a single base mismatch. Hereby, we can assert the specificity of each probe by quantifying the relative binding of the target sequence between the PM and MM probe.



(b) a) Two-channel and b) Single-channel microarray platforms. This infographic (reproduced from [Daw23, Fig. 6]) illustrates the technical protocol differences between single-channel and two-channel microarray technologies.

Figure F.6: A brief overview of micro-array chips to quantify the transcriptome.

8

neutralisation Neutralisation is the process by which antibodies prevent viral infection of a host cell or necrosis due to toxins circulating in blood, by binding to surface proteins, thus preventing the virus from entering body cells [Cam+20, Figure 20a, Chapter 43]. On the other hand, *opsonisation* fends off bacteria invasion, not by stopping the infection, but rather by promoting phagocytosis (Figure 43.20b): the two antigen-binding sites can aggregate foreign substances, easing their engulfment, while the other end-tail acts as a marker for macropahes and neutrophiles [Cam+20, Figure 20b, Chapter 43] 353

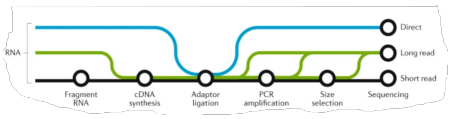
Oxford Nanopore After library preparation and adaptor ligation which includes a subsequent step of attaching motor proteins, individual DNA templates are loaded into a flowcell and dock with nanopores (tiny holes dug into the flow membrane).
The motor protein ensures the good translocation of the RNA strand, namely that it is well-threaded through the nanopore. As each nucleotide passes through, it temporarily blocks the nanopore, causing a specific and noticeable change in the electrical current. Precisely, the speed at which the current is blocked indicates the type of nucleotide, while its duration corresponds to the nucleotide's position within the DNA strand.

Oxford Nanopore sequencing protocol can generate long reads of 1 to 10 kb (kilobases) size, however, it tends to have a higher error rate. By keeping the native RNA structure (no prior conversion into cDNA is required, see key RNA library), it has revolutionized personalised medicine and microbiome studies, allowing the detection of larger structural variants(see [MA23], [LHM21], [**maitra_etal12**] and [LGN16]). 356

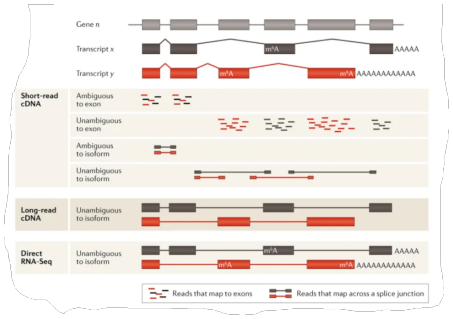
Pacific Biosciences Pacific Bio comprises two main stages. First, thousands of DNA templates are first coupled with DNA polymerase and tethered on the bottom of a nanowell. Then, a miniature camera underneath each well captures in the second step the sequential extension of nucleotides to the DNA templates, by measuring the resulting fluorescent reactions: when a base pairs with the DNA template, its signal's intensity increases. PacBio is tailored for the generation of long, preserved reads (see [Moh+22], [Nak20], [Gro+16] and [RA15] for details and concrete biological applications, and [hae11] for a Youtube tutorial introduction). 356

phagocytosis Phagocytosis, originating from the Greek term "cell eater", is the intricate process through which phagocytes internalise and digest foreign intruders and dead cells. After a first recognition stage, the phagocyte encapsulates the intruder in the phagosome, then fuses within its cytoplasm with the lysosome organelle, whose enzymes cut into several pieces the foreign molecules. 22

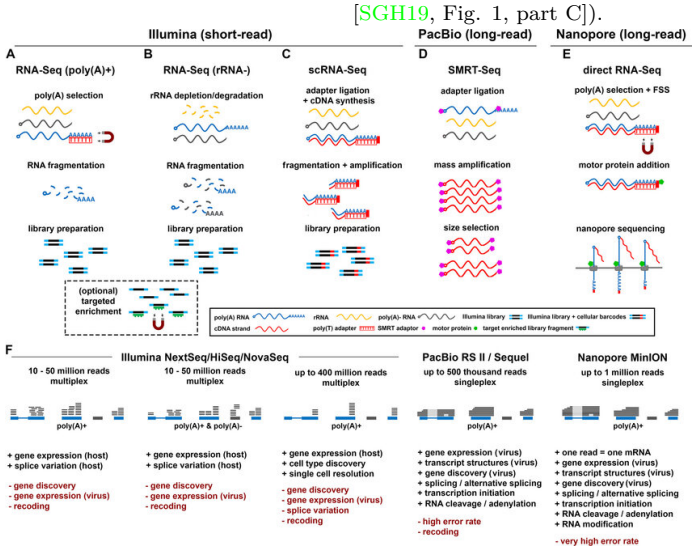
RNA library Briefly, the library, in the Bioinformatics fields, refers to the total collection of reads generated by a sequencing platform.
The protocol to generate this collection of sequences largely differs depending on the sequencing platform and the size and nature of reads generated:



(a) Ecosystem of library preparation. Presented is a summary of library preparation techniques for the most common RNA-seq methods, classified with respect to the protocol preparation and the resulting sequencing opening frame into short-read (black line), long-read c(omplementary)DNA (green) or direct long-read (blue). Reproduced from [SGH19, Fig. 1, part A]).



(b) Comparison of short-read, long-read and direct RNA-seq analysis: More complex and detailed can be captured as we move from short-read cDNA sequencing to long-read methods that can directly sequence isoforms. Indeed, detecting isoforms with short-read cDNA sequencing is hampered by unclear mapped reads, especially when exons are shared between isoforms. Long-read cDNA methods, by returning the complete isoform in a single stage, largely alleviate these issues. Reproduced from [SGH19, Fig. 1, part C]).

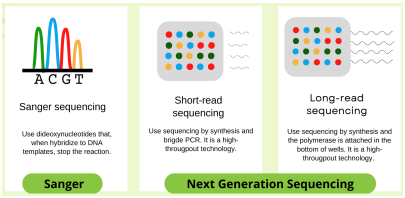


(c) Comprehensive comparison of RNA sequencing applications, reproduced from [DMW19, Fig. 1].

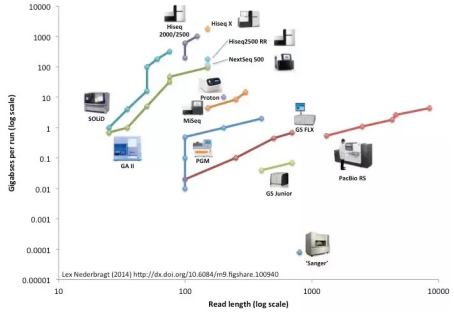
Figure F.7: Pros and cons of short-read (Illumina) or long-read-based (PacBio and Nanopore) sequencing platforms.

The biological application of each sequencing platform will depend notably on the quality, the number and the averaged size of reads within ([SGH19]). 9

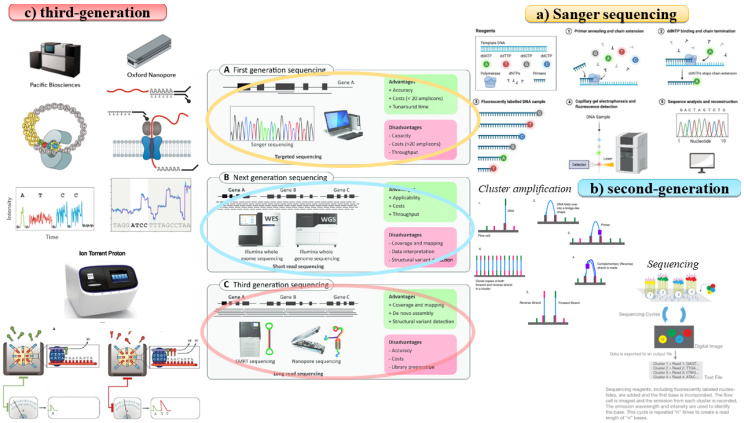
RNASeq Sequencer platform All three sequencing platform generation are briefly reviewed in Figure below:



(a) Sequencing techniques. We generally classify sequencing techniques in the *Sanger sequencing* with low throughput analysis and *Next-Generation Sequencing* with much higher throughput analysis, cheaper process and often increased sequencing quality. It is further possible to set apart NGS methods focusing on short read from those sequencing long reads (reproduced from [Gal23, Fig. 1]).



(b) Comparison of Sequencing methods. This scatter plot (reproduced from [DrS17, Fig. 1]) illustrates the relationship between the average length of processed sequences (representing the reading frame) on the x-axis and the total genome size that can be simultaneously analysed on the y-axis.



(c) RNASeq generation: We illustrate key technical concepts underlying three distinct RNASeq technologies: briefly, the first generation refers to the Sanger sequencing, the second generation to the Illumina sequencing, reviewed in details in Section 1.2.2 and the "third generation" refers to a range of recent sequencing methods that share the same objective, namely sequencing long reads: Pacific Biosciences, Ion Torrent Proton and Oxford Nanopore. The central figure is reproduced from [dBru+21, Fig .3], the Sanger workflow from [Shr22, Fig .1], the Illumina principle from [DML23] and the third generation from [SGH19, Fig. 1].

Figure F.8: An overview of sequencing techniques, by generation..

It is further common to separate ngs technologies according to the average length of the reads output by the method (short-reads or long-read sequencing, [Cle22]). 9

Sanger sequencing Sanger sequencing, also known as the *chain termination method*, was first developed in 1977. Briefly, it relies on the joint presence of modified nucleotides called dideoxynucleotides (ddNTPs) with a specific DNA polymerase able to resist to high temperatures. The ddNTPs lack a 3' hydroxyl group that block the transcription of the DNA strand.
By ranking the reads generated by increasing size and identify for each of them the terminal nucleotide, namely identified by the ddNTP incorporated in the original DNA sequence, the sequence of the target DNA can be determined.
Historically, classifying the reads require gel electrophoresis and manual annotation, but this task has since been automated through an analyser device. In addition, modern Sanger sequencing

alleviates the integration of ddNTPs into the original genome, since the four different reactions, each corresponding to the addition of one of the four dideoxynucleotides, can be performed in one single reaction (see also video [**quickbiochemistrybasics19**], Wikipedia page [**Sanger23**] and first original mention to this technique in [**sanger_etal77**]).
Interestingly, while Sanger sequencing tends to lag behind other NGS technologies, by generating longer DNA reads and upholding a minimal error rate (base calling accuracy close to 99.99%, [**shendure_ji08**]), it prevails other contemporary NGS methods, particularly in public health endeavours like decoding the spike protein of SARS-CoV-2 ([**daniels_etal21**], [Lam+12]). 356
serendipity Serendipity refers to fortunate discoveries performed by chance or accident. 318

Bibliography

[Mac67] J. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, Jan. 1, 1967. URL: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

[Cri70] Francis Crick. "Central Dogma of Molecular Biology". In: *Nature* (Aug. 1970). ISSN: 1476-4687. DOI: 10.1038/227561a0. URL: https://www.nature.com/articles/227561a0.

[HK70] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* (Feb. 1, 1970). ISSN: 0040-1706. DOI: 10.1080/00401706.1970.10488634. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* (Sept. 1977). ISSN: 00359246. DOI: 10.1111/j.2517-6161.1977.tb01600.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1977.tb01600.x.

[Bis81] N. P. Bishun. "The Role of Cytogenetic Tests in Detection and Prevention of Cancer". In: *Journal of Surgical Oncology* (1981). ISSN: 1096-9098. DOI: 10.1002/jso.2930180311. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jso.2930180311.

[Rou85] Peter Rousseeuw. "Multivariate Estimation With High Breakdown Point". In: *Mathematical Statistics and Applications Vol. B* (Jan. 1, 1985). ISSN: 978-94-010-8901-2. DOI: 10.1007/978-94-009-5438-0_20.

[Alt+90] Stephen F. Altschul et al. "Basic Local Alignment Search Tool". In: *Journal of Molecular Biology* (Oct. 5, 1990). ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

[Lok+90] Michael R. Loken et al. "Establishing Optimal Lymphocyte Gates for Immunophenotyping by Flow Cytometry". In: *Cytometry* (1990). ISSN: 1097-0320. DOI: 10.1002/cyto.990110402. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.990110402.

[MR93] XIAO-LI MENG and Donald Rubin. “Maximum Likelihood Estimation via the ECM Algorithm: A General Framework”. In: *Biometrika* (June 1, 1993). DOI: 10.1093/biomet/80.2.267.

[Nom+94] Tsutomu. Nomizu et al. “Determination of Calcium Content in Individual Biological Cells by Inductively Coupled Plasma Atomic Emission Spectrometry”. In: *Analytical Chemistry* (Oct. 1, 1994). ISSN: 0003-2700. DOI: 10.1021/ac00091a004. URL: https://doi.org/10.1021/ac00091a004.

[Thu+96] M. Thurnher et al. “Human Renal-Cell Carcinoma Tissue Contains Dendritic Cells”. In: *International Journal of Cancer* (Sept. 27, 1996). ISSN: 0020-7136. DOI: 10.1002/(SICI)1097-0215(19960927)68:1<1::AID-IJC1>3.0.CO;2-V.

[Tib96] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996). ISSN: 0035-9246. URL: https://www.jstor.org/stable/2346178.

[Col+98] Francis S. Collins et al. “New Goals for the U.S. Human Genome Project: 1998-2003”. In: *Science* (Oct. 23, 1998). DOI: 10.1126/science.282.5389.682. URL: https://www.science.org/doi/full/10.1126/science.282.5389.682.

[Tak+99] Osamu Takeuchi et al. “Differential Roles of TLR2 and TLR4 in Recognition of Gram-Negative and Gram-Positive Bacterial Cell Wall Components”. In: *Immunity* (Oct. 1999). ISSN: 10747613. DOI: 10.1016/S1074-7613(00)80119-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S1074761300801193.

[Wan+00] Xiao-Bo Wang et al. “Cell Separation by Dielectrophoretic Field-flow-fractionation”. In: *Analytical chemistry* (Feb. 15, 2000). ISSN: 0003-2700. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2726255/.

[Gew+01] Andrew T. Gewirtz et al. “Cutting Edge: Bacterial Flagellin Activates Basolaterally Expressed TLR5 to Induce Epithelial Proinflammatory Gene Expression1”. In: *The Journal of Immunology* (Aug. 15, 2001). ISSN: 0022-1767. DOI: 10.4049/jimmunol.167.4.1882. URL: https://doi.org/10.4049/jimmunol.167.4.1882.

[Le +01] Agnes Le Bon et al. “Type I Interferons Potently Enhance Humoral Immunity and Can Promote Isotype Switching by Stimulating Dendritic Cells In Vivo”. In: *Immunity* (Apr. 1, 2001). ISSN: 1074-7613. DOI: 10.1016/S1074-7613(01)00126-1. URL: https://www.sciencedirect.com/science/article/pii/S1074761301001261.

[TD01] Linus Torvalds and David Diamond. *Just for Fun: The Story of an Accidental Revolutionary*. HarperBusiness, 2001. 262 pp. ISBN: 978-1-58799-151-6.

[BSA02] Craig H. Bassing, Wojciech Swat, and Frederick W. Alt. “The Mechanism and Regulation of Chromosomal V(D)J Recombination”. In: *Cell* (Apr. 19, 2002). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/S0092-8674(02)00675-X. URL: https://www.cell.com/cell/abstract/S0092-8674(02)00675-X.

[CJC02] X. Chen, Z. L. Ji, and Y. Z. Chen. “TTD: Therapeutic Target Database”. In: *Nucleic Acids Research* (Jan. 1, 2002). ISSN: 0305-1048. DOI: 10.1093/nar/30.1.412. URL: https://doi.org/10.1093/nar/30.1.412.

