Développements méthodologiques pour la caractérisation de l'hétérogénéité tumorale

Clémentine Decamps

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Clémentine DECAMPS

Thèse dirigée par **Daniel JOST,** CR, Ecole Normale Supérieure Lyon,
et co-encadrée par **Magali Richard,** CR, Université Grenoble Alpes

préparée au sein du **Laboratoire TIMC**
dans **l'École Doctorale Igénieurie pour la Santé, la Cognition et l'Environnement**

Développements méthodologiques pour la caractérisation de l'hétérogénéité tumorale

Thèse soutenue publiquement le **25 Novembre 2021**,
devant le jury composé de :

Madame Sophie ACHARD
Directrice de recherche, CNRS, Présidente du jury
Madame Marie DE TAYRAC
PU-PH, Université de Rennes 1 / CHU de Rennes, Rapporteure
Madame Anaïs BAUDOT
Directrice de recherche, CNRS, Rapporteure
Monsieur Daniel JOST
Chargé de recherche, CNRS, Directeur de thèse
Monsieur Laurent GUYON
Chercheur en EPIC, CEA, Examinateur
Monsieur William RITCHIE
Directeur de recherche, CNRS, Examinateur
et
Madame Magali RICHARD
Chargée de recherche, CNRS, Encadrante et membre invitée
Madame Yuna BLUM
Chargée de recherche, CNRS, Membre invitée

À Léa Rôle, dite Lilith,
Je sais que tu aurais aimé lire ce manuscrit (et surtout venir au pot).
Tu nous manques beaucoup.

Remerciements

En cette fin de thèse, et ce début de manuscrit, je voudrais remercier tous ceux qui au cours de ces dernières années m'ont dit "ah ouais heu ben... bon courage.." en apprenant que je faisais une thèse, mais surtout tous ceux grâce à qui ce fut possible d'y arriver :

Avant toute chose, l'ensemble de mon jury pour m'avoir fait l'honneur d'accepter ce rôle ; mes rapporteures Anaïs Baudot et Marie de Tayrac, William Ritchie, Sophie Achard, Laurent Guyon et Yuna Blum. Je suis vraiment heureuse de pouvoir profiter de vos retours sur mon travail ! Laurent, Sophie, un merci particulier pour votre disponibilité tout au long de ma thèse à travers la participation à mon comité de suivi, et Yuna encore au-delà.

Évidemment, Magali et Daniel, pour m'avoir accueillie en M2, puis encouragée dans ce projet de thèse. Pour le temps accordé, les nombreux échanges enrichissants, et avoir su gérer ce difficile équilibre de l'encadrement réussi, en me guidant vers l'indépendance sans négliger le soutien dont j'avais besoin. J'espère que vos prochains doctorants seront nombreux et auront la chance de bénéficier d'autant de conseils et de bienveillance que moi !

Je tiens aussi à remercier le Pr. Christophe Tatout, qui fut le premier à m'encourager vers la recherche, et à semer la graine de l'idée de faire une thèse dans mon esprit.

Enfin je remercie toute l'équipe BCM-devenue-MAGE pour cet environnement de travail agréable et tous les séminaires partagés. Notamment l'équipe de Magali : Yasmina, Slim, et les différents stagiaires que nous avons accueillis, Clément et Basile, mes co-bureaux (même s'ils ne parlent que de tennis), Laura et Amandine+++ pour tous ces moments précieux partagés en dehors du labo !

La thèse n'est pas un simple diplôme ou un travail comme un autre, elle a, pour moi, un énorme impact sur la vie privée. En ce sens, je remercie tous mes proches pour m'avoir soutenue quand je détestais la science, et avoir partagé mes joies quand la recherche ne me paraissait finalement pas si mal (je ne me prononcerai pas sur la proportion de chacun de ces deux états) :

Mes parents, pour avoir accepté que je mette la fac en 1er vœu sur APB même si "c'est pour les branleurs" (papa - propos niés depuis), pour leur confiance et leur soutien infinis, pour m'avoir permis de devenir une personne heureuse. Marceau pour ces 24 années à grandir ensemble, et tout le reste de ma famille pour leur amour.

Puis, en vrac, sans ordre de préférence, mes amis,

La méta, car là on ne parle pas de soutien pendant la thèse mais pendant toutes mes études et ma vie d'adulte, Lilith, Matsou, Harty, Wawa, Uffh, Maelg, Cornet, Symfo, Mackay, Roys, Zomzom, Mad, Glen, etc. Toutes les heures à jouer, à discuter, ces weekends et vacances IRL, j'espère qu'on continuera longtemps à grandir ensemble, à partager nos dramas, nos photos de nourriture et notre amitié si précieuse.

Arnaud, Clément, Ragondin, Jules et Jonathan, cheh jsuis la seule à avoir fini le cycle universitaire j'espère que vous m'appellerez Dr toute ma vie. Blague à part, votre soutien est relatif mais votre amitié essentielle, et au moins vous achèterez une maison avant moi sans doctorat.

Le groupe ACNH, sans qui le confinement aurait paru bien plus long, Anaïs et Agathe pour ces précieuses soirées à créer autre chose que du code.

Un paragraphe particulier pour mes amis doctorants, rencontrés sur le serveur Discord de Mathilde Maillard. Personne ne comprend la thèse à moins d'en faire une, ou d'en avoir fait une. Pouvoir partager ses doutes, ses questionnements, ses peines et ses joies, dans une communauté aussi bienveillante et compréhensive, c'est clairement ce qu'il me fallait pour terminer cette thèse en bons termes avec elle-même ! Je pourrais continuer longtemps sur toutes les super personnes qui fréquentent cette communauté et la font vivre, mais mes préférées c'est le rush final promo 2021. Des mois à rédiger ensemble, à vivre chaque heure éveillée avec vous, à retenir sa respiration à chaque soutenance. Dr Pauline, Dr Julien, Dr Cathou, Mumu (+QI), Isa, Hélène, Marie, Barnabé, Camille, Charlotte, Malina, Olivia, (jvais oublier des gens c'est sûr), JVS AIME.

Merci aussi à tous ceux que j'oublie ici mais qui m'ont aidée d'une façon ou d'une autre à traverser ces années, je vous aime pas moins que les autres, promis !

Enfin, pour finir, je remercie mes merveilleuses collègues de télétravail, sans lesquelles ces deux dernières années de thèse sous confinements auraient été terriblement ennuyeuses, Lune et Arcas (Figure 1).

Figure 1 - Lune (à droite) et Arcas (à gauche).

Résumé

Le cancer est une maladie complexe où chaque tumeur est différente, aussi bien au niveau des mutations génétiques présentes, de l'expression des gènes, ou de la composition en types cellulaires. Ces nombreux facteurs jouent un rôle déterminant dans la réponse aux traitements et la survie des patients. Durant ma thèse, j'ai développé des méthodes innovantes basées sur différents types de données omiques afin de mieux caractériser l'hétérogénéité tumorale à ces différentes échelles, permettant ainsi une meilleure compréhension individuelle des tumeurs, avec à terme la possibilité de mettre en place des stratégies de médecine personnalisée.

La première partie de ma thèse s'est concentrée sur le développement d'une méthode d'analyse différentielle personnalisée. Cette méthode permet d'inférer les gènes dérégulés à l'échelle d'une tumeur unique à partir de données RNAseq. A l'aide de cette méthode, j'ai caractérisé l'hétérogénéité inter-tumorale (entre individus) dans le cancer du poumon. Nous avons ainsi pu isoler des gènes d'intérêts, comme des gènes "super-conservés" (dont l'expression ne varie jamais entre les patients) ou bien des gènes systématiquement dérégulés, et également découvrir de nouveaux sous-types tumoraux liés à la survie des patients ainsi qu'inférer les biomarqueurs associés.

Dans la deuxième partie, je me suis intéressé à l'hétérogénéité intra-tumorale, c'est-à-dire à la composition en types cellulaires d'une tumeur donnée, qui contient à la fois des cellules cancéreuses et des cellules du microenvironnement (cellules immunitaires, stroma, cellules saines). Le but est ici d'inférer cette composition à partir de données omiques (RNAseq et méthylation de l'ADN) obtenues sur la tumeur entière. Pour cela, nous avons organisé un premier data challenge multi-disciplinaire pour explorer différentes méthodes de déconvolution dites sans référence qui se basent sur les données de méthylation et évaluer l'intérêt du pré-traitement des données. Nous en avons tiré une analyse comparative des méthodes existantes, ainsi qu'un guide des bonnes pratiques quant à l'utilisation de la déconvolution pour inférer l'hétérogénéité tumorale. Nous avons ensuite exploré l'intégration de deux types de données (RNA seq et méthylation) dans un second data challenge qui a notamment débouché sur la mise en place d'une plateforme de benchmarking.

Enfin, la troisième partie de ma thèse fait le lien entre les deux premières. En associant la déconvolution et l'analyse différentielle, j'ai développé un algorithme qui permet d'inférer la dérégulation des gènes dans les cellules cancéreuses

uniquement, et d'associer cette dérégulation à la composition du microenvironnement tumoral. Après l'étude poussée des différents paramètres de la méthode sur des jeux de simulations, cette approche devrait permettre de déterminer des gènes dont la dérégulation dans le cancer joue un rôle dans l'organisation du microenvironnement tumoral, ouvrant ainsi la porte à la possibilité d'identifier de nouveaux biomarqueurs de l'hétérogénéité qui pourraient être en lien avec la virulence du cancer, et ainsi globalement d'obtenir une meilleure compréhension des mécanismes biologiques impliqués.

Acronymes

A | C | F | I | L | M | N | R | T

A

ACP Analyse en Composantes Principales. 87, 88, 96, 97, 104–108, 110, 112, 115, 116, 152

ADNm Méthylation de l'ADN (correspond aux données du méthylome). 11, 72, 73, 75, 76, 91, 95, 116, 121, 122, 124, 126, 128, 134, 151, 168, 173, 177, 189, 211

C

CF Counfonding Factors, en français facteurs de confusion. 104, 110, 112

F

FDR False Discovery Rate, en français taux de fausse découverte. 18, 36–38, 40, 50, 95

FPR False Positive Rate, en français taux de faux positifs. 29, 30, 32, 36, 37, 39

FS Feature Selection, en français sélection des sondes. 109, 110, 113, 126

I

ICA Independant Component Analysis, en français Analyse en Composantes Indépendantes. 71, 87, 88, 126, 127, 168

L

LUAD LUng ADenocarcinoma, en français adénocarcinome pulmonaire. 8, 9, 29, 30, 41–48, 82, 83, 115, 116, 119, 151, 166, 169, 170, 173, 174, 181

LUSC LUng Squamous cell Carcinoma, en français carcinome épidermoïde ou carcinome à cellules squameuses pulmonaire. 8, 9, 29, 30, 41–45, 48, 82, 83, 115, 116, 119

M

MAE Mean Absolute Error, en français Erreur absolue moyenne. 84–86, 108, 135, 192

N

NMF Non-negative Matrix Factorization, en français Factorisation matricielle non négative. 71, 72, 76, 77, 85, 86, 127, 168, 169

R

RFE La méthode RefFreeEwas. 92

RNA-seq Séquençage de l'ARN (correspond aux données du transcriptome). 10–12, 29, 33, 39, 48, 50, 52, 70, 115, 116, 119, 121, 122, 124, 126–128, 133, 134, 151, 166, 168, 177, 211, 214

T

TCGA The Cancer Genome Atlas. 9, 17, 18, 29, 41, 46, 48, 50, 53, 54, 60, 79, 80, 82–84, 115, 116, 135, 137, 151, 166, 169

TPR True Positive Rate, en français taux de vrais positifs. 29, 30, 32, 36, 37, 39, 50, 61

Sommaire

Introduction générale

1 Contexte biologique

1.1 Médecine personnalisée

Un des grands défis de la médecine actuelle est de tendre vers un traitement personnalisé, adapté au profil individuel de chaque patient. De nombreuses études ont en effet montré que la diversité génétique et génomique entre les individus avait un impact important sur la sensibilité aux maladies et sur la réponse aux traitements [1, 2].

C'est particulièrement vrai pour le cancer, où l'hétérogénéité entre les patients est si importante que chaque tumeur peut être vue comme une maladie indépendante, avec une réponse variable aux traitements [3]. Grâce aux séquenceurs de nouvelle génération, on peut désormais analyser de grandes cohortes de patients et mieux comprendre les différentes hétérogénéités inter-patients. Ces analyses ont par exemple déjà permis de découvrir des biomarqueurs associés au pronostic ou à la réponse au traitement [4, 5], et ouvrent la voie à une médecine plus précise, où les données génétiques et moléculaires sont intégrées dans le diagnostic et la prise en charge médicale.

Ces traitements ciblés et adaptés à la tumeur sont prometteurs et donnent de meilleurs résultats que les traitements génériques [6]. Cependant, l'application à grande échelle de stratégies médicales personnalisées nécessite entre autres le développement de méthodes et d'outils robustes pour caractériser les différentes hétérogénéités existantes et identifier au niveau individuel les changements moléculaires et les dérégulations associées.

1.2 Hétérogénéité tumorale

Dans le contexte du cancer, l'hétérogénéité tumorale s'exprime à de nombreux niveaux : entre deux tumeurs bien sûr (de deux personnes différentes, ou entre une tumeur d'origine et sa métastase), mais également au sein d'une même tumeur, à la fois au niveau de la composition cellulaire et au niveau moléculaire, ainsi que dans des dimensions spatiales et temporelles.

Pour illustrer cette hétérogénéité inter-tumorale, on peut prendre comme exemple les cancers du poumon, qui sont d'abord classés en sous-types. Historiquement, les différents sous-types de tumeurs ont été identifiés par histologie, c'est-à-dire en regardant directement l'aspect des cellules par microscopie. Cette classification permettait de relier ces sous-types tumoraux, inférés morphologiquement, avec d'autres variables cliniques, comme les perspectives de survie. L'utilisation grandissante du séquençage sur ces tumeurs a par la suite conduit à une reclassification de celles-ci en fonction des mutations de leur génome [7]. Cela a permis une meilleure caractérisation des tumeurs [8] et une prédiction améliorée de la survie [9]. Il est à noter que ces tumeurs sont aussi classifiables selon d'autres critères ; certaines études proposent par exemple d'utiliser l'expression des gènes [10], ou une approche multi-dimensionnelle regroupant différents types d'informations [11] pour former des groupes au sein d'un même sous-type tumoral.

Les origines de l'hétérogénéité moléculaire dans le cancer du poumon, résumées dans une revue publiée en 2019 [12] à travers la figure 1, sont très diverses. On retrouve d'abord une part importante d'anomalies génétiques, à plusieurs niveaux, qui vont du changement d'un seul nucléotide de l'ADN à des ré-arrangements chromosomiques. Il y a également des causes épigénétiques aux changements d'expression des gènes, avec des modifications de la méthylation de l'ADN, des changements de la structure de la chromatine et des modifications des histones.

Cette hétérogénéité moléculaire se retrouve également à l'échelle intra-tumorale car différentes populations de cellules cancéreuses, issues de différentes cellules-souches cancéreuses avec chacune leurs propres mutations, peuvent cohabiter dans une tumeur [13]. De surcroît, les cellules cancéreuses sont en connexion étroite avec les cellules environnantes, appelées les cellules du micro-environnement tumoral. Le micro-environnement peut également varier au sein de la tumeur, et avec lui des paramètres comme l'acidité et l'oxygénation du milieu, qui sélectionnent différentes lignées cancéreuses et peuvent influer sur la résistance aux traitements [14].

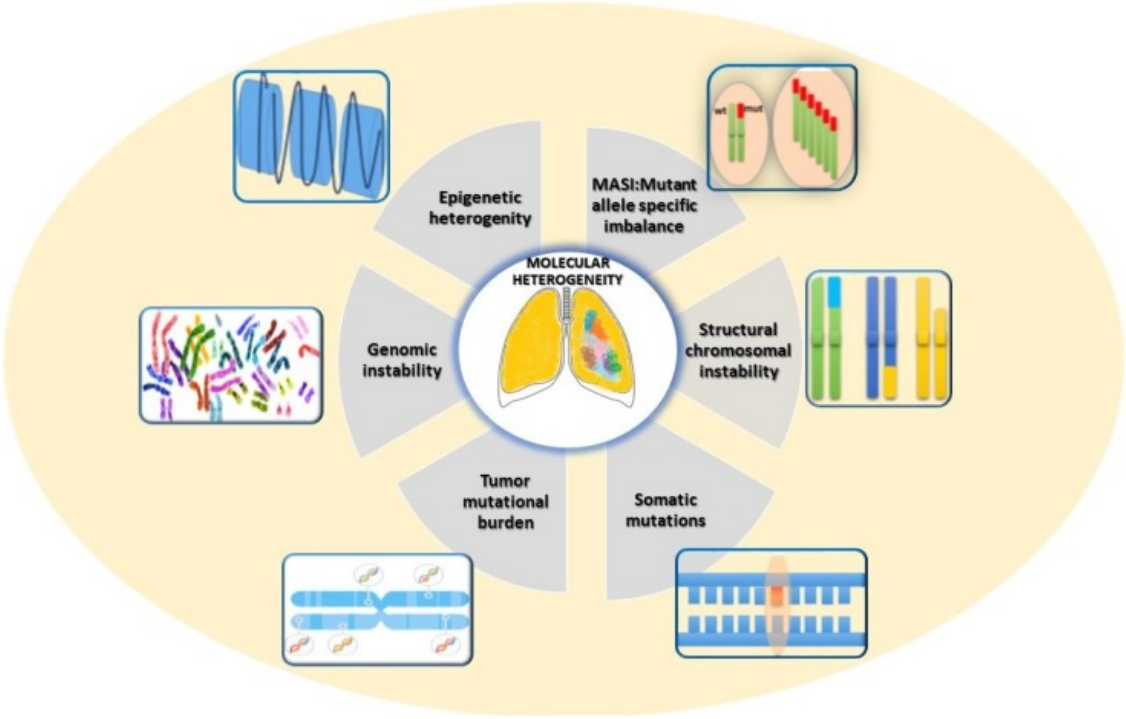


FIGURE 1 – **Les différents mécanismes à l'origine de l'hétérogénéité moléculaire dans le cancer du poumon.** Figure publiée par Z. Marino et al, 2019 [12]. Les causes de l'hétérogénéité moléculaires sont variées, allant d'une instabilité génomique générale, avec des mutations au niveau des nucléotides, mais aussi des réarrangements chromosomiques et des sélections d'allèles mutants, jusqu'à une hétérogénéité épigénétique.

À cela, s'ajoutent deux dimensions : comme la tumeur mute en permanence, cela implique une hétérogénéité temporelle (les cellules cancéreuses vont évoluer dans le temps) et spatiale (dans une même tumeur, tous les secteurs ne vont pas être identiques) [15].

Ces exemples à différentes échelles illustrent bien la complexité (et l'ampleur de la tâche) de la caractérisation de l'hétérogénéité tumorale. Au cours de ma thèse, nous allons considérer en particulier deux niveaux d'hétérogénéité : celle entre deux tumeurs indépendantes (inter-tumorale), et l'hétérogénéité d'une tumeur donnée à travers sa composition en types cellulaires (intra-tumorale).

2. Problématique

Actuellement, de nombreuses données moléculaires sont disponibles et ont été utilisées à des fins de pronostic et de diagnostic. Cependant, les méthodes d'analyse appliquées sur ces données ne prennent généralement pas en compte l'hétérogénéité tumorale, ce qui peut induire un biais dans l'interprétation des résultats. Dans ce contexte, il est donc nécessaire de développer des méthodes qui permettent d'avoir accès à des **informations précises sur la nature de la tumeur à l'échelle individuelle et qui intègrent cette problématique d'hétérogénéité**.

Nous nous sommes particulièrement intéressés aux trois questions suivantes :

1. Comment estimer l'expression différentielle à l'échelle individuelle ?
2. Comment estimer la composition du micro-environnement tumoral ?
3. Comment étudier la régulation génétique du micro-environnement tumoral ?

Objectif global

Pour répondre à ces questions, mon objectif global est de développer des méthodes prenant en compte la variabilité entre les échantillons, et donc capables d'apporter des informations à l'échelle de la tumeur unique, et non plus uniquement à celle de la population. Appliquées sur les données existantes, ces méthodes permettront de mieux comprendre les mécanismes du cancer et d'identifier des biomarqueurs de survie ou de réponse au traitement. S'ouvre alors la possibilité d'une application clinique, avec l'adaptation de la prise en charge du patient en fonction du résultat des analyses génomiques.

Analyse différentielle

Plus spécifiquement, en lien avec la première question, la première partie de mon travail s'est concentrée sur le développement d'une méthode d'analyse différentielle permettant de connaître les gènes dérégulés à l'échelle individuelle.

En effet, les méthodes classiques d'analyse différentielle sont peu adaptées aux enjeux du cancer et à l'hétérogénéité inter-tumorale, où chaque échantillon présente une forte dérégulation de son transcriptome [16], car elles se basent généralement sur une comparaison à l'échelle populationnelle, ou sur l'appariement des échantillons.

Composition en types cellulaires

En lien avec la deuxième question, nous nous sommes ensuite intéressés à la caractérisation de la composition cellulaire des échantillons tumoraux. Beaucoup d'échantillons séquencés sont composés d'un mélange de types cellulaires, et les données obtenues contiennent donc le signal des cellules cancéreuses, mais également de ce qui est appelé le micro-environnement tumoral : un mélange de cellules saines, de cellules interagissant avec le cancer et de cellules immunitaires. La caractérisation de ce micro-environnement tumoral est très importante, car il joue un rôle à la fois dans la régulation, et dans la promotion du développement du cancer, ainsi que dans la résistance aux traitements [17]. Afin de retrouver la proportion de chacun des types cellulaires, et le signal correspondant aux cellules cancéreuses, il est possible d'appliquer des techniques de déconvolution de matrices. Des méthodes ont été développées dans ce sens, mais il y a un manque de comparaisons claires entre elles. De plus, aucune étude n'a encore été menée sur l'intérêt du pré-traitement des données et de la prise en compte des facteurs de confusion pour faciliter la déconvolution, ni sur la possibilité d'intégration multi-omique.

Régulation de l'expression en lien avec le micro-environnement

Dans la dernière partie de ma thèse, liée à la troisième question, encore à un stade plus exploratoire, la problématique est de caractériser la dérégulation du transcriptome dans les cellules tumorales qui est reliée à la composition du micro-environnement. Concrètement, nous allons isoler le signal des cellules cancéreuses "purifiées" à l'aide des méthodes de la deuxième partie, puis utiliser la méthode d'analyse différentielle développée dans la première partie pour inférer les gènes dérégulés. L'objectif est de trouver des gènes dont la dérégulation

serait un biomarqueur de l'hétérogénéité intra-tumorale et de la composition en types cellulaires, ce qui permettrait de classifier les tumeurs plus facilement.

3. Modèles d'études

Nos méthodes vont êtres appliquées à deux types de cancer : le cancer du poumon et le cancer du pancréas.

3.1 Le cancer du poumon

Nous allons particulièrement nous intéresser au cancer du poumon, qui présente deux avantages dans le cadre d'un développement méthodologique : il est bien caractérisé et de nombreux jeux de données sont disponibles. Son étude présente également un fort intérêt de santé publique puisqu'il est toujours la première cause de mortalité relative au cancer dans le monde [18].

Sous-types étudiés.

Concernant la caractérisation du cancer du poumon, les tumeurs sont généralement classifiées en trois grandes histologies : cancers non à petites cellules, cancers à petites cellules et carcinoïdes. Les cancers non à petites cellules représentent 85% des cas, et sont eux-mêmes divisés en trois autres niveaux d'histologies. Pour nos analyses, nous allons utiliser les deux histologies les plus communes, qui dérivent de cellules épithéliales.

— **L'adénocarcinome pulmonaire (LUAD)** est le cancer du poumon le plus courant. Il présente un très haut taux de mutations somatiques et de réarrangements génomiques, ce qui rend notamment compliquée l'identification de ses mutations d'origine [19].

— **Le carcinome épidermoïde pulmonaire (LUSC)**, qui dérive de cellules squameuses, a également un haut taux de dérégulation, mais sur des gènes différents ; en effet, les traitements pour LUAD ne fonctionnent pas

sur LUSC [20].

LUAD et LUSC sont globalement très différents, par exemple les altérations génétiques de LUSC ressemblent plus aux autres carcinomes épidermoïdes qu'à celles de LUAD, et les cibles thérapeutiques ne sont pas les mêmes [21].

Données utilisées.

The Cancer Genome Atlas. Nous allons nous baser en grande partie sur "The Cancer Genome Atlas" (TCGA). Le TCGA est un programme lancé en 2006, de manière conjointe entre deux instituts américains : le NCI (institut national du cancer) et le NHGRI (institut national de recherche sur le génome humain). Aujourd'hui, le TCGA représente une énorme base de données contenant des milliers de données omiques de différents types moléculaires, sur 33 types de cancer différents, ainsi que sur des tissus sains appariés [site du TCGA].

Cohorte Brambilla. Le choix du cancer du poumon a également été dicté par une collaboration locale entre notre équipe et celle de l'anatomopathologiste Elisabeth Brambilla du CHU de Grenoble, responsable d'une large cohorte de données du cancer du poumon [22]. Ces données pourront servir à valider les résultats obtenus sur la cohorte du TCGA.

3.2 Le cancer du pancréas.

Dans un second temps, nous avons noué une collaboration avec l'équipe de l'anatomopathologiste Jérome Cros travaillant sur le cancer du pancréas, ce qui nous a menés à l'utiliser comme modèle d'étude pour certaines parties de ma thèse. Le cancer du pancréas est un cancer très agressif, il est souvent détecté tard et il est notamment caractérisé par le plus faible taux de survie à 5 ans (9% en 2020) [18]. Le micro-environnement joue un rôle particulièrement important dans ce cancer, mais son rôle exact est encore mal caractérisé [23]. Il est à noter que l'essentiel de cette collaboration ne fait pas partie de mon travail de thèse et ne sera pas présentée ici.

4. Nature des données biologiques

La caractérisation moléculaire des différents niveaux d'hétérogénéité tumorale passe par l'étude et la quantification de différentes molécules biologiques, regroupées sous le terme "omiques". Ces données omiques englobent plusieurs échelles : ADN (génomique), ARN (transcriptomique), protéines (protéomique), etc. Au cours de ma thèse, nous allons nous intéresser en particulier à deux types de données : le transcriptome et le méthylome.

4.1 Transcriptome

Les méthodes d'analyses transcriptomiques permettent d'obtenir une quantification des ARN messagers (ARNm) présents dans un échantillon. Ces données reflètent donc l'expression des gènes, ce qui est une information biologiquement très explicite, qui permet par exemple de comprendre le rôle de certains gènes en regardant les différences entre deux conditions expérimentales.