[Mur+02] Paul D. Murray et al. “Cellular Sources and Targets of IFN-γ-Mediated Protection against Viral Demyelination and Neurological Deficits”. In: *European journal of immunology* (Mar. 2002). ISSN: 0014-2980. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319413/.

[FYA03] Tracey C. Fleischer, Ui Jeong Yun, and Donald E. Ayer. "Identification and Characterization of Three New Components of the mSin3A Corepressor Complex". In: *Molecular and Cellular Biology* (May 2003). ISSN: 0270-7306. DOI: 10.1128/MCB.23.10.3456-3467.2003.

[Lam+03] Justin Lamb et al. "A Mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer". In: *Cell* (Aug. 8, 2003). ISSN: 0092-8674. DOI: 10.1016/s0092-8674(03)00570-1.

[SD03] Xavier Sagaert and Christiane De Wolf-Peeters. "Classification of B-cells According to Their Differentiation Status, Their Micro-Anatomical Localisation and Their Developmental Lineage". In: *Immunology Letters* (Dec. 15, 2003). ISSN: 0165-2478. DOI: 10.1016/j.imlet.2003.09.007. URL: https://www.sciencedirect.com/science/article/pii/S0165247803002268.

[AT04a] Shizuo Akira and Kiyoshi Takeda. "Toll-like Receptor Signalling". In: *Nature Reviews Immunology* (July 2004). ISSN: 1474-1741. DOI: 10.1038/nri1391. URL: https://www.nature.com/articles/nri1391.

[AT04b] Ted T. Ashburn and Karl B. Thor. "Drug Repositioning: Identifying and Developing New Uses for Existing Drugs". In: *Nature Reviews. Drug Discovery* (Aug. 2004). ISSN: 1474-1776. DOI: 10.1038/nrd1468.

[Bir+04] Ewan Birney et al. "An Overview of Ensembl". In: *Genome Research* (Jan. 5, 2004). ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.1860604. URL: https://genome.cshlp.org/content/14/5/925.

[CM04] Vladimir Cherkassky and Yunqian Ma. "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression". In: *Neural Networks: The Official Journal of the International Neural Network Society* (Jan. 2004). ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00169-2.

[Gau+04] Laurent Gautier et al. "Affy—Analysis of Affymetrix GeneChip Data at the Probe Level". In: *Bioinformatics* (Feb. 12, 2004). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg405. URL: https://doi.org/10.1093/bioinformatics/btg405.

[IBB04] Jan Ihmels, Sven Bergmann, and Naama Barkai. "Defining Transcription Modules Using Large-Scale Gene Expression Data". In: *Bioinformatics* (Sept. 1, 2004). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth166. URL: https://doi.org/10.1093/bioinformatics/bth166.

[RE04] Yvonne Reiss and Britta Engelhardt. "FACS Analysis of Endothelial Cells". In: *Methods in Endothelial Cell Biology*. Ed. by Hellmut G. Augustin. Springer Lab Manuals. Berlin, Heidelberg: Springer, 2004. ISBN: 978-3-642-18725-4. DOI: 10.1007/978-3-642-18725-4_15. URL: https://doi.org/10.1007/978-3-642-18725-4_15.

[AH05] Ahsan Abdullah and Amir Hussain. "Biclustering Gene Expression Data in the Presence of Noise". In: *Artificial Neural Networks: Biological Inspirations – ICANN 2005*. Ed. by Włodzisław Duch et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005. ISBN: 978-3-540-28754-4. DOI: 10.1007/11550822_95.

[Dur+05] Steffen Durinck et al. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". In: *Bioinformatics* (2005).

[FC05] F. Franceschini and I. Cavazzana. "Anti-Ro/SSA and La/SSB Antibodies". In: *Autoimmunity* (Feb. 1, 2005). ISSN: 0891-6934. DOI: 10.1080/08916930400022954. URL: https://doi.org/10.1080/08916930400022954.

[GKT05] Olga Georgieva, Frank Klawonn, and Katharina Tschumitschew. “Noise Clustering via Dynamic Data Assigning Assessment.” In: Jan. 1, 2005.

[Huh+05] Dongeun Huh et al. “Microfluidics for Flow Cytometric Analysis of Cells and Particles”. In: *Physiological Measurement* (Feb. 2005). ISSN: 0967-3334. DOI: 10.1088/0967-3334/26/3/R02. URL: https://dx.doi.org/10.1088/0967-3334/26/3/R02.

[Pla05] Leonidas C. Platanias. “Mechanisms of Type-I- and Type-II-interferon-mediated Signalling”. In: *Nature Reviews Immunology* (May 2005). ISSN: 1474-1741. DOI: 10.1038/nri1604. URL: https://www.nature.com/articles/nri1604.

[SC05] Michael R. Speicher and Nigel P. Carter. “The New Cytogenetics: Blurring the Boundaries with Molecular Biology”. In: *Nature Reviews Genetics* (Oct. 2005). ISSN: 1471-0064. DOI: 10.1038/nrg1692. URL: https://www.nature.com/articles/nrg1692.

[Sub+05] Aravind Subramanian et al. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles”. In: *Proceedings of the National Academy of Sciences of the United States of America* (Oct. 25, 2005). ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

[ZH05] Hui Zou and Trevor Hastie. “Regularization and Variable Selection via the Elastic Net”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (2005). ISSN: 1369-7412. URL: https://www.jstor.org/stable/3647580.

[Kuo+06] Winston Patrick Kuo et al. “A Sequence-Oriented Comparison of Gene Expression Measurements across Different Hybridization-Based Technologies”. In: *Nature Biotechnology* (July 2006). ISSN: 1546-1696. DOI: 10.1038/nbt1217. URL: https://www.nature.com/articles/nbt1217.

[Lam+06] Justin Lamb et al. “The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease”. In: *Science* (Sept. 29, 2006). DOI: 10.1126/science.1132939. URL: https://www.science.org/doi/full/10.1126/science.1132939.

[Pat+06] Tucker A. Patterson et al. “Performance Comparison of One-Color and Two-Color Platforms within the Microarray Quality Control (MAQC) Project”. In: *Nature Biotechnology* (Sept. 2006). ISSN: 1546-1696. DOI: 10.1038/nbt1242. URL: https://www.nature.com/articles/nbt1242.

[RV06] PETER J. ROUSSEEUW and KATRIEN VAN DRIESSEN. “Computing LTS Regression for Large Data Sets”. In: *Data Mining and Knowledge Discovery* (Jan. 1, 2006). ISSN: 1573-756X. DOI: 10.1007/s10618-005-0024-4. URL: https://doi.org/10.1007/s10618-005-0024-4.

[Sch+06] Walter Schubert et al. “Analyzing Proteome Topology and Function by Automated Multidimensional Fluorescence Microscopy”. In: *Nature Biotechnology* (Oct. 2006). ISSN: 1087-0156. DOI: 10.1038/nbt1250.

[Sev+06] Marco Severgnini et al. “Strategies for Comparing Gene Expression Profiles from Different Microarray Platforms: Application to a Case-Control Experiment”. In: *Analytical Biochemistry* (June 1, 2006). ISSN: 0003-2697. DOI: 10.1016/j.ab.2006.03.023.

[DM07] Sean Davis and Paul Meltzer. “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and Bioconductor”. In: *Bioinformatics* (2007).

[Ger+07] Mark B. Gerstein et al. “What Is a Gene, Post-ENCODE? History and Updated Definition”. In: *Genome Research* (Jan. 6, 2007). ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.6339607. URL: https://genome.cshlp.org/content/17/6/669.

[Goë+07] V Goëb et al. “Clinical Significance of Autoantibodies Recognizing Sjögren’s Syndrome A (SSA), SSB, Calpastatin and Alpha-Fodrin in Primary Sjögren’s Syndrome”. In: *Clinical and Experimental Immunology* (May 2007). ISSN: 0009-9104. DOI: 10.1111/j.1365-2249.2007.03337.x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1868868/.

[JLR07] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods”. In: *Biostatistics* (Jan. 1, 2007). ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. URL: https://doi.org/10.1093/biostatistics/kxj037.

[Lam07] Justin Lamb. “The Connectivity Map: A New Tool for Biomedical Research”. In: *Nature Reviews. Cancer* (Jan. 2007). ISSN: 1474-175X. DOI: 10.1038/nrc2044.

[Blo+08] Vincent D. Blondel et al. “Fast Unfolding of Communities in Large Networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* (Oct. 2008). ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[Cha+08] Damien Chaussabel et al. “A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus”. In: *Immunity* (July 18, 2008). ISSN: 1097-4180. DOI: 10.1016/j.immuni.2008.05.012.

[MS08] Misako Matsumoto and Tsukasa Seya. “TLR3: Interferon Induction by Double-Stranded RNA Including Poly(I:C)”. In: *Advanced Drug Delivery Reviews*. Toll-like Receptor and Pattern Sensing for Evoking Immune Response (Apr. 29, 2008). ISSN: 0169-409X. DOI: 10.1016/j.addr.2007.11.005. URL: https://www.sciencedirect.com/science/article/pii/S0169409X07003833.

[OC08] Martin L. Olsson and Henrik Clausen. “Modifying the Red Cell Surface: Towards an ABO-universal Blood Supply”. In: *British Journal of Haematology* (2008). ISSN: 1365-2141. DOI: 10.1111/j.1365-2141.2007.06839.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2141.2007.06839.x.

[Wis+08] David S. Wishart et al. “DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets”. In: *Nucleic Acids Research* (Jan. 1, 2008). ISSN: 0305-1048. DOI: 10.1093/nar/gkm958. URL: https://doi.org/10.1093/nar/gkm958.

[Aba+09] Luca Abatangelo et al. “Comparative Study of Gene Set Enrichment Methods”. In: *BMC bioinformatics* (Sept. 2, 2009). ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-275. URL: https://doi.org/10.1186/1471-2105-10-275.

[Ilj+09] Kristiina Iljin et al. “High-Throughput Cell-Based Screening of 4910 Known Drugs and Drug-like Small Molecules Identifies Disulfiram as an Inhibitor of Prostate Cancer Cell Growth”. In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* (Oct. 1, 2009). ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-09-1035.

[Kau+09] Audrey Kauffmann et al. “Importing ArrayExpress datasets into R/Bioconductor”. In: *Bioinformatics* (2009).

[Li+09] Heng Li et al. “The Sequence Alignment/Map Format and SAMtools”. In: *Bioinformatics* (Aug. 15, 2009). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. URL: https://doi.org/10.1093/bioinformatics/btp352.

[LdCL09] Giampaolo Luiz Libralon, André Carlos Ponce de Leon Ferreira de Carvalho, and Ana Carolina Lorena. "Pre-Processing for Noise Detection in Gene Expression Classification Data". In: *Journal of the Brazilian Computer Society* (Mar. 2009). ISSN: 1678-4804. DOI: 10.1007/BF03192573. URL: https://journal-bcs.springeropen.com/articles/10.1007/BF03192573.

[Tan+09] Fuchou Tang et al. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell". In: *Nature Methods* (May 2009). ISSN: 1548-7105. DOI: 10.1038/nmeth.1315. URL: https://www.nature.com/articles/nmeth.1315.

[WGS09] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics". In: *Nature reviews. Genetics* (Jan. 2009). ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/.

[WL09] Brian T. Wilhelm and Josette-Renée Landry. "RNA-Seq—Quantitative Measurement of Expression through Massively Parallel RNA-sequencing". In: *Methods*. Global Approaches to Study Gene Regulation (July 1, 2009). ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2009.03.016. URL: https://www.sciencedirect.com/science/article/pii/S1046202309000632.

[Wri09] Daniel Wright. "Ten Statisticians and Their Impacts for Psychologists". In: *Perspectives on Psychological Science* (Nov. 1, 2009). DOI: 10.1111/j.1745-6924.2009.01167.x.

[AH10] Simon Anders and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data". In: *Genome Biology* (Oct. 27, 2010). ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106. URL: https://doi.org/10.1186/gb-2010-11-10-r106.

[And10] Simon Andrews. *Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*. A quality control tool for high throughput sequence data. 2010. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[HK10] Thomas J. Hardcastle and Krystyna A. Kelly. "baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data". In: *BMC Bioinformatics* (Aug. 10, 2010). ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-422. URL: https://doi.org/10.1186/1471-2105-11-422.

[Jea+10] Marine Jeanmougin et al. "Should We Abandon the T-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies". In: *PLOS ONE* (Sept. 3, 2010). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0012336. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012336.

[MK10] Matthias Meyer and Martin Kircher. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing". In: *Cold Spring Harbor Protocols* (Jan. 6, 2010). ISSN: 1940-3402, 1559-6095. DOI: 10.1101/pdb.prot5448. URL: http://cshprotocols.cshlp.org/content/2010/6/pdb.prot5448.

[RMS10] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data". In: *Bioinformatics* (Jan. 1, 2010). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: https://doi.org/10.1093/bioinformatics/btp616.

[Tra+10] Cole Trapnell et al. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation”. In: *Nature Biotechnology* (May 2010). ISSN: 1546-1696. DOI: 10.1038/nbt.1621. URL: https://www.nature.com/articles/nbt.1621.

[Cal11] The Regents of the University of California. *English: Overlapping Reads from PET Form Contigs; Contigs and Gaps of Known Length Form Scaffolds.* Nov. 28, 2011. URL: https://commons.wikimedia.org/wiki/File:PET_contig_scaffold.png.

[Che+11] Zhongxue Chen et al. “Statistical Methods on Detecting Differentially Expressed Genes for RNA-seq Data”. In: *BMC Systems Biology* (Dec. 23, 2011). ISSN: 1752-0509. DOI: 10.1186/1752-0509-5-S3-S1. URL: https://doi.org/10.1186/1752-0509-5-S3-S1.

[Gra+11] Gregory R. Grant et al. “Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)”. In: *Bioinformatics* (Sept. 15, 2011). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr427. URL: https://doi.org/10.1093/bioinformatics/btr427.

[hae11] haemse, director. *Single Molecule Real Time Sequencing - Pacific Biosciences.* Oct. 21, 2011. URL: https://www.youtube.com/watch?v=v8p4ph2MAvI.

[Hug+11] JP Hughes et al. “Principles of Early Drug Discovery”. In: *British Journal of Pharmacology* (Mar. 2011). ISSN: 0007-1188. DOI: 10.1111/j.1476-5381.2010.01127.x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058157/.

[Kuh+11] Alexandre Kuhn et al. “Population-Specific Expression Analysis (PSEA) Reveals Molecular Changes in Diseased Brain”. In: *Nature Methods* (Nov. 2011). ISSN: 1548-7105. DOI: 10.1038/nmeth.1710. URL: https://www.nature.com/articles/nmeth.1710.

[LYC11] Jae-Hak Lee, Hana Yi, and Jongsik Chun. “rRNASelector: A Computer Program for Selecting Ribosomal RNA Encoding Sequences from Metagenomic and Metatranscriptomic Shotgun Libraries”. In: *The Journal of Microbiology* (Aug. 1, 2011). ISSN: 1976-3794. DOI: 10.1007/s12275-011-1213-z. URL: https://doi.org/10.1007/s12275-011-1213-z.

[MO11] John H. Malone and Brian Oliver. “Microarrays, Deep Sequencing and the True Measure of the Transcriptome”. In: *BMC Biology* (May 31, 2011). ISSN: 1741-7007. DOI: 10.1186/1741-7007-9-34. URL: https://doi.org/10.1186/1741-7007-9-34.

[Mar11] Marcel Martin. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads”. In: *EMBnet.journal* (May 2, 2011). ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200.

[NF11] Marko Nagode and Matija Fajdiga. “The REBMIX Algorithm and the Univariate Finite Mixture Estimation”. In: *Communications in Statistics—Theory and Methods* (Mar. 1, 2011). DOI: 10.1080/03610920903480890.