Une des premières technologies à "haut débit" à avoir permis d'obtenir ces données d'expression des gènes est celle des micro-puces à ADN (microarrays) développées en 1999 [24]. Le principe est de transcrire les ARNm en ADN complémentaires marqués par fluorescence. Les ADN complémentaires se fixent ensuite sur des puces à ADN, et c'est le niveau de fluorescence de chaque puce qui permet de quantifier l'ARNm correspondant au gène. Un inconvénient de cette technique est le design des puces, qui limite le nombre de régions possibles à étudier.

En 2001, le premier génome humain a été séquencé par les équipes du "Human genome project" [25] et du International Human Genome Sequencing Consortium [26], ce qui marque l'essor des technologies de séquençages nouvelles générations (NGS). Pour le transcriptome, c'est la technologie du RNA-seq, appliquée pour la première fois sur la levure en 2008 [27], qui permet d'avoir accès à l'expression de tout le génome. Les ARNm sont transcrits en ADN complémentaires puis séquencés pour obtenir des "reads", ou lectures. Ces "reads" sont ensuite alignés sur un génome de référence pour obtenir un nombre de "reads" par gène, qui représente l'expression quantitative du gène [28].

Ces méthodes expérimentales sont généralement appliquées sur des échantillons mélangés ("bulk"), ce qui signifie qu'on obtient l'expression moyenne d'un gène dans une population de cellules. Il est cependant possible d'appliquer le RNA-seq à l'échelle d'une cellule unique, on parle alors de scRNA-seq, pour "single cell RNA-seq" [29]. Le scRNA-seq est très prometteur, mais présente encore des inconvénients : le séquençage de cellules uniques est compliqué, très coûteux, et nécessite une grande amplification de l'ADN qui entraîne des biais techniques particuliers [30]. Au cours de ma thèse, les méthodes que nous développons sont donc prévues pour une application à des données "bulk". De plus, de très nombreuses données ont été déjà générées avec ces technologies et leur exploitation peut encore apporter de nouvelles informations.

4.2 Méthylome

La méthylation de l'ADN (ADNm) est une marque épigénétique, ce qui signifie qu'elle est réversible, transmissible et qu'elle régule l'expression des gènes sans modifier la séquence nucléotidique [31]. Concrètement, il s'agit de l'ajout d'un groupement méthyle sur une base cytosine. Chez l'homme, la méthylation de l'ADN se retrouve principalement au niveau des séquences CpG, c'est-à-dire une cytosine suivie d'une guanine [32].

Le lien entre expression des gènes et méthylation n'est pas trivial, c'est un processus complexe qui n'est pas encore pleinement caractérisé [33], même si l'hyper-méthylation de la région promotrice d'un gène est plutôt un marqueur de la répression de son expression [34].

Pour identifier les bases méthylées expérimentalement, on passe par une première étape de traitement au bisulfite. Le bisulfite est un composé chimique qui transforme les cytosines non méthylées en uracile, puis en thymine. Comme pour l'étude du transcriptome, il existe une technique par micro-puces et une technique par séquençage de tout le génome.

Dans la technique "Illumina beadchip" dont sont issues nos données, les puces permettent de mesurer l'hybridation de fragments génomiques à des sondes avec une fluorescence différente en fonction de la présence d'une cytosine ou d'une thymine. Cela permet alors de quantifier la proportion de cytosines

s'étant transformées en thymine à une position donnée, et donc la proportion de méthylation dans l'échantillon. Il est à noter qu'il existe plusieurs versions de cette technique : la première comportait environ 27k sondes [35], la seconde 450k [36] et la plus récente atteint 850k sondes [37], réparties sur tout le génome. Ces sondes ont été sélectionnées pour être dans des régions d'intérêt : notamment les régions promotrice des gènes, et les îlots riches en CpG.

La technique Bisulfite-seq (BS-seq) est quant à elle basée sur le séquençage haut débit et permet d'avoir accès à la méthylation de l'ensemble des sites CpG, révélant ainsi plusieurs dizaines de millions de cytosines méthylées chez l'homme [38]. Cependant, l'utilisation du BS-seq reste encore assez minoritaire car les données sont complexes à obtenir et à analyser. Comme pour le RNA-seq, le BS-seq est généralement appliqué à des échantillons mélangés regroupant plusieurs cellules, mais l'application à l'échelle unique est possible, on parle alors de scBS-seq [39]. Durant ma thèse, nous utiliserons seulement des données issues des Illumina beadchip qui sont beaucoup plus courantes.

Organisation du manuscrit

Le manuscrit va être organisé en trois grandes parties correspondant aux trois problématiques. Nous allons d'abord aborder la caractérisation des différences d'expression au niveau d'une tumeur individuelle (partie I), puis chercher à inférer la composition cellulaire d'un échantillon mélangé (partie II). L'objectif final est de réussir à détecter les gènes dérégulés dans des cellules cancéreuses qui ont un lien avec la composition du micro-environnement tumoral (partie III).

Chaque partie du manuscrit contient une introduction détaillée des enjeux de chaque question et des outils pré-existants ainsi qu'une présentation des résultats et une discussion détaillée. Les articles correspondant sont donnés en annexe. Ces outils ont été développés en R [40], un langage de programmation particulièrement adapté pour l'analyse de données, très utilisé en bioinformatique et permettant de diffuser nos outils simplement, sous la forme de packages. Les vignettes correspondantes sont également fournies en annexe.

Première partie

Analyse différentielle à l'échelle individuelle

Introduction

Le premier niveau d'hétérogénéité dans le cancer est celui entre les patients : chaque tumeur peut être vue comme une maladie indépendante, très différente d'une autre tumeur provenant du même type d'organe, avec une réponse propre aux traitements. Comme abordé dans l'introduction générale, cette hétérogénéité est multi-factorielle : mutations génétiques, changements épigénétiques, composition en cellules différentes, etc. Dans tous les cas, la conséquence est la même : les gènes dérégulés sont différents dans chaque tumeur. Afin d'améliorer la compréhension de cette hétérogénéité inter-tumorale et la sensibilité des études faites à ce sujet, il est donc nécessaire de développer des méthodes fiables permettant d'identifier les dérégulations existant au niveau individuel.

Dans l'analyse des données transcriptomiques, beaucoup de méthodes bioinformatiques et statistiques existent déjà pour identifier les gènes dérégulés à l'échelle de la population. Ces méthodes dites « fold-change» consistent à comparer la distribution de l'expression des gènes entre deux conditions (typiquement cas versus contrôles) en y appliquant un test statistique. C'est par exemple le cas de DESeq2 [41] où, pour chaque gène, la distribution d'expression dans une population est modélisée par une distribution binomiale négative permettant d'inférer la moyenne et la variance. Un test de Wald est ensuite effectué pour déterminer si les distributions dans les 2 populations sont différentes, et donc pour inférer si le gène est dérégulé. La méthode edgeR [42], également très utilisée, modélise quant à elle la distribution par un modèle de Poisson. Enfin, il existe aussi limma [43] qui applique un modèle linéaire pour chaque gène, puis différentes méthodes statistiques comme des variations du test de Student pour en étudier l'expression. Ces méthodes permettent de trouver les gènes qui sont

en moyenne différentiellement exprimés entre deux conditions avec une bonne précision, et ont déjà permis d'obtenir des résultats robustes expérimentalement [44].

Cependant, elles ne donnent pas d'information au niveau individuel et une dérégulation plus rare, présente dans seulement une petite proportion des individus, risque donc de ne pas être détectée. Par ailleurs, elles sont généralement très sensibles aux effets dits «de batch», c'est-à-dire aux variations dues aux aléas expérimentaux (conditions expérimentales variables comme la température de la pièce, la personne qui fait l'expérience, le lot de réactif...), qui, sans bonne correction, peuvent générer des faux positifs ou faire disparaître des effets de sous-populations intéressants [45]. Différentes méthodes de normalisations sont généralement appliquées sur les données pour en réduire l'incidence (par exemple, DESeq et edgeR intègrent cette étape), mais la normalisation peut générer de nouveaux biais et également perturber le signal biologique [46] [47].

D'autres types de méthodes sont donc nécessaires pour résoudre ces problèmes et capturer la dérégulation spécifique au niveau de l'individu sans être sensible aux bruits expérimentaux. Plusieurs méthodes ont été développées dans ce sens (elles sont détaillées dans une revue parue en 2017 [48]).

Certaines d'entre elles, comme DEGseq [49], NOISeq [50] ou GFOLD [51] utilisent des échantillons appariés (de la même étude, ou de références) pour effectuer l'analyse différentielle ; ces échantillons doivent avoir plusieurs réplicats pour DEGseq et NOISeq. Dans le cas du cancer, ces échantillons appariés sont souvent assez rares et les réplicats aussi, il n'y a généralement qu'un seul jeu de données par biopsie. De plus, l'inconvénient de ces méthodes est que pour un petit nombre de réplicats, il est difficile d'établir si la différence d'expression provient vraiment d'une dérégulation ou s'il s'agit de l'hétérogénéité intrinsèque (différence de stade du cycle cellulaire par exemple), voir de biais techniques. Enfin, la pertinence d'appariement des échantillons pour le cancer n'est pas établie. Une étude multi-plateformes s'est ainsi intéressée à la classification moléculaire de 3 500 échantillons de douze types de cancers différents, habituellement définis en fonction de leur tissu d'origine. Les cancers classifiés

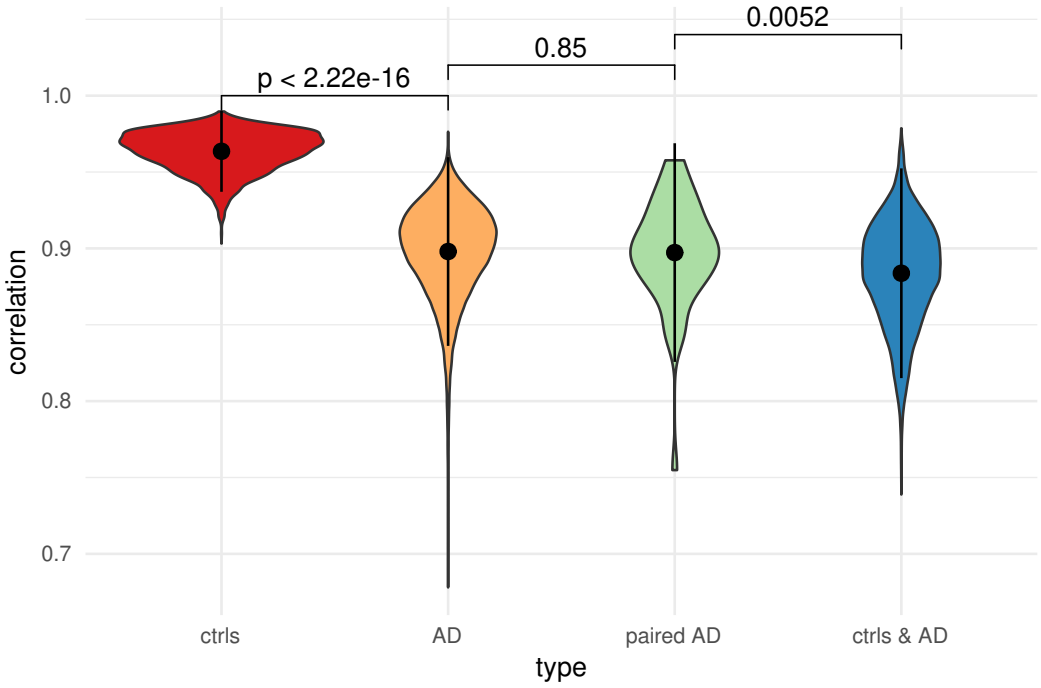


FIGURE 2 – **Distribution des corrélations dans les adénocarcinomes de TCGA.** La corrélation de Spearman a été calculée pour plusieurs combinaisons d'échantillons : entre chaque contrôles deux à deux (ctrls, n = 4656 paires), entre chaque cancer deux à deux (AD, n = 103.285 paires), entre échantillons contrôle et cancer venant du même patient (paired AD, n = 48 paires) et entre des tumeurs et des contrôles non appariés (ctrls & AD, n = 44.135 paires). La p-valeur correspond au test de Mann-Whitney.

sans à priori d'origine ont convergé en 11 grands groupes, dont seulement 5 discriminent l'organe [5]. Plus concrètement, dans les cohortes d'adénocarcinomes pulmonaires TCGA (voir partie 4 de l'introduction), la corrélation moyenne entre les tumeurs et tissus sains appariés est similaire à celle entre les tumeurs de deux patients différents, et à peine plus grande que celle entre une tumeur et un tissu sain non apparié (voir figure 2).

Par ailleurs, il existe une méthode d'analyse différentielle individuelle basée sur les rangs : Rankcomp [52, 53]. Dans un premier temps, les valeurs d'expression des gènes sont converties en rang dans chaque échantillon contrôle, puis les gènes sont appariés pour établir des paires ayant un ordre relatif stable. Dans un

second temps, un test de Fisher est utilisé pour détecter les gènes dérégulés dans les échantillons tumoraux de manière individuelle, en testant si le nombre d'inversions de paires est significatif. La méthode Rankcomp a montré des résultats intéressants pour définir des facteurs de risque ou de réponse au traitement [54, 55, 56] et elle permet de s'affranchir des problèmes de normalisation en se basant sur les rangs internes aux échantillons. Cependant, elle entraîne également un très fort taux de faux positifs. Ainsi, en calculant le taux de fausse découverte (FDR) obtenus sur les simulations contenues dans leur article [52], on retrouve un taux de FDR variant entre 20 et 50%. Dans nos tests sur d'autres simulations indépendantes, le taux de fausse découverte atteint est de 20 %.

Face à ces constats, nous avons décidé de développer notre propre méthode d'analyse différentielle personnalisée, appelée Penda pour PersoNalized Differential Analysis, que je vais détailler dans cette partie. Comme RankComp, cette méthode se base sur les rangs des gènes dans les échantillons contrôles pour inférer les dérégulations dans chaque échantillon tumoral individuellement. Nous avons également développé une méthode de simulation de dérégulation des gènes pour pouvoir tester les différents paramètres de Penda et la comparer à d'autres méthodes (RankComp, et DESeq2 utilisée de manière individuelle). Nous avons ensuite appliqué cette méthode à deux cohortes de cancers du poumon issues du consortium TCGA et nous avons détecté de nouvelles histologies moléculaires associées à la survie ainsi que de nouveaux biomarqueurs. Ce travail est publié dans un article de recherche dans PLoS Computational Biology visible en annexe 1 (dans la section 7.2.4). Je présenterai ensuite l'exploration des limites de la méthode à travers l'analyse d'échantillons sans contrôles provenant d'un autre type de cancer : le glioblastome.

Chapitre 1

Principe et implémentation de la méthode Penda

1.1 Présentation de la méthode Penda

La méthode Penda est une méthode d'analyse différentielle basée sur les rangs. Elle se déroule en deux étapes : établir les rangs relatifs entre gènes dans les échantillons contrôles, puis, à partir de ceux-ci, établir la dérégulation dans chaque échantillon tumoral.

1.1.1 Rang dans les contrôles (listes L et H)

Pour établir les rangs de référence, notre méthode se base sur l'ordre relatif des gènes dans les échantillons contrôles. Pour chaque gène *g*, deux listes sont établies : la liste *L(g)*, des gènes ayant une expression inférieure au gène *g* (Lower) et la liste *H(g)* des gènes ayant une expression supérieure au gène *g* (Higher). L'ordre relatif entre ces gènes doit être stable dans une large proportion des échantillons contrôles, fixée par l'utilisateur, généralement à 100%, pour qu'il soit ajouté aux différentes listes. Ce n'est pas l'expression des gènes qui importe, mais bien l'ordre relatif entre *g* et le gène testé (voir la figure 1.1).

Pour éviter d'utiliser comme référence des gènes avec un niveau d'expression très différent du gène ciblé *g*, on fixe ensuite une taille de liste maximale avec le

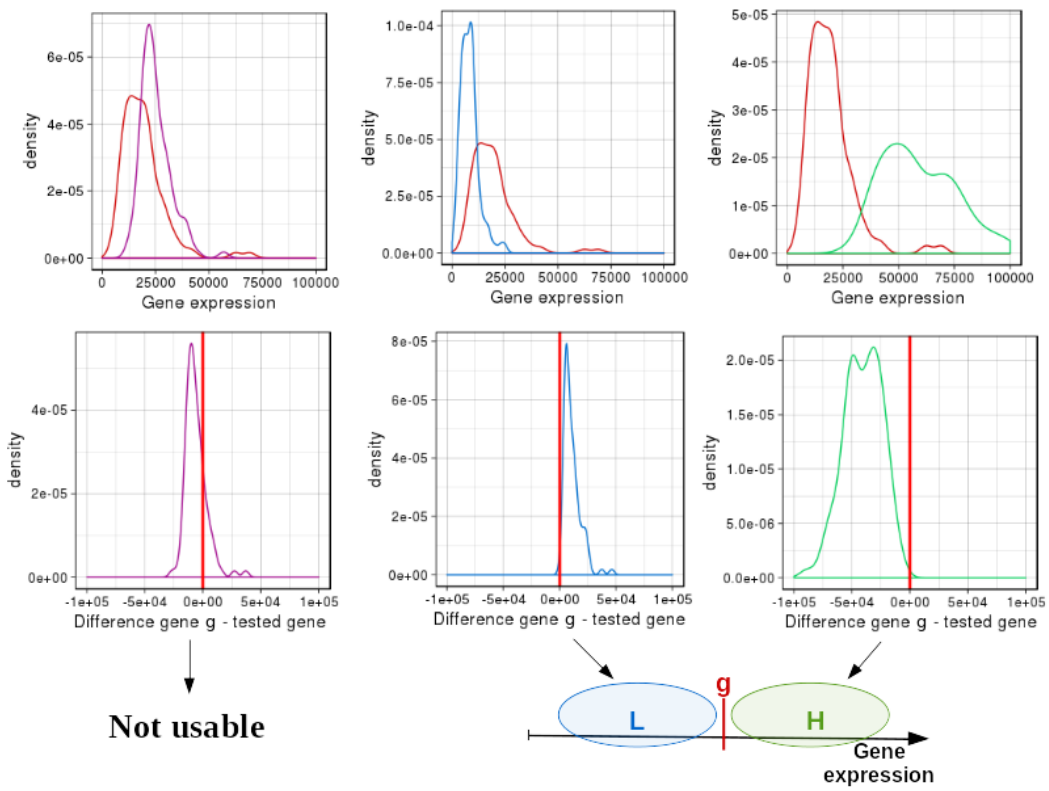


FIGURE 1.1 – **Illustration de la création des listes de références Penda.** (Haut) Distributions de l'expression (counts RNA-seq) de 3 gènes qui sont comparés au gène *g* (rouge) dans les contrôles. (Bas) Distribution de la différence d'expression dans un échantillon entre *g* et le gène testé. Le gène violet n'a pas un rang relatif stable par rapport au gène *g*, il a parfois un niveau d'expression supérieur, et parfois un niveau inférieur, il n'est donc pas utilisé. Les gènes bleus et verts ont un rang relatif stable par rapport au gène *g*, ils sont systématiquement moins exprimés ou plus exprimés que le gène *g* dans les contrôles, ils peuvent donc servir de référence dans les listes *L(g)* et *H(g)*.

paramètre *l*. Seuls les *l* gènes dont la médiane de l'expression dans les contrôles est la plus proche de celle du gène d'intérêt *g* sont conservés dans chaque liste *L(g)* et *H(g)*. Ce critère permet de conserver dans chaque liste les gènes qui seront les plus sensibles à un faible changement d'expression, et donc de détecter des dérégulations qui ne sont potentiellement pas prises en compte par les autres méthodes d'analyse différentielle n'incluant pas cette étape.

1.1.2 Test de dérégulation

Une fois les listes *L(g)* et *H(g)* établies à partir des échantillons contrôle, l'analyse de dérégulation se déroule pour chaque tumeur de manière indépendante (figure 1.2). Chaque gène *g* est analysé un à un, en comparant sa position relative dans la tumeur par rapport aux gènes des listes *L(g)* et *H(g)*. On s'intéresse aux gènes qui changent de liste entre l'échantillon tumoral et les échantillons contrôles. Les gènes qui sont dans la liste *L(g)* chez les contrôles et qui ont une expression supérieure à *g* (up) dans la tumeur sont appelés *Lu(g)*, à l'inverse les gènes de la liste *H(g)* qui ont une expression inférieure à *g* (down) dans la tumeur sont nommés *Hd(g)*. C'est la proportion de ces gènes changeant de position (*Lu(g)* et *Hd(g)*) qui est comparée à un seuil (ou threshold) fixé par l'utilisateur, et qui détermine si *g* est dérégulé ou non.

$$Si \frac{|Lu(g)|}{|L(g)|} > threshold \Rightarrow \text{g est sous-exprimé dans la tumeur}$$

$$Si \frac{|Hd(g)|}{|H(g)|} > threshold \Rightarrow \text{g est sur-exprimé dans la tumeur}$$

L'étape du test implique quelques subtilités et cas particuliers qui sont détaillés dans la partie suivante.

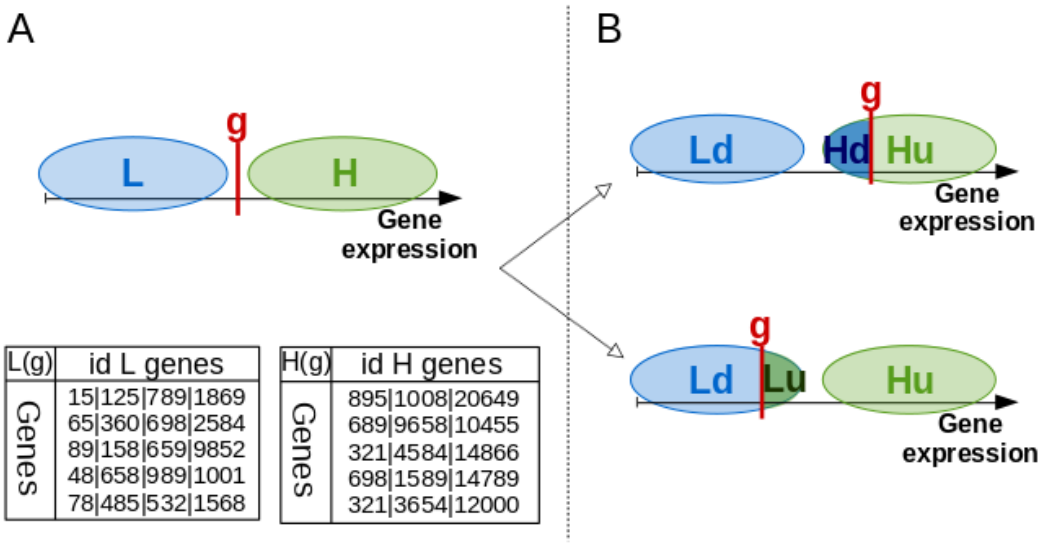


FIGURE 1.2 – **Procédure du test Penda.** Panneau **A** : les gènes ont été classés en fonction de leur expression dans les contrôles. Pour chaque gène *g*, on a formé les listes de référence *L(g)* des gènes ayant un niveau d'expression inférieur, et *H(g)* des gènes ayant un niveau d'expression supérieure. Ces listes sont conservées sous la forme de matrices (en bas), seul l'identifiant des gènes est conservé, on ne garde pas la valeur dans les contrôles. Panneau **B** : pour une tumeur donnée, le rang de *g* dans cette tumeur est comparée aux rangs des gènes *L(g)* et *H(g)* établis dans les contrôles. On définit 4 nouveaux sous-ensembles de gènes : *Ld(g)* les gènes *L(g)* toujours inférieurs à *g*, *Hu(g)* les gènes *H(g)* toujours supérieurs à *g*, *Hd(g)* les gènes *H(g)* désormais moins exprimés que le gène *g*, et *Lu(g)* les gènes *L(g)* désormais plus exprimés que *g*. Pour qu'un gène soit considéré comme dérégulé, la proportion de *Hd(g)* ou de *Lu(g)* doit dépasser un seuil fixé.

1.1.3 Subtilités du test de dérégulation

Itérations

Les gènes détectés comme dérégulés ne constituent à priori plus une bonne référence dans les listes *L* et *H* car le fait de les retrouver dans une liste *Lu* ou *Hd* peut simplement être dû à leur propre dérégulation. Penda fonctionne donc de façon itérative, en recommençant l'analyse de dérégulation (voir ci-dessus) en supprimant des listes *L* et *H* les gènes inférés comme dérégulés à l'itération précédente, jusqu'à obtenir une convergence sur l'identité des gènes dérégulés. Parfois, cette liste ne se stabilise pas, et typiquement oscille entre deux états. Dans ce cas, le nombre d'itérations maximal est fixé par l'utilisateur : s'il est atteint, c'est l'ensemble des gènes dérégulés à l'itération n et à l'itération n-1 qui est considéré dérégulé. Ce processus est détaillé dans la figure 1.3.

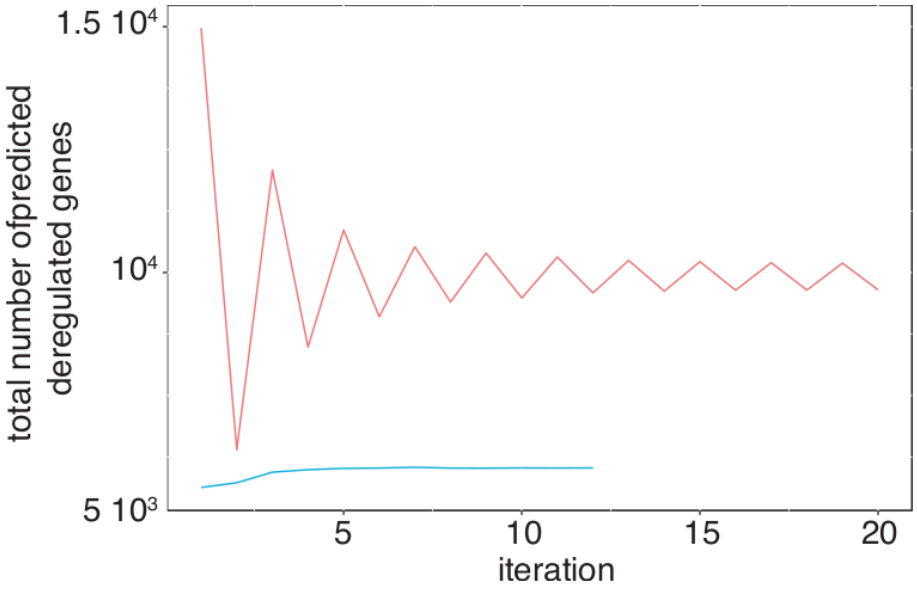


FIGURE 1.3 – **Nombre de gènes dérégulés à chaque itération Penda.** La méthode Penda est appliquée sur un même échantillon avec deux seuils de test : 0,1 (ligne rouge), et 0,4 (ligne bleue). Pour le seuil de 0,4, la liste de gènes dérégulés est rapidement stabilisée et le test s'arrête à l'itération 12. Pour le seuil de 0,1, le nombre de gènes dérégulés ne parvient pas à une stabilisation car la liste oscille à chaque itération. Une fois l'itération maximale n = 20 atteinte, ce sera donc l'union des gènes dérégulés à n = 20 et à n = 19 qui sera conservée. Figure issue du papier.