[Ng+11] Wan-Fai Ng et al. “United Kingdom Primary Sjogren’s Syndrome Registry–a United Effort to Tackle an Orphan Rheumatic Disease”. In: *Rheumatology (Oxford, England)* (Jan. 2011). ISSN: 1462-0332. DOI: 10.1093/rheumatology/keq240.

[SL11] Richard Simard and Pierre L'Ecuyer. "Computing the Two-Sided Kolmogorov-Smirnov Distribution". In: *Journal of Statistical Software* (Mar. 9, 2011). ISSN: 1548-7660. DOI: 10.18637/jss.v039.i11. URL: https://doi.org/10.18637/jss.v039.i11.

[Tar+11] Sonia Tarazona et al. "NOIseq: A RNA-seq Differential Expression Method Robust for Sequencing Depth Biases". In: *EMBnet.journal* (2011). ISSN: 2226-6089. DOI: 10.14806/ej.17.B.265. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/265.

[WB11] Mathew W. Wright and Elspeth A. Bruford. "Naming 'Junk': Human Non-Protein Coding RNA (ncRNA) Gene Nomenclature". In: *Human Genomics* (Jan. 1, 2011). ISSN: 1479-7364. DOI: 10.1186/1479-7364-5-2-90. URL: https://doi.org/10.1186/1479-7364-5-2-90.

[AH12] Simon Anders and Wolfgang Huber. "Differential Expression of RNA-Seq Data at the Gene Level – the DESeq Package". In: (2012).

[dSou12] Natalie de Souza. "The ENCODE Project". In: *Nature Methods* (Nov. 2012). ISSN: 1548-7105. DOI: 10.1038/nmeth.2238. URL: https://www.nature.com/articles/nmeth.2238.

[Eck+12] Joseph R. Ecker et al. "ENCODE Explained". In: *Nature* (Sept. 2012). ISSN: 1476-4687. DOI: 10.1038/489052a. URL: https://www.nature.com/articles/489052a.

[Gue+12] M. Guedj et al. "A Refined Molecular Taxonomy of Breast Cancer". In: *Oncogene* (Mar. 1, 2012). ISSN: 1476-5594. DOI: 10.1038/onc.2011.301.

[Guo12] Xiao-Qiang Guo. "[Discoverer of genetic principle for antibody diversity–Susumu Tonegawa]". In: *Yi Chuan = Hereditas* (Nov. 2012). ISSN: 0253-9772. DOI: 10.3724/sp.j.1005.2012.01501.

[KNT12] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. "SortMeRNA: Fast and Accurate Filtering of Ribosomal RNAs in Metatranscriptomic Data". In: *Bioinformatics* (Dec. 1, 2012). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts611. URL: https://doi.org/10.1093/bioinformatics/bts611.

[Lam+12] Hugo Y. K. Lam et al. "Performance Comparison of Whole-Genome Sequencing Platforms". In: *Nature Biotechnology* (Jan. 2012). ISSN: 1546-1696. DOI: 10.1038/nbt.2065. URL: https://www.nature.com/articles/nbt.2065.

[LS12] Ben Langmead and Steven L. Salzberg. "Fast Gapped-Read Alignment with Bowtie 2". In: *Nature Methods* (Apr. 2012). ISSN: 1548-7105. DOI: 10.1038/nmeth.1923. URL: https://www.nature.com/articles/nmeth.1923.

[MLR12] Jeffrey Charles Miecznikowski, Song Liu, and Xing Ren. "Statistical Modeling for Differential Transcriptome Analysis Using RNA-Seq Technology". In: *Journal of Solid Tumors* (Aug. 21, 2012). ISSN: 1925-4075. DOI: 10.5430/jst.v2n5p33. URL: https://www.sciedu.ca/journal/index.php/jst/article/view/1305.

[San+12] Philippe Sanseau et al. "Use of Genome-Wide Association Studies for Drug Repositioning". In: *Nature Biotechnology* (Apr. 10, 2012). ISSN: 1546-1696. DOI: 10.1038/nbt.2151.

[Sîr+12] Alina Sîrbu et al. "RNA-Seq vs Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering". In: *PLOS ONE* (Dec. 10, 2012). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0050986. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0050986.

[Smi+12] Steven B. Smith et al. “Identification of Common Biological Pathways and Drug Targets across Multiple Respiratory Viruses Based on Human Host Gene Expression Analysis”. In: *PloS One* (2012). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0033174.

[TS12] Todd J. Treangen and Steven L. Salzberg. “Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions”. In: *Nature Reviews Genetics* (Jan. 2012). ISSN: 1471-0064. DOI: 10.1038/nrg3117. URL: https://www.nature.com/articles/nrg3117.

[Wal+12] Jewell N. Walters et al. “Regulation of Human Microsomal Prostaglandin {{E}} Synthase-1 by {{IL-1β}} Requires a Distal Enhancer Element”. In: *The Biochemical Journal* (Apr. 15, 2012). ISSN: 1470-8728. DOI: 10.1042/BJ20111801.

[WWL12] Liguo Wang, Shengqin Wang, and Wei Li. “RSeQC: Quality Control of RNA-seq Experiments”. In: *Bioinformatics* (Aug. 15, 2012). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts356. URL: https://doi.org/10.1093/bioinformatics/bts356.

[Yu+12] Guangchuang Yu et al. “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *OMICS: A Journal of Integrative Biology* (2012). DOI: 10.1089/omi.2011.0118.

[And+13] Simon Anders et al. “Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor”. In: *Nature Protocols* (Sept. 2013). ISSN: 1750-2799. DOI: 10.1038/nprot.2013.099. URL: https://www.nature.com/articles/nprot.2013.099.

[Cao+13] Mengfei Cao et al. “Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks”. In: *PLOS ONE* (Oct. 23, 2013). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076339. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0076339.

[Dil+13] Marie-Agnès Dillies et al. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis”. In: *Briefings in Bioinformatics* (Nov. 1, 2013). ISSN: 1467-5463. DOI: 10.1093/bib/bbs046. URL: https://doi.org/10.1093/bib/bbs046.

[Dob+13] Alexander Dobin et al. “STAR: Ultrafast Universal RNA-seq Aligner”. In: *Bioinformatics* (Jan. 1, 2013). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: https://doi.org/10.1093/bioinformatics/bts635.

[EL13] Eli Eisenberg and Erez Y. Levanon. “Human Housekeeping Genes, Revisited”. In: *Trends in genetics: TIG* (Oct. 2013). ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.05.010. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(13)00089-9.

[Gau13] R. Gaujoux. “An Introduction to Gene Expression Deconvolution and the CellMix Package A Comprehensive Framework for Gene Expression Deconvolution”. In: *undefined* (2013). URL: https://www.semanticscholar.org/paper/An-introduction-to-gene-expression-deconvolution-A-Gaujoux/980b8ac01435d2faa76eb1e0bc94e0a83b27b7a3.

[Ger+13] Michael J. Gerdes et al. “Highly Multiplexed Single-Cell Analysis of Formalin-Fixed, Paraffin-Embedded Cancer Tissue”. In: *Proceedings of the National Academy of Sciences of the United States of America* (July 16, 2013). ISSN: 0027-8424. DOI: 10.1073/pnas.1300136110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718135/.

[GE13] Norbert Gleicher and Uri Elkayam. “Preventing Congenital Neonatal Heart Block in Offspring of Mothers with Anti-SSA/Ro and SSB/La Antibodies: A Review of Published Literature and Registered Clinical Trials”. In: *Autoimmunity Reviews* (Sept. 1, 2013). ISSN: 1568-9972. DOI: 10.1016/j.autrev.2013.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1568997213000773.

[GM13] David Golan and Paul Medvedev. “Using State Machines to Model the Ion Torrent Sequencing Process and to Improve Read Error Rates”. In: *Bioinformatics (Oxford, England)* (July 1, 2013). DOI: 10.1093/bioinformatics/btt212.

[Got+13] Jacques-Eric Gottenberg et al. “Serum Levels of Beta2-Microglobulin and Free Light Chains of Immunoglobulins Are Associated with Systemic Disease Activity in Primary Sjögren’s Syndrome. Data at Enrollment in the Prospective ASSESS Cohort”. In: *PloS One* (2013). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0059868.

[Har+13] Traver Hart et al. “Finding the Active Genes in Deep RNA-seq Gene Expression Studies”. In: *BMC genomics* (Nov. 11, 2013). ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-778.

[Ju+13] Wenjun Ju et al. “Defining Cell-Type Specificity at the Transcriptional Level in Human Disease”. In: *Genome Research* (Nov. 2013). ISSN: 1549-5469. DOI: 10.1101/gr.155697.113.

[Kim+13] Daehwan Kim et al. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions”. In: *Genome Biology* (Apr. 25, 2013). ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-4-r36. URL: https://doi.org/10.1186/gb-2013-14-4-r36.

[Li+13] He Li et al. “Interferons in Sjögren’s Syndrome: Genes, Mechanisms, and Effects”. In: *Frontiers in Immunology* (2013). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2013.00290.

[LX13] Yi Li and Xiaohui Xie. “A Mixture Model for Expression Deconvolution from RNA-seq in Heterogeneous Tissues”. In: *BMC bioinformatics* (2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-S5-S11.

[Maz+13] Elie Maza et al. “Comparison of Normalization Methods for Differential Gene Expression Analysis in RNA-Seq Experiments”. In: *Communicative & Integrative Biology* (Nov. 9, 2013). ISSN: null. DOI: 10.4161/cib.25849. URL: https://doi.org/10.4161/cib.25849.

[Mei+13] Felix Meissner et al. “Direct Proteomic Quantification of the Secretome of Activated Immune Cells”. In: *Science* (Apr. 26, 2013). DOI: 10.1126/science.1232578. URL: https://www.science.org/doi/10.1126/science.1232578.

[Pic+13] Simone Picelli et al. “Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells”. In: *Nature Methods* (Nov. 2013). ISSN: 1548-7105. DOI: 10.1038/nmeth.2639.

[RJN13] Andrea Rau, Florence Jaffrézic, and Grégory Nuel. “Joint Estimation of Causal Effects from Observational and Intervention Gene Expression Data”. In: *BMC Systems Biology* (Oct. 31, 2013). ISSN: 1752-0509. DOI: 10.1186/1752-0509-7-111. URL: https://doi.org/10.1186/1752-0509-7-111.

[Slo+13] Roman Sloutsky et al. “Accounting for Noise When Clustering Biological Data”. In: *Briefings in Bioinformatics* (July 1, 2013). ISSN: 1467-5463. DOI: 10.1093/bib/bbs057. URL: https://doi.org/10.1093/bib/bbs057.

[Suz+13] Yukiya Suzuki et al. “Involvement of Mincle and Syk in the Changes to Innate Immunity after Ischemic Stroke”. In: *Scientific Reports* (Nov. 11, 2013). ISSN: 2045-2322. DOI: 10.1038/srep03177.

[WZ13] Zhong-Yi Wang and Hong-Yu Zhang. “Rational Drug Repositioning by Medical Genetics”. In: *Nature Biotechnology* (Dec. 2013). ISSN: 1546-1696. DOI: 10.1038/nbt.2758. URL: https://www.nature.com/articles/nbt.2758.

[Xu+13] Xiao Xu et al. “Parallel Comparison of Illumina RNA-Seq and Affymetrix Microarray Platforms on Transcriptomic Profiles Generated from 5-Aza-Deoxy-Cytidine Treated HT-29 Colon Cancer Cells and Simulated Datasets”. In: *BMC Bioinformatics* (June 28, 2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-S9-S1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3697991/.

[Zhu+13] Yuelin Zhu et al. “SRAdb: query and use public next-generation sequencing data from within R”. In: *BMC bioinformatics* (2013).

[Ang+14] Michael Angelo et al. “Multiplexed Ion Beam Imaging of Human Breast tumours”. In: *Nature Medicine* (Apr. 2014). ISSN: 1546-170X. DOI: 10.1038/nm.3488. URL: https://www.nature.com/articles/nm.3488.

[BLU14] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data”. In: *Bioinformatics* (Aug. 1, 2014). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170.

[Cla+14] Marcus R. Clark et al. “Orchestrating B Cell Lymphopoiesis through Interplay of IL-7 Receptor and Pre-B Cell Receptor Signalling”. In: *Nature reviews. Immunology* (Feb. 2014). ISSN: 1474-1733. DOI: 10.1038/nri3570. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276135/.

[Don+14] Xiaowen Dong et al. “Clustering on Multi-Layer Graphs via Subspace Analysis on Grassmann Manifolds”. In: *IEEE Transactions on Signal Processing* (Feb. 2014). ISSN: 1941-0476. DOI: 10.1109/TSP.2013.2295553.

[GR14] Georg Gasteiger and Alexander Y. Rudensky. “Interactions between Innate and Adaptive Lymphocytes”. In: *Nature Reviews Immunology* (Sept. 2014). ISSN: 1474-1741. DOI: 10.1038/nri3726. URL: https://www.nature.com/articles/nri3726.

[Gie+14] Charlotte Giesen et al. “Highly Multiplexed Imaging of tumour Tissues with Subcellular Resolution by Mass Cytometry”. In: *Nature Methods* (Apr. 2014). ISSN: 1548-7105. DOI: 10.1038/nmeth.2869. URL: https://www.nature.com/articles/nmeth.2869.

[Isl+14] Saiful Islam et al. “Quantitative Single-Cell RNA-seq with Unique Molecular Identifiers”. In: *Nature Methods* (Feb. 2014). ISSN: 1548-7105. DOI: 10.1038/nmeth.2772. URL: https://www.nature.com/articles/nmeth.2772.

[Law+14] Charity W. Law et al. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-seq Read Counts”. In: *Genome Biology* (Feb. 3, 2014). ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29. URL: https://doi.org/10.1186/gb-2014-15-2-r29.

[Li+14] Jing Li et al. “RNA-Seq Analysis Pipeline Based on Oshell Environment”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Sept. 2014). ISSN: 1557-9964. DOI: 10.1109/TCBB.2014.2321156.

[LSS14] Yang Liao, Gordon K. Smyth, and Wei Shi. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features". In: *Bioinformatics* (Apr. 1, 2014). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656. URL: https://doi.org/10.1093/bioinformatics/btt656.

[LHA14] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2". In: *Genome Biology* (Dec. 5, 2014). ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8.

[Man+14] Kirk J Mantione et al. "Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq". In: *Medical Science Monitor Basic Research* (Aug. 23, 2014). ISSN: 2325-4394. DOI: 10.12659/MSMBR.892101. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152252/.

[MVS14] Juan Manuel Moreno-Moya, Felipe Vilella, and Carlos Simón. "MicroRNA: Key Gene Expression Regulators". In: *Fertility and Sterility* (June 1, 2014). ISSN: 0015-0282. DOI: 10.1016/j.fertnstert.2013.10.042. URL: https://www.sciencedirect.com/science/article/pii/S0015028213032056.

[Mor+14] D. L. Morris et al. "MHC Associations with Clinical and Autoantibody Manifestations in European SLE". In: *Genes and Immunity* (Apr. 2014). ISSN: 1476-5470. DOI: 10.1038/gene.2014.6.

[Mur+14] Susana Murteira et al. "Drug Reformulations and Repositioning in the Pharmaceutical Industry and Their Impact on Market Access: Regulatory Implications". In: *Journal of Market Access & Health Policy* (2014). ISSN: 2001-6689. DOI: 10.3402/jmahp.v2.22813.

[PG14] Menake E. Piyasena and Steven W. Graves. "The Intersection of Flow Cytometry with Microfluidics and Microfabrication". In: *Lab on a Chip* (Feb. 17, 2014). ISSN: 1473-0189. DOI: 10.1039/C3LC51152A. URL: https://pubs.rsc.org/en/content/articlelanding/2014/lc/c3lc51152a.

[Sha+14] Alex K. Shalek et al. "Single-Cell RNA-seq Reveals Dynamic Paracrine Control of Cellular Variation". In: *Nature* (June 2014). ISSN: 1476-4687. DOI: 10.1038/nature13437. URL: https://www.nature.com/articles/nature13437.