Méthode du centile

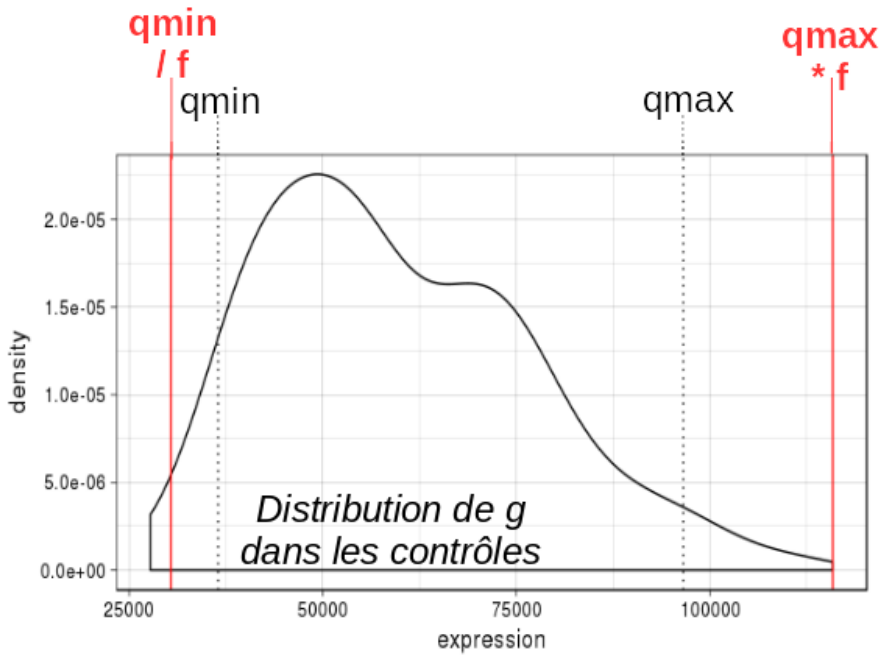


FIGURE 1.4 – **La méthode du centile.** Quand les listes de référence dans les contrôles ne sont pas disponibles, la dérégulation est inférée par la méthode du centile. La valeur du gène g dans la tumeur est comparée à sa distribution dans les contrôles : $qmin$ correspond au centile c fixé par l'utilisateur, $qmax$ au centile $(1-c)$. Ensuite, le facteur est appliqué pour définir les bornes rouges ($qmin/f$ et $qmax*f$). Le gène g dans la tumeur est dérégulé si sa valeur d'expression dépasse une de ces bornes.

Il arrive parfois qu'une des listes $L(g)$ ou $H(g)$ soit vide. C'est par exemple le cas quand beaucoup de gènes ont été détectés comme dérégulés dans l'échantillon à l'itération précédente et ont donc été retirés des listes (voir ci-dessus), ou quand l'expression du gène est parmi les plus faibles ou les plus fortes de l'échantillon. Dans ce cas, c'est la méthode du centile (ou percentile) qui est utilisée pour évaluer la dérégulation dans la direction où la liste est vide. Cette méthode détecte le gène g comme étant dérégulé dans la tumeur s'il a une valeur aberrante par rapport à la distribution de l'expression de g dans les échantillons contrôles (figure 1.4).

En pratique, l'utilisateur fixe la valeur du centile c et un facteur f. $qmin$

correspond au centile(c) de la distribution du gène dans les échantillons sains, $qmax$ au centile $(1-c)$. Avec cette méthode, si la liste $L(g)$ est vide, le gène est détecté sous-exprimé dans la tumeur si son expression est inférieure à $qmin$ divisé par le facteur. Si la liste $H(g)$ est vide, il est considéré sur-exprimé si son expression est supérieure à $qmax$ multiplié par le facteur. L'ajout d'un facteur par rapport à l'utilisation du simple centile permet de s'éloigner de la distribution normale du gène dans les contrôles.

$$\text{Si } expr(g)_{tum} < \frac{qmin(expr(g)_{ctrl})}{f} \Rightarrow \text{g est sous-exprimé}$$

$$\text{Si } expr(g)_{tum} > qmax(expr(g)_{ctrl}) * f \Rightarrow \text{g est sur-exprimé}$$

Concrètement, plus le centile est petit, et plus le facteur est grand, plus l'expression du gène dans la tumeur devra être éloignée de la distribution d'expression dans les contrôles pour être détecté dérégulé.

1.1.4 Simulations

Afin d'évaluer l'impact des différents paramètres de la méthode Penda, de permettre à l'utilisateur de choisir le plus adapté, et de comparer entre elles les méthodes d'analyse différentielle existantes, nous avons développé une manière de simuler la dérégulation des gènes en s'inspirant d'échantillons réels.

La méthode de simulation développée pour Penda se base sur un set d'échantillons contrôles et un set d'échantillons tumoraux. L'objectif est de s'inspirer des différences réelles entre les deux pour simuler au mieux un échantillon tumoral en partant d'un échantillon contrôle. Il y a donc deux étapes, la première étant d'inférer les paramètres de dérégulation, et la seconde de les appliquer.

Détermination des paramètres de dérégulation

L'idée de départ de notre méthode de simulation est que des gènes ayant une expression proche auront une amplitude et une probabilité de dérégulation proche. Les gènes sont donc regroupés en fonction de leur niveau d'expression dans les contrôles pour déterminer les différents paramètres de simulation : la

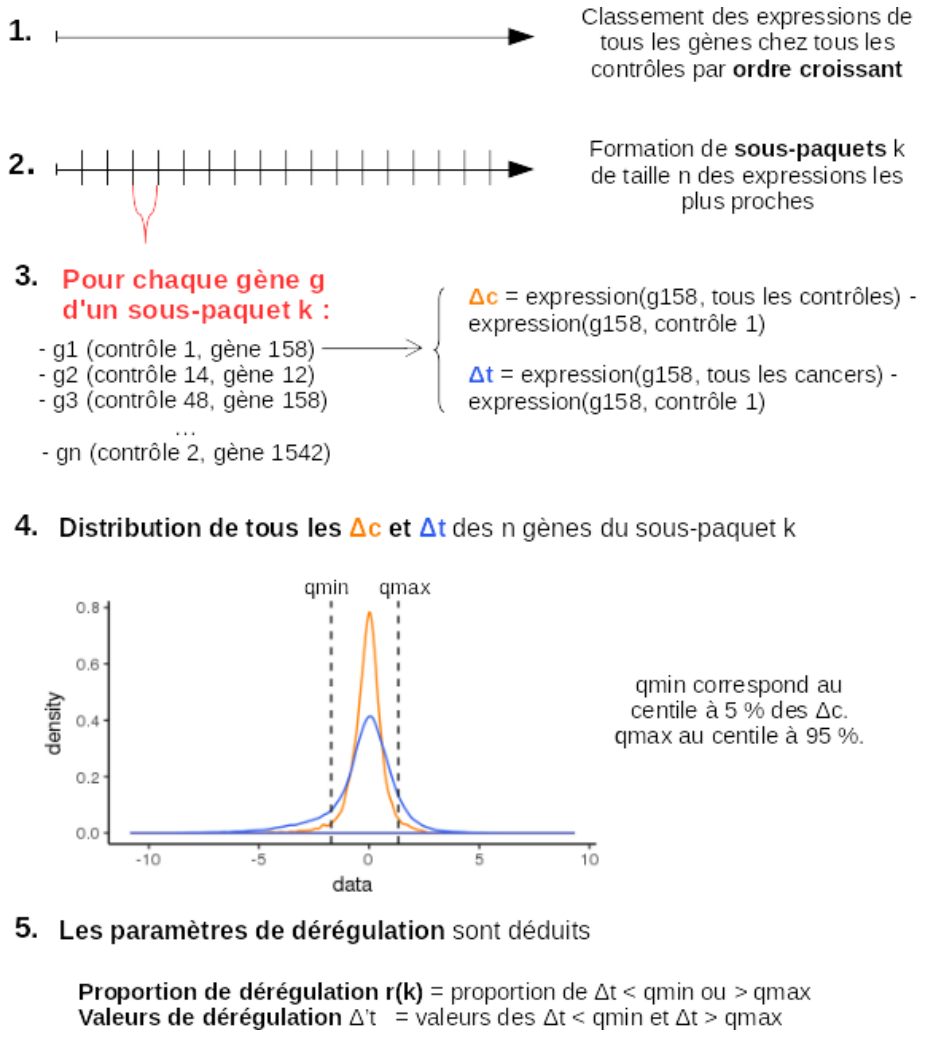


FIGURE 1.5 – **Les 5 étapes de la simulation utilisée dans Penda.** 1. Les gènes des contrôles sont triés par expression croissante, tous échantillons confondus. 2. Ils sont ensuite regroupés en paquets de taille fixe n par expression proche. 3. Pour chaque paquet k, $\Delta c(k)$ et $\Delta t(k)$ sont calculées, correspondant pour chaque gène du paquet à la différence entre son expression dans les autres contrôles et tumeurs et son expression dans le groupe. 4. La distribution des $\Delta c(k)$ et $\Delta t(k)$ permet de fixer des paramètres de dérégulation (voir texte). 5. Ces paramètres sont stockés pour chaque groupe et serviront de référence pour déréguler les gènes.

probabilité d'être dérégulé, et les valeurs possibles de cette dérégulation. Ces paramètres de dérégulation serviront ensuite à simuler des tumeurs à partir d'échantillons contrôles : les échantillons contrôles qui serviront de base aux échantillons simulés sont donc retirés du jeu de données.

Dans un premier temps, tous les autres échantillons contrôles sont mélangés, et l'ensemble des valeurs d'expressions de tous leurs gènes est classé par ordre croissant (figure 1.5, étape 1). Ces expressions sont ensuite regroupées par paquets de taille fixe n (100 par défaut), en suivant l'ordre croissant, et en regroupant donc des valeurs d'expression proches (figure 1.5, étape 2). La formation de ces paquets permet de gagner en puissance statistique par rapport à l'utilisation des distribution des gènes un par un, en particulier quand les jeux de contrôles comportent peu d'échantillons.

Pour chaque paquet k, et pour chaque valeur d'expression le composant, on calcule alors deux métriques. $\Delta c(k)$ correspond à la différence entre l'expression du gène dans tous les autres contrôles et la valeur d'expression du gène dans le paquet. $\Delta t(k)$ est l'équivalent en comparant cette fois à la valeur du gène dans toutes les tumeurs (figure 1.5, étape 3).

La distribution de tous les $\Delta c(k)$ permet ensuite de fixer des seuils : les valeurs entre le centile à 5% et celui à 95% sont considérées comme des différences d'expression normales. Les valeurs de $\Delta t(k)$ sortant de ces bornes sont considérées comme des dérégulations dues au cancer, elles sont conservées dans une liste $\Delta t'(k)$ pour être utilisées dans la prochaine étape des simulations. On définit le ratio $r(k)$ comme nombre d'éléments dans $\Delta t'(k)$ divisé par le nombre dans $\Delta t(k)$. Il correspond à la probabilité qu'un gène ayant une expression dans un contrôle comprise dans le groupe k soit dérégulé dans une tumeur (figure 1.5, étape 4).

À l'issue de cette étape, il y a donc un ensemble de sous-groupes, chacun défini par des valeurs minimales et maximales d'expression dans ce groupe, une probabilité $r(k)$ de dérégulation dans le cancer, et un ensemble $\Delta t'(k)$ de valeurs de dérégulation possible (figure 1.5, étape 5).

Simulation de la dérégulation

La dérégulation est simulée en partant des lignées contrôles qui n'ont pas été utilisées pour inférer les paramètres de dérégulation. Pour chaque gène, la probabilité de dérégulation $r(k)$ est tirée du groupe k correspondant à sa valeur d'expression. Si le gène est dérégulé, la valeur de dérégulation est tirée parmi les $\Delta t'(k)$, avec comme condition que l'expression finale ne doit pas être inférieure à 0. En résumé, pour le gène g d'expression $expr(g)$, la valeur de dérégulation $dereg(g)$ est choisie de la façon suivante :

$$dereg(g) \in \Delta t'(k), \text{tel que } expr(g) + dereg(g) > 0$$

La simulation de la dérégulation se veut donc réaliste et adaptée aux données, puisque les paramètres sont fixés en fonction de la valeur d'expression du gène et des observations sur les données soumises en entrée. Par ailleurs, le principe des simulations, basé sur le regroupement des gènes d'expression proche, n'est utilisé par aucune des méthodes d'analyse différentielle utilisées ensuite, ce qui nous permet donc de les comparer sur cette base sans introduire de biais.

Métriques d'évaluation

Les simulations permettent d'évaluer les résultats des méthodes d'analyse différentielle en comparant les gènes inférés comme dérégulés à ceux réellement dérégulés à travers une matrice de confusion classique : vrais positifs (True Positive - TP) les gènes détectés dérégulés correctement, vrais négatifs (True Negative, TN) les gènes détectés non-dérégulés correctement, faux positifs (False Positive, FP) les gènes détectés dérégulés à tort, et faux négatifs (False Negative, FN) les gènes dont la dérégulation n'est pas détectée.

Ces valeurs permettent de calculer trois métriques qui sont utilisées pour tracer des courbes ROC (aussi appelées fonctions d'efficacité du récepteur) :

$$\text{Le taux de vrais positifs : } TPR = \frac{TP}{TP + FN}$$

$$\text{Le taux de faux positifs : } FPR = \frac{FP}{FP + TN}$$

Le taux de fausse découverte : $FDR = \frac{FP}{TP + FP}$

1.2 Paramètres Penda

À l'aide de ces simulations, nous avons pu tester l'influence des différents paramètres de la méthode Penda, mais aussi proposer des fonctions permettant aux utilisateurs de fixer les paramètres optimaux adaptés à leurs données. Dans cette partie, je vais détailler l'impact des différents paramètres de la méthode Penda sur une "simulation type" basée sur les cancers du poumon non à petites cellules du TCGA (voir partie 4 de l'introduction), puis je vais également expliquer comment l'utilisateur peut choisir les paramètres de la méthode Penda pour optimiser les résultats obtenus sur ses propres données.

Plus concrètement, en faisant varier le seuil du test Penda, nous pouvons réaliser des courbes ROC qui permettent d'étudier la variation des différents paramètres de la méthode, en visualisant leur impact sur le TPR et le FPR. Ces courbes permettent également de choisir le paramètre optimal, soit visuellement comme ce sera le cas dans cette partie, soit en calculant des métriques comme l'aire sous la courbe.

1.2.1 Simulations LUAD et LUSC

Dix tumeurs ont été simulées à partir des données RNA-seq de la base de données TCGA des deux principaux types de cancers non à petites cellules : LUAD (adénocarcinome) et LUSC (carcinome épidermoïde). Seules les tumeurs primaires ont été conservées, les réplicats ont également été retirés, ce qui nous donne une cohorte de 1026 échantillons : 455 tumeurs LUAD, 473 tumeurs LUSC et 98 échantillons sains. Nous avons sélectionné les gènes codants pour des protéines et n'ayant pas une valeur d'expression nulle dans tous les échantillons, ce qui nous permet de conserver 18 143 gènes.

Les dix tumeurs ont été simulées à partir d'échantillons sains comme décrit dans la section 1.1.4, en se basant sur les 88 échantillons sains restants et les

tumeurs LUAD et LUSC mélangées. On obtient environ 30% de gènes dérégulés (moyenne des $r(k)$).

1.2.2 Impact de la taille l des listes L et H

Le paramètre l correspond au nombre de gènes conservés dans les listes L et H. Fixer le nombre de gènes dans les listes a plusieurs avantages : normaliser l'effet du seuil Penda entre les gènes (en effet, le seuil Penda est une proportion, il peut donc y avoir un biais si les tailles de listes sont très différentes entre les gènes), éviter de prendre comme référence des gènes à l'expression très éloignée du gène étudié (seuls les gènes dont l'expression est la plus proche du gène g étudié sont conservés, ils constituent donc des références plus sensibles) et également réduire les temps d'analyse et les ressources nécessaires pour les calculs. Dans les simulations basées sur LUAD et LUSC, on peut voir que la taille de liste permettant de maximiser le TPR en minimisant le FPR est une taille de 30 (figure 1.6, panneau a). Des listes de références trop petites ($l < 10$) ou trop grandes ($l > 300$) ont toutes les deux un impact négatif sur les résultats en diminuant le nombre de vrais positifs. Dans le cas d'une liste de départ très petite, la taille diminue encore au cours des itérations (voir partie 1.1.3) ce qui peut entraîner l'utilisation de la méthode du centile de manière plus fréquente, de plus une petite liste est très sensible aux changements de proportions du seuil (le pas entre 60% et 50% est plus petit dans un échantillon de 10 que dans un échantillon de 100). Dans le cas d'une liste de gènes très grande, c'est l'effet inverse, il y a une perte de sensibilité car les gènes utilisés comme référence ont une expression trop éloignée pour détecter de faibles variations d'expression. Dans la suite des analyses, c'est $l = 30$ qui est choisi.

1.2.3 Impact des propriétés des données

Dans cette partie, nous allons traiter de la taille des données, c'est-à-dire du nombre de gènes et d'échantillons étudiés, qui sont des paramètres importants car ils influent sur la qualité des distributions et des rangs qui sont utilisés à chaque étape (simulation, référence L et H dans les contrôles, test...). Nous allons

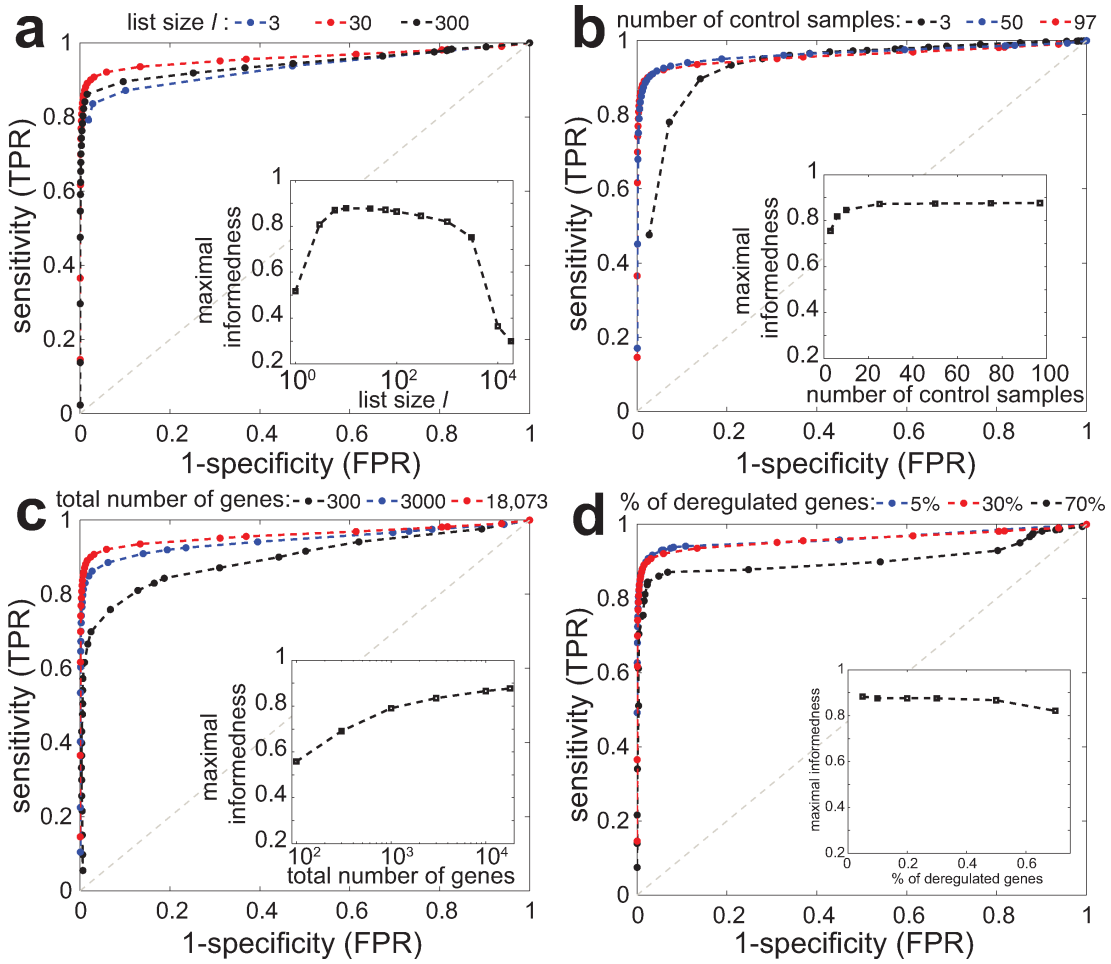


FIGURE 1.6 – **Impact de différents paramètres sur les résultats de la méthode Penda.** Les courbes ROC sont calculées pour plusieurs valeurs des paramètres de la méthode Penda en faisant varier le seuil de la méthode. L'encart montre l'information maximale (maximal informedness) qui représente la valeur maximale de la différence TPR-FPR pour chaque courbe ROC. Paramètres étudiés : (a) la taille des listes L et H, (b) le nombre d'échantillons contrôles, (c) le nombre de gènes, (d) la proportion des gènes dérégulés lors de la simulation. Détail des simulations partie 1.2.1. Figure issue du papier.

également étudier l'impact du taux moyen de dérégulation des lignées tumorales simulées.

Le premier paramètre à être testé est le nombre d'échantillons contrôles utilisés pour constituer les listes L et H (figure 1.6, panneau b). La méthode est beaucoup moins efficace pour 3 contrôles, mais donne des résultats semblables entre 50 et 97 contrôles. La baisse de précision pour 3 contrôles est prévisible puisque le rang relatif entre les gènes est beaucoup plus fiable quand le nombre d'échantillons est grand. Le nombre de contrôles a également un fort impact quand on utilise la méthode des centiles à la place de la méthode Penda, la distribution de l'expression d'un gène sur 3 patients a très peu de sens statistique. Cependant, la méthode reste assez robuste et on obtient dans tous les cas un rapport TPR / FPR satisfaisant, même pour un nombre limité de contrôles : pour 3 contrôles on atteint ainsi presque 0,8 de TPR pour une FPR inférieur à 0,1.

Nous avons également testé l'impact du nombre de gènes, qui joue un rôle crucial au moment de l'établissement des listes L et H. Cette fois, plus il y a de gènes plus la méthode Penda donne de bons résultats (figure 1.6, panneau c). Effectivement, plus il y a de gènes au total, plus les listes L et H pourront inclure des gènes proches du gène étudié et plus elles seront définies précisément et constitueront de bonnes références. C'est ce qu'on observe à partir de 3 000 gènes.

Enfin, nous avons comparé l'efficacité de la méthode Penda pour différentes proportions de gènes dérégulés (figure 1.6, panneau d). Les résultats restent très stables et très efficaces pour 5 ou 30% de gènes dérégulés, puis diminuent légèrement au delà de 60% de dérégulation. Ces résultats sont logiques car une grande proportion de gènes dérégulés influe sur la pertinence des listes de référence L et H : elles seront rapidement "vidées" de leurs éléments lors des itérations successives de Penda (voir partie 1.1.3), et la méthode des centiles, moins efficace (voir partie 1.4.2 ci-dessous), sera plus souvent utilisée.

Globalement, Penda est robuste aux variations des données, tant au niveau du nombre d'échantillons, que du nombre total de gènes ou de la proportion globale de gènes dérégulés.

1.3 Implémentation du package R

La méthode Penda a été implémentée en R et est disponible sur Github. Elle est composée de nombreuses fonctions, permettant de réaliser les différentes étapes du test.

1.3.1 Pré-traitement

L'étape de pré-traitement, réalisée par la fonction *penda : :make_dataset*, inclut des fonctions de tri facultatives et une étape nécessaire pour transformer les données (RNA-seq des échantillons contrôles et tumoraux) au format "penda dataset".

Expression faible

Les gènes dont l'expression est très faible dans la quasi-totalité des échantillons peuvent être assimilés à du bruit, leur information a de fortes chances de ne pas être intéressante et ils peuvent poser problème au moment du calcul des listes de référence L et H.

L'option "detectlowvalue" de la fonction *penda : :make_dataset* permet donc de supprimer les gènes sous un certain seuil d'expression dans plus d'une certaine proportion des échantillons. La proportion des échantillons est un paramètre entré par l'utilisateur, par défaut elle est de 99%. Cette proportion doit être respectée indépendamment dans les échantillons contrôles et dans les échantillons tumoraux, afin de ne pas perdre l'information d'une sur-expression d'un gène dans certaines tumeurs.

La valeur du seuil de faible expression peut au choix être fixée directement par l'utilisateur, ou calculée de manière automatique par Penda de deux façons. Si le paramètre bimod est FALSE, on se place dans le cadre d'une distribution d'expression suivant une loi normale, la valeur seuil de faible expression est donc fixée simplement par le centile à 10% de toutes les valeurs d'expression (voir la figure 1.7, partie A). Si le paramètre bimod est TRUE, on se place dans le cadre d'une distribution d'expression bimodale. Dans ce cas là, la distribution des expressions des gènes est composée de deux lois normales : une correspondant

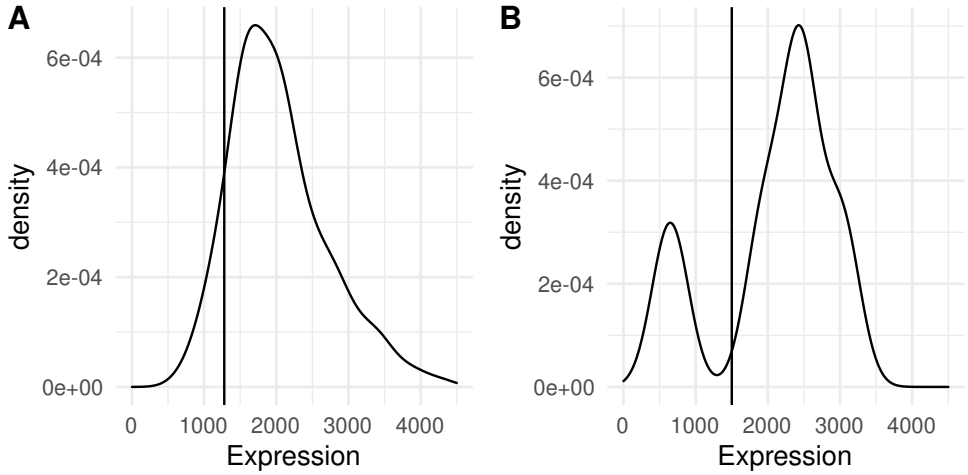


FIGURE 1.7 – **Choix du seuil de faible expression en fonction de la distribution des données.** En A, les données suivent une distribution gaussienne, le seuil de faible valeur est fixé par défaut au quantile à 10% des données (barre verticale). En B, les données suivent une distribution bimodale, le seuil de faible valeur est fixé pour couper entre les deux gaussiennes (barre verticale).

aux "faibles valeurs" et une considérée comme étant des valeurs d'expression significatives (voir la figure 1.7, partie B). On utilise la fonction *normalmixEM* du package R mixtools [57] pour déterminer le seuil entre les deux courbes par un algorithme d'espérance-maximisation.