[SK14] Jan Stuchlý and Tomáš Kalina. "Analyses of Large Flow Cytometry Datasets". In: *Cytometry Part A* (2014). ISSN: 1552-4930. DOI: 10.1002/cyto.a.22431. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.22431.

[Su+14] Zhenqiang Su et al. "A Comprehensive Assessment of RNA-seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium". In: *Nature Biotechnology* (Sept. 2014). ISSN: 1546-1696. DOI: 10.1038/nbt.2957. URL: https://www.nature.com/articles/nbt.2957.

[Tra+14] Cole Trapnell et al. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells". In: *Nature Biotechnology* (Apr. 2014). ISSN: 1546-1696. DOI: 10.1038/nbt.2859. URL: https://www.nature.com/articles/nbt.2859.

[Wag+14] James R. Wagner et al. "The Relationship between DNA Methylation, Genetic and Expression Inter-Individual Variation in Untransformed Human Fibroblasts". In: *Genome Biology* (Feb. 20, 2014). ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r37. URL: https://doi.org/10.1186/gb-2014-15-2-r37.

[Wan+14] Bo Wang et al. “Similarity Network Fusion for Aggregating Data Types on a Genomic Scale”. In: *Nature Methods* (Mar. 2014). ISSN: 1548-7105. DOI: 10.1038/nmeth.2810. URL: https://www.nature.com/articles/nmeth.2810.

[YYB14] Chun Yu, Weixin Yao, and Xue Bai. “Robust Linear Regression: A Review and Comparison”. Apr. 24, 2014. URL: http://arxiv.org/abs/1404.6274.

[ZH14] A. J. Zajac and L. E. Harrington. “Immune Response to Viruses: Cell-Mediated Immunity”. In: *Reference Module in Biomedical Sciences*. Elsevier, Jan. 1, 2014. ISBN: 978-0-12-801238-3. DOI: 10.1016/B978-0-12-801238-3.02604-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128012383026040.

[Zho+14] Quan Zhou et al. “A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing”. Sept. 5, 2014. URL: http://arxiv.org/abs/1409.1976.

[APH15] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. “HTSeq—a Python Framework to Work with High-Throughput Sequencing Data”. In: *Bioinformatics* (Jan. 15, 2015). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638. URL: https://doi.org/10.1093/bioinformatics/btu638.

[Blo15] RNA-Seq Blog. *RPKM, FPKM and TPM, Clearly Explained | RNA-Seq Blog*. July 22, 2015. URL: https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/.

[Che+15] Kok Hao Chen et al. “RNA Imaging. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells”. In: *Science (New York, N.Y.)* (Apr. 24, 2015). ISSN: 1095-9203. DOI: 10.1126/science.aaa6090.

[FFF15] H. Christina Fan, Glenn K. Fu, and Stephen P. A. Fodor. “Combinatorial Labeling of Single Cells for Gene Expression Cytometry”. In: *Science* (Feb. 6, 2015). DOI: 10.1126/science.1258367. URL: https://www.science.org/doi/full/10.1126/science.1258367.

[GMP15] Vladimir Gligorijevi, No?l Malod-Dognin, and Nata?a Pr?ulj. “Patient-Specific Data Fusion for Cancer Stratification and Personalised Treatment”. In: *Biocomputing 2016*. WORLD SCIENTIFIC, Nov. 18, 2015. ISBN: 978-981-4749-40-4. DOI: 10.1142/9789814749411_0030. URL: https://www.worldscientific.com/doi/abs/10.1142/9789814749411_0030.

[Hub+15] W. Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature Methods* (2015). URL: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

[LC15] Félix López and Eulogio Cruz. “Literature Review about Neo4j Graph Database as a Feasible Alternative for Replacing RDBMS”. In: *Industrial Data* (Dec. 24, 2015). DOI: 10.15381/idata.v18i2.12106.

[Men+15] Jörg Menche et al. “Uncovering Disease-Disease Relationships through the Incomplete Interactome”. In: *Science* (Feb. 20, 2015). DOI: 10.1126/science.1257601. URL: https://www.science.org/doi/10.1126/science.1257601.

[New+15] Aaron Newman et al. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles”. In: *Nature methods* (Mar. 30, 2015). DOI: 10.1038/nmeth.3337.

[Pai+15] Hyojung Paik et al. “Repurpose Terbutaline Sulfate for Amyotrophic Lateral Sclerosis Using Electronic Medical Records”. In: *Scientific Reports* (Mar. 5, 2015). ISSN: 2045-2322. DOI: 10.1038/srep08580. URL: https://www.nature.com/articles/srep08580.

[Per+15] Mihaela Pertea et al. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads”. In: *Nature Biotechnology* (Mar. 2015). ISSN: 1546-1696. DOI: 10.1038/nbt.3122. URL: https://www.nature.com/articles/nbt.3122.

[PY15] Emma Pierson and Christopher Yau. “ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis”. In: *Genome Biology* (Nov. 2, 2015). ISSN: 1474-760X. DOI: 10.1186/s13059-015-0805-z. URL: https://doi.org/10.1186/s13059-015-0805-z.

[RA15] Anthony Rhoads and Kin Fai Au. “PacBio Sequencing and Its Applications”. In: *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments (Oct. 1, 2015). ISSN: 1672-0229. DOI: 10.1016/j.gpb.2015.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1672022915001345.

[Rit+15] Matthew E. Ritchie et al. “Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies”. In: *Nucleic Acids Research* (Apr. 20, 2015). ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. URL: https://doi.org/10.1093/nar/gkv007.

[RRA15] Michael D. Rosenblum, Kelly A. Remedios, and Abul K. Abbas. “Mechanisms of Human Autoimmunity”. In: *The Journal of Clinical Investigation* (June 1, 2015). ISSN: 0021-9738. DOI: 10.1172/JCI78088. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4518692/.

[SHZ15] David Steinberg, Geoffrey Horwitz, and Daphne Zohar. “Building a Business Model in Digital Medicine”. In: *Nature Biotechnology* (Sept. 2015). ISSN: 1546-1696. DOI: 10.1038/nbt.3339.

[Ver+15a] Montse Verdu et al. “Cross-Reactivity of EGFR Mutation-specific Immunohistochemistry Assay in HER2-positive tumours”. In: *Applied Immunohistochemistry & Molecular Morphology* (Sept. 2015). ISSN: 1541-2016. DOI: 10.1097/PAI.0000000000000129. URL: https://journals.lww.com/appliedimmunohist/abstract/2015/09000/cross_reactivity_of_egfr_mutation_specific.4.aspx.

[Ver+15b] Chris Verschoor et al. “An Introduction to Automated Flow Cytometry Gating Tools and Their Implementation”. In: *Frontiers in immunology* (Aug. 18, 2015). DOI: 10.3389/fimmu.2015.00380.

[Zha+15] Wenqian Zhang et al. “Comparison of RNA-seq and Microarray-Based Models for Clinical Endpoint Prediction”. In: *Genome Biology* (June 25, 2015). ISSN: 1465-6906. DOI: 10.1186/s13059-015-0694-1. URL: https://doi.org/10.1186/s13059-015-0694-1.

[Ake+16] Bronwen L. Aken et al. “The Ensembl Gene Annotation System”. In: *Database* (Jan. 1, 2016). ISSN: 1758-0463. DOI: 10.1093/database/baw093. URL: https://doi.org/10.1093/database/baw093.

[Als+16] Abdulraheem Alshareef et al. “The Use of Cellular Thermal Shift Assay (CETSA) to Study Crizotinib Resistance in ALK-expressing Human Cancers”. In: *Scientific Reports* (Sept. 19, 2016). ISSN: 2045-2322. DOI: 10.1038/srep33710.

[Bak16] Monya Baker. “1,500 Scientists Lift the Lid on Reproducibility”. In: *Nature* (May 1, 2016). ISSN: 1476-4687. DOI: 10.1038/533452a. URL: https://www.nature.com/articles/533452a.

[Bra+16] Nicolas L. Bray et al. “Near-Optimal Probabilistic RNA-seq Quantification”. In: *Nature Biotechnology* (May 2016). ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: https://www.nature.com/articles/nbt.3519.

[CCZ16] Caterina Catalanotto, Carlo Cogoni, and Giuseppe Zardo. "MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions". In: *International Journal of Molecular Sciences* (Oct. 2016). ISSN: 1422-0067. DOI: 10.3390/ijms17101712. URL: https://www.mdpi.com/1422-0067/17/10/1712.

[Con+16] Ana Conesa et al. "A Survey of Best Practices for RNA-seq Data Analysis". In: *Genome Biology* (Jan. 26, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. URL: https://doi.org/10.1186/s13059-016-0881-8.

[GT16] Andreas V. Goules and Athanasios G. Tzioufas. "Primary {{Sjögren}}'s Syndrome: {{Clinical}} Phenotypes, Outcome and the Development of Biomarkers". In: *Autoimmunity Reviews* (July 2016). ISSN: 1873-0183. DOI: 10.1016/j.autrev.2016.03.004.

[Gro+16] Paul Groot-Kormelink et al. "High Throughput Random Mutagenesis and Single Molecule Real Time Sequencing of the Muscle Nicotinic Acetylcholine Receptor". In: *PLOS ONE* (Sept. 20, 2016). DOI: 10.1371/journal.pone.0163129.

[Gru+16] Fred Gruber et al. "Bayesian Network Models of Multiple Myeloma: Drivers of High Risk and Durable Response". In: *Blood* (Dec. 2, 2016). ISSN: 0006-4971. DOI: 10.1182/blood.V128.22.4406.4406. URL: https://ashpublications.org/blood/article/128/22/4406/101366/Bayesian-Network-Models-of-Multiple-Myeloma.

[Hod+16] Rachel A. Hodos et al. "In Silico Methods for Drug Repurposing and Pharmacology". In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* (May 2016). ISSN: 1939-005X. DOI: 10.1002/wsbm.1337.

[HV16] Yu-Han Huang and Christopher R. Vakoc. "A Biomarker Harvest from One Thousand Cancer Cell Lines". In: *Cell* (July 28, 2016). ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.07.010.

[Ior+16] Francesco Iorio et al. "A Landscape of Pharmacogenomic Interactions in Cancer". In: *Cell* (July 28, 2016). ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.06.017.

[16] *La guerre du générique : Illustration avec le cas Servier et le Périndopril*. Ecole de Guerre Economique. Dec. 5, 2016. URL: https://www.ege.fr/infoguerre/2016/12/la-guerre-du-generique-illustration-avec-le-cas-servier-et-le-perindopril.

[LT16] Serena Liu and Cole Trapnell. "Single-Cell Transcriptome Sequencing: Recent Advances and Remaining Challenges". In: *F1000Research* (Feb. 17, 2016). ISSN: 2046-1402. DOI: 10.12688/f1000research.7223.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758375/.

[LGN16] Hengyun Lu, Francesca Giordano, and Zemin Ning. "Oxford Nanopore MinION Sequencing and Genome Assembly". In: *Genomics, Proteomics & Bioinformatics*. SI: Big Data and Precision Medicine (Oct. 1, 2016). ISSN: 1672-0229. DOI: 10.1016/j.gpb.2016.05.004. URL: https://www.sciencedirect.com/science/article/pii/S1672022916301309.

[Maz16] Elie Maza. "In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design". In: *Frontiers in Genetics* (2016). ISSN: 1664-8021. URL: https://www.frontiersin.org/articles/10.3389/fgene.2016.00164.

[Nos16] Nicola Nosengo. "Can You Teach Old Drugs New Tricks?" In: *Nature* (June 1, 2016). ISSN: 1476-4687. DOI: 10.1038/534314a. URL: https://www.nature.com/articles/534314a.

[Ron+16] J. Ronholm et al. "Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing". In: *Clinical Microbiology Reviews* (Oct. 2016). ISSN: 0893-8512. DOI: 10.1128/CMR.00056-16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5010751/.

[SB16] Jack W. Scannell and Jim Bosley. "When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis". In: *PLOS ONE* (Feb. 10, 2016). Ed. by Mauro Gasparini. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0147215. URL: https://dx.plos.org/10.1371/journal.pone.0147215.

[Scr+16] Luca Scrucca et al. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". In: *The R journal* (Aug. 2016). ISSN: 2073-4859. DOI: 10.32614/RJ-2016-021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/.

[Stå+16] Patrik L. Ståhl et al. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics". In: *Science (New York, N.Y.)* (July 1, 2016). ISSN: 1095-9203. DOI: 10.1126/science.aaf2403.

[Var+16] Hugo Varet et al. "SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data". In: *PLOS ONE* (June 9, 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0157022. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157022.

[Wil+16] Mark D. Wilkinson et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* (Mar. 15, 2016). ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://www.nature.com/articles/sdata201618.

[BW17] Alan N. Baer and Brian Walitt. "Sjögren Syndrome and Other Causes of Sicca in the Older Adult". In: *Clinics in geriatric medicine* (Feb. 2017). ISSN: 0749-0690. DOI: 10.1016/j.cger.2016.08.007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5125547/.

[CDL17] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. "RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool". In: *PLOS ONE* (Dec. 21, 2017). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0190152. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190152.

[De +17] Glynnis De Greve et al. "Endotype-Driven Treatment in Chronic Upper Airway Diseases". In: *Clinical and Translational Allergy* (July 12, 2017). ISSN: 2045-7022. DOI: 10.1186/s13601-017-0157-8. URL: https://doi.org/10.1186/s13601-017-0157-8.

[DrS17] Dr.Samanthi. *Difference Between NGS and Sanger Sequencing*. Compare the Difference Between Similar Terms. Feb. 20, 2017. URL: https://www.differencebetween.com/difference-between-ngs-and-vs-sanger-sequencing/.

[Eve+17] Celine Everaert et al. "Benchmarking of RNA-sequencing Analysis Workflows Using Whole-Transcriptome RT-qPCR Expression Data". In: *Scientific Reports* (May 8, 2017). ISSN: 2045-2322. DOI: 10.1038/s41598-017-01617-3. URL: https://www.nature.com/articles/s41598-017-01617-3.

[GRT17a] Daniel Greene, Sylvia Richardson, and Ernest Turro. "ontologyX: A Suite of R Packages for Working with Ontological Data". In: *Bioinformatics* (Apr. 1, 2017). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw763. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386138/.

[GRT17b] Daniel Greene, Sylvia Richardson, and Ernest Turro. "ontologyX: a suite of R packages for working with ontological data". In: *Bioinformatics* (2017).

[Guc17] Xavier Guchet. "Médecine personnalisée versus médecine de la personne : une fausse alternative". In: *Lato Sensu: Revue de la Société de philosophie des sciences* (2017). ISSN: 2295-8029. DOI: 10.20416/lsrsps.v4i2.813. URL: https://ojs.uclouvain.be/index.php/latosensu/article/view/3343.

[Kos+17] Gautier Koscielny et al. "Open Targets: A Platform for Therapeutic Target Identification and Validation". In: *Nucleic Acids Research* (Jan. 4, 2017). ISSN: 0305-1048. DOI: 10.1093/nar/gkw1055. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210543/.

[Li+17] Huamin Li et al. "Gating Mass Cytometry Data by Deep Learning". In: *Bioinformatics* (Nov. 1, 2017). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx448. URL: https://doi.org/10.1093/bioinformatics/btx448.

[Mus+17] Aliyu Musa et al. "A Review of Connectivity Map and Computational Approaches in Pharmacogenomics". In: *Briefings in Bioinformatics* (Jan. 9, 2017). ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbw112. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw112.

[PST17] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. "Software for Molecular Docking: A Review". In: *Biophysical Reviews* (Apr. 2017). ISSN: 1867-2450. DOI: 10.1007/s12551-016-0247-1.

[Pat+17] Rob Patro et al. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression". In: *Nature Methods* (Apr. 2017). ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. URL: https://www.nature.com/articles/nmeth.4197.