Valeurs NA

Les valeurs NA (valeur inconnue) sont assez courantes dans les données biologiques et proviennent souvent d'erreurs lors du séquençage. Il arrive parfois que certains gènes ou que certains échantillons aient beaucoup de valeurs inconnues, ce qui peut poser problème au moment de l'analyse. L'option "detectNA" de la fonction *penda : :make_dataset* permet de supprimer les échantillons et les gènes qui ont une proportion de valeurs inconnues supérieures à un seuil, fixé par l'utilisateur ou par défaut à 99%.

Format de données "penda dataset"

Le format de données utilisé par la méthode Penda est facile à reproduire sans passer par la fonction de pré-traitement si ça se révèle nécessaire. Il suffit de trier les gènes par leur médiane dans les échantillons contrôles, puis de constituer une liste de deux matrices, data_ctrl avec les échantillons contrôles et data_case avec les échantillons tumoraux. La fonction de pré-traitement permet aussi de renvoyer un résumé permettant ensuite de générer automatiquement un paragraphe de matériel et méthodes sur les paramètres utilisés et les tris effectués.

1.3.2 Listes L et H

Le calcul des listes de références dans les contrôles est exécuté par la fonction *compute_lower_and_higher_lists*. Cette fonction est implémentée en C++ à travers le package Rcpp [58], ce qui permet de diminuer significativement les temps d'exécution. Elle prend en paramètre le nombre de gènes à conserver dans chaque liste (30 par défaut) et la proportion de contrôles dans lesquels le rang doit être stable (99% par défaut).

1.3.3 Test Penda

L'étape du test Penda est implémentée en trois fonctions imbriquées.

La fonction *regulation_test* effectue le test au niveau d'un seul gène, dans une seule tumeur, pour une seule itération. Elle renvoie -1 si le gène est sous-exprimé, 0 s'il n'a pas bougé et 1 s'il est sur-exprimé.

Au niveau suivant, *sample_test* effectue le test pour tous les gènes d'un échantillon. Les gènes dérégulés à l'itération n-1 sont retirés des listes L et H de l'itération n, jusqu'à stabilisation des gènes détectés dérégulés. La fonction renvoie deux vecteurs true/false correspondant aux gènes sous-exprimés et aux gènes sur-exprimés.

Enfin au niveau utilisateur, c'est la fonction *penda_test* qui est utilisée. Elle exécute la fonction *sample_test* pour tous les échantillons, et renvoie les matrices de résultats Penda : deux matrices avec en colonne les échantillons, en ligne les

gènes, et respectivement 0 ou 1 si le gène est dérégulé. La matrice D contient l'information des gènes sous-exprimés, et la matrice U l'information des gènes sur-exprimés

1.3.4 Simulation et tests de paramètres

Simulations

La simulation de la dérégulation des gènes inspirée des données réelles est effectuée avec la fonction *complex_simulation*. La fonction *results_simulation* permet de calculer simplement les différentes métriques d'erreur, et la fonction *draw_results* d'en tracer l'histogramme. À partir de celles-ci, d'autres fonctions permettent de choisir les paramètres de la méthode des centiles et le seuil Penda à partir des simulations.

Choix des paramètres de la méthode des centiles

Les paramètres de la méthode du centile ne sont pas les plus importants à l'échelle de l'analyse globale, car elle n'est pas censée être utilisée fréquemment. Cependant, ils peuvent avoir un impact significatif dans le cas d'usages un peu extrêmes de Penda, par exemple sur des données très différentes des contrôles.

Des fonctions sont implémentées dans le package Penda pour permettre facilement de varier les seuils et d'obtenir les courbes ROC associées (voir la figure 1.8). La fonction *penda : : choose_quantile* fait tourner la méthode des centiles pour toutes les combinaisons de factor_values et quantile_values en entrée, et calcule pour chaque combinaison les métriques FPR, FDR et TPR. Dans un second temps, la fonction *penda : : select_quantile_param* permet de sélectionner directement les paramètres maximisant le TPR pour un FDR maximal fixé par l'utilisateur.

Choix du Seuil du test (threshold)

Le seuil est le paramètre le plus important de la méthode. Avec un grand seuil, les changements de rang doivent être plus importants, avec un petit seuil la méthode sera plus sensible à une dérégulation faible.

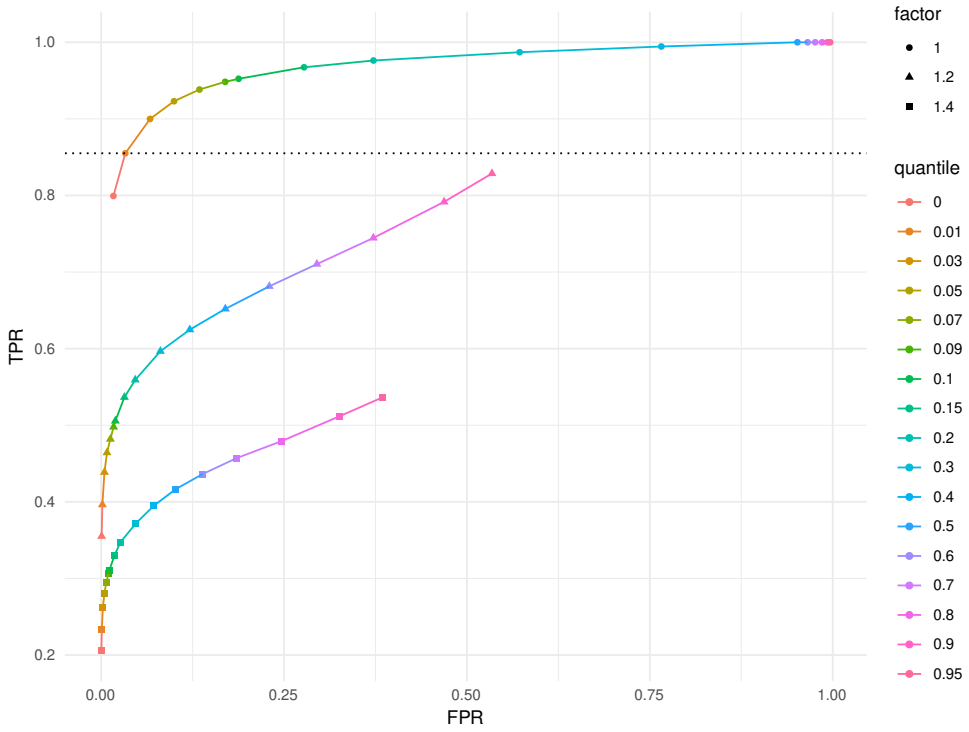


FIGURE 1.8 – **Variation des paramètres de la méthode des centiles.** Les courbes ROC sont obtenues par la fonction *choose_quantile* du package Penda, en faisant le test à de multiples reprises en variant le facteur et le quantile. Ces valeurs permettent de choisir les meilleurs paramètres, ici pour un FDR maximal de 0,1 on obtient factor = 1 et quantile = 0,01.

La principale conséquence est que le seuil doit absolument être choisi en fonction des données, des fonctions semblables à celles des paramètres de la méthode des centiles sont donc implémentées. La fonction *penda : : choose_threshold* permet de faire tourner Penda pour une liste de seuils fixée, et de calculer pour chacun les métriques FPR, FDR et TPR (voir la figure 1.9). Dans un second temps, la fonction *penda : : select_quantile_param* permet de sélectionner automatiquement le seuil maximisant le TPR pour un FDR maximal fixé par l'utilisateur. Cette deuxième partie peut être remplacée par d'autres critères, comme un choix visuel sur une courbe ROC.

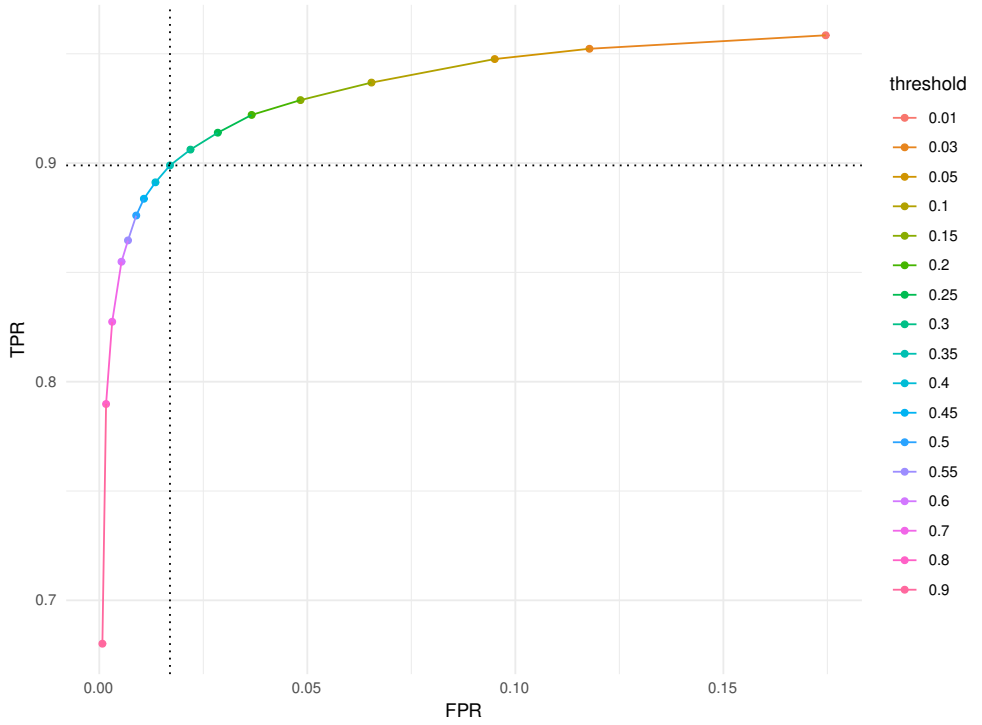


FIGURE 1.9 – **Variation des résultats de Penda sur une simulation pour différents seuils du test.** Cette courbe ROC est obtenue par la fonction *choose_threshold* du package Penda, en faisant le test à de multiples reprises en variant le seuil du test. Ces résultats permettent de choisir le meilleur paramètres : ici pour un FDR maximal de 0,05 on obtient le seuil de 0,35.

1.3.5 Vignettes et documentation

Le package Penda est entièrement documenté, deux vignettes sont disponibles pour permettre aux utilisateurs de prendre en main simplement la méthode. La vignette "simulation" (en annexe 2, voir section 7.2.4) permet de réaliser l'étape de simulation réaliste sur ses propres données, puis de choisir les paramètres de la méthode. La vignette "penda" (en annexe 3, voir section 7.2.4) permet ensuite d'appliquer le test en fonction des paramètres choisis.

1.4 Comparaison avec les méthodes existantes

À partir des simulations (voir partie 1.2.1), nous avons comparé Penda avec les méthodes d'analyses individuelles existantes : les deux versions de RankComp [52] [53], la méthode des centiles, et DESeq2 [41] qui est développée pour les analyses populationnelles mais qui est utilisé ici pour l'analyse individuelle.

1.4.1 Utilisation des méthodes

Les versions 1 et 2 de Rankcomp ont été lancées à travers le logiciel en C "Relative Expression Ordering Analysis (REOA)" développé par les créateurs de la méthode et disponible sur leur Github, avec les paramètres par défaut.

DESeq2 a été lancé en R avec les paramètres par défaut, en comparant une seule tumeur simulée aux échantillons contrôles. Étant donné l'absence de réplicats pour la tumeur, on peut ici utiliser les valeurs de dépendances variance / moyenne calculées pour les contrôles comme paramètres pour la tumeur [59].

1.4.2 Résultats

Les méthodes ont été testées pour 3 conditions expérimentales, sur les 97 échantillons contrôles avec ou sans normalisations, et sur un sous-set de 10 contrôles (figure 1.10). L'absence de normalisation a été simulée en multipliant les valeurs de comptage du RNA-seq des échantillons contrôles et tumoraux par un facteur aléatoire compris entre 1 et 5. Tous les gènes d'un même échantillon ont été multipliés par le même facteur.

Pour chacune de ces conditions, c'est Penda qui permet d'obtenir le meilleur ratio entre TPR et FPR, en particulier dans la zone avec un faible taux de faux positifs. De manière surprenante, les méthodes RankComp obtiennent de meilleurs résultats pour 10 contrôles (partie b de la figure) que pour 97 contrôles (partie a de la figure), sans dépasser les résultats de Penda.

Enfin, sur des données non-normalisées avec un effet de lot (batch effect) simulé (partie c de la figure), l'efficacité de la méthode du centile se retrouve fortement diminuée. Les méthodes basées sur les rangs restent robustes, de même

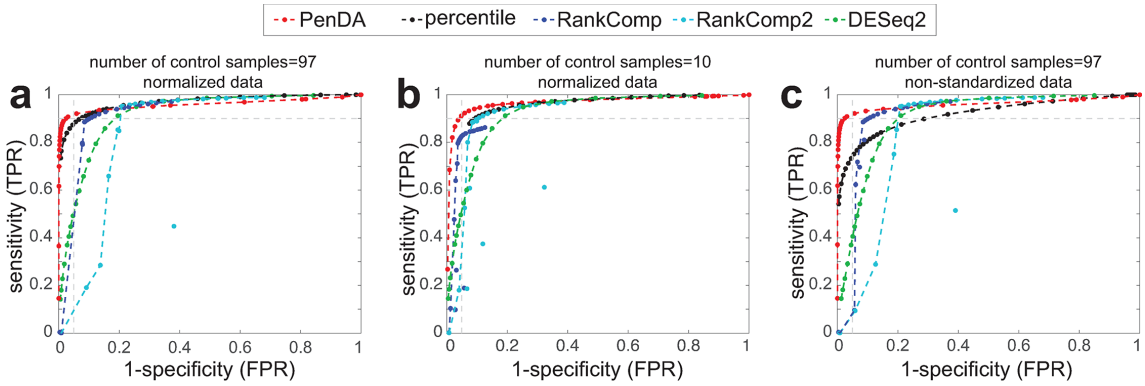


FIGURE 1.10 – **Comparaison entre les méthodes d'analyses différentielles personnalisées.** Les 5 méthodes ont été exécutées sur des simulations avec différentes conditions expérimentales : a. 97 contrôles sur des données normalisées, b. 10 contrôles sur des données normalisées, c. 97 contrôles sur des données non-normalisées. Les seuils de détection des gènes dérégulés ont été variés dans chaque méthode afin de tracer les courbes ROC (le seuil pour Penda, la valeur du centile pour la méthode percentile, le niveau de FDR pour RankComp et le log2 du fold-change pour DESeq2). Figure issue du papier.

que DESeq2 qui est lancé avec sa propre routine de normalisation intégrée.

Il est à noter ici que les faibles résultats de DESeq2 s'expliquent en grande partie par le fait qu'il n'est pas développé pour cet usage d'analyse individuelle et que ses statistiques ne sont pas fiables pour comparer une seule tumeur à une cohorte de contrôles.

1.5 Conclusion

La méthode Penda a démontré de très bons résultats sur les données simulées de manière réaliste, et se révèle ici plus efficace que les méthodes d'analyse différentielle individuelle pré-existantes. Les vignettes et l'implémentation sous la forme d'un package R la rendent facilement accessible et réutilisable par la communauté scientifique.

Chapitre 2

Applications biologiques de Penda

Dans la précédente partie, nous avons montré que la méthode Penda permettait d'obtenir de très bons résultats sur des simulations, et qu'elle se plaçait au-dessus des autres méthodes d'analyse différentielle individuelle. Dans ce chapitre, nous allons explorer plusieurs applications de Penda. D'abord l'application détaillée aux cancers non à petites cellules LUAD et LUSC à travers les cohortes publiques TCGA et la cohorte Brambilla du CHU de Grenoble publiée dans le papier, qui a permis de trouver de nouveaux biomarqueurs de survie et de classification. Ensuite, une application assez générale à différents types de cancers TCGA, afin de visualiser les variations entre eux. Et enfin, l'exploration des limites de la méthode Penda avec son application sur des cultures cellulaires de glioblastomes ne comportant pas d'échantillons contrôles.

2.1 Application aux cancers du poumon

2.1.1 Matériel et méthodes

La méthode Penda a été appliquée aux cohortes LUAD [19] et LUSC [60] du TCGA avec les paramètres choisis sur des simulations, comme décrit dans la section 1.3.4 : listes L et H de 30 gènes, centile de 0,02 et facteur de 1,2 pour la méthode des centiles et seuil de 0,3 pour Penda. Il y avait 928 échantillons tumoraux (455 LUAD et 473 LUSC) pour 98 échantillons contrôles. 18 143 gènes

ont été analysés.

2.1.2 Résultats bruts de Penda

Les résultats obtenus par Penda sont représentés figure 2.1. On obtient un nombre de gènes dérégulés différent dans chaque tumeur, cette proportion variant de 3 à 61% dans LUAD, avec en moyenne 33% de gènes détectés dérégulés, et de 0,4 à 55% dans LUSC avec en moyenne 45% de dérégulation (panneau a). Globalement, on observe plus de gènes sous-exprimés que de gènes sur-exprimés.

Les gènes ayant un profil de dérégulation différent entre LUAD et LUSC sont intéressants car ils montrent que notre méthode infère des dérégulations spécifiques à chaque type de cancer. Avec un test statistique (test z à deux proportions), on en retrouve 5 346 significativement différemment sous-exprimés entre les deux cohortes (panneau b) et 5616 pour la sur-expression (panneau c).

Afin de valider biologiquement ces gènes différentiellement dérégulés, on s'est intéressés à deux mécanismes biologiques dont la dérégulation est connue dans ces cancers du poumon : la différenciation squameuse pour LUSC [60] et les récepteurs tyrosine kinase fréquemment dérégulés dans LUAD [19] (les ronds et carrés sur les figures 2.1 b et c). Nous avons retrouvé les mêmes biomarqueurs dérégulés que ceux identifiés dans de précédentes analyses au niveau populationnel [61] [62] : pour LUSC une sur-expression des gènes SOX2 [63] et TP63 [64], pour LUAD le gène ERBB2 [65].

2.1.3 Analyse des profils de dérégulation

Les dérégulations des gènes stables entre les tumeurs sont intéressantes ; les gènes constamment dérégulés à travers les tumeurs peuvent être impliqués dans la tumorigenèse ou dans la propagation tumorale, alors que les gènes qui ne sont jamais dérégulés peuvent être essentiels au développement cancéreux.

Pour chaque gène, nous avons donc regardé la proportion de tumeurs dans lesquelles il était sur- ou sous-exprimé (figure 2.2, panneaux a et b). La première chose qu'on observe, c'est que la majorité des gènes sont dérégulés dans une seule direction au sein d'un type tumoral, sur- ou sous-expression mais rarement les deux types de dérégulation dans des tumeurs différentes. On peut noter que

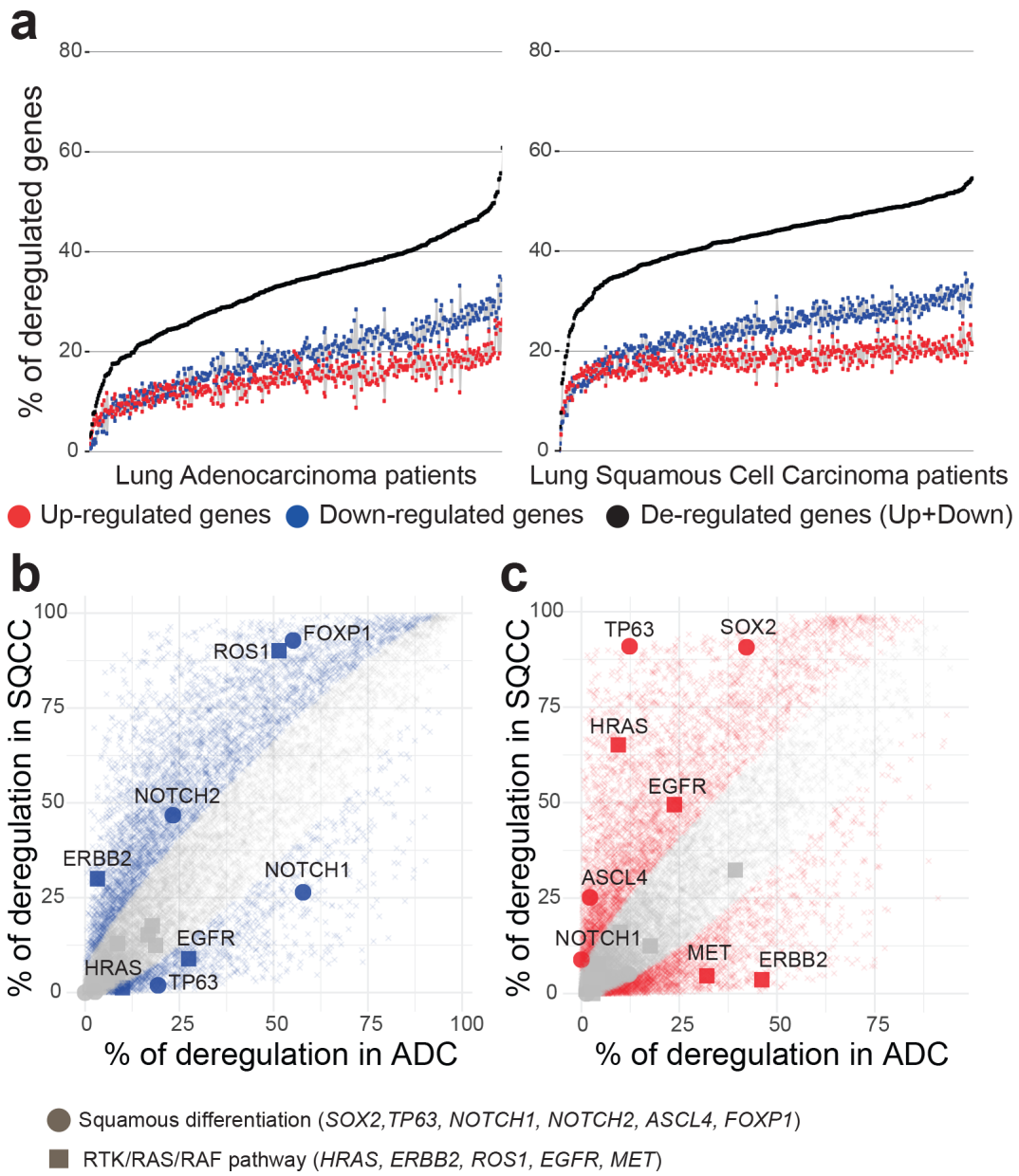


FIGURE 2.1 – **Représentation des résultats de la méthode Penda sur les cohortes LUAD et LUSC du TCGA.** En (a) on voit la proportion de gènes dérégulés dans les patients, pour LUAD à gauche et pour LUSC à droite, en rouge pour la sur-expression, en bleu pour la sous-expression, et en noir pour le total. En (b) et (c) on voit la proportion de dérégulation de chaque gène entre les deux type de cancers, avec LUAD en abscisse et LUSC en ordonnée, respectivement pour la sous et la sur-expression. Les points colorés sont significativement différents entre les deux types de cancer. Figure issue du papier.

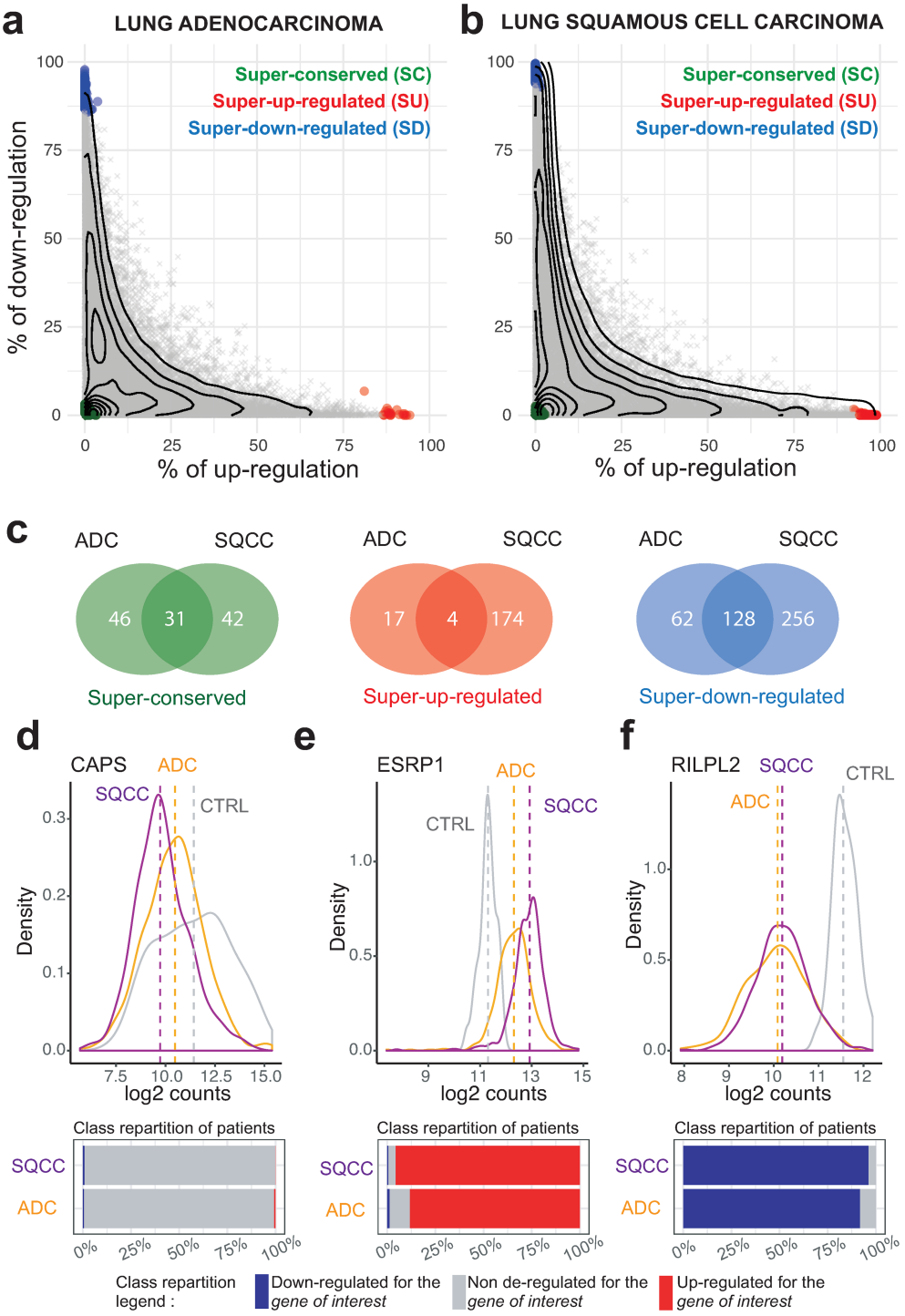


FIGURE 2.2 – **Les différents profils de dérégulation.** En a et b, on voit pour chaque gène la proportion de tumeurs où il est sous ou sur-exprimé, à gauche pour LUAD et à droite pour LUSC. Les profils extrêmes : gènes jamais dérégulés (verts), toujours sur-exprimés (rouges) ou toujours sous-exprimés (bleus) sont particulièrement intéressants, leur intersection entre les deux cancers est représentée en c. En d, e et f, les distributions d'expression dans les différentes cohortes de trois de ces gènes sont détaillées, CAPS (d) n'est jamais dérégulé, ESRP1 (e) est sur-exprimé dans la majorité des tumeurs, et RILPL2 (f) est sous-exprimé dans la majorité des tumeurs. Figure issue du papier.

le sens de dérégulation n'est pas systématiquement conservé entre les tumeurs LUAD et LUSC.

À partir d'un test de Student entre la dérégulation moyenne de chaque gène par rapport à la dérégulation moyenne de tous les gènes, nous avons identifié les gènes appartenant aux trois types de catégories extrêmes : super-conservés (SC) en vert, super-sur-exprimés (SU) en rouge et super-sous-exprimés (SD) en bleu. Ces gènes ne sont pas forcément communs aux deux types de cancer (panneau c), on en retrouve une grande partie spécifique à chacune des deux histologies. Trois de ces gènes ont été étudiés plus en détail comme exemple.

CAPS est un gène SC (panneau d), il code la calcyphosine, une protéine de liaison avec le calcium. Quand on regarde uniquement la distribution de l'expression normalisée du gène dans les deux types de tumeurs et dans les contrôles, les moyennes semblent différentes, pourtant le gène n'est jamais détecté dérégulé par Penda à l'échelle individuelle. ESRP1 est un gène SU (panneau e), c'est un régulateur d'épissage spécifique au type cellulaire épithélial qui a déjà été identifié comme étant sur-exprimé dans d'autres cancers [66], et même comme marqueur de mauvais pronostic pour le cancer de la prostate [67]. De son côté, RILPL2 (panneau f) est un gène SD qui code une protéine d'interaction avec le lysosome. Dans le cancer du sein, sa sur-expression est associée à une diminution de la prolifération tumorale et des métastases [68]. Quand on regarde les distributions des expression des gènes SU et SD, elles se chevauchent entre tumeurs et contrôles, pourtant la tendance de dérégulation globale est très nette dans les résultats individuels calculés par Penda.

Ces résultats montrent de nouveau une cohérence entre les gènes détectés par Penda et la littérature, mais aussi que l'information individuelle apportée est importante même quand la dérégulation est présente dans toutes les tumeurs, car les approches populationnelles ne permettent pas d'avoir l'information du caractère systématique de la dérégulation ou de la non-dérégulation.

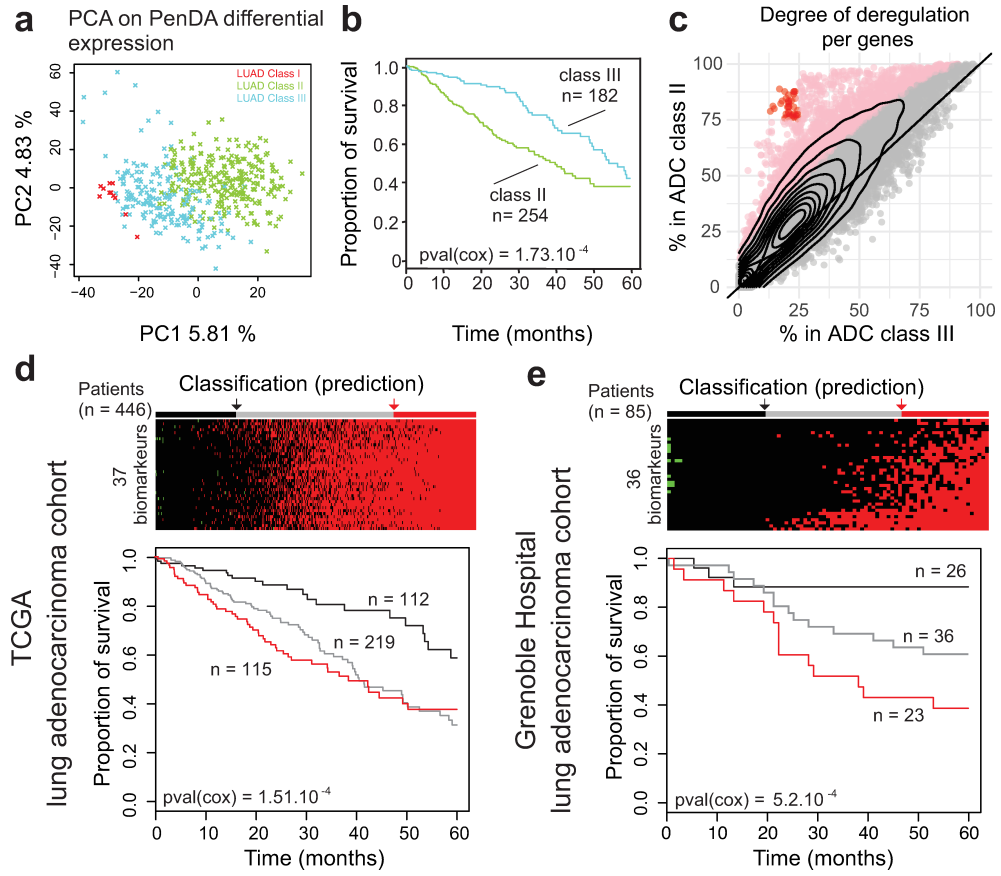


FIGURE 2.3 – **Nouvelle classification et nouveaux biomarqueurs LUAD.** En a, le scree plot de l'analyse en composantes principales des résultats Penda obtenus sur les cancers LUAD du TCGA. Les couleurs sont les trois classes séparées par clustering hiérarchique. En b, les courbes de survie des deux principaux sous-types. En c, le nuage de points des dérégulations par gène entre les deux principaux sous-types, les points rouges sont les 37 gènes fortement dérégulés dans la classe II (> 75%) et faiblement dérégulés dans la classe III (<25%). En d, les 37 gènes sont utilisés comme biomarqueurs pour diviser les patients LUAD en trois catégories selon leur niveau de dérégulation. Les courbes de survie sont ensuite tracées pour les trois catégories. En e, même chose mais sur la cohorte indépendante du CHU de Grenoble. Figure issue du papier.

2.1.4 Sous-types et biomarqueurs LUAD

Une première étape de classification hiérarchique des cancers LUAD en fonction de leurs dérégulations Penda a permis d'obtenir trois sous-classes. Ces sous-classes sont également bien séparées par les deux premières classes d'analyse en composante principale (figure 2.3, panneau a). De façon intéressante, cette classification ne correspond pas à celle habituellement décrite dans la littérature, ni à des paramètres cliniques comme le stade tumoral. On choisit de particulièrement se pencher sur les deux sous-classes majoritaires (II, 254 tumeurs et III, 182 tumeurs) afin d'étudier leurs caractéristiques.

Dans un premier temps, une régression de Cox est réalisée sur ces deux groupes (avec le package R *survival* [69]), et on obtient une différence de survie à 5 ans significative entre les 2 classes, la classe III ayant un meilleur pronostic (panneau b). Pour comprendre quels gènes influent sur cette différence, on regarde pour chaque gène la proportion de dérégulation dans chacune des deux classes (panneau c). En moyenne, les gènes sont beaucoup plus dérégulés dans la classe II que dans la classe III, sans que ce soit en lien avec le stade tumoral, l'âge ou le sexe des patients. En particulier, 37 gènes sont dérégulés dans plus de 75% des tumeurs de type II et dans moins de 25% des tumeurs de type III.

Ces gènes sont donc testés comme prédicteurs de la survie. Pour cela, un score est calculé en fonction des valeurs de dérégulation obtenues par Penda : 0 si le gène n'est pas dérégulé, -1 si il est sous-exprimé, +1 si il est sur-exprimé. Les quartiles de la distribution des scores dans les 446 tumeurs LUAD permettent de définir des seuils pour diviser la cohorte en trois catégories : dans la première, les patients sous le 1er quartile, soit un score inférieur à 4, dans la deuxième tous les patients entre le 1er et le 3ème quartile, soit avec un score entre 4 et et 34, et dans la troisième catégorie les patients avec un score supérieur à 34. Les régressions de Cox sur ces trois groupes sont ensuite réalisées, et montrent une différence significative de survie, avec un mauvais pronostic pour les patients ayant un score élevé (panneau d).

Dans un second temps, ces biomarqueurs sont validés sur une cohorte indépendante de 85 cancers LUAD provenant du CHU de Grenoble [22] (panneau e). Un gène a été retiré de l'analyse car l'expression des gènes de cette seconde

cohorte a été obtenue avec des micro-puces à ARN, et non du RNA-seq comme pour les données TCGA, l'expression de ce gène n'était donc pas connue. Les quartiles découpant les catégories sont cette fois de 0 et 15, mais la différence de survie prédite est toujours significative. L'analyse différentielle de Penda permet donc d'identifier des biomarqueurs robustes.

2.1.5 Conclusion

Cette analyse détaillée des cancers LUAD et LUSC par notre méthode nous permet d'affirmer que les résultats obtenus par Penda sont biologiquement pertinents. Les résultats obtenus grâce à une analyse différentielle individuelle peuvent être utilisés de plusieurs façons, ici nous avons montré que l'étude des gènes "super-conservés" ou "super-dérégulés" entre les tumeurs avait un potentiel intérêt biologique comme cibles thérapeutiques, mais aussi que les motifs de dérégulation au sein d'un sous-type cancéreux pouvait révéler de nouveaux biomarqueurs.

Il est également intéressant de noter que les résultats obtenus sur le transcriptome de la cohorte TCGA, séquencée par RNA-seq, sont reproductibles sur la cohorte du CHU obtenue avec des micro-puces. Les résultats obtenus par Penda sont donc robustes à travers différentes bases de données et technologies de séquençage.

Nous allons maintenant aborder une analyse pan-cancéreuse rapide afin de voir si Penda permet d'obtenir des résultats spécifiques aux différents types de données, avant de nous pencher sur un cas d'utilisation extrême : des données sans contrôles.

2.2 Vue d'ensemble des cancers TCGA

Afin d'avoir une idée des résultats Penda sur des cohortes différentes, nous avons fait tourner la méthode sur l'ensemble des cancers TCGA ayant plus de 10 contrôles.

	BLCA	BRCA	COAD	ESCA	HNSC	KICH	KIRC	KIRP
Nombre de contrôles	19	111	41	11	44	24	72	32
Nombre de tumeurs	408	1064	458	162	502	65	531	289
quant	0	0	0,01	0	0,01	0,09	0	0
fquant	1,4	1	1	1,4	1,2	1,2	1	1
seuil	0,9	0,4	0,35	0,9	0,9	0,7	0,3	0,35
Nb de gènes retirés	5181	5476	5615	5358	3411	5572	5668	5248

	LIHC	LUAD	LUSC	PRAD	READ	STAD	THCA	UCEC
Nombre de contrôles	50	59	49	52	10	32	58	35
Nombre de tumeurs	374	515	501	496	167	375	510	544
quant	0,05	0,1	0,05	0,01	0	0,03	0,03	0,05
fquant	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2
seuil	0,6	0,7	0,8	0,8	0,9	0,8	0,8	0,6
Nb de gènes retirés	4633	5286	5258	3865	5972	5285	6152	4893

Tableau 2.1 – **Résumé des cancers étudiés et des paramètres déterminés par la méthode Penda.** Les paramètres de la méthode des centiles quant et fquant et le seuil du test Penda ont été choisis automatiquement par les vignettes. Les gènes retirés correspondent aux gènes détectés comme faiblement exprimés dans plus de 99% des contrôles et des tumeurs. BLCA correspond au carcinome urothélial [70], BRCA au carcinome du sein [71], COAD au carcinome colorectal [72], ESCA au carcinome œsophagien [73], HNSC au carcinome à cellules squameuses de la tête et du cou [74], KICH au carcinome rénal chromophobe [75], KIRC au carcinome renal a cellule claire [75], KIRP au carcinome rénal papillaire [75], LIHC au carcinome hépatocellulaire, LUAD à l'adénocarcinome pulmonaire [19], LUSC au carcinome pulmonaire à cellules squameuses [60], PRAD à l'adénocarcinome de la prostate [76], READ à l'adénocarcinome du rectum [72], STAD à l'adénocarcinome de l'estomac [77], THCA au carcinome thyroïdien et UCEC au carcinome de l'endomètre [78].

2.2.1 Méthode

Les données RNA-seq du TCGA ont été normalisées par DEseq2 [41], puis les vignettes Penda ont été appliquées avec les paramètres par défaut. Le pré-traitement a permis de retirer les gènes dont 99% de l'expression dans les contrôles et dans les cancers étaient sous un seuil déterminé automatiquement par Penda pour une distribution bimodale (voir partie 1.3.1). Trois simulations sont ensuite réalisées avec des listes L et H de taille maximale l = 60, une taille qui permet de maximiser l'information sans être trop sensible aux différences entre les jeux de données. Les paramètres de la méthode du centile sont choisis pour maximiser le TPR pour un FDR maximal de 0,1. Le seuil du test est choisi pour maximiser le TPR pour un FDR maximal de 0,05. Le nombre d'échantillons, les paramètres sélectionnés et le nombre de gènes filtrés à l'étape de pré-traitement pour chacune des cohortes est résumé dans le tableau 2.1.

Dans un second temps, la méthode Penda a été lancée sur toutes les tumeurs avec les paramètres fixés sur les simulations. La méthode DESeq2 a également été exécutée avec les paramètres par défaut au niveau populationnel.

2.2.2 Résultats de l'analyse de dérégulation

La première chose que l'on peut remarquer dans l'analyse pan-cancéreuse de Penda est que les paramètres de la méthode du centile et le seuil Penda sélectionnés par les vignettes diffèrent beaucoup entre les types de cancer (tableau 2.1). Ce sont des résultats attendus puisque la méthode s'adapte aux données, à leur distribution et à leur hétérogénéité pour définir les paramètres.

On s'intéresse ensuite aux résultats de dérégulation sur tous les cancers calculés par Penda et DEseq (figure 2.4). Côté Penda, on voit pour chaque patient la proportion totale de gènes dérégulés (noir), la proportion de gènes sous-exprimés (bleus) et la proportion de gènes sur-exprimés (rouge). Les droites horizontales grises correspondent à la proportion de gènes dérégulés détectés par DESeq2 dans chaque cohorte, en gris foncé si on conserve les gènes avec une p-valeur inférieure à 0,05, en gris clair sous le seuil de 0,01.

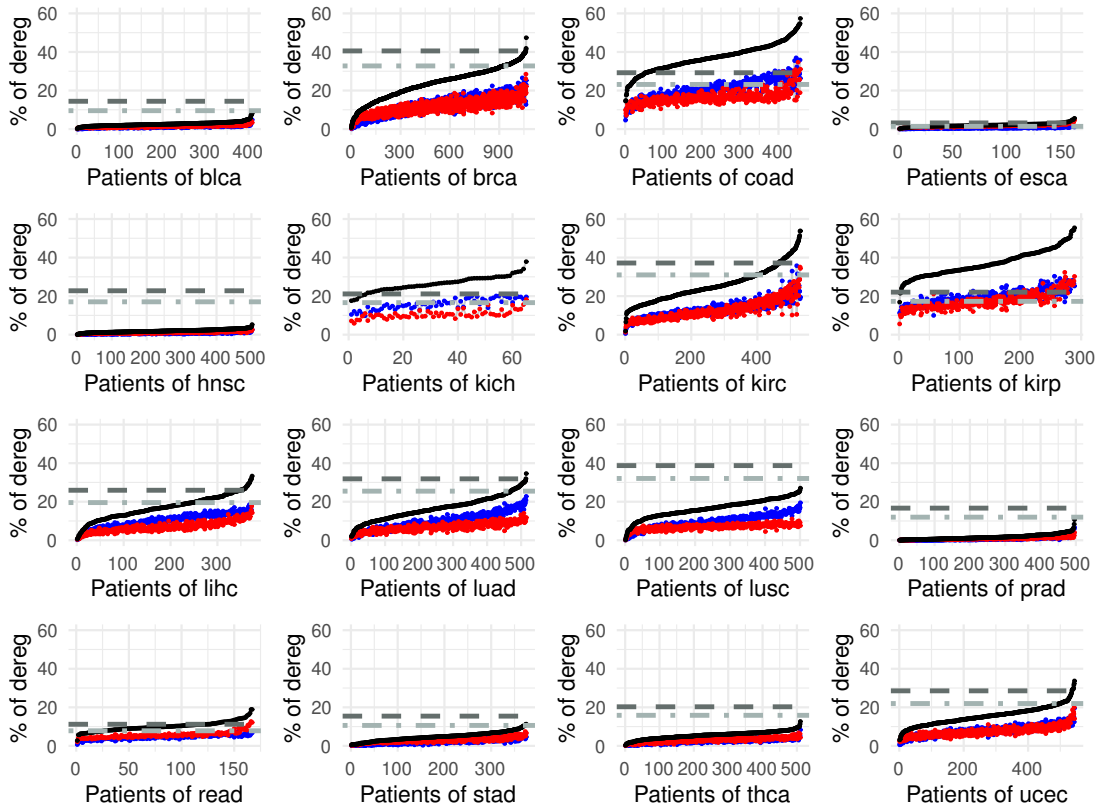


FIGURE 2.4 – **Proportions de gènes dérégulées par patients pour les cohortes TCGA.** Les patients de chaque cohorte sont ordonnés par nombre de gènes dérégulés croissant. La courbe noire correspond à la proportion totale de gènes dérégulés, la bleue aux gènes sous-exprimés et la rouge aux gènes sur-exprimés. La ligne gris foncé correspond à la proportion de gènes détectés comme dérégulés par DESeq2 avec un seuil de p-valeur de 0,05. La courbe gris clair avec un seuil de 0,01.

On observe des profils de dérégulation très variables selon les cancers. Certains ont une proportion de gènes dérégulés très faible, comme le cancer œsophagien (ESCA) avec en médiane 2% de gènes dérégulés ou le cancer rectal (READ) avec 1% de dérégulation médiane, dont les résultats sont en accord avec DESeq2, mais également comme le cancer urothélial (BLCA) avec 2,5% de dérégulation médiane pour Penda et 14% pour DESeq2. Ces résultats sont cohérents avec le seuil très restrictif choisi pour le test Penda (généralement 0,9).

D'autres cancers ont beaucoup de gènes dérégulés, là encore la consistance entre les résultats Penda et ceux de DESeq2 varie. Le cancer avec la proportion de dérégulation la plus haute est le même pour Penda et DESeq2 : le carcinome colorectal (COAD) avec 37% de médiane pour Penda, 29% au seuil de p-valeur de 0,05 pour DEseq2.

Ces résultats mériteraient des investigations plus en profondeur afin de déterminer le lien entre les dérégulations détectées par Penda et par DEseq2, ainsi que l'influence des paramètres sélectionnés sur les simulations. En effet, le nombre de gènes détectés dérégulés par Penda est directement dépendant du seuil choisi lors des simulations, et il est donc difficile de tirer des conclusions. Néanmoins, la variabilité des résultats entre les cancers est un point encourageant, les résultats Penda étant bien spécifiques aux différentes cohortes.

2.3 Application de Penda à des cultures cellulaires

Dans le cadre d'une collaboration avec Annabelle Bellasta, chercheuse à l'Inserm, nous avons essayé d'appliquer Penda dans des conditions non-optimales. L'objectif du travail de son équipe est de développer un modèle mathématique des dérégulations génétiques de plusieurs réseaux de gènes à l'échelle individuelle. Leurs données sont composées du séquençage RNA-seq de 20 échantillons, correspondant à des cultures cellulaires de différentes lignées de glioblastome, ainsi que d'une liste de gènes d'intérêt.

Le glioblastome est un cancer du cerveau extrêmement hétérogène, avec une énorme variabilité entre les patients [79]. De plus, la provenance des cellules d'origine est encore débattue au sein de la communauté scientifique [80]. La

première difficulté pour l'application de Penda était donc l'absence de d'échantillons contrôle adaptés ; en effet, leur étude ne comporte aucun contrôle et, on le comprend facilement, les échantillons sains de cerveau sont difficiles à obtenir. Le second enjeu était que jusqu'ici, tous les tests de Penda étaient effectués sur des échantillons "bulk" composés de tumeurs prélevées puis broyées, nous n'avions jamais essayé d'analyser des échantillons de culture cellulaire.

2.3.1 Choix des contrôles

Le choix des contrôles était l'étape la plus importante. Dans un premier temps, nous avons essayé de réunir plusieurs cohortes de données contrôles : deux fois 121 échantillons de cerveaux sains provenant de deux donneurs différents, référencés H0351.2001 et H0351.2002 sur le portail "Allen Brain Atlas" (https ://human.brain-map.org/), les échantillons du TCGA composés de 5 contrôles et 168 cancer [81] et 116 échantillons provenant d'une étude sur Alzheimer [82], avec dans l'idée que les dérégulations provoquées par Alzheimer sont négligeables comparées à celles médiées par le cancer.

Toutes ces données ont été normalisées avec DESeq2 [41]. Sur la figure 2.5, on peut voir la distribution des pseudo-logs (log(expression + 1)) pour chacune des lignées. On peut voir une différence de distribution entre les lignées du glioblastome (courbes noires) et les autres, notamment sur l'intensité du pic à 0 et sur le plateau autour de 5. On s'attendrait plutôt à des distributions globales d'expression semblables entre contrôles et tumeurs, comme c'est par exemple le cas entre les cancers et les contrôles TCGA (courbes vertes). Cet écart interroge sur la pertinence de prendre comme contrôles les données issues de biopsies, en effet le protocole expérimental est très différent et les résultats préliminaires obtenus par Penda avec ces contrôles indiquent une dérégulation anormalement élevée.

Après concertation avec les biologistes de l'Inserm, nous avons finalement choisi une cohorte d'astrocytes cultivés comme référence [83]. Cette fois, le protocole expérimental est pertinent, et les distributions sont proches (voir figure 2.6). En revanche ,la cohorte se compose d'uniquement 12 échantillons

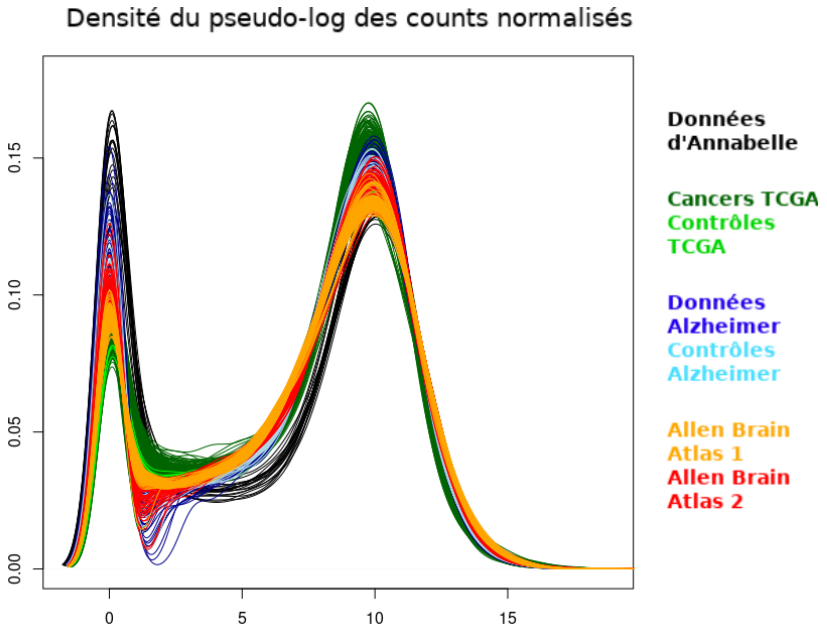


FIGURE 2.5 – **Distribution de l'expression des gènes dans les différentes cohortes après normalisation.** Les données d'expression sont transformées par pseudo-log afin de réduire l'échelle pour faciliter la visualisation. Chaque ligne correspond à un échantillon, en noir les lignées glioblastome d'intérêt, en vert les échantillons TCGA, en bleu ceux Alzheimer et en rouge/orange les données du Allen Brain Atlas.

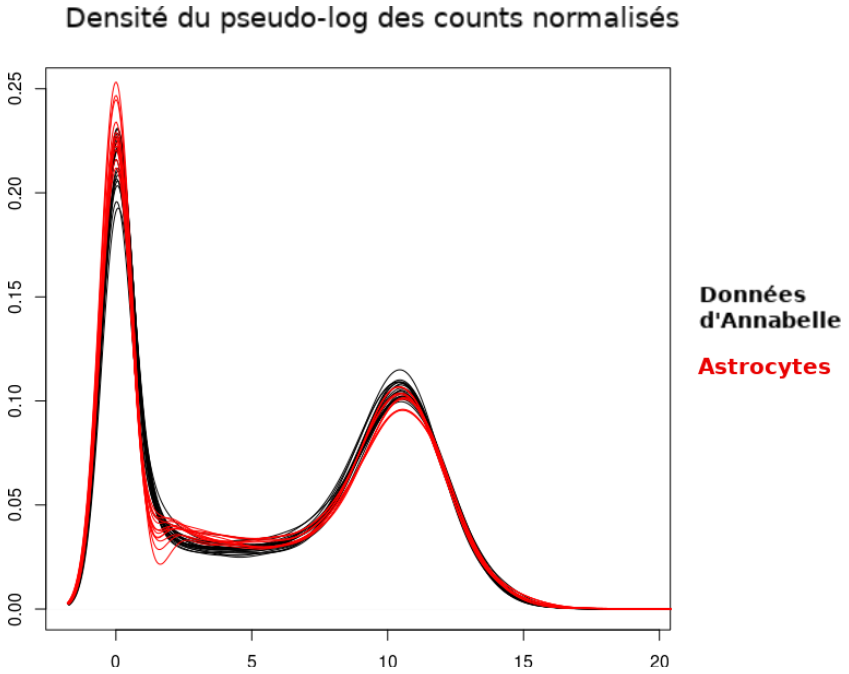


FIGURE 2.6 – **Distribution de l'expression des gènes dans les cas et les contrôles après normalisation.** Les données d'expression sont transformées par pseudo-log afin de réduire l'échelle pour faciliter la visualisation. Chaque ligne correspond à un échantillon, en noir les lignées glioblastome d'intérêt, et en rouge les lignées d'astrocytes choisies comme contrôles.

contrôles, en réalité trois réplicats de quatre lignées différentes, ce qui est très peu. De plus, les astrocytes ne constituent pas une référence parfaite car on ne peut pas parler de cellules saines dont seraient issues les cellules du glioblastome.