[PF17] J. M. B. Prieto and M. J. B. Felippe. "Development, Phenotype, and Function of Non-Conventional B Cells". In: *Comparative Immunology, Microbiology and Infectious Diseases* (Oct. 1, 2017). ISSN: 0147-9571. DOI: 10.1016/j.cimid.2017.08.002. URL: https://www.sciencedirect.com/science/article/pii/S0147957117300723.

[Qiu+17] Xiaojie Qiu et al. "Reversed Graph Embedding Resolves Complex Single-Cell Trajectories". In: *Nature Methods* (Oct. 2017). ISSN: 1548-7105. DOI: 10.1038/nmeth.4402.

[Rai17] Johannes Rainer. *EnsDb.Hsapiens.v86: Ensembl based annotation package*. 2017.

[Sub+17] Aravind Subramanian et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles". In: *Cell* (Nov. 2017). ISSN: 00928674. DOI: 10.1016/j.cell.2017.10.049. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867417313090.

[Tas+17] Shinya Tasaki et al. "Multiomic Disease Signatures Converge to Cytotoxic CD8 T Cells in Primary Sjögren's Syndrome". In: *Annals of the Rheumatic Diseases* (Aug. 1, 2017). ISSN: 0003-4967, 1468-2060. DOI: 10.1136/annrheumdis-2016-210788. URL: https://ard.bmj.com/content/76/8/1458.

[vdBri+17] Susanne C. van den Brink et al. "Single-Cell Sequencing Reveals Dissociation-Induced Gene Expression in Tissue Subpopulations". In: *Nature Methods* (Sept. 29, 2017). ISSN: 1548-7105. DOI: 10.1038/nmeth.4437.

[YHO17] De Yang, Zhen Han, and Joost J. Oppenheim. "ALARMINS AND IMMUNITY". In: *Immunological reviews* (Nov. 2017). ISSN: 0105-2896. DOI: 10.1111/imr.12577. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5699517/.

[Yat+17] Bethan Yates et al. "Genenames.Org: The HGNC and VGNC Resources in 2017". In: *Nucleic Acids Research* (Jan. 4, 2017). ISSN: 0305-1048. DOI: 10.1093/nar/gkw1033. URL: https://doi.org/10.1093/nar/gkw1033.

[Zhe+17] Grace X. Y. Zheng et al. "Massively Parallel Digital Transcriptional Profiling of Single Cells". In: *Nature Communications* (Jan. 16, 2017). ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: https://www.nature.com/articles/ncomms14049.

[And+18] Christophe Andrieu et al. *On the Utility of Metropolis-Hastings with Asymmetric Acceptance Ratio*. Mar. 26, 2018. DOI: 10.48550/arXiv.1803.09527. URL: http://arxiv.org/abs/1803.09527. preprint.

[Bar+18] Guillermo Barturen et al. "Moving towards a Molecular Taxonomy of Autoimmune Rheumatic Diseases". In: *Nature Reviews. Rheumatology* (Jan. 24, 2018). ISSN: 1759-4804. DOI: 10.1038/nrrheum.2017.220.

[Chi18] Marie Chion. "Développement de Nouvelles Méthodologies Statistiques Pour l'analyse de Données de Protéomique Quantitative". These en préparation. Strasbourg, 2018. URL: http://www.theses.fr/s269600.

[Cor+18] MacIntosh Cornwell et al. "VIPER: Visualization Pipeline for RNA-seq, a Snakemake Workflow for Efficient and Complete RNA-seq Analysis". In: *BMC Bioinformatics* (Apr. 12, 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2139-9. URL: https://doi.org/10.1186/s12859-018-2139-9.

[Cot18] Paul Cottrell. *Advantages and Drawbacks of Next Generation Sequencing*. Feb. 5, 2018. DOI: 10.2139/ssrn.3183340. URL: https://papers.ssrn.com/abstract=3183340. preprint.

[FT18] Francesca Finotello and Zlatko Trajanoski. "Quantifying tumour-Infiltrating Immune Cells from Transcriptomics Data". In: *Cancer immunology, immunotherapy: CII* (July 2018). ISSN: 1432-0851. DOI: 10.1007/s00262-018-2150-z.

[GA18] Amir Giladi and Ido Amit. "Single-Cell Genomics: A Stepping Stone for Future Immunology Discoveries". In: *Cell* (Jan. 11, 2018). ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.11.011.

[Gol+18] Yury Goltsev et al. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging". In: *Cell* (Aug. 9, 2018). ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.07.010.

[Hua+18] Hongbin Huang et al. "Reverse Screening Methods to Search for the Protein Targets of Chemopreventive Compounds". In: *Frontiers in Chemistry* (2018). ISSN: 2296-2646. URL: https://www.frontiersin.org/articles/10.3389/fchem.2018.00138.

[KR18] Bernd Klaus and Stefanie Reisenauer. "An End to End Workflow for Differential Gene Expression Using Affymetrix Microarrays". In: *F1000Research* (July 3, 2018). DOI: 10.12688/f1000research.8967.2.

[Lam+18] Diether Lambrechts et al. "Phenotype Molding of Stromal Cells in the Lung tumour Microenvironment". In: *Nature Medicine* (Aug. 2018). ISSN: 1546-170X. DOI: 10.1038/s41591-018-0096-5. URL: https://www.nature.com/articles/s41591-018-0096-5.

[LA18] Amanda J. Lee and Ali A. Ashkar. "The Dual Nature of Type I and Type II Interferons". In: *Frontiers in Immunology* (2018). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02061.

[Lee+18] Junseok Lee et al. "Ensemble Modeling for Sustainable Technology Transfer". In: *Sustainability* (July 2, 2018). DOI: 10.3390/su10072278.

[Lim+18] Jeffrey Chun Tatt Lim et al. "An Automated Staining Protocol for Seven-Colour Immunofluorescence of Human Tissue Sections for Diagnostic and Prognostic Use". In: *Pathology* (Apr. 2018). ISSN: 1465-3931. DOI: 10.1016/j.pathol.2017.11.087.

[MJ18] Chenchen Ma and Tieming Ji. "Detecting Differentially Expressed Genes for Syndromes by Considering Change in Mean and Dispersion Simultaneously". In: *BMC Bioinformatics* (Sept. 20, 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2354-4. URL: https://doi.org/10.1186/s12859-018-2354-4.

[McK18] Katherine M. McKinnon. "Flow Cytometry: An Overview". In: *Current Protocols in Immunology* (2018). ISSN: 1934-368X. DOI: 10.1002/cpim.40. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpim.40.

[Mos+18] Sara Mostafavi et al. "A Molecular Network of the Aging Human Brain Provides Insights into the Pathology and Cognitive Decline of Alzheimer's Disease". In: *Nature neuroscience* (June 2018). ISSN: 1097-6256. DOI: 10.1038/s41593-018-0154-9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6599633/.

[Pet+18] Florent Petitprez et al. "Quantitative Analyses of the tumour Microenvironment Composition and Orientation in the Era of Precision Medicine". In: *Frontiers in Oncology* (2018). ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00390.

[RP18] Pongali Raghavendra and Thammineni Pullaiah. "Chapter 7 - Pathogen Identification Using Novel Sequencing Methods". In: *Advances in Cell and Molecular Diagnostics*. Ed. by Pongali Raghavendra and Thammineni Pullaiah. Academic Press, Jan. 1, 2018. ISBN: 978-0-12-813679-9. DOI: 10.1016/B978-0-12-813679-9.00007-5. URL: https://www.sciencedirect.com/science/article/pii/B9780128136799000075.

[Sar18] Ballantyne Sarah. *The Link Between Cancer and Autoimmune Disease*. The Paleo Mom. Mar. 19, 2018. URL: https://www.thepaleomom.com/the-link-between-cancer-and-autoimmune-disease/.

[Seg+18] Aude I. Segaliny et al. "Functional TCR T Cell Screening Using Single-Cell Droplet Microfluidics". In: *Lab on a Chip* (Dec. 4, 2018). ISSN: 1473-0189. DOI: 10.1039/C8LC00818C. URL: https://pubs.rsc.org/en/content/articlelanding/2018/lc/c8lc00818c.

[SGA18] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. "Overview of Next Generation Sequencing Technologies". In: *Current protocols in molecular biology* (Apr. 2018). ISSN: 1934-3639. DOI: 10.1002/cpmb.59. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020069/.

[SN18] Jang-il Sohn and Jin-Wu Nam. "The Present and Future of de Novo Whole-Genome Assembly". In: *Briefings in Bioinformatics* (Jan. 1, 2018). ISSN: 1477-4054. DOI: 10.1093/bib/bbw096. URL: https://doi.org/10.1093/bib/bbw096.

[TH18] Zhesen Tan and Jennifer M. Heemstra. "High-Throughput Measurement of Small-Molecule Enantiopurity by Using Flow Cytometry". In: *ChemBioChem* (2018). ISSN: 1439-7633. DOI: 10.1002/cbic.201800341. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201800341.

[Tau+18] Janis M. Taube et al. “Implications of the tumour Immune Microenvironment for Staging and Therapeutics”. In: *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* (Feb. 2018). ISSN: 1530-0285. DOI: 10.1038/modpathol.2017.156.

[VPG18] Filippo Veglia, Michela Perego, and Dmitry Gabrilovich. “Myeloid-Derived Suppressor Cells Coming of Age”. In: *Nature Immunology* (Feb. 2018). ISSN: 1529-2916. DOI: 10.1038/s41590-017-0022-x. URL: https://www.nature.com/articles/s41590-017-0022-x.

[VD18] Stephan Vilgis and Hans-Peter Deigner. “Chapter 5 - Sequencing in Precision Medicine”. In: *Precision Medicine*. Ed. by Hans-Peter Deigner and Matthias Kohl. Academic Press, Jan. 1, 2018. ISBN: 978-0-12-805364-5. DOI: 10.1016/B978-0-12-805364-5.00005-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128053645000056.

[ZA18] Weizhuang Zhou and Russ B. Altman. “Data-Driven Human Transcriptomic Modules Determined by Independent Component Analysis”. In: *BMC Bioinformatics* (Sept. 17, 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2338-4. URL: https://doi.org/10.1186/s12859-018-2338-4.

[AS19] Beatriz Aguilar-Bravo and Pau Sancho-Bru. “Laser Capture Microdissection: Techniques and Applications in Liver Diseases”. In: *Hepatology International* (Mar. 1, 2019). ISSN: 1936-0541. DOI: 10.1007/s12072-018-9917-3. URL: https://doi.org/10.1007/s12072-018-9917-3.

[BR19] Josephine Bageritz and Gianmarco Raddi. “Single-Cell RNA Sequencing with Drop-Seq”. In: *Single Cell Methods: Sequencing and Proteomics*. Ed. by Valentina Proserpio. Methods in Molecular Biology. New York, NY: Springer, 2019. ISBN: 978-1-4939-9240-9. DOI: 10.1007/978-1-4939-9240-9_6. URL: https://doi.org/10.1007/978-1-4939-9240-9_6.

[Bec+19] Etienne Becht et al. “Reverse-Engineering Flow-Cytometry Gating Strategies for Phenotypic Labelling and High-Performance Cell Sorting”. In: *Bioinformatics* (Jan. 15, 2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty491. URL: https://doi.org/10.1093/bioinformatics/bty491.

[BFM19] Thomas A. Blair, Andrew L. Frelinger, and Alan D. Michelson. “35 - Flow Cytometry”. In: *Platelets (Fourth Edition)*. Ed. by Alan D. Michelson. Academic Press, Jan. 1, 2019. ISBN: 978-0-12-813456-6. DOI: 10.1016/B978-0-12-813456-6.00035-7. URL: https://www.sciencedirect.com/science/article/pii/B9780128134566000357.

[BC19] Anna D. Broido and Aaron Clauset. “Scale-Free Networks Are Rare”. In: *Nature Communications* (Mar. 4, 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-08746-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6399239/.

[CKB19] Feixiong Cheng, István A. Kovács, and Albert-László Barabási. “Network-Based Prediction of Drug Combinations”. In: *Nature Communications* (Mar. 13, 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-09186-x. URL: https://www.nature.com/articles/s41467-019-09186-x.

[DMW19] Daniel Depledge, Ian Mohr, and Angus Wilson. “Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes”. In: *Journal of virology* (Jan. 1, 2019). DOI: 10.1128/JVI.01342-18.

[Eng+19] Chee-Huat Linus Eng et al. “Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+”. In: *Nature* (Apr. 2019). ISSN: 1476-4687. DOI: 10.1038/s41586-019-1049-y. URL: https://www.nature.com/articles/s41586-019-1049-y.

[Fin+19a] Francesca Finotello et al. “Molecular and Pharmacological Modulators of the tumour Immune Contexture Revealed by Deconvolution of RNA-seq Data”. In: *Genome Medicine* (May 24, 2019). ISSN: 1756-994X. DOI: 10.1186/s13073-019-0638-6. URL: https://doi.org/10.1186/s13073-019-0638-6.

[Fin+19b] Francesca Finotello et al. “Next-Generation Computational Tools for Interrogating Cancer Immunity”. In: *Nature Reviews Genetics* (Dec. 2019). ISSN: 1471-0064. DOI: 10.1038/s41576-019-0166-7. URL: https://www.nature.com/articles/s41576-019-0166-7.

[Gri+19] Sofia Grigoriadou et al. “B Cell Depletion with Rituximab in the Treatment of Primary Sjögren’s Syndrome: What Have We Learnt?” In: *Clinical and Experimental Rheumatology* (2019). ISSN: 0392-856X.

[Kim+19] Daehwan Kim et al. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype”. In: *Nature Biotechnology* (Aug. 2019). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0201-4. URL: https://www.nature.com/articles/s41587-019-0201-4.

[Lah+19] Amal Lahiani et al. “Generalising Multistain Immunohistochemistry Tissue Segmentation Using End-to-End Colour Deconvolution Deep Neural Networks”. In: *IET Image Processing* (2019). ISSN: 1751-9667. DOI: 10.1049/iet-ipr.2018.6513. URL: https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2018.6513.

[Li+19] Yupeng Li et al. “A Bayesian Gene Network Reveals Insight into the JAK-STAT Pathway in Systemic Lupus Erythematosus”. In: *PLOS ONE* (Dec. 2019). Ed. by Gang Han. DOI: 10.1371/journal.pone.0225651.

[Mal+19] Noël Malod-Dognin et al. “Towards a Data-Integrated Cell”. In: *Nature Communications* (Feb. 18, 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-08797-8. URL: https://www.nature.com/articles/s41467-019-08797-8.

[MO19] H. Carlo Maurer and Kenneth P. Olive. “Laser Capture Microdissection on Frozen Sections for Extraction of High-Quality Nucleic Acids”. In: *Pancreatic Cancer: Methods and Protocols*. Ed. by Gloria H. Su. Methods in Molecular Biology. New York, NY: Springer, 2019. ISBN: 978-1-4939-8879-2. DOI: 10.1007/978-1-4939-8879-2_23. URL: https://doi.org/10.1007/978-1-4939-8879-2_23.

[Pus+19] Sudeep Pushpakom et al. “Drug Repurposing: Progress, Challenges and Recommendations”. In: *Nature Reviews. Drug Discovery* (Jan. 2019). ISSN: 1474-1784. DOI: 10.1038/nrd.2018.168.

[RKA19] Juri Reimand, Raivo Kolde, and Tambet Arak. *gProfileR: Interface to the g:Profiler Toolkit*. 2019. URL: https://CRAN.R-project.org/package=gProfileR.

[RG19] Edward S. Rice and Richard E. Green. “New Approaches for Genome Assembly and Scaffolding”. In: *Annual Review of Animal Biosciences* (2019). DOI: 10.1146/annurev-animal-020518-115344. URL: https://doi.org/10.1146/annurev-animal-020518-115344.