2.3.2 Stratégie de validation croisée

Étant donné le faible nombre de contrôles, nous avons choisi de mettre en place une stratégie de validation croisée pour s'assurer de la robustesse des résultats Penda. Nous avons lancé dix analyses indépendantes, en utilisant à chaque fois seulement dix des douze contrôles choisis aléatoirement de façon à ce qu'ils soient différents entre chaque analyse. Les deux contrôles restants à chaque fois sont mélangés aux tumeurs analysées afin d'estimer le nombre de faux positifs.

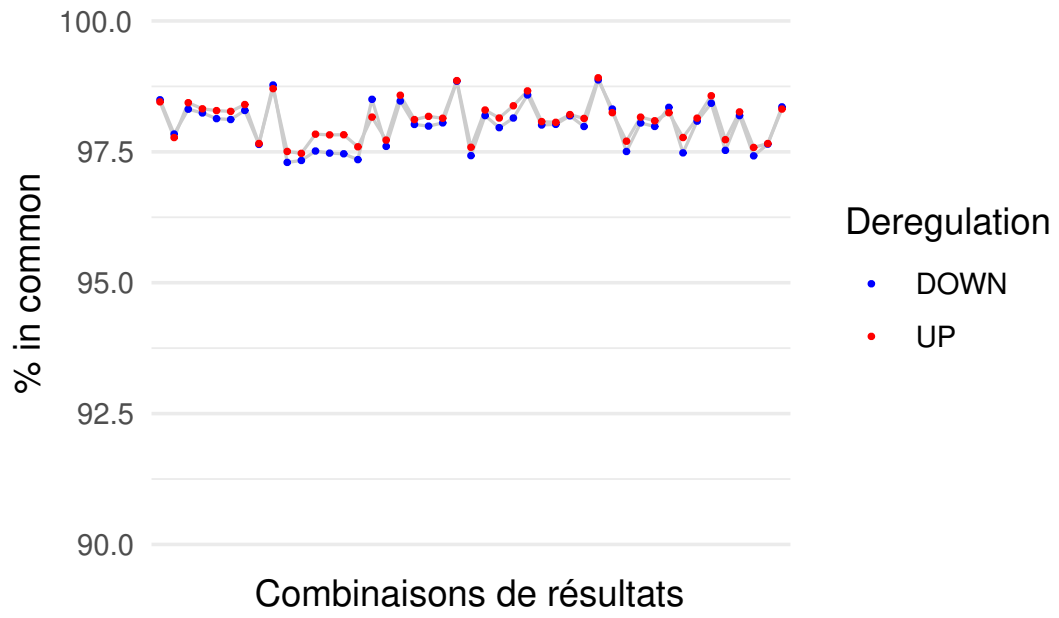


FIGURE 2.7 – **Proportion de résultats Penda en commun pour chaque combinaison d'analyses.** Dix analyses Penda ont été réalisées avec un sous-set de 10 contrôles. Les résultats sont ensuite comparés deux à deux, chaque point correspond à une des 45 combinaisons. En rouge les gènes sur-exprimés, en bleu ceux sous-exprimés.

Finalement, nous avons obtenu pour chaque analyse autour de 35% de gènes dérégulés, ce taux de dérégulation correspond aux attentes des biologistes. Les résultats ont ensuite été combinés deux à deux entre chaque analyse pour comparer la proportion de gènes détectés dérégulés en commun. Ces résultats sont visibles en figure 2.7, on voit qu'il y a systématiquement plus de 97% de cohérence entre deux analyses. Concernant les résultats sur les échantillons contrôles mis de côté (visibles sur la figure 2.8), on détecte à chaque fois une centaine de gènes dérégulés, sur 22 445 gènes analysés, le taux de faux positif est donc très faible.

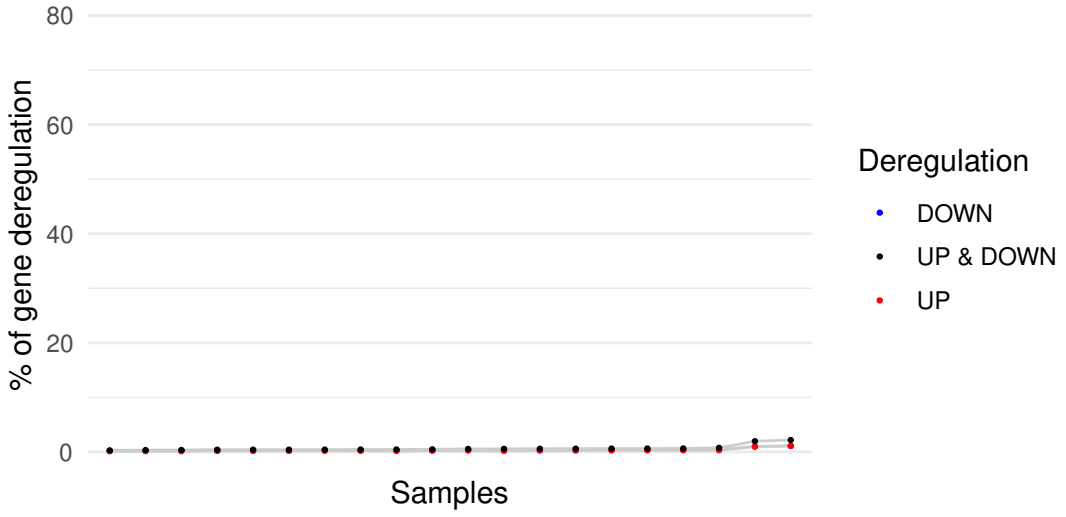


FIGURE 2.8 – **Résultats de l'analyse Penda sur les échantillons contrôles.** Pour chacune des analyses Penda en validation croisée, on réalise le test sur les 2 contrôles ne servant pas à établir les rangs de référence. Chaque point correspond à un de ces contrôles. En rouge les gènes sur-exprimés, en bleu ceux sous-exprimés.

2.3.3 Résultats

Nous avons donc conservé les gènes dérégulés communs aux dix analyses Penda, ce qui nous donne en moyenne 30% de dérégulation par échantillon. Le but de l'analyse était de regarder les gènes d'intérêts pour Annabelle Bellasta dont la dérégulation varie entre les lignées cellulaires, ce sont les résultats qu'on

observe sur la figure 2.9.

On retrouve bien des gènes différents dans chaque patient, avec parfois un sens de dérégulation spécifique et parfois un mélange entre sur- et sous-expression entre les patients. Ces résultats sont désormais exploités par l'équipe de l'Inserm.

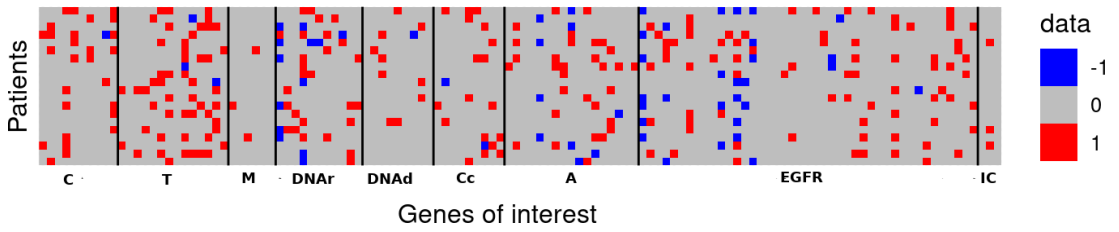


FIGURE 2.9 – **Résultats de dérégulation Penda pour chaque patient.** Les 20 lignées des différents patients sont représentées en ligne. Chaque colonne correspond à un gène d'intérêt fourni par nos collaborateurs, les catégories sont séparées par des traits verticaux : C, les gènes associés au facteur de transcription CLOCK et à l'horloge circadienne, T les gènes associés au transport, M les gènes associés au métabolisme, DNAr les gènes associés à la réparation de l'ADN, DNAd les gènes associés à la réponse aux dommages de l'ADN, Cc les gènes associés au cycle cellulaire, A les gènes associés à l'Apoptose, EGFR pour le réseau de gènes EGFR, un récepteur de croissance important dans le cancer et IC pour les gènes impliqués dans le système immunitaire. Une case bleue implique une sous-expression du gène pour le patient donné, rouge sur-expression, grise pas de changement.

Discussion et conclusion

Dans cette première grande partie, nous avons développé une nouvelle méthode pour l'analyse différentielle personnalisée : Penda, qui est basée sur les rangs dans des échantillons contrôles pour inférer la dérégulation dans un échantillon tumoral.

Penda est dans la lignée de la méthode RankComp : les deux méthodes d'analyse différentielle sont basées sur les rangs, et utilisent dans un premier temps les contrôles pour définir les références, avant d'analyser ensuite les échantillons tumoraux de manière individuelle. Les rangs sont établis différemment, RankComp se base sur des paires de gènes d'ordre stable, alors que Penda calcule des listes de références pour chaque gène, en conservant uniquement les gènes les plus proches et donc les plus sensibles à un faible changement. Le test est également différent, RankComp effectue un test exact de Fisher pour savoir si les paires qui impliquent le gène étudié s'inversent, alors que Penda se base sur un seuil représentant la proportion de changement. Enfin, l'influence des autres gènes est également traitée différemment : RankComp retire les paires concernées, Penda aussi, mais avec un système d'itérations pour arriver à une stabilisation des gènes dérégulés.

Dans notre travail, nous avons montré que Penda était plus robuste et plus efficace que les méthodes pré-existantes sur des simulations. Son implémentation sous la forme d'un package R, avec différentes vignettes simples à utiliser pour adapter les paramètres à des données différentes, facilite sa diffusion et sa prise en main.

En appliquant Penda à des données réelles, nous avons pu produire des résultats biologiquement pertinents, par exemple en identifiant 37 gènes dont la dérégulation constitue un biomarqueur de survie dans l'adénocarcinome du poumon. Nous avons également vu que les résultats de Penda variaient en fonction des profils des données et dépendaient donc de la situation.

Limites de la méthode

Fiabilité des échantillons contrôles

La principale faiblesse de la méthode Penda est la nécessité d'avoir des échantillons contrôles de bonne qualité pour produire des résultats fiables. Si les gènes des échantillons contrôles ont des niveaux d'expression très variables par rapport aux échantillons analysés par le test Penda, tous les gènes risquent d'être détectés comme étant dérégulés, et la méthode Penda s'appliquera sur des listes L et H très faibles, ou uniquement via la méthode des centiles, ce qui entraînera beaucoup de faux positifs et de faux négatifs. Il ne faut pas non plus négliger que les échantillons contrôles du TCGA correspondent à des tissus adjacents, qui ne sont pas tumoraux mais pas forcément parfaitement purs non plus, et possiblement affectés par les dérégulations du cancer. Pouvoir croiser l'information de plusieurs échantillons est donc essentiel.

Ainsi, la méthode Penda est vraiment pertinente pour des jeux de données comportant beaucoup de contrôles, ce qui permet à la fois d'avoir des références robustes pour les listes L et H, mais aussi d'appliquer la méthode des centiles sur une distribution significative.

Plusieurs stratégies ont été testées pour palier à ce problème. Si le jeu de données contrôle contient un petit nombre d'échantillons, on peut appliquer ce qui a été fait dans la partie précédente pour le glioblastome, c'est-à-dire effectuer l'ensemble du test Penda plusieurs fois sur des sous-sets différents de contrôles, et conserver l'intersection des gènes dérégulés. Cette stratégie augmente la robustesse des gènes détectés, mais augmente aussi potentiellement le nombre de faux négatifs puisqu'on rajoute une contrainte à la détection de la

dérégulation.

Quand il n'est vraiment pas possible d'obtenir plus d'un échantillon contrôle, une version alternative a été développée pour tester Penda : penda1ctrl. Ces fonctions sont également implémentées dans le package Github. La fonction *compute_lower_and_higher_lists_1ctrl* permet d'établir les listes de rangs de référence en laissant une marge entre l'expression du gène g étudié et l'expression des gènes utilisés pour les listes (en excluant les x gènes les plus proches, avec x fixé par l'utilisateur), afin d'être moins sensible à une différence d'expression très faible. La fonction *penda_test_1ctrl* permet ensuite d'appliquer le test Penda sans la méthode du centile, qui n'a pas de sens sans distribution d'expression dans des contrôles. Si un gène n'a pas de liste L, il ne peut pas être détecté sous-exprimé, inversement, s'il n'a pas de liste H, pas de sur-expression. Cette version de la méthode permet d'appliquer Penda dans des conditions extrêmes, mais n'est pas conseillée car elle induit des biais (voir figure 2.10).

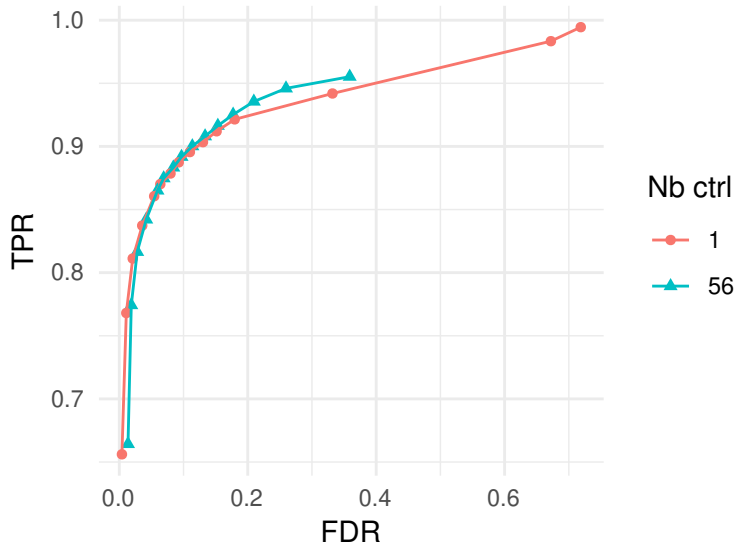


FIGURE 2.10 – **Comparaison des résultats entre les deux versions de Penda.** Les deux versions de Penda (Penda classique, et Penda adapté pour un seul contrôle) ont tourné pour plusieurs seuil sur un jeu de simulations. A seuil équivalent, la version avec 1 seul contrôle (rouge) a un TPR légèrement plus bas. La différence est particulièrement notable pour les faibles seuils (à droite de la courbe).

Enfin, un faible nombre de contrôles impacte également la pertinence des simulations intégrées dans Penda pour choisir les paramètres. Quand on a peu de contrôles, par exemple n = 10, on conseillera de simuler et d'analyser 10 simulations indépendantes, calculées chacune sur un seul contrôle, en utilisant les 9 autres contrôles pour établir les listes L et H. On pourra alors croiser les paramètres obtenus pour chacune de ces simulations et choisir le plus pertinent. Quand on a un seul contrôle et qu'il n'est pas possible d'effectuer des simulations, on peut tester plusieurs seuils en s'interrogeant plutôt sur la proportion de gènes dérégulés qu'on s'attend à obtenir à l'issue de l'analyse, et fixer notre choix ainsi.

Nombre et expression des gènes

On a vu que Penda était également sensible au nombre de gènes considérés. Pour établir des listes de référence fiables pour un gène g, il faut des gènes ayant une expression proche de g, qui seront des marqueurs fiables même pour un changement de rangs faible. La taille des données est donc très importante, car plus il y a de gènes, plus il est probable qu'il y ai un gène d'expression proche, capable de détecter un changement.

Un autre biais possible est le fait que les gènes dont les niveaux d'expression sont trop faibles soient assimilés à du bruit et retirés dans l'étape de prétraitement, ce qui pourrait également faire perdre de l'information. Si les données utilisées sont susceptibles d'avoir des gènes d'intérêt très faiblement exprimés dans tous les échantillons, il est donc important de fixer le seuil de faible expression avec soin, voir de retirer uniquement les gènes ayant une expression nulle dans tous les échantillons.

Format des résultats

Par défaut, Penda renvoie une information binaire, sur-exprimé : oui/non, sous-exprimé : oui/non. Ce format peut déconcerter les utilisateurs ayant l'habitude de recevoir une information analogique continue d'amplitude de dérégulation, associée à une p-valeur de certitude du test, comme ce qu'on obtient avec les méthodes par fold-change.

Le résultat Penda dépend bien d'une information numérique d'amplitude de dérégulation, puisque c'est la proportion de gènes de référence ayant changé de rang qu'on étudie, et qui est comparée au seuil du test. Cependant, puisque ce seuil influe les changements de listes à chaque itération, il n'est pas trivial de renvoyer directement cette proportion à l'utilisateur. La décision actuelle est donc de conserver ce format de réponse binaire. Si l'utilisateur veut faire varier le seuil de détection pour obtenir plus ou moins de gènes dérégulés, il suffit de relancer la méthode avec un autre paramètre de seuil.

Perspectives

Approfondir les dérégulations

Une application intéressante de Penda pourrait être d'utiliser les autres types de données biologiques disponibles pour approfondir l'étude des dérégulations. Par exemple des données de CNV (Copy Number Variation), qui correspondent à la duplication de certains gènes sur le génome, sont parfois disponibles et peuvent être la cause de la sur-expression de certains gènes.

L'étude des dérégulations de gènes groupés par leur fonction biologique a été rapidement abordée, mais pourrait également justifier un travail supplémentaire, pour identifier des gènes systématiquement co-dérégulés par exemple.

Il en est de même pour la partie d'analyse pan-cancer, qui mériterait d'être approfondie pour caractériser les gènes systématiquement conservés ou dérégulés à travers toutes les cohortes. Ce travail n'a pas encore été effectué par manque de temps.

Appliquer Penda à d'autres données

Dans cette partie, nous nous sommes concentrés sur l'étude de données transcriptomiques issues du cancer. Cependant, la méthode Penda pourrait être appliquée à d'autres situations, puisqu'il suffit d'avoir deux jeux de données : l'un considéré comme contrôle et l'autre qu'on voudrait comparer de manière individuelle au premier. L'application à d'autres maladies semble logique, mais on

pourrait également imaginer des applications en dehors du secteur médical, par exemple étudier les variations d'expression des gènes d'une même espèce entre deux conditions environnementales ou bien analyser des gènes spécifiquement exprimés dans certains tissus.

De la même façon, le type de données étudiées pourrait varier. Ici, nous analysons uniquement des données de transcriptome qui correspondent à une quantité d'ARN au moment du séquençage, mais n'importe quelle variable quantitative pourrait fonctionner. Des tests ont été effectués pour appliquer Penda aux données de méthylation de l'ADN, où la valeur obtenue correspond à la probabilité pour une cytosine donnée d'être méthylée. La limite était pour nous le nombre très important de sondes à étudier (typiquement 450 000 sondes au lieu de 20 000 gènes), et la complexité d'interprétation des résultats puisqu'une hyper-méthylation peut être associée à la fois à une sous- ou une sur-expression. Cependant, une analyse croisée par PenDA des données de méthylation et d'expression dans un même patient ouvrirait la possibilité d'étudier plus facilement la corrélation entre dérégulation épigénétique et transcriptomique, dans le cancer par exemple.

Conclusion

Penda est un outil puissant pour analyser les gènes dérégulés à l'échelle d'une tumeur par rapport à un ensemble d'échantillons contrôles. Les perspectives sont nombreuses : nouveaux biomarqueurs, nouvelles cibles thérapeutiques, meilleure classification des tumeurs, etc. Cependant, la source de la dérégulation en elle-même n'est pas expliquée.

À l'échelle intra-tumorale, de nombreux facteurs influent sur la dérégulation des gènes, l'un d'entre eux est la composition cellulaire. En effet, les échantillons analysés dans cette partie ne correspondent pas à des lignées tumorales pures, mais à des tumeurs broyées, composées de cellules du cancer, mais aussi d'autres types cellulaires comme ceux du tissu adjacent de l'organe sain, ou encore des cellules immunitaires.

Dans la seconde partie de ma thèse, nous allons nous intéresser à cette hété-

rogénéité intra-tumorale et chercher à quantifier la part des cellules cancéreuses et la part du micro-environnement dans les tumeurs. Enfin, dans la troisième partie, nous allons combiner ces deux informations en cherchant les gènes dont la dérégulation de l'expression dans le type tumoral est associée à la composition du micro-environnement.

Deuxième partie

Déconvolution de la composition tumorale

Introduction

Dans la partie précédente, nous avons vu un nouvel outil permettant d'analyser l'hétérogénéité inter-tumorale en identifiant les gènes différentiellement dérégulés dans une tumeur donnée. Une des causes de cette variabilité observée est la composition de la tumeur en elle-même : dans cette partie nous allons aborder ces enjeux.

Une tumeur est un ensemble complexe de cellules. Elle est composée de cellules cancéreuses à différents niveaux de différenciation et de mutation, mais aussi de cellules conjonctives, de cellules immunitaires, et des autres tissus composant l'organe d'origine. L'ensemble des cellules non-cancéreuses présentes autour et dans la tumeur est appelé micro-environnement [84]. Ce micro-environnement joue un rôle crucial dans l'initiation et le développement du cancer. Il influe sur la réponse au traitement, la virulence du cancer et donc la survie des patients et leurs chances de rémission [85]. Ce lien est complexe et peut à la fois tempérer ou avantager le cancer : par exemple, les fibroblastes associés au cancer (CAF) peuvent être reliés à la fois à une modération de la prolifération des cellules cancéreuses, et à la fois à une résistance accrue des tumeurs à la chimiothérapie [86]. Ce point est aussi illustré par les cellules immunitaires, qui peuvent à la fois combattre les tumeurs et dont l'activation peut donc être un objectif thérapeutique (pour l'immunothérapie notamment), et à la fois être un marqueur de prolifération et de métastases dans certaines tumeurs infiltrées [87]. La caractérisation du micro-environnement tumoral représente donc un enjeu important dans l'étude du cancer [88].

Les méthodes d'analyses "single-cell" qui permettraient d'isoler les différents types cellulaires et d'obtenir l'information pour une seule cellule sont très complexes et coûteuses et ne sont pas encore applicables dans le cadre clinique où une caractérisation rapide de la tumeur est nécessaire [89, 90]. En réalité, dans la routine clinique, les tumeurs extraites par les chirurgiens sont généralement broyées avant d'être analysées. L'information obtenue par les méthodes d'analyses courantes sur des échantillons mélangés "bulk" (fluorescence par puce ou RNA-seq, voir partie 4 de l'introduction) ne correspond donc pas à celle d'un type cellulaire pur, ni à celle de la lignée tumorale isolée, mais se présente sous la forme d'un signal convolué, composé du signal des différents types cellulaires présents dans la tumeur pondérés par leur proportion au sein de l'échantillon (voir figure 2.11). Mathématiquement, cela veut dire que les données D_i d'un patient i sont égales à :

$$D_i = \sum_{j=1}^{k} A_{i,j} T_j \tag{2.1}$$

avec T_j le profil pur pour le type cellulaire j, $A_{i,j}$ la proportion du type j dans l'échantillon du patient i et K le nombre total de types cellulaires différents.

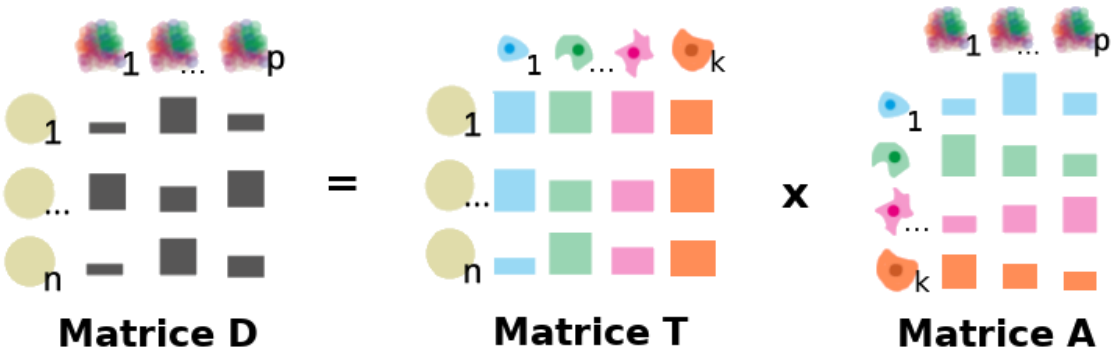


FIGURE 2.11 – **Représentation de la décomposition de la matrice complexe obtenue par séquençage.** La matrice D contient les valeurs moyennes de n sondes pour p patients. Elle peut se décrire par le produit de deux sous-matrices : la matrice T, qui correspond aux profils purs pour chacun des k types cellulaires, et la matrice A, qui contient les proportions des k types cellulaires dans les p échantillons.

Des approches expérimentales telles que l'immunohistochimie [91] ou la cytométrie en flux [92] permettent de retrouver la composition d'un échantillon complexe, mais présupposent la connaissance des types recherchés et l'existence de marqueurs moléculaires associés, là où dans le cas du cancer il est difficile de savoir avec précision ce que contient la tumeur. Une approche computationnelle directement à partir de l'échantillon mélangé représente donc une bonne alternative pour inférer la composition d'une tumeur.

Afin de caractériser cette composition, il est donc nécessaire de passer par une étape de décomposition du signal, appelée déconvolution. Concrètement, il faut inférer les matrices T et/ou A à partir de D (Figure 2.11). Cette problématique est très large, et peut être abordée sous différents angles : par exemple vis-à-vis du choix du type de données étudiées dans D (expression des gènes, méthylation, etc.), ou encore en fonction de la présupposition ou non des types cellulaires composant l'échantillon, c'est à dire fixer à priori la matrice T (méthodes supervisées, basées sur une référence) ou inférer T en même temps que A (méthodes non-supervisées ou reference-free). Les méthodes de déconvolution sont généralement en deux parties : un algorithme de factorisation permet d'initialiser A et/ou T, puis les matrices sont optimisées pour minimiser la distance entre D et $A * T$.

Du côté des approches utilisant le transcriptome, une revue parue en 2018 liste une cinquantaine de méthodes de déconvolutions existantes sans pour autant identifier de méthode de référence [93].

Les méthodes non supervisées se basent principalement sur des méthodes de factorisation de matrice comme l'analyse en composantes indépendantes (ICA) ou la factorisation non-négative (NMF) [94]. L'ICA est par exemple utilisée par le package R deconICA [95], spécialement conçu pour l'extraction de la composante immunitaire des tumeurs. C'est une méthode de factorisation généralement utilisée pour la réduction de dimension qui se base sur la contrainte que les composantes déconvoluées (la matrice T pour nous) doivent s'éloigner le plus possible d'une distribution gaussienne [96]. La NMF, utilisée par exemple par la méthode Decoder [97], se base cette fois sur la contrainte de déconvoluer des

matrices uniquement positives.

Comme exemple de méthode semi-supervisée, on peut citer WISP [98], où la matrice T est pré-définie grâce à des profils moléculaires purs, et où la matrice A est ensuite calculée pour minimiser la distance entre D et $A * T$ tout en gardant des valeurs positives et la somme des proportions dans un échantillon égale à 1.

Enfin, d'autres méthodes se concentrent plutôt sur des mesures d'enrichissement de certains types cellulaires à l'aide de biomarqueurs, comme ESTIMATE [99], sans inférer les proportions exactes des différentes composantes.

Pour les approches utilisant le méthylome, il existe principalement trois méthodes n'utilisant pas de référence pour les types cellulaires : RefFreeEwas [100], MeDeCom [101] et EDec [102]. Ces trois méthodes se basent sur la NMF pour la factorisation, avec quelques variations sur l'implémentation de l'algorithme. Par exemple, MeDeCom utilise la validation croisée pour optimiser les résultats de la déconvolution en la reproduisant sur des sous-échantillons. EDec de son côté propose de faire de la semi-supervision avec des listes de biomarqueurs des différents types cellulaires.

Pour les méthodes basées sur des références, on peut citer Epidish [103] qui utilise une liste de sondes par type cellulaire pour inférer A à partir d'un mélange D. Même si moins de méthodes ont actuellement été développées pour les données de méthylome, l'ADNm est un bon biomarqueur des types cellulaires, et les applications sont prometteuses [104].

Si des méthodes de déconvolution non supervisée existent, il n'y a pas encore de comparaisons entre elles et il parait difficile de déterminer la meilleure méthode. De plus, ces études négligent certains aspects : prise en compte des facteurs de confusion, pré-filtrage des données pour optimiser les résultats, ou encore intégration multi-omique.

Dans ma thèse, j'ai choisi de me focaliser sur les méthodes reference-free car les tumeurs sont des ensembles de cellules complexes, il est difficile d'inférer à l'avance l'ensemble des types cellulaires qui la composeront et de trouver des lignées de référence pour chacun d'entre eux. Nous avons d'abord choisi de s'introduire à la problématique sous l'angle restreint des méthodes se basant sur

des données de méthylation de l'ADN (ADNm) en organisant un data challenge. À l'issue de celui-ci, une analyse comparative des méthodes existantes a été réalisée, avec comme objectif d'étudier l'effet du pré-traitement des données (pré-sélection des sondes, en lien ou non avec la présence de facteurs de confusion). Enfin, un deuxième data challenge a été réalisé sur l'idée d'améliorer les résultats de la déconvolution en croisant plusieurs types de données : expression des gènes et méthylation de l'ADN. L'intégration multi-omique est un enjeu de recherche important, car multiplier les sources d'information pour un même échantillon permettrait de diminuer le bruit de chacune des méthodes et d'améliorer le signal d'intérêt.

Chapitre 3

Exploration de la déconvolution des données d'ADNm à travers un data challenge

Dans cette partie, j'ai étudié les différentes méthodes existantes permettant de déconvoluer les types cellulaires à partir de données d'ADNm sans référence, et j'ai analysé si une étape de pré-traitement permettait d'améliorer ou non ces résultats. La méthylation de l'ADN a été choisie comme premier pas dans la déconvolution car c'est une marque épigénétique stable, avec un profil spécifique pour chaque type cellulaire [105, 106], et qui est donc adaptée à la question de la déconvolution.

En revanche, l'ADNm est sensible à différentes variables biologiques comme le sexe ou l'âge, et à des facteurs d'exposition comme l'alimentation ou le tabac [107]. Or, l'impact de ces facteurs sur l'efficacité de la déconvolution n'avait que très peu été étudié. Nous avons donc réalisé un data challenge afin de répondre à cette question, car ce format permettait de comparer facilement les méthodes existantes et de mettre en place un travail collaboratif et multi-disciplinaire enrichissant.

Je vais d'abord présenter les principes et enjeux de l'organisation d'un data challenge, puis expliquer la simulation des données utilisées et enfin les résultats obtenus.

3.1 Principe du data challenge

Le data challenge est un format assez innovant d'ateliers collaboratifs (ou workshops), où les participants sont soumis à une question complexe par les organisateurs. L'idée est ensuite de répondre au mieux à la problématique par équipes, sous la forme d'une compétition avec un classement des méthodes fonctionnant le mieux sur des données « benchmarkées » fournies par les organisateurs et passant par leur évaluation automatique selon des critères fixés à l'avance.

3.1.1 La problématique

Notre data challenge a été organisé par l'équipe sous la direction de Magali Richard (*description du data challenge sur notre site internet*). Le but du data challenge était double : d'une part offrir aux participants une introduction à la déconvolution non supervisée des données de méthylation de l'ADN en R, et d'autre part évaluer si ces méthodes pouvaient être améliorées en rajoutant une étape de pré-traitement sur les données.

Pour la partie introductive du data challenge, chacune des trois méthodes reference-free pour l'ADNm existantes a pu être présentée aux participants par un de ses développeurs : E. Andres Houseman pour EDec [102], Eugène Lurie pour RefFreeEwas [100] et Pavlo Lutsik pour MeDeCom [101]. Ces trois méthodes reposent sur le même principe de déconvolution : la factorisation non-négative des matrices (ou non-negative matrix factorization, NMF). La NMF est un algorithme permettant de résoudre $D = T * A$ (voir Figure 2.11) en estimant T et A à partir de D. Une des deux matrices (T ou A) est d'abord initialisée, puis la deuxième est estimée pour minimiser la distance entre $T * A$ et D. Ensuite, le procédé est itératif, on recalcule l'autre matrice pour minimiser la distance, en passant de T à A à chaque étape. Les trois méthodes diffèrent toutes sur l'initialisation de la première matrice (aléatoire ou par clustering, sur T ou sur A), ainsi que par des contraintes différentes sur les résultats. En outre, EDec se base sur une pré-sélection des sondes d'intérêt dans la littérature, et MeDeCom

	RefFreeEwas *(E.A. Houseman & al., 2016)*	**EDec (stage 1)** *(V.Onuchic & al., 2016)*	**MeDeCom** *(P.Lutski & al., 2017)*
Algorithme de déconvolution	Factorisation de matrice non-négative	Factorisation de matrice non-négative	Factorisation de matrice non-négative
Initialisation	T (clustering hiérarchique)	A aléatoire	Multiples A aléatoires
Validation croisée	/	/	Oui
Pré-sélection des sondes dans la littérature	/	Oui	/
Contraintes	T et A entre 0 et 1 Somme des proportions <1	T et A entre 0 et 1 Somme des proportions <1	T et A entre 0 et 1 Somme des proportions <1

Tableau 3.1 – **Résumé des principales différences entre les méthodes de déconvolution sans référence des données de méthylation de l'ADN.** Les trois méthodes sont basées sur une NMF pour la déconvolution, mais toutes ne commencent pas l'initialisation sur la même matrice. MeDeCom a la spécificité de proposer de la validation croisée (cross-validation) alors qu'EDec implique une phase de présélection des sondes dans la littérature. Enfin, les méthodes incluent dans l'algorithme de déconvolution des contraintes sur les résultats : A et T doivent être entre 0 et 1, et la somme des proportions dans un échantillon inférieure à 1. MeDeCom rajoute une contrainte pour que les valeurs de méthylation de T soient proches de 0 ou de 1 (aucune méthylation ou 100 % méthylé).

applique une cross-validation sur ses paramètres (voir 3.1).

Étant donné que le pré-traitement des données dans les problématiques de déconvolution était encore largement inexploré, alors même que de nombreux facteurs de confusion autre que le type cellulaire ont un impact sur la méthylation de l'ADN (génétique, biologique, conditions environnementales, effets expérimentaux…) [107], nous avons demandé aux participants d'imaginer des façons de prendre en compte ces variables expérimentales, par exemple en corrigeant leurs effets ou en filtrant les sondes affectées.

3.1.2 Les participants au data challenge

Dans notre cas, c'est une trentaine de participants qui ont été réunis à Aussois (France) du 10 au 14 Décembre 2018 pour ce data challenge. Un des gros atouts de ce format est la diversité des participants, qui permet de mélanger

différentes disciplines (bio-informatique, biologie, mathématiques, statistiques, informatique...), différents statuts (chercheurs, jeunes chercheurs, étudiants...) et différentes origines (universités et instituts de toute l'Europe) au sein d'une même équipe, et donc de créer très rapidement des liens. En plus des liens crées par le travail collaboratif en équipe, les relations entre participants sont accentuées par le lieu de vie commun (restaurant, bar, chambres doubles et salles de réunion au sein du même bâtiment), de nombreux échanges sont possibles entre participants d'une même équipe ou non, et de solides collaborations peuvent se créer au cours du séjour.

3.1.3 Le contenu du data challenge

Après l'introduction aux méthodes existantes par leurs développeurs à travers des séminaires de recherche, le reste de la semaine était consacré à la pratique et au data challenge en lui-même. À partir de données complexes de méthylation d'ADN simulées en amont (voir section 3.2), les participants devaient tester et développer différentes méthodes de déconvolution. Leurs résultats étaient ensuite comparés aux matrices d'origine, et un score ainsi qu'un classement leurs étaient attribués en fonction de la pertinence du résultat (voir partie 3.2.4 pour le détail du score). Un premier data challenge plus simple permettait de s'introduire aux méthodes, puis un second introduisant des facteurs de confusion dans les matrices a permis d'explorer si leur prise en compte par un pré-traitement des données permettait d'améliorer les performances des algorithmes de déconvolution.

La plateforme choisie pour soumettre les méthodes développées et établir les classements de manière dynamique était Codalab (*lien vers la plateforme*) qui permet de créer des challenges personnalisés de façon assez simple.

3.2 Simulation des données complexes de méthylation de l'ADN

La simulation des données fut un enjeu important du data challenge, car c'est à partir de celles-ci que les méthodes développées par les différentes équipes vont être évaluées et classées. Dans notre cas, la matrice D contenant la méthylation moyenne pour chaque sonde et chaque patient a été simulée par le produit de la matrice T contenant le profil de méthylation pour chaque type cellulaire, et la matrice A contenant les proportions des différents types cellulaires pour chacun des patients (voir la figure 2.11). Le score des participants a été calculé uniquement sur la matrice A, d'une part car la composition du micro-environnement est généralement le principal enjeu des études sur l'hétérogénéité tumorale, et d'autre part car le nombre de lignes de la matrice T calculée est amené à beaucoup changer en fonction des choix de sélection de sondes, et ne permet donc pas une comparaison équilibrée entre les participants.

Des matrices différentes ont été simulées pour la première et la deuxième phase du data challenge, le principal changement étant que les échantillons de la première sont composés de seulement trois types cellulaires mélangés, alors qu'il y en a cinq dans la seconde. Pour la deuxième partie, nous avons aussi rajouté différents effets confondants, reliés ou non aux données cliniques.

3.2.1 Simulation de la matrice des profils de méthylation T

La matrice des profils de méthylation est composée directement de lignées publiées dans la littérature. Nous avons choisi de simuler des tumeurs pulmonaires car nous étions déjà familiers des données du TCGA sur les cancers du poumon (adénocarcinome et carcinome à cellules squameuses, voir partie 4 de l'introduction générale) pour les avoir analysées avec Penda.

Les lignées composant T ont été sélectionnées dans la base de données publiques GEO [108], en choisissant des lignées humaines, de préférence masculines (quand le sexe était précisé). Elles sont listées dans le tableau 3.2.

Ces données de méthylation ont été mesurées par puces 27k ou 450k (voir partie 4 de l'introduction générale). Pour le data challenge, nous n'avons retenu

	Épithélial cancer	Mésenchyme cancer	Épithélial sain	Fibroblaste sain	Lymphocyte T
Partie 1	*GSM1560911*	/	/	GSM1354675	GSM1641098
Partie 2	GSM1560930	GSM1560925	GSM2743808	GSM1354676	GSM1641099

Tableau 3.2 – **Récapitulatif des lignées cellulaires utilisées pour simuler la matrice T dans ce data challenge.** Les lignées de cellules épithéliales et mésenchymateuses cancéreuses proviennent de l'étude GSE63940 qui séquence en 27k différentes lignées d'adénocarcinome pulmonaire. La lignée épithéliale sain provient de la série GSE102726 qui contient les séquences en 450k de petites cellules épithéliales respiratoires humaines. Les lignées de fibroblastes sains proviennent de la série GSE56074 qui contient les séquences en 27k de lignées commerciales de fibroblastes pulmonaires. Enfin, les lymphocytes T proviennent de l'étude GSE67170, il s'agit de lymphocytes T du sang périphérique de tumeurs du foie, séquencés en 450k.

dans nos jeux de données que les sondes qui sont communes aux deux technologies de séquençage et qui sont détectées dans au moins un échantillon TCGA, car ils seront utilisés pour calculer les facteurs de confusion (voir partie 3.2.3) : c'est à dire 23 381 sondes. Pour la première partie du data challenge, la matrice T sera donc composée des valeurs brutes de ces sondes pour les 3 lignées cellulaires utilisées. Pour la deuxième, elle dépendra des valeurs ajustées des 5 lignées pour tenir compte des facteurs de confusion (voir section 3.2.3).

3.2.2 Simulation de la matrice de proportion A

Les matrices A sont simulées à l'aide d'une distribution de Dirichlet, permettant de représenter la loi de probabilité pour des variables multinomiales.

Mathématiquement, la loi de Dirichlet se décrit par sa densité de probabilité définie par :

$$f(x_1,\dots,x_K;\alpha_1,\dots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

avec K le nombre de types cellulaires, x les proportions (comprises entre 0 et 1, avec une somme inférieure à 1) et $\mathrm{B}(\alpha)$ la fonction beta multinomiale définie

par :

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \alpha = (\alpha_1, \ldots, \alpha_K).$$

La valeur moyenne $\tilde{\alpha}_i$ de x_i est donnée par :

$$\tilde{\alpha}_i \equiv \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k} = \frac{\alpha_i}{\alpha_0}, \qquad \alpha_0 = \sum_{i=1}^{K} \alpha_i.$$

Et sa variance par :

$$\frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\alpha_0 + 1}$$

On voit donc que pour des proportions moyennes $\{\tilde{\alpha}_i\}$ fixées, $\alpha 0$ va jouer sur la variance.

Concrètement, la distribution de Dirichlet permet ainsi de "couper" aléatoirement la tumeur en K parties représentants les types cellulaires, avec chacune une proportion moyenne représentée par le paramètre $\tilde{\alpha}_i$ et une variabilité autour de cette proportion dépendant de $\alpha 0$.

	Épithélial cancer	**Mésenchyme cancer**	**Épithélial sain**	**Fibroblaste sain**	**Lymphocyte T**
Partie 1	*70%*	/	/	20%	10%
Partie 2	60%	10%	15%	10%	5%

Tableau 3.3 – **Récapitulatif des proportions en types cellulaires des simulations du data challenge.** Les proportions dans les échantillons varient ensuite autour de ces valeurs grâce à la distribution de Dirichlet.

C'est la fonction rdirichlet du package R bmixture [109] qui est utilisée pour cette étape. Pour le premier data challenge, les proportions de 100 patients sont simulées en fixant en moyenne ($\tilde{\alpha}$) 20 % de fibroblastes, 70 % de type épithélial cancéreux et 10 % de lymphocytes T. Pour le deuxième data challenge, les proportions moyennes sont 10 % de fibroblastes, 60 % d'épithélial cancéreux, 5 % de lymphocytes T, 15 % d'épithélial sain et 10 % de mésenchyme cancéreux (voir tableau 3.3). Ces proportions sont inspirées des types cellulaires habituel-

lement retrouvés dans les tumeurs pulmonaires, avec une majorité de cellules cancéreuses mais également des tissus conjonctifs habituels et une infiltration immunitaire.

3.2.3 Facteurs de confusion

Pour la deuxième partie du challenge, visant à étudier la prise en compte de l'effet des facteurs de confusion, nous voulions simuler des données cliniques proches de la réalité. En nous inspirant de celles du TCGA, nous avons créé un groupe expérimental composé à la fois de variables assimilées à des facteurs de confusion modifiant les valeurs de méthylation et de variables n'ayant pas d'impact sur la méthylation, comme on retrouve dans les jeux de données réels. Nous avons donc simulé plusieurs types de facteurs de confusion, certains directement sur la matrice T comme l'âge et le sexe, et certains sur la matrice D pour simuler des effets "expérimentaux" comme les effets de batch (biais de mesures liés au centre où a eu lieu l'analyse) ainsi qu'un bruit résiduel gaussien de mesure. L'ensemble de la procédure de simulation est résumée dans la figure 3.1.

Facteurs de confusion sur T

L'intégration des effets des facteurs de confusion s'est basée sur l'analyse des données TCGA des cancers LUAD et LUSC, le principe étant de partir de la matrice T "brute" des 23 381 sondes pour les 5 lignées cellulaires et d'ajuster la valeur de certaines sondes pour tenir compte d'un biais sur l'âge et le sexe.

Pour le sexe, nous avons effectué une régression linéaire sur les cohortes LUAD et LUSC, de manière indépendante, ce qui nous a permis d'identifier un nombre de sondes de méthylation dont la valeur est statistiquement associée au sexe et de définir l'amplitude de dérégulation à simuler sur chaque sonde. Nous avons ensuite testé plusieurs seuils de détection pour l'association du sexe et des sondes en variant la p-valeur du modèle (le seuil de significativité) et la robustesse (sondes impactées significativement dans les deux cancers ou dans un seul des deux). En regardant l'impact sur des méthodes de factorisation de base (voir partie 3.2.4), nous avons finalement choisi un seuil de p-valeur de 0,01

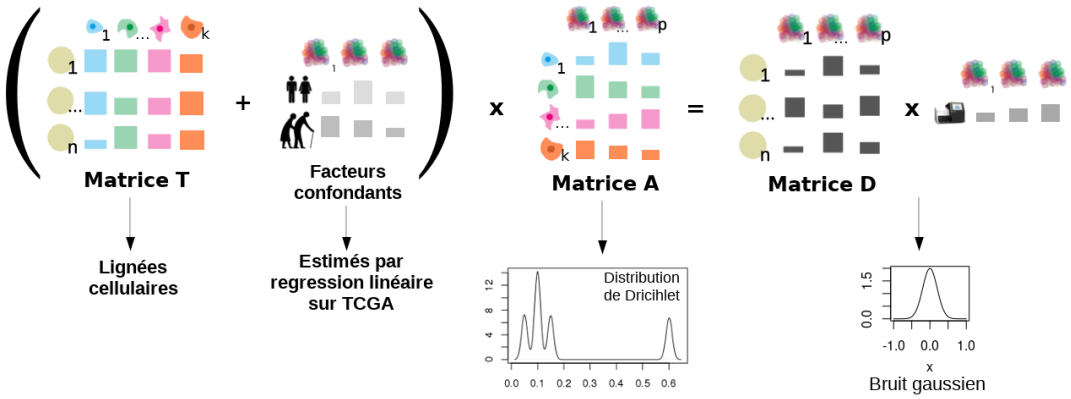


FIGURE 3.1 – **Résumé de la procédure de simulations pour la partie 2 du data challenge.** La matrice T est générée à partir de lignées cellulaires publiques, puis un effet pour le sexe et l'âge calculé par régression linéaire sur les données de cancer du poumon TCGA est rajouté à chaque échantillon en fonction des données cliniques du patient. La matrice A des proportions en types cellulaire est calculée par une distribution de Dirichlet. Ensuite, un facteur en fonction de la plaque de séquençage est appliqué sur la matrice D, puis enfin un bruit gaussien suivant une loi normale.

dans LUAD ou LUSC, ce qui représentait 586 sondes soit environ 6 % des sites. Pour la valeur de dérégulation, nous avons choisi de considérer le sexe masculin comme référence des lignées. Nous avons ensuite utilisé la moyenne des résidus de la régression linéaire entre LUAD et LUSC, allant de -0,6 à 0,4, comme valeur de dérégulation à appliquer aux matrices T des échantillons provenant d'une femme.

L'effet de l'âge a également été calculé par régression linéaire sur la cohorte TCGA. 113 sites ont été identifiés en fixant un seuil par défaut de 0,05 dans LUAD et LUSC. L'âge de la cohorte TCGA variant de 25 à 87 ans, nous avons choisi une même amplitude d'âge pour nos patients. Le calcul d'un modèle de régression linéaire pour chacun des 113 sites identifiés a permis de créer un modèle de prédiction pour la valeur d'une sonde de n'importe quel âge tiré dans nos données. Cette valeur de dérégulation a été associée au type cellulaire épithélial contrôle. Nous avons ensuite calculé un ratio entre les types cellulaires pour chaque sonde afin de l'extrapoler aux lignées.

Arrivé à cette étape, il y a donc une matrice T différente pour chaque patient en fonction de son âge et de son sexe (voir figure 3.1).

Facteurs de confusion sur D

Pour simuler la variabilité expérimentale à travers le "batch effect", nous avons choisi de rajouter un facteur sur la matrice D en fonction de la plaque de séquençage. Pour chaque échantillon, la plaque a été tirée au hasard parmi 22 possibilités, et la valeur finale de méthylation de toutes les sondes a été multipliée par ce facteur variant entre 0,77 et 1,77, lui-aussi inspiré des données TCGA. Enfin, un bruit gaussien suivant une loi $N(0, 0, 2^2)$ a été rajouté à toutes les valeurs de la matrice D (voir figure 3.1). Les données de méthylation ont ensuite été seuillées pour conserver une valeur comprise entre 0 et 1.

Le reste des données cliniques qui étaient fournies aux participants du data challenge a été tiré au hasard dans les données TCGA sans rajouter de lien avec la méthylation (stade tumoral, exposition à la cigarette, rémission, poids, ethnie, traitement, etc.). Nous avons ainsi choisi au final 22 variables expérimentales, dont 3 seulement reliées aux valeurs de méthylation (âge, sexe et plaque). L'identité de ces 3 variables était évidemment cachée aux participants.

3.2.4 Calcul du score et choix des paramètres de simulation

Évaluation des résultats

Pour évaluer les méthodes développées par les participants, nous avons établi plusieurs scores. Comme expliqué précédemment, il était compliqué d'évaluer la matrice T des profils de méthylation purs car ses dimensions sont susceptibles de beaucoup varier entre les méthodes de présélection des sondes. Nous avons donc choisi d'évaluer l'erreur sur la matrice de proportions A à travers deux métriques : l'erreur absolue moyenne (A- MAE, norme L1) et la racine-carrée de

l'erreur quadratique moyenne (A-RMSE, norme L2) :

$$A-MAE = \frac{\sum_{i=1}^{p}\sum_{j=1}^{K}\left|A_{i,j}^{th} - A_{i,j}^{pred}\right|}{p \times K} \quad \text{et} \quad A-RMSE = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{K}(A_{i,j}^{th} - A_{i,j}^{pred})^2}{p \times K}}$$

avec A^{th} la matrice A théorique et A^{pred} la matrice prédite. Nous avons constaté que ces deux mesures établissaient le même ordre entre les méthodes, les analyses comparatives qui suivent se concentrent donc sur la métrique A-MAE : plus l'erreur est faible, mieux la méthode a réussi à déconvoluer la matrice des proportions.

Une étape importante de l'évaluation est également de définir la correspondance entre les types d'origine et les types obtenus. En effet, les lignes de la matrice A^{pred} inférée qui correspondent aux différents types cellulaires ne sont pas forcément ordonnées de la même manière que dans la matrice de référence A^{th}. Pour éviter ce biais de comparaison, nous utilisons le package R combinat [110] qui permet de générer toutes les permutations possibles des lignes de la matrice A^{pred}. Toutes les A-MAE correspondantes sont calculées, et la permutation retenue est celle minimisant l'erreur globale.

Choix des paramètres pour le data challenge

Afin tester nos simulations avant le data challenge, nous avons réalisé un pré-test en faisant varier plusieurs paramètres comme le nombre de types cellulaires, l'écart-type du bruit gaussien ou la proportion de sondes affectées par le sexe pour déterminer leurs impacts et adapter la difficulté des simulations. Pour chacune des matrices, trois méthodes ont été testées : RefFreeEwas utilisé de manière naïve, et deux méthodes basées directement sur la factorisation non-négative (NMF) de R, avec ou sans retrait des sondes reliées au sexe (figure 3.2).

Nous avons pu constater que les méthodes semblaient être assez robustes. Par exemple, la MAE variait très peu entre le résultat de la déconvolution sur une matrice à 3 ou à 5 types cellulaires, même si un fort bruit gaussien diminuait l'efficacité des méthodes. Mais ce qui importait surtout pour le data challenge c'est que l'ordre relatif entre les méthodes pour une condition donnée reste

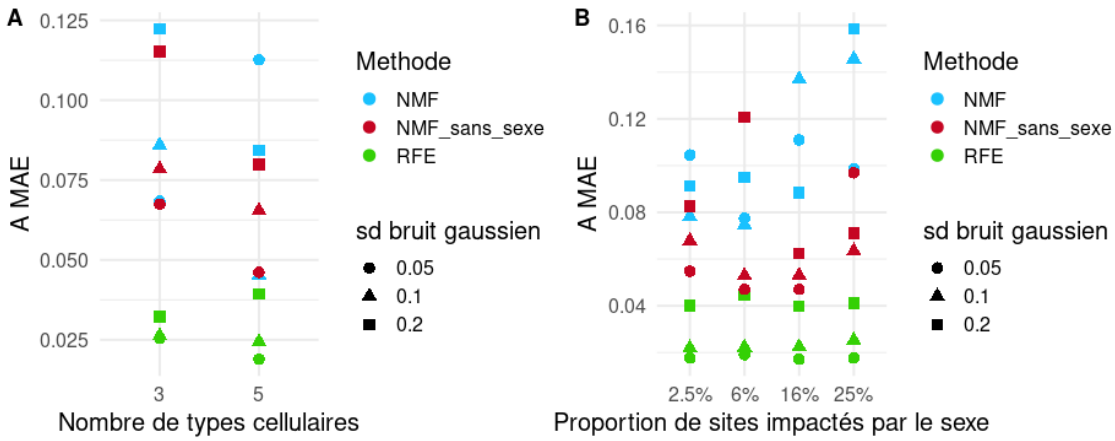


FIGURE 3.2 – **Test naïf de différents paramètres de simulation fait avant le 1er data challenge.** Pour trois méthodes simples de déconvolution (NMF, NMF avec retrait des sondes reliées au sexe, RFE : RefFreeEwas), l'erreur moyenne absolue de déconvolution (A-MAE) a été calculée pour plusieurs paramètres de simulations : 3 ou 5 types cellulaires, trois écarts-types de bruit gaussien, quatre versions de dérégulation en lien avec le sexe).

stable quels que soient les paramètres testés, et c'est ce qu'on observe ici avec RefFreeEwas plus efficace que la NMF avec pré-sélection des sondes, elle-même plus efficace qu'une NMF naïve. Le nombre de types cellulaires n'ayant pas un grand impact dans ce test, il a été décidé d'utiliser les deux, n = 3 pour la première partie et n = 5 pour la seconde. L'écart-type du bruit gaussien a été fixé à 0,2 (c'est à dire un bruit suivant la loi $N(0, 0,2^2)$) pour que la déconvolution ne soit pas trop simple. Pour la proportion de sites affectés par le facteur de confusion du sexe, c'est une valeur moyenne de 6%, qui nous semblait plus biologiquement pertinent, qui a été conservée (voir partie 3.2.3).