[Rod+19] Samuel G. Rodriques et al. “Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution”. In: *Science* (Mar. 29, 2019). ISSN: 1095-9203. DOI: 10.1126/science.aaw1219. URL: https://www.science.org/doi/10.1126/science.aaw1219.

[Sar+19] Mimosa Sarma et al. “A Diffusion-Based Microfluidic Device for Single-Cell RNA-seq”. In: *Lab on a Chip* (Mar. 27, 2019). ISSN: 1473-0189. DOI: 10.1039/C8LC00967H. URL: https://pubs.rsc.org/en/content/articlelanding/2019/lc/c8lc00967h.

[SK19] Tobias Siems and Lisa Koeppel. *A Note on the Metropolis-Hastings Acceptance Probabilities for Mixture Spaces*. Aug. 2, 2019. DOI: 10.48550/arXiv.1808.00789. URL: http://arxiv.org/abs/1808.00789. preprint.

[Sta+19] Janet Staats et al. “Guidelines for Gating Flow Cytometry Data for Immunological Assays”. In: *Immunophenotyping: Methods and Protocols*. Ed. by Jr McCoy J. Philip. Methods in Molecular Biology. New York, NY: Springer, 2019. ISBN: 978-1-4939-9650-6. DOI: 10.1007/978-1-4939-9650-6_5. URL: https://doi.org/10.1007/978-1-4939-9650-6_5.

[SGH19] Rory Stark, Marta Grzelak, and James Hadfield. “RNA Sequencing: The Teenage Years”. In: *Nature Reviews Genetics* (Nov. 2019). ISSN: 1471-0064. DOI: 10.1038/s41576-019-0150-2. URL: https://www.nature.com/articles/s41576-019-0150-2.

[Stu+19] Gregor Sturm et al. “Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology”. In: *Bioinformatics (Oxford, England)* (July 15, 2019). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz363.

[Vic+19] Sanja Vickovic et al. “High-Definition Spatial Transcriptomics for in Situ Tissue Profiling”. In: *Nature Methods* (Oct. 2019). ISSN: 1548-7105. DOI: 10.1038/s41592-019-0548-y. URL: https://www.nature.com/articles/s41592-019-0548-y.

[Yan19] Xin-She Yang. “7 - Support Vector Machine and Regression”. In: *Introduction to Algorithms for Data Mining and Machine Learning*. Ed. by Xin-She Yang. Academic Press, Jan. 1, 2019. ISBN: 978-0-12-817216-2. DOI: 10.1016/B978-0-12-817216-2.00014-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128172162000144.

[Aba+20] Semagn Mekonnen Abate et al. “Rate of Intensive Care Unit Admission and Outcomes among Patients with Coronavirus: A Systematic Review and Meta-analysis”. In: *PLoS ONE* (July 10, 2020). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0235653. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7351172/.

[Ack+20] Maximilian Ackermann et al. “Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19”. In: *New England Journal of Medicine* (July 9, 2020). ISSN: 0028-4793. DOI: 10.1056/NEJMoa2015432. URL: https://www.nejm.org/doi/10.1056/NEJMoa2015432.

[Amb20] Ambry Genetics, director. *Illumina Sequencing Overview: Library Prep to Data Analysis | Webinar | Ambry Genetics*. Aug. 3, 2020. URL: https://www.youtube.com/watch?v=6jf_6STEnI4.

[Bec+20] Noam D. Beckmann et al. “Multiscale Causal Networks Identify VGF as a Key Regulator of Alzheimer’s Disease”. In: *Nature Communications* (Aug. 7, 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-17405-z. URL: https://www.nature.com/articles/s41467-020-17405-z.

[Bic20] Antoine Bichat. “Discovering multi-scale metagenomic signatures through hierarchical organization of species.” PhDthesis. Université Paris-Saclay, Dec. 9, 2020. URL: https://theses.hal.science/tel-03121038.

[Bla+20] Daniel Blanco-Melo et al. *SARS-CoV-2 Launches a Unique Transcriptional Signature from in Vitro, Ex Vivo, and in Vivo Systems*. Mar. 24, 2020. DOI: 10.1101/2020.03.24.004655. URL: https://www.biorxiv.org/content/10.1101/2020.03.24.004655v1. preprint.

[Bru+20] Elspeth A. Bruford et al. "Guidelines for Human Gene Nomenclature". In: *Nature Genetics* (Aug. 2020). ISSN: 1546-1718. DOI: 10.1038/s41588-020-0669-3. URL: https://www.nature.com/articles/s41588-020-0669-3.

[Cam+20] Neil A. Campbell et al. *Biology: A Global Approach*. Pearson Education Limited, May 14, 2020. 1512 pp. ISBN: 978-1-292-34163-7.

[Car+20] Julien Carvelli et al. "Association of COVID-19 Inflammation with Activation of the C5a–C5aR1 Axis". In: *Nature* (Dec. 2020). ISSN: 1476-4687. DOI: 10.1038/s41586-020-2600-6. URL: https://www.nature.com/articles/s41586-020-2600-6.

[CH20] Nicholas Chavkin and Karen Hirschi. "Single Cell Analysis in Vascular Biology". In: *Frontiers in Cardiovascular Medicine* (Mar. 31, 2020). DOI: 10.3389/fcvm.2020.00042.

[Che+20] Wei-Ting Chen et al. "Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease". In: *Cell* (Aug. 1, 2020). ISSN: 1097-4172. DOI: 10.1016/j.cell.2020.06.038. URL: https://doi.org/10.1016/j.cell.2020.06.038.

[Col20] Flo colado Colabs. *Mass Cytometry*. flow. 2020. URL: https://flow.ucsf.edu/mass-cytometry.

[Cor+20] Luis A. Corchete et al. "Systematic Comparison and Assessment of RNA-seq Procedures for Gene Expression Quantitative Analysis". In: *Scientific Reports* (Nov. 12, 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-76881-x. URL: https://www.nature.com/articles/s41598-020-76881-x.

[Dia20] CD Creative Diagnostics. *Innate and Adaptive Immunity - Creative Diagnostics*. 2020. URL: https://www.creative-diagnostics.com/innate-and-adaptive-immunity.htm.

[Don+20] Li Dong et al. "Semi-CAM: A Semi-Supervised Deconvolution Method for Bulk Transcriptomic Data with Partial Marker Gene Information". In: *Scientific Reports* (Mar. 25, 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-62330-2. URL: https://www.nature.com/articles/s41598-020-62330-2.

[Ewe+20] Philip A. Ewels et al. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines". In: *Nature Biotechnology* (Mar. 2020). ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x. URL: https://www.nature.com/articles/s41587-020-0439-x.

[Fa+20] Cobos Fa et al. "Comprehensive Benchmarking of Computational Deconvolution of Transcriptomics Data". In: (Jan. 10, 2020). DOI: 10.1101/2020.01.10.897116. URL: https://europepmc.org/article/ppr/ppr108248.

[G20] Jenna G. *Next-Generation Sequencing (NGS) Overview | iRepertoire, Inc.* iRepertoire. Mar. 10, 2020. URL: https://irepertoire.com/ngs-overview-from-sample-to-sequencer-to-results/.

[Gon+20] Wenbin Gong et al. "Preliminary Exploration of the Potential of Spliceosome-Associated Protein 130 for Predicting Disease Severity in Crohn's Disease". In: *Annals of the New York Academy of Sciences* (Feb. 2020). ISSN: 1749-6632. DOI: 10.1111/nyas.14240.

[Ho+20] Jessica Sook Yuin Ho et al. *Topoisomerase 1 Inhibition Therapy Protects against SARS-CoV-2-induced Inflammation and Death in Animal Models*. Dec. 1, 2020. DOI: 10.1101/2020.12.01.404483. URL: https://www.biorxiv.org/content/10.1101/2020.12.01.404483v1. preprint.

[Idr+20] Mohammed M. Idris et al. *Downregulation of Defensin Genes in SARS-CoV-2 Infection*. Sept. 23, 2020. DOI: 10.1101/2020.09.21.20195537. URL: https://www.medrxiv.org/content/10.1101/2020.09.21.20195537v1. preprint.

[Kea+20] Sarah M Keating et al. "SBML Level 3: An Extensible Format for the Exchange and Reuse of Biological Models". In: *Molecular Systems Biology* (Aug. 2020). ISSN: 1744-4292. DOI: 10.15252/msb.20199110. URL: https://www.embopress.org/doi/full/10.15252/msb.20199110.

[Kra20] Michał Krassowski. *ComplexUpset*. 2020. DOI: 10.5281/zenodo.3700590. URL: https://doi.org/10.5281/zenodo.3700590.

[Law+20] Nathan Lawlor et al. "V-SVA: An R Shiny Application for Detecting and Annotating Hidden Sources of Variation in Single-Cell RNA-seq Data". In: *Bioinformatics* (June 11, 2020). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa128. URL: https://doi.org/10.1093/bioinformatics/btaa128.

[Lei+20] Haoyun Lei et al. "tumour Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data". In: *Journal of Computational Biology* (Apr. 2020). DOI: 10.1089/cmb.2019.0302. URL: https://www.liebertpub.com/doi/10.1089/cmb.2019.0302.

[Liu+20] Kaixiong Liu et al. "Spliceosome-Associated Protein 130: A Novel Biomarker for Idiopathic Pulmonary Fibrosis". In: *Annals of Translational Medicine* (Aug. 2020). ISSN: 2305-5839. DOI: 10.21037/atm-20-4404. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7475450/.

[Mae+20] Evelyne Maes et al. "FACS-Based Proteomics Enables Profiling of Proteins in Rare Cell Populations". In: *International Journal of Molecular Sciences* (Jan. 2020). ISSN: 1422-0067. DOI: 10.3390/ijms21186557. URL: https://www.mdpi.com/1422-0067/21/18/6557.

[Mar+20] Yosra Marnissi et al. "Majorize-Minimize Adapted Metropolis-Hastings Algorithm". In: *IEEE Transactions on Signal Processing* (Mar. 2020). DOI: 10.1109/TSP.2020.2983150. URL: https://hal.science/hal-01909153.

[MM20] Miriam Merad and Jerome C. Martin. "Pathological Inflammation in Patients with COVID-19: A Key Role for Monocytes and Macrophages". In: *Nature Reviews Immunology* (June 2020). ISSN: 1474-1741. DOI: 10.1038/s41577-020-0331-4. URL: https://www.nature.com/articles/s41577-020-0331-4.

[Mol+20] M. Mollaei et al. "The Intrinsic and Extrinsic Elements Regulating Inflammation". In: *Life Sciences* (Nov. 1, 2020). ISSN: 0024-3205. DOI: 10.1016/j.lfs.2020.118258. URL: https://www.sciencedirect.com/science/article/pii/S0024320520310109.

[Nak20] Lilian Nakakawa. *PacBio SMRT - THIRD GENERATION SEQUENCING TECHNIQUE*. 2020. URL: https://www.slideshare.net/MuundaMudenda1/pacbio-smrt-third-generation-sequencing-technique.

[NSH20] PhD Nikhil Rao, PhD Sheila Clark, and Olivia Habern. "Bridging Genomics and Tissue Pathology". In: *Genetic Engineering & Biotechnology News* (Feb. 7, 2020). DOI: 10.1089/gen.40.02.16. URL: https://www.liebertpub.com/doi/10.1089/gen.40.02.16.

[PSE20] Omar Pacha, Mary Alice Sallman, and Scott E. Evans. "COVID-19: A Case for Inhibiting IL-17?" In: *Nature Reviews Immunology* (June 2020). ISSN: 1474-1741. DOI: 10.1038/s41577-020-0328-z. URL: https://www.nature.com/articles/s41577-020-0328-z.

[Qiu20] Peng Qiu. "Embracing the Dropouts in Single-Cell RNA-seq Analysis". In: *Nature Communications* (Mar. 3, 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-14976-9. URL: https://www.nature.com/articles/s41467-020-14976-9.

[Ren+20] Ziyou Ren et al. *Information-Theory-Based Benchmarking and Feature Selection Algorithm Improve Cell Type Annotation and Reproducibility of Single Cell RNA-seq Data Analysis Pipelines*. Nov. 9, 2020. DOI: 10.1101/2020.11.02.365510. URL: https://www.biorxiv.org/content/10.1101/2020.11.02.365510v3. preprint.

[RTA20] Domenico Ribatti, Roberto Tamma, and Tiziana Annese. "Mast Cells and Angiogenesis in Multiple Sclerosis". In: *Inflammation Research* (Nov. 1, 2020). ISSN: 1420-908X. DOI: 10.1007/s00011-020-01394-2. URL: https://doi.org/10.1007/s00011-020-01394-2.

[Rot+20] Annika Roth et al. *LL-37 Fights SARS-CoV-2: The Vitamin D-Inducible Peptide LL-37 Inhibits Binding of SARS-CoV-2 Spike Protein to Its Cellular Receptor Angiotensin Converting Enzyme 2 In Vitro*. Dec. 4, 2020. DOI: 10.1101/2020.12.02.408153. URL: https://www.biorxiv.org/content/10.1101/2020.12.02.408153v2. preprint.

[Sch+20] Alexandra Schnell et al. "The Yin and Yang of Co-Inhibitory Receptors: Toward Anti-tumour Immunity without Autoimmunity". In: *Cell Research* (Apr. 2020). ISSN: 1748-7838. DOI: 10.1038/s41422-020-0277-x. URL: https://www.nature.com/articles/s41422-020-0277-x.

[Sny+20] Michael P. Snyder et al. "Perspectives on ENCODE". In: *Nature* (July 2020). ISSN: 1476-4687. DOI: 10.1038/s41586-020-2449-8. URL: https://www.nature.com/articles/s41586-020-2449-8.

[Sou+20] Mohamed Soudy et al. "UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase)". In: *Journal of Proteomics* (2020). ISSN: 1874-3919. DOI: 10.1016/j.jprot.2019.103613. URL: https://www.sciencedirect.com/science/article/pii/S1874391919303859.

[Sri+20] Avi Srivastava et al. "Alignment and Mapping Methodology Influence Transcript Abundance Estimation". In: *Genome Biology* (Sept. 7, 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-02151-8. URL: https://doi.org/10.1186/s13059-020-02151-8.

[War+20] Stefanie Warnat-Herresthal et al. "Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics". In: *iScience* (Jan. 24, 2020). ISSN: 2589-0042. DOI: 10.1016/j.isci.2019.100780. URL: https://www.sciencedirect.com/science/article/pii/S2589004219305255.

[Xu+20] Yungang Xu et al. “scIGANs: Single-Cell RNA-seq Imputation Using Generative Adversarial Networks”. In: *Nucleic Acids Research* (Sept. 4, 2020). ISSN: 0305-1048. DOI: 10.1093/nar/gkaa506. URL: https://doi.org/10.1093/nar/gkaa506.

[YGN20] Chika Yokota, Daniel Gyllborg, and Mats Nilsson. “In Situ Sequencing for RNA Analysis in Tissue Sections”. In: (Feb. 2, 2020). URL: https://www.protocols.io/view/in-situ-sequencing-for-rna-analysis-in-tissue-sect-bb2giqbw.

[YJW20] Alexander P. Young, Daniel J. Jackson, and Russell C. Wyeth. “A Technical Review and Guide to RNA Fluorescence in Situ Hybridization”. In: *PeerJ* (Mar. 19, 2020). ISSN: 2167-8359. DOI: 10.7717/peerj.8806. URL: https://peerj.com/articles/8806.

[AT21] Hananeh Aliee and Fabian J. Theis. “AutoGeneS: Automatic Gene Selection Using Multi-Objective Optimization for RNA-seq Deconvolution”. In: *Cell Systems* (July 21, 2021). ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.05.006. URL: https://www.sciencedirect.com/science/article/pii/S2405471221001927.

[Bel+21] Carolina L. Bellera et al. “Can Drug Repurposing Strategies Be the Solution to the COVID-19 Crisis?” In: *Expert Opinion on Drug Discovery* (June 2021). ISSN: 1746-045X. DOI: 10.1080/17460441.2021.1863943.

[Bra+21] Bryony Braschi et al. “The Risks of Using Unapproved Gene Symbols”. In: *The American Journal of Human Genetics* (Oct. 7, 2021). ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2021.09.004. URL: https://www.sciencedirect.com/science/article/pii/S0002929721003402.

[CMR21] Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. “The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances”. In: *Frontiers in Ecology and Evolution* (Mar. 31, 2021). DOI: 10.3389/fevo.2021.588292. URL: https://hal.sorbonne-universite.fr/hal-03215628.

[Coh+21] Ariella T. Cohain et al. “An Integrative Multiomic Network Model Links Lipid Metabolism to Glucose Regulation in Coronary Artery Disease”. In: *Nature Communications* (Jan. 22, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-020-20750-8.

[Col+21] John J. Cole et al. “Searchlight: Automated Bulk RNA-seq Exploration and Visualisation Using Dynamically Generated R Scripts”. In: *BMC Bioinformatics* (Aug. 19, 2021). ISSN: 1471-2105. DOI: 10.1186/s12859-021-04321-2. URL: https://doi.org/10.1186/s12859-021-04321-2.

[Cri+21] Etienne Crickx et al. “Rituximab-Resistant Splenic Memory B Cells and Newly Engaged Naive B Cells Fuel Relapses in Patients with Immune Thrombocytopenia”. In: *Science Translational Medicine* (Apr. 14, 2021). DOI: 10.1126/scitranslmed.abc3961. URL: https://www.science.org/doi/10.1126/scitranslmed.abc3961.

[Cur+21] Sandra Curras-Alonso et al. “Spatial Transcriptomics for Respiratory Research and Medicine”. In: *European Respiratory Journal* (July 1, 2021). ISSN: 0903-1936, 1399-3003. DOI: 10.1183/13993003.04314-2020. URL: https://erj.ersjournals.com/content/58/1/2004314.

[dBru+21] Suzanne de Bruijn et al. “The Impact of Modern Technologies on Molecular Diagnostic Success Rates, with a Focus on Inherited Retinal Dystrophy and Hearing Loss”. In: *International Journal of Molecular Sciences* (Mar. 14, 2021). DOI: 10.3390/ijms22062943.

[Des+21] Emiko Desvaux et al. “Network-Based Repurposing Identifies Anti-Alarmins as Drug Candidates to Control Severe Lung Inflammation in COVID-19”. In: *PLOS ONE* (2021). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0254374. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254374.

[Det21] Philipp Dettmer. *Immune: The bestselling book from Kurzgesagt - a gorgeously illustrated deep dive into the immune system*. 1er édition. Hodder & Stoughton, Nov. 2, 2021. 368 pp. ISBN: 978-1-5293-6068-4.

[DC21] Aliou Dia and Ian H. Cheeseman. “Single-Cell Genome Sequencing of Protozoan Parasites”. In: *Trends in Parasitology* (Sept. 1, 2021). ISSN: 1471-4922. DOI: 10.1016/j.pt.2021.05.013. URL: https://www.sciencedirect.com/science/article/pii/S1471492221001379.

[Fis+21] Giulia Fiscon et al. “SAveRUNNER: A Network-Based Algorithm for Drug Repurposing and Its Application to COVID-19”. In: *PLOS Computational Biology* (Feb. 5, 2021). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008686. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008686.

[Fri+21] Miguel Fribourg et al. “CyTOF-Enabled Analysis Identifies Class-Switched B Cells as the Main Lymphocyte Subset Associated With Disease Relapse in Children With Idiopathic Nephrotic Syndrome”. In: *Frontiers in Immunology* (2021). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2021.726428.

[GCS21] Vibhav Gautam, Sourav Chatterjee, and Ananda K. Sarkar. “Single Cell Type Specific RNA Isolation and Gene Expression Analysis in Rice Using Laser Capture Microdissection (LCM)-Based Method”. In: *Rice Genome Engineering and Gene Editing: Methods and Protocols*. Ed. by Anindya Bandyopadhyay and Roger Thilmony. Methods in Molecular Biology. New York, NY: Springer US, 2021. ISBN: 978-1-07-161068-8. DOI: 10.1007/978-1-0716-1068-8_18. URL: https://doi.org/10.1007/978-1-0716-1068-8_18.

[Gau21] Marine Gauthier. “Méthodes Statistiques Pour l’analyse Différentielle de Données RNA-seq En Masse et En Cellule Unique Appliquées En Immunologie”. These de doctorat. Bordeaux, Dec. 2, 2021. URL: https://www.theses.fr/2021BORD0304.

[Gre21a] Daniel Greene. *ontologyPlot: Visualising Sets of Ontological Terms*. 2021. URL: https://CRAN.R-project.org/package=ontologyPlot.

[Gre21b] Daniel Greene. *ontologySimilarity: Calculating Ontological Similarities*. 2021. URL: https://CRAN.R-project.org/package=ontologySimilarity.

[Han+21] Wenkai Han et al. *Self-Supervised Contrastive Learning for Integrative Single Cell RNA-seq Data Analysis*. July 27, 2021. DOI: 10.1101/2021.07.26.453730. URL: https://www.biorxiv.org/content/10.1101/2021.07.26.453730v1. preprint.

[HS21] Yuanhua Huang and Guido Sanguinetti. “Uncertainty versus Variability: Bayesian Methods for Analysis of scRNA-seq Data”. In: *Current Opinion in Systems Biology* (Dec. 1, 2021). ISSN: 2452-3100. DOI: 10.1016/j.coisb.2021.100375. URL: https://www.sciencedirect.com/science/article/pii/S245231002100069X.

[Jan+21] Danielle Janosevic et al. “The Orchestrated Cellular and Molecular Responses of the Kidney to Endotoxin Define a Precise Sepsis Timeline”. In: *eLife* (Jan. 15, 2021). Ed. by Jos WM van der Meer et al. ISSN: 2050-084X. DOI: 10.7554/eLife.62270. URL: https://doi.org/10.7554/eLife.62270.

[Jin+21] Chong Jin et al. “Cell-Type-Aware Analysis of RNA-seq Data”. In: *Nature Computational Science* (Apr. 2021). ISSN: 2662-8457. DOI: 10.1038/s43588-021-00055-6. URL: https://www.nature.com/articles/s43588-021-00055-6.

[Kim+21] Kicheol Kim et al. “Cell Type-Specific Transcriptomics Identifies Neddylation as a Novel Therapeutic Target in Multiple Sclerosis”. In: *Brain: A Journal of Neurology* (Mar. 3, 2021). ISSN: 1460-2156. DOI: 10.1093/brain/awaa421. URL: https://www.mdpi.com/2218-273X/11/8/1161.

[Lee+21] Charles Lee et al. “Three Decades of the Human Genome Organization”. In: *American Journal of Medical Genetics Part A* (2021). ISSN: 1552-4833. DOI: 10.1002/ajmg.a.62512. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.62512.

[LHM21] Bo Lin, Jianan Hui, and Hongju Mao. “Nanopore Technology and Its Applications in Gene Sequencing”. In: *Biosensors* (July 2021). ISSN: 2079-6374. DOI: 10.3390/bios11070214. URL: https://www.mdpi.com/2079-6374/11/7/214.

[Lon+21] Sophia K. Longo et al. “Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics”. In: *Nature Reviews Genetics* (Oct. 2021). ISSN: 1471-0064. DOI: 10.1038/s41576-021-00370-8. URL: https://www.nature.com/articles/s41576-021-00370-8.

[Luo+21] Junwei Luo et al. “A Comprehensive Review of Scaffolding Methods in Genome Assembly”. In: *Briefings in Bioinformatics* (Sept. 1, 2021). ISSN: 1477-4054. DOI: 10.1093/bib/bbab033. URL: https://doi.org/10.1093/bib/bbab033.

[MAL21] KAVITA MALI. *Everything You Need to Know about Linear Regression!* Analytics Vidhya. Oct. 4, 2021. URL: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/.

[21] *Mediator, France’s Biggest Modern Health Scandal, Explained*. Hindustan Times. Mar. 30, 2021. URL: https://www.hindustantimes.com/world-news/mediator-france-s-biggest-modern-health-scandal-explained-101617078561403.html.

[Mys+21] Vijayashree Mysore et al. “FcγR Engagement Reprograms Neutrophils into Antigen Cross-Presenting Cells That Elicit Acquired Anti-tumour Immunity”. In: *Nature Communications* (Aug. 9, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-24591-x. URL: https://www.nature.com/articles/s41467-021-24591-x.

[Ngu21] Julie Nguyen. *Ion Torrent Sequencing*. Nov. 24, 2021. URL: https://apollo-institute.org/ion-torrent-sequencing/.

[Pfi+21] Ulrich Pfisterer et al. “Single-Cell Sequencing in Translational Cancer Research and Challenges to Meet Clinical Diagnostic Needs”. In: *Genes, Chromosomes and Cancer* (2021). ISSN: 1098-2264. DOI: 10.1002/gcc.22944. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.22944.

[Pic+21] Milan Picard et al. “Integration Strategies of Multi-Omics Data for Machine Learning Analysis”. In: *Computational and Structural Biotechnology Journal* (Jan. 1, 2021). ISSN: 2001-0370. DOI: 10.1016/j.csbj.2021.06.030. URL: https://www.sciencedirect.com/science/article/pii/S2001037021002683.

[Rao+21] Anjali Rao et al. “Exploring Tissue Architecture Using Spatial Transcriptomics”. In: *Nature* (Aug. 2021). ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9. URL: https://www.nature.com/articles/s41586-021-03634-9.

[Sch+21] Coursen W. Schneider et al. “A Novel Use of Romiplostim for SARS-CoV-2–Induced Thrombocytopenia”. In: *Journal of Pediatric Hematology/Oncology* (Aug. 2021). ISSN: 1077-4114. DOI: 10.1097/MPH.0000000000001961. URL: https://journals.lww.com/jpho-online/Abstract/2021/08000/A_Novel_Use_of_Romiplostim_for_SARS_CoV_2_induced.20.aspx.

[Sim+21] Joël Simoneau et al. “Current RNA-seq Methodology Reporting Limits Reproducibility”. In: *Briefings in Bioinformatics* (Jan. 1, 2021). ISSN: 1477-4054. DOI: 10.1093/bib/bbz124. URL: https://doi.org/10.1093/bib/bbz124.

[Sor+21] Perrine Soret et al. “A New Molecular Classification to Drive Precision Treatment Strategies in Primary Sjögren’s Syndrome”. In: *Nature Communications* (June 10, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-23472-7. URL: https://www.nature.com/articles/s41467-021-23472-7.

[Sos+21] Olukayode A. Sosina et al. “Strategies for Cellular Deconvolution in Human Brain RNA Sequencing Data”. In: (Aug. 4, 2021). DOI: 10.12688/f1000research.50858.1. URL: https://f1000research.com/articles/10-750.

[Sum+21] Tofael Ahmed Sumon et al. “A Revisit to the Research Updates of Drugs, Vaccines, and Bioinformatics Approaches in Combating COVID-19 Pandemic”. In: *Frontiers in Molecular Biosciences* (2021). ISSN: 2296-889X. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2020.585899.

[Tab21] Marco Taboga. “Markov Chain Monte Carlo (MCMC) Methods | Introduction and Explanation”. Lectures on probability theory and mathematical statistics. 2021. URL: https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo.

[VK21] Jure Vogrinc and Wilfrid S. Kendall. “Counterexamples for Optimal Scaling of Metropolis–Hastings Chains with Rough Target Densities”. In: *The Annals of Applied Probability* (Apr. 2021). ISSN: 1050-5164, 2168-8737. DOI: 10.1214/20-AAP1612. URL: https://projecteuclid.org/journals/annals-of-applied-probability/volume-31/issue-2/Counterexamples-for-optimal-scaling-of-MetropolisHastings-chains-with-rough-target/10.1214/20-AAP1612.full.

[XL21] Nan Miles Xi and Jingyi Jessica Li. “Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data”. In: *Cell Systems* (Feb. 17, 2021). ISSN: 2405-4720. DOI: 10.1016/j.cels.2020.11.008.

[Yal+21] Demet Yalcin Kehribar et al. “The Receptor for Advanced Glycation End Product (RAGE) Pathway in COVID-19”. In: *Biomarkers* (Feb. 17, 2021). ISSN: 1354-750X. DOI: 10.1080/1354750X.2020.1861099. URL: https://doi.org/10.1080/1354750X.2020.1861099.

[Zam+21] Carme Zambrana et al. “Network Neighbors of Viral Targets and Differentially Expressed Genes in COVID-19 Are Drug Target Candidates”. In: *Scientific Reports* (Sept. 23, 2021). ISSN: 2045-2322. DOI: 10.1038/s41598-021-98289-x. URL: https://www.nature.com/articles/s41598-021-98289-x.

[Zha+21] Yingdong Zhao et al. “TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository”. In: *Journal of Translational Medicine* (June 22, 2021). ISSN: 1479-5876. DOI: 10.1186/s12967-021-02936-w. URL: https://doi.org/10.1186/s12967-021-02936-w.

[ZW21] Tianhao Zhou and Jian Wang. "Laser Capture Microdissection of Vascular Endothelial CellsVascular Endothelial Cells from Frozen Heart Tissues". In: *Cardiovascular Development: Methods and Protocols*. Ed. by Xu Peng and Warren E. Zimmer. Methods in Molecular Biology. New York, NY: Springer US, 2021. ISBN: 978-1-07-161480-8. DOI: 10.1007/978-1-0716-1480-8_12. URL: https://doi.org/10.1007/978-1-0716-1480-8_12.

[Cal+22] Andrew B. Caldwell et al. "Endotype Reversal as a Novel Strategy for Screening Drugs Targeting Familial Alzheimer's Disease". In: *Alzheimer's & Dementia* (2022). ISSN: 1552-5279. DOI: 10.1002/alz.12553. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.12553.

[Car22a] Vince Carey. *ontoProc: processing of ontologies of anatomy, cell lines, and so on*. 2022.

[Car22b] Marc Carlson. *org.Hs.eg.db: Genome wide annotation for Human*. 2022.

[Cha+22] Chiranjib Chakraborty et al. "A Detailed Overview of Immune Escape, Antibody Escape, Partial Vaccine Escape of SARS-CoV-2 and Their Emerging Variants With Escape Mutations". In: *Frontiers in Immunology* (Feb. 9, 2022). ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.801522. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8863680/.

[Cle22] ClevaLab, director. *Next Generation Sequencing - A Step-By-Step Guide to DNA Sequencing*. Dec. 4, 2022. URL: https://www.youtube.com/watch?v=WKAUtJQ69n8.

[CW22] Fortran original by Leo Breiman and Adele Cutler and R. port by Andy Liaw and Matthew Wiener. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. Version 4.7-1.1. May 23, 2022. URL: https://cran.r-project.org/web/packages/randomForest/index.html.

[Dav22] Sean Davis. *GEOquery: Get data from NCBI Gene Expression Omnibus (GEO)*. 2022.

[Dey+22] Igor V. Deyneko et al. "Modeling and Cleaning RNA-seq Data Significantly Improve Detection of Differentially Expressed Genes". In: *BMC Bioinformatics* (Nov. 16, 2022). ISSN: 1471-2105. DOI: 10.1186/s12859-022-05023-z. URL: https://doi.org/10.1186/s12859-022-05023-z.

[DH22] Steffen Durinck and Wolfgang Huber. *biomaRt: Interface to BioMart databases (i.e. Ensembl)*. 2022.

[Gre22] Daniel Greene. *ontologyIndex: Reading Ontologies into R*. 2022. URL: https://CRAN.R-project.org/package=ontologyIndex.

[Gue+22] Mickaël Guedj et al. "Industrializing AI-powered Drug Discovery: Lessons Learned from the Patrimony Computing Platform". In: *Expert Opinion on Drug Discovery* (Aug. 3, 2022). ISSN: 1746-0441. DOI: 10.1080/17460441.2022.2095368. URL: https://doi.org/10.1080/17460441.2022.2095368.

[He+22] Shanshan He et al. "High-Plex Imaging of RNA and Proteins at Subcellular Resolution in Fixed Tissue by Spatial Molecular Imaging". In: *Nature Biotechnology* (Dec. 2022). ISSN: 1546-1696. DOI: 10.1038/s41587-022-01483-z. URL: https://www.nature.com/articles/s41587-022-01483-z.

[Hej+22] Boris P. Hejblum et al. *Neglecting Normalization Impact in Semi-Synthetic RNA-seq Data Simulation Generates Artificial False Positives*. May 11, 2022. DOI: 10.1101/2022.05.10.490529. URL: https://www.biorxiv.org/content/10.1101/2022.05.10.490529v1. preprint.

[Kas+22a] Makoto Kashima et al. *RNA-Seq Data Analysis for Planarian with Tensor Decomposition-Based Unsupervised Feature Extraction*. Mar. 6, 2022. DOI: 10.1101/2021.06.15.448531. URL: https://www.biorxiv.org/content/10.1101/2021.06.15.448531v2. preprint.

[Kas+22b] Aditya Kashyap et al. "Quantification of tumour Heterogeneity: From Data Acquisition to Metric Generation". In: *Trends in Biotechnology* (June 1, 2022). ISSN: 0167-7799, 1879-3096. DOI: 10.1016/j.tibtech.2021.11.006. URL: https://www.cell.com/trends/biotechnology/abstract/S0167-7799(21)00267-5.

[KES22] Audrey Kauffmann, Ibrahim Emam, and Michael Schubert. *ArrayExpress: Access the ArrayExpress Microarray Database at EBI and build Bioconductor data structures: ExpressionSet, AffyBatch, NChannelSet*. 2022.

[Lee+22] Jeffrey T. Leek et al. *sva: Surrogate Variable Analysis*. 2022.

[Li+22] Yumei Li et al. "Exaggerated False Positives by Popular Differential Expression Methods When Analyzing Human Population Samples". In: *Genome Biology* (Mar. 15, 2022). ISSN: 1474-760X. DOI: 10.1186/s13059-022-02648-4. URL: https://doi.org/10.1186/s13059-022-02648-4.

[LAH22] Michael Love, Simon Anders, and Wolfgang Huber. *DESeq2: Differential gene expression analysis based on the negative binomial distribution*. 2022. URL: https://github.com/mikelove/DESeq2.

[mer22] merck. *Immunohistochimie*. 2022. URL: https://www.sigmaaldrich.com/FR/fr/applications/protein-biology/immunohistochemistry.

[Moh+22] Abde Aliy Mohammed et al. "Pacific Bioscience Sequence Technology: Review". In: *International Journal of Veterinary Science and Research* (Mar. 29, 2022). DOI: 10.17352/ijvsr.000108. URL: https://peertechzpublications.com/articles/IJVSR-8-208.php.

[Mor+22] Martin Morgan et al. *SummarizedExperiment: SummarizedExperiment container*. 2022. URL: https://Bioconductor.org/packages/SummarizedExperiment.

[phi22] philippe. *Why 90% of Clinical Drug Development Fails and How to Improve It?* Pharma Excipients. Mar. 9, 2022. URL: https://www.pharmaexcipients.com/news/90-of-drugs-fail-clinical-trials/.

[Shr22] Aastha Shrestha. *DNA Sequencing - Sanger Sequencing Method • Microbe Online*. Microbe Online. Nov. 26, 2022. URL: https://microbeonline.com/dna-sequencing-sanger-sequencing-method/.

[Smy+22] Gordon Smyth et al. "Limma: Linear Models for Microarray Data". In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by Robert Gentleman et al. Statistics for Biology and Health. New York, NY: Springer, 2022. ISBN: 978-0-387-29362-2. DOI: 10.18129/B9.bioc.limma. URL: https://Bioconductor.org/packages/limma/.

[SM22] Mohamed Soudy and Ali Mostafa. *UniprotR: Retrieving Information of Proteins from Uniprot*. 2022. URL: https://github.com/Proteomicslab57357/UniprotR.

[Tai+22] Noha Samir Taibe et al. "Progress, Pitfalls, and Path Forward of Drug Repurposing for COVID-19 Treatment". In: *Therapeutic Advances in Respiratory Disease* (2022). ISSN: 1753-4666. DOI: 10.1177/17534666221132736.

[Tan22] Lin Tang. "Sequencing Single Cells without Killing". In: *Nature Methods* (Oct. 2022). ISSN: 1548-7105. DOI: 10.1038/s41592-022-01648-3. URL: https://www.nature.com/articles/s41592-022-01648-3.

[Yu22] Guangchuang Yu. *clusterProfiler: A universal enrichment tool for interpreting omics data*. 2022.

[Zai+22] Aleksandr Zaitsev et al. "Precise Reconstruction of the TME Using Bulk RNA-seq and a Machine Learning Algorithm Trained on Artificial Transcriptomes". In: *Cancer Cell* (Aug. 8, 2022). ISSN: 1535-6108, 1878-3686. DOI: 10.1016/j.ccell.2022.07.006. URL: https://www.cell.com/cancer-cell/abstract/S1535-6108(22)00319-1.

[ZD22] Jack Zhu and Sean Davis. *SRAdb: A compilation of metadata from NCBI SRA and tools*. 2022. URL: http://gbnci.abcc.ncifcrf.gov/sra/.

[Che23] James Chell. *Spatial Transcriptomics*. In: *Wikipedia*. Aug. 17, 2023. URL: https://en.wikipedia.org/w/index.php?title=Spatial_transcriptomics&oldid=1170879947.

[Chi23] Julien Chiquet. *COMPUTO*. 2023. URL: https://computo.sfds.asso.fr/.

[Daw23] Noor Dawany. "Large-Scale Integration of Microarray Data: Investigating the Pathologies of Cancer and Infectious Diseases". In: (Aug. 18, 2023).

[DeB23] Mike DeBerardine. *BRGenomics: Tools for the Efficient Analysis of High-Resolution Genomics Data*. Version 1.12.0. Bioconductor version: Release (3.17), 2023. DOI: 10.18129/B9.bioc.BRGenomics. URL: https://Bioconductor.org/packages/BRGenomics/.

[DGvM23] Richard Dimelow, William R. Gillespie, and Andre van Maurik. "Population Model-Based Analysis of the Memory B-cell Response Following Belimumab Therapy in the Treatment of Systemic Lupus Erythematosus". In: *CPT: Pharmacometrics & Systems Pharmacology* (2023). ISSN: 2163-8306. DOI: 10.1002/psp4.12919. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/psp4.12919.

[DML23] DMLapato. *Illumina Dye Sequencing*. In: *Wikipedia*. Aug. 18, 2023. URL: https://en.wikipedia.org/w/index.php?title=Illumina_dye_sequencing&oldid=1170974333.

[Gal23] Adriana Gallego. "Overview of Sequencing Techniques". 2023. URL: https://goldbio.com/articles/article/Overview-Sequencing-Techniques.

[hos23] The ottawa hospital. *How To CyToF - The OHRI Guide to Mass Cytometry*. 2023. URL: https://www.ohri.ca/cytof/.

[ill23] compagny illumina. "An Introduction to Next-Generation Sequencing Technology". In: (2023).

[Liu23] Yunze Liu. *genekitr: Gene Analysis Toolkit*. 2023. URL: https://www.genekitr.fun/.

[Lou23] Thomas Louis. "Cross-Sectional Study | Definition, Uses & Examples". Online website. Scribbr. June 2023. URL: https://www.scribbr.com/methodology/cross-sectional-study.

[MA23] Morgan MacKenzie and Christos Argyropoulos. "An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations". In: *Micromachines* (Feb. 2023). ISSN: 2072-666X. DOI: 10.3390/mi14020459. URL: https://www.mdpi.com/2072-666X/14/2/459.

[Mal+23] Noël Malod-Dognin et al. “A Phenotype Driven Integrative Framework Uncovers Molecular Mechanisms of a Rare Hereditary Thrombophilia”. In: *PLOS ONE* (Apr. 25, 2023). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0284084. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0284084.

[Mau+23] Daniele Mauro et al. “UBE2L3 Regulates TLR7-induced B Cell Autoreactivity in Systemic Lupus Erythematosus”. In: *Journal of Autoimmunity* (Apr. 1, 2023). ISSN: 0896-8411. DOI: 10.1016/j.jaut.2023.103023. URL: https://www.sciencedirect.com/science/article/pii/S089684112300032X.

[Pag+23] Hervé Pagès et al. *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. 2023. URL: https://Bioconductor.org/packages/AnnotationDbi.

[Sch+23] Chantal Scheepbouwer et al. “NORMSEQ: A Tool for Evaluation, Selection and Visualization of RNA-Seq Normalization Methods”. In: *Nucleic Acids Research* (July 5, 2023). ISSN: 0305-1048. DOI: 10.1093/nar/gkad429. URL: https://doi.org/10.1093/nar/gkad429.

[Sli23] SlifertheRyeDragon. *Spatial Transcriptomics*. In: *Wikipedia*. Aug. 17, 2023. URL: https://en.wikipedia.org/w/index.php?title=Spatial_transcriptomics&oldid=1170879947.

[Twy+23] J. Twynam-Perkins et al. “An Innovative Strategy for Personalised Medicine in a CFSPID Case That Evolved with Time”. In: *Paediatric Respiratory Reviews* (June 17, 2023). ISSN: 1526-0542. DOI: 10.1016/j.prrv.2023.06.001. URL: https://www.sciencedirect.com/science/article/pii/S1526054223000374.

[Xen+23] Alexandros Xenos et al. “Integrated Data Analysis Uncovers New COVID-19 Related Genes and Potential Drug Re-Purposing Candidates”. In: *International Journal of Molecular Sciences* (Jan. 11, 2023). ISSN: 1422-0067. DOI: 10.3390/ijms24021431.

[Xu+23] Xinlan Xu et al. “Short Text Classification Based on Hierarchical Heterogeneous Graph and LDA Fusion”. In: *Electronics* (Jan. 2023). ISSN: 2079-9292. DOI: 10.3390/electronics12122560. URL: https://www.mdpi.com/2079-9292/12/12/2560.

[ZR23] Kangmei Zhao and Seung Yon Rhee. “Interpreting Omics Data with Pathway Enrichment Analysis”. In: *Trends in Genetics* (Apr. 1, 2023). ISSN: 0168-9525. DOI: 10.1016/j.tig.2023.01.003. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(23)00018-5.

[] *Poster Presentations – ECCB2022*. URL: https://eccb2022.org/poster-presentations/.

[] *Scientific Image and Illustration Software | BioRender*. URL: https://www.biorender.com/.

[Yad+] Vivek Yadav et al. “Repositioning of Drugs as a Promising Strategy to Fight COVID-19”. In: *Coronaviruses* (). URL: https://www.eurekaselect.com/article/112823.

[YCh] YCharts. *Pfizer’s Expiring Viagra Patent Adversely Affects Other Drugmakers Too*. Forbes. URL: https://www.forbes.com/sites/investor/2013/12/20/pfizers-expiring-viagra-patent-adversely-affects-other-drugmakers-too/.

Application of Multivariate Gaussian Convolution and Mixture Models for Identifying Key Biomarkers Underlying Variability in Transcriptomic Profiles and the Diversity of Therapeutic Responses

Abstract

The diversity of phenotypes and conditions observed within the human species is driven by multiple intertwined biological processes. However, in the context of personalized medicine and the treatment of increasingly complex, systemic, and heterogeneous diseases, it is crucial to develop approaches that comprehensively capture the complexity of the biological mechanisms underlying the variability in biological profiles. This spans from the individual level to the cellular level, encompassing tissues and organs. Such granularity and precision are essential for clinicians, biologists, and statisticians to understand the underlying causes of the diversity in responses to clinical treatments and predict potential adverse effects.

This manuscript primarily focuses on two biological entities of interest, namely transcriptome profiles and immune cell populations, for dissecting the diversity of disease outcomes and responses to treatment observed across individuals. The introductory section provides a comprehensive overview on the intertwined mechanisms controlling the activity and abundance of these inputs, and subsequently details standard physical methods for quantifying them in real-world conditions.
To comprehensively address the intricate multi-layered organization of biological systems, we considered two distinct resolution scopes in this manuscript. At the lowest level of granularity, referred to in this manuscript as an "endotype" we examine variations in the overall bulk expression profiles across individuals. To account for the unexplained variability observed among patients sharing the same disease, we introduce an underlying latent discrete factor. To identify the unobserved subgroups characterized by this hidden variable, we employ a mixture model-based approach, assuming that each individual transcriptomic profile is sampled from a multivariate Gaussian distribution.
Subsequently, we delve into a bigger layer of complexity, by integrating the cellular composition of heterogeneous tissues. Specifically, we discuss various deconvolution techniques designed to estimate the ratios of cellular populations, contributing in unknown proportions to the total observed bulk transcriptome. We then introduce an independent deconvolution algorithm, `DeCovarT`, which demonstrates improved accuracy in delineating highly correlated cell types by explicitly incorporating the co-expression network structures of each purified cell type.

Keywords: gaussian mixture models, cellular deconvolution, transcriptome pipeline, drug repositioning

Application de modèles de convolution et de mélange gaussiens pour l'identification des biomarqueurs clés sous-jacents à la variabilité des profils transcriptomiques et à la diversité des réponses thérapeutiques

Abrégé

La diversité des phénotypes et des conditions observées au sein de l'espèce humaine est le résultat de multiples processus biologiques interdépendants. Cependant, dans le contexte de la médecine personnalisée et du traitement de maladies de plus en plus complexes, systématiques et hétérogènes, il est crucial de développer des approches qui capturent de manière exhaustive la complexité des mécanismes biologiques sous-jacents à la variabilité des profils biologiques. Cela s'étend du niveau individuel au niveau cellulaire, englobant les tissus et les organes. Une telle précision et une telle granularité sont essentielles pour que les cliniciens, les biologistes et les statisticiens comprennent les causes sous-jacentes de la diversité des réponses aux traitements cliniques et puissent prédire d'éventuels effets indésirables.

Afin d'aborder de manière exhaustive la complexité hiérarchique et stratifiée des systèmes biologiques, nous avons considéré deux niveaux d'étude dans ce manuscrit. Au niveau de granularité le plus bas, désigné dans ce manuscrit sous le terme "endotype", nous examinons les processus conduisant aux variations observées ans les profils d'expression transcriptomiques entre individus. Notamment, pour tenir compte de la variabilité non expliquée observée entre patients affectés par la même maladie, nous introduisons une variable latente discrète. Pour identifier les sous-groupes non observés, dépendant de cette variable cachée, nous utilisons des modèles de mélange probabilistes, en supposant que chaque profil transcriptomique individuel est échantillonné à partir d'une distribution gaussienne multivariée, dont les paramètres ne peuvent pas être directement estimés dans la population générale.
Ensuite, nous nous intéressons à un niveau de complexité supplémentaire, en passant en revue les méthodes canoniques permettant d'estimer la composition des tissus, souvent très hétérogènes, au sein d'un même individu. Plus précisément, nous discutons de diverses techniques de déconvolution conçues pour estimer les ratios de populations cellulaires, ces dernières contribuant en proportions inconnues au profil transcriptomique global mesuré. Nous présentons ensuite notre propre algorithme de déconvolution, nommé "DeCovarT", qui offre une précision améliorée de la délimitation de populations cellulaires fortement corrélées, en incorporant explicitement les réseaux de co-expression propres à chaque type cellulaire purifié.

Mots clés : modèles de mélange gaussiens, déconvolution cellulaire, filière de traitement de données transcriptomiques, repositionnement de médicaments