3.3 Résultats du data challenge

La première phase du data challenge étant une initiation à la déconvolution pour les participants, nous nous intéressons surtout aux résultats de la deuxième phase.

3.3.1 Méthodes obtenues

À l'issue de la deuxième phase du data challenge, nous avons récupéré les meilleurs scripts des participants que j'ai décomposé en deux grandes parties : les méthodes de pré-traitement et les méthodes de déconvolution. Parmi les méthodes de pré-traitement, deux sous-catégories se distinguent : la prise en compte des facteurs de confusion en corrigeant ou supprimant les sondes concernées, et la sélection des sondes sur d'autres critères comme celui des sondes les plus variables. La liste des différentes méthodes pour les deux parties se trouve ci-dessous (pour une description plus précise, voir la partie 4.1) :

Méthodes de pré-traitement :

1. Retrait des sondes affectées par les facteurs de confusion détectées par une régression linéaire sur toutes les variables cliniques.
2. Correction de l'effet du sexe détecté par ICA (Independent Component Analysis), puis sélection des sondes les plus contributives par ICA.
3. Retrait des sondes affectées par les facteurs de confusion détectées par une régression linéaire sur toutes les variables cliniques, puis sélection des sondes les plus contributives par Analyse en Composante Principale (ACP).
4. Sélection des sondes les plus informatives dans la littérature (infloci).
5. Retrait des sondes situées sur les chromosomes X et Y, retrait des sondes affectées par le sexe détectées par une régression linéaire, sélection des sondes les plus variables.
6. Pas de pré-traitement.

Méthodes de déconvolution :

1. RefFreeEwas, avec les paramètres par défaut, puis une normalisation pour fixer la somme des proportion dans un patient à 1.
2. RefFreeEwas, avec initialisation personnalisée de la matrice T par clustering hiérarchique en distance euclidienne.
3. RefFreeEwas, avec initialisation de la matrice T par clustering hiérarchique en distance de Manhattan.
4. RefFreeEwas, avec initialisation personnalisée de la matrice T par clustering hiérarchique sur une ICA.
5. Résultat d'EDec (non reproductible).
6. RefFreeEwas avec les paramètres par défaut.

Globalement, les facteurs de confusion sont souvent pris en compte en détectant par régression linéaire les sondes reliées aux variables cliniques. C'est le sexe qui est le plus souvent pris en compte par les participants (par 4 des 6 méthodes) car c'est le facteur de confusion le plus évident. La sélection des sondes les plus informatives peut se faire par ICA (méthode 2) ou par ACP (méthode 3) pour sélectionner les sondes les plus contributives, mais aussi sur la variance (méthode 5), ou par recherche bibliographique (méthode 4).

Pour la déconvolution, la majorité des participants a choisi d'utiliser la méthode RefFreeEwas. Cela s'explique par la première phase exploratoire du data challenge, où elle s'est révélée la plus simple d'utilisation et la plus efficace sur les jeux de données utilisés. Les paramètres de la méthode, notamment l'initialisation de la matrice T, varient légèrement entre les groupes, mais il s'agit majoritairement de choix par tâtonnement justifiés par la maximisation du score obtenu. EDec a été utilisé pendant le data challenge uniquement par son développeur, mais de manière non-reproductible en faisant tourner la méthode en local sur son ordinateur et en ne soumettant que la matrice de résultat (le script utilisé n'est pas disponible).

3.3.2 Analyse des résultats des méthodes

Les scores obtenus par les participants pendant le data challenge sont assez serrés, et les choix faits relèvent parfois d'un test aléatoire des paramètres, en choisissant le meilleur résultat à posteriori. Afin d'aller un peu plus loin dans l'analyse des méthodes obtenues, nous avons dans un premier temps utilisé la simulation de base du data challenge et simulé 10 matrices D différentes à partir de celle-ci, avec des bruits gaussiens différents mais de même amplitude, pour évaluer si les croisements de méthodes entre pré-traitement et déconvolution permettaient d'améliorer les résultats.

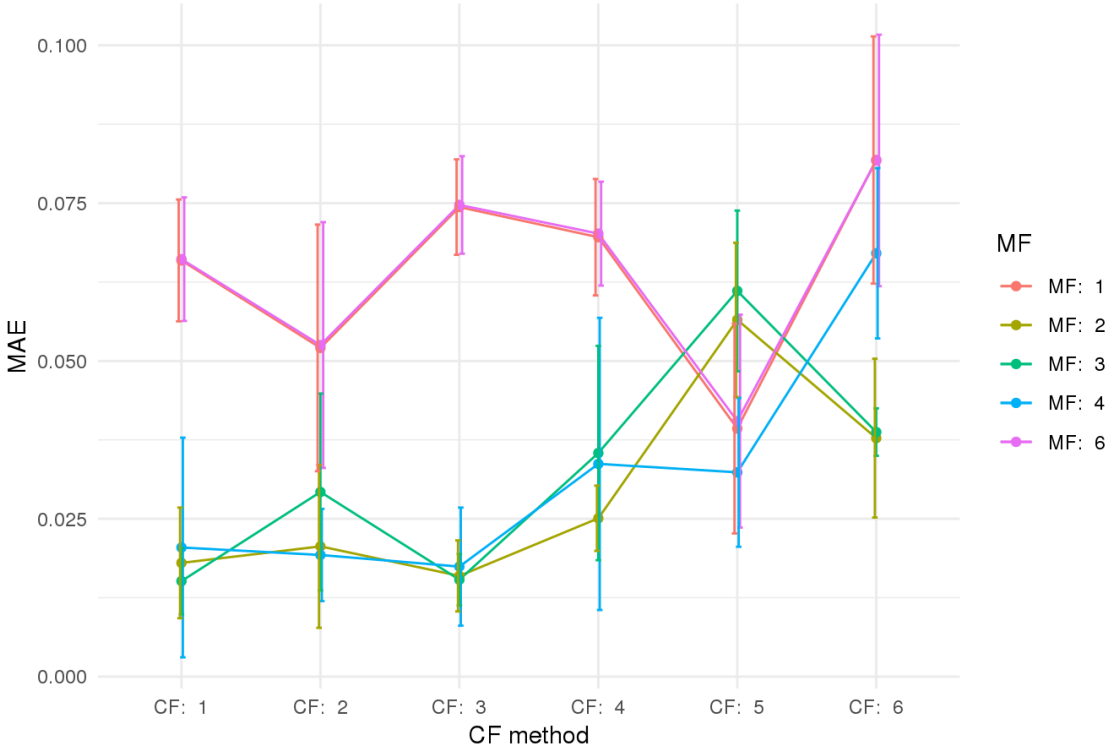


FIGURE 3.3 – **Résultats du croisement des méthodes de pré-traitement et de déconvolution obtenues.** Les 6 méthodes de pré-traitement (abrégé ici CF pour "confounding factor") ont été croisées avec les 5 méthodes de déconvolution reproductibles (abrégée ici MF pour "Matrix Factorisation") et appliquées à 10 réalisations de matrice D. Pour chacune, l'erreur absolue moyenne sur A sur ces 10 réalisations est calculée.

Les résultats obtenus sont visibles sur la figure 3.3, chaque ligne représentant une des méthodes de déconvolution. On observe deux grands groupes : les méthodes 1 et 6 d'une part avec une erreur globalement plus haute quelle que soit la méthode de pré-traitement utilisée, et les méthodes 2, 3 et 4 d'autre part, avec des résultats proches. Comme attendu, les paramètres par défauts de RefFreeEwas (méthodes 1 et 6) semblent donc donner des résultats moins bons que les paramètres optimisés basés sur les résultats du data challenge. En abscisse on voit l'impact des différentes méthodes de pré-traitement : les méthodes 1 à 4 présentent des résultats similaires, la méthode 5 de présélection des sondes dans la littérature regroupe les erreurs des différentes méthodes de déconvolution autour d'une même valeur, mais l'absence de pré-traitement (méthode 6) augmente globalement l'erreur quelle que soit la méthode de déconvolution. La prise en compte des facteurs de confusion et la présélection des sondes semblent donc être une étape d'intérêt. Ce qui ressort également, c'est qu'entre les dix réalisations de bruit de même amplitude (barres d'erreur), le résultat de la déconvolution peut beaucoup varier et les méthodes de déconvolution ne semblent pas si robustes que ça.

Ces résultats préliminaires sont encourageants, notamment au sujet du prétraitement des données, mais ils montrent aussi que les meilleures méthodes du data challenge sont largement basées sur la sur-optimisation par rapport aux simulations utilisées. Ainsi, il semble nécessaire de mieux comprendre l'impact des paramètres de simulation, leur influence sur les méthodes de déconvolution et de mettre en place une manière bien plus systématique de comparer les différentes méthodes de pré-traitement et de déconvolution.

Chapitre 4

Analyse comparative des méthodes de déconvolution

La data challenge a permis de découvrir les méthodes de déconvolution et les problématiques associées, mais pas vraiment d'obtenir une méthode se démarquant des autres. Dans cette partie, nous avons souhaité mettre en place une comparaison systématique des méthodes existantes de déconvolution reference-free de l'ADNm, et établir si le pré-traitement des données pouvait améliorer le résultat. Cette comparaison se basant sur des données simulées, nous avons également voulu explorer l'impact de tous les paramètres de simulation.

Le pipeline de déconvolution se déroule en plusieurs étapes : pré-traitement des données (en lien ou non avec les facteurs de confusion), choix du nombre de types cellulaires (partie peu abordée au cours du data challenge, mais essentielle sur de vraies données) et la déconvolution en elle-même.

Cette partie, qui correspond à l'élaboration d'un guide clair sur la déconvolution reference free de l'ADNm et à la comparaison des méthodes et des paramètres de simulation, a été publiée dans un article de recherche dans BMC Bioinformatics dont je suis première auteure (disponible en annexe 7.2.4) et implémentée dans un package R nommé MeDePir (voir partie 4.5.2).

4.1 Choix des méthodes

L'idée étant de constituer un pipeline composé de blocs interchangeables pour tester les différentes combinaisons de méthodes, il a fallu choisir celles utilisées pour chaque étape. C'est donc trois méthodes de déconvolution, trois méthodes de sélection des sondes informatives, une méthode optionnelle de correction des facteurs de confusion et quatre méthodes pour le choix du nombre de composantes à déconvoluer qui ont été sélectionnées.

4.1.1 Déconvolution

Étant donné la grande variabilité du comportement des méthodes de déconvolution observée en fonction des paramètres de simulations pendant le data challenge, nous avons décidé de tester les trois méthodes de déconvolution établies avant celui-ci : EDec, MeDeCom et RefFreeEwas (voir tableau 3.1).

EDec

La méthode EDec est lancée avec la fonction *run_edec_stage*1 avec les paramètres par défaut. Comme décrit précédemment, la particularité de cette méthode est d'utiliser une liste de sondes pré-sélectionnées dans la littérature faisant ainsi d'EDec une sorte de méthode semi-supervisée. Cependant, nous avons choisi de dissocier l'étape de sélection des sondes afin de comparer les méthodes sur la même base. La fonction est donc exécutée avec toutes les sondes en entrée, sans le filtrage pré-intégré des sondes.

RefFreeEwas (RFE)

L'algorithme RefFreeEwas, régulièrement abrégé RFE dans la suite du manuscrit, est exécuté avec la fonction *RefFreeCellMix* sur 9 itérations. Par défaut, la matrice T est initialisée par clustering hiérarchique sur T en utilisant une méthode de distance euclidienne, mais RefFreeEwas propose aussi d'autres méthodes d'initialisation. Afin d'évaluer leurs effets, nous avons testé les résultats de la déconvolution avec trois méthodes implémentées dans RefFreeEwas pour

initialiser la matrice T : le clustering hiérarchique sur D avec une distance euclidienne ou une distance de Manhattan (fonction *RefFreeEWAS : :RefFreeCellMixInitialize* en changeant le paramètre dist.method) et la décomposition en valeurs singulières (SVD) sur D avec la fonction *RefFreeEWAS : :RefFreeCellMixInitializeBySVD*. Nous avons également testé l'utilisation de la vraie matrice T simulée (avant ajout des facteurs de confusion) comme initialisation.

Nous avons donc fait tourner la déconvolution avec ces quatre méthodes d'initialisation sur dix réalisations de bruit de dix matrices D simulées différentes (soit 100 matrices D), puis nous avons calculé la moyenne de l'erreur absolue moyenne des matrices A obtenues (Figure 4.1).

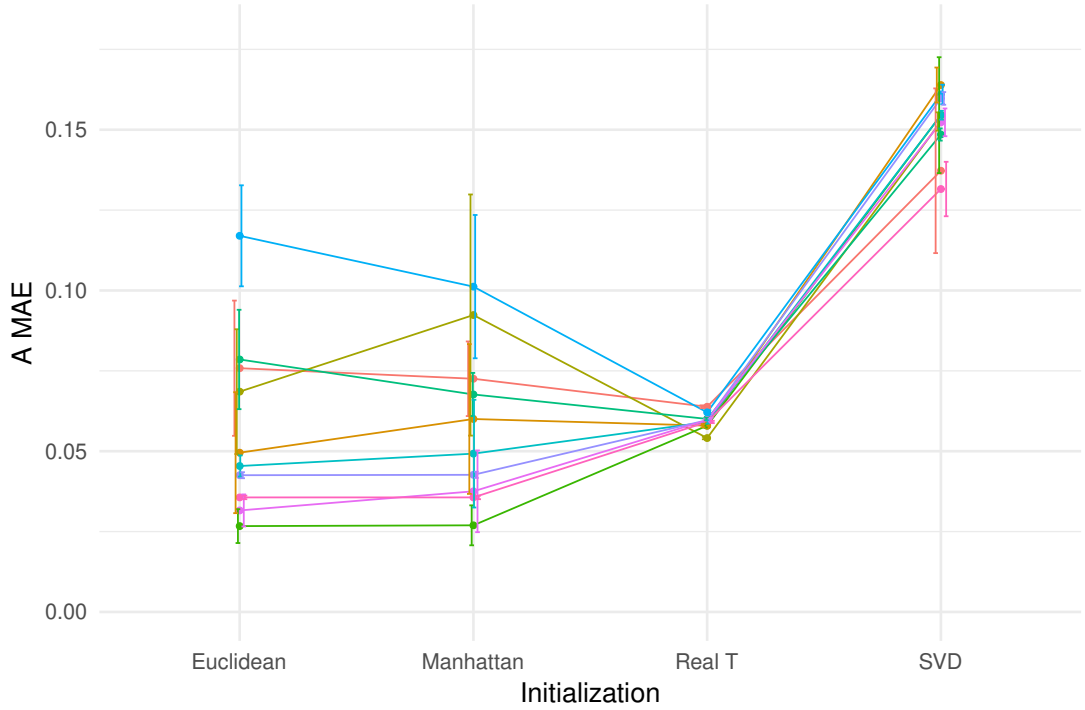


FIGURE 4.1 – **Impact de l'initialisation de RefFreeEwas sur la déconvolution.** La déconvolution a été réalisée sur dix fois dix matrices D simulées avec les paramètres par défaut (voir section 4.2). Chaque couleur correspond à une matrice D réalisée sur la base d'une matrice A différente (pour une matrice T unique), la barre d'erreur correspond à dix réalisations de bruit. Figure issue du papier.

Il apparaît que les résultats sont assez variables en fonction de la simulation, mais que les méthodes d'initialisation utilisant le clustering hiérarchique (Euclidean et Manhattan) donnent des résultats moyens semblables entre elles, parfois même meilleurs que l'utilisation de la vraie matrice T. Dans cet exemple, la méthode par SVD donne des résultats significativement moins bons. Pour la suite des analyses avec RefFreeEwas, on utilisera donc uniquement la méthode par défaut de classification hiérarchique basée sur la distance euclidienne.

MeDeCom (MDC)

La méthode MeDeCom, qui est utilisée grâce à la fonction *runMeDeCom* de leur package, comporte beaucoup de paramètres puisque l'étape de validation croisée est largement personnalisable. Nous avons défini la majorité des paramètres de la même façon que les auteurs de MeDeCom dans leur vignette de présentation du package : 10 initialisations aléatoires de la matrice A (NINIT = 10), 10 groupes pour la validation croisée (NFOLDS = 10), 300 itérations maximum pour calculer A et T (ITERMAX = 300).

Un autre des paramètres de la méthode, appelé λ permet de régler la contrainte sur les résultats de la matrice T. En effet, la matrice T est contrainte pour que ses valeurs soient proches de 0 ou de 1, ce qui correspond aux valeurs de méthylation couramment observées, mais le niveau de contrainte doit être choisi avec précaution pour pouvoir améliorer les résultats [101]. Le paramètre λ est donc choisi lui-aussi par validation croisée, en essayant à chaque fois les paramètres conseillés par les développeurs de la méthode (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0). C'est la fonction *getStatistics* qui permet ensuite de choisir le paramètre λ ayant la plus faible erreur de validation croisée (Cross-Validation Error, CVE), et donc de sélectionner les résultats de déconvolution correspondants.

4.1.2 Pré-traitement

L'étape de pré-traitement se compose de deux sous-parties : la correction des facteurs de confusion et le choix de sondes informatives.

Prise en compte des facteurs de confusion (Counfonding Factor, CF)

La première étape est la correction des facteurs de confusion. En effet, la méthylation de l'ADN est propre à chaque type cellulaire, mais dépend également de nombreux autres facteurs qui peuvent être biologiques comme l'âge ou le sexe mais aussi dépendant de l'exposition comme l'alimentation ou la cigarette [107]. Ces facteurs ne sont pas pris en compte dans les méthodes récentes de déconvolution des types cellulaires à partir des données d'ADNm.

Une seule méthode a été retenue pour cette étape. Une régression linéaire est effectuée sur les sondes pour chaque variable du groupe expérimental. Les p-valeurs obtenues sont corrigées avec la méthode Benjamini-Hochberg [111] grâce à la fonction R *p.adjust*, puis les sondes qui sont associées à un facteur de confusion avec un seuil de FDR (taux de fausse découverte) inférieur à 0,15 sont éliminées. Ce seuil a été fixé de manière assez arbitraire, cependant des tests ultérieurs ont montré la faible incidence de ce choix (voir la figure 4.2).

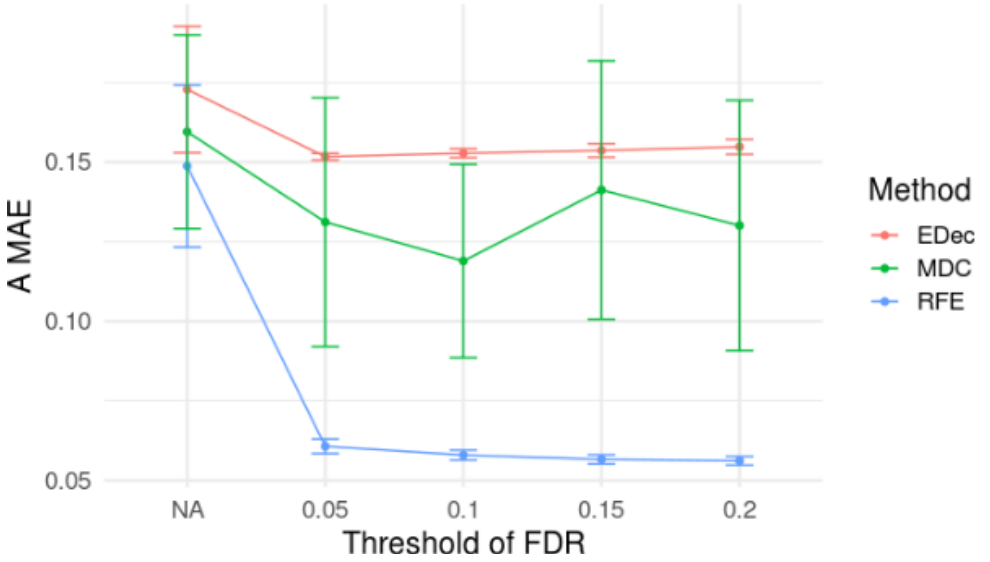


FIGURE 4.2 – **Impact du seuil de FDR pour la détection des facteurs de confusion.** La déconvolution a été réalisée sur une matrice D simulée, avec 10 réalisations de bruit (barre d'erreur). Chaque couleur correspond à méthode de déconvolution : EDec en rouge, MDC = MeDeCom en vert et RFE = RefFreeEwas en bleu. Pour un seuil de FDR de 0,05 on garde en moyenne 22 540 sondes, 21 998 sondes pour un seuil de 0,1 21 489 sondes pour un seuil de 0,15 et 20 950 sondes pour un seuil de 0,2. Figure issue du papier.

Sélection des sondes informatives (Feature Selection, FS)

La deuxième étape de pré-traitement consiste à sélectionner les sondes les plus informatives, c'est-à-dire celles qui sont le plus susceptibles de varier entre les types cellulaires et donc de faciliter leur différenciation. L'idée est ici double : d'une part, analyser uniquement ces sondes pourrait améliorer la déconvolution en diminuant le bruit ; d'autre part, analyser moins de sondes devrait diminuer la puissance de calcul nécessaire à la déconvolution et donc aussi le temps d'analyse. Trois méthodes de sélection des sondes ont été testées.

a. Sélection selon la variance. La méthode sur la variance (FS variance) consiste simplement à sélectionner les sondes ayant la plus grande variance entre les échantillons. Dans cette analyse, ce sont les sondes ayant une variance supérieure à 0,2 qui sont conservées. Cette valeur de 0,2 est un choix arbitraire. Un désavantage de cette méthode à seuil fixé est de ne pas conserver toujours le même nombre de sondes entre les analyses. La fonction implémentée dans la version finale du pipeline, dans le package Medepir (voir partie 4.5.2), permet une alternative : l'utilisateur peut fixer le nombre de sondes les plus variables qui seront conservées.

b. Sélection par ACP. La deuxième méthode fait partie de celles développées par les participants pendant le data challenge, il s'agit de conserver les sondes fortement corrélées (p-valeur < 0,1) avec les quatre premières composantes d'une analyse en composante principale (ACP) sur la matrice D. Ces sondes discriminant bien les données, elles sont donc particulièrement informatives. Cette méthode utilise la fonction *big_SVD* du package bigstatsr [112] pour effectuer l'ACP, puis une régression linéaire est effectuée entre les sondes et les composantes pour calculer la corrélation. Enfin, un test du χ^2 permet de sélectionner les sondes ayant une p-valeur inférieure à 0,1.

c. Sélection par à priori biologique (infloci). Cette méthode de sélection de sondes surnommée Infloci pour "informative loci" est une liste de sondes sélectionnées par les développeurs de EDec pendant le data-challenge pour appliquer

leur méthode. Ces sondes ont été trouvées dans la littérature comme particulièrement différentiellement méthylées entre les types cellulaires des tumeurs pulmonaires, et devraient donc faciliter la déconvolution de T.

4.1.3 Choix du nombre de types cellulaires (k)

Le choix du nombre de composantes, qui sont ici les types cellulaires, est un paramètre essentiel de la déconvolution (voir partie 4.3). Dans la partie suivante correspondant à l'analyse comparative des méthodes de déconvolution, nous appliquerons toutes les méthodes avec le vrai nombre de composantes correspondant à nos simulations. Le test de cette étape du pipeline est donc à part, quatre méthodes ont été testées.

Analyse en Composantes Principales (ACP)

Une des méthodes les plus intuitives pour déterminer le nombre de composantes est l'ACP. Si on part du principe que la méthylation dépend uniquement des types cellulaires, les principaux axes de l'analyse en composante principale de la matrice D devraient chacun correspondre à un type cellulaire, à l'exception du premier axe qui en discrimine deux. En affichant le scree plot représentant la valeur propre de chacune des composantes, on peut visuellement choisir le bon nombre de composantes selon la règle de Cattel [113] qui préconise de choisir le nombre de composants avant le coude (+ 1 pour le premier axe) (voir figure 4.6 D).

Clustering hiérarchique

Pour cette méthode, on commence par calculer la matrice de distance de Manhattan entre les sondes, puis on applique un clustering hiérarchique sur le résultat. On calcule ensuite le coefficient de silhouette $s_{sil}(k)$ pour différents nombres k de clusters. Cette métrique représente la différence moyenne entre la distance typique entre les points d'un même cluster et la distance typique avec les points des autres groupes [114]. Plus cette différence est élevée plus le découpage des groupes est efficace. Plus précisément, pour un point i dans

un cluster C_i ayant une distance moyenne d_g avec les points du cluster le plus proche et une distance moyenne d_v avec les autres points du cluster C_i, on peut définir :

$$s(i) = \frac{d_v(i) - d_g(i)}{\max(d_g(i), d_v(i))}$$

$s_{sil}(k)$ est la moyenne de $s(i)$ sur tous les points du système. En affichant $s_{sil}(k)$ en fonction de k sur une courbe, on choisit graphiquement le nombre optimal de groupes : le plus grand nombre de groupes qui permette de conserver un bon score (voir plus loin la figure 4.6, panneau A).

Validation croisée (MDC)

Le package MeDeCom propose une méthode pour déterminer k par validation croisée d'une façon semblable à celle utilisée pour déterminer leur paramètre λ (voir partie 4.1.1). Cette fois, c'est le paramètre λ qui est fixé, et la déconvolution est effectuée pour différentes valeurs de k. De nouveau, c'est le paramètre minimisant l'erreur de validation croisée (CVE) qui est sélectionné.

Bootstrap (RFE)

La méthode RefFreeEwas propose de choisir k par bootstrap avec la fonction *RefFreeCellMixArrayDevianceBoots*. Le principe est assez semblable à la méthode de cross-validation de MeDeCom ; la déconvolution est effectuée sur des sous-échantillons pour différentes valeurs de k, puis c'est la déviance du bootstrap qui est calculée. On choisit ensuite le k minimisant cette déviance par visualisation graphique (voir plus loin la figure 4.6, panneau B).

4.2 Étude des paramètres de simulations

4.2.1 Méthodologie

Avant de tester les différentes méthodes de pré-traitement des données, nous voulions mieux caractériser l'impact des paramètres de simulation sur les résultats des méthodes de déconvolution. Nous avons choisi de garder par défaut

les valeurs de paramètres utilisées durant la deuxième phase data challenge, puis de les faire varier un par un. Dans cette partie, nous n'avons pas considéré les facteurs de confusion et nous utilisons pour k le vrai nombre de composantes utilisées pour nos simulations ($k = 5$).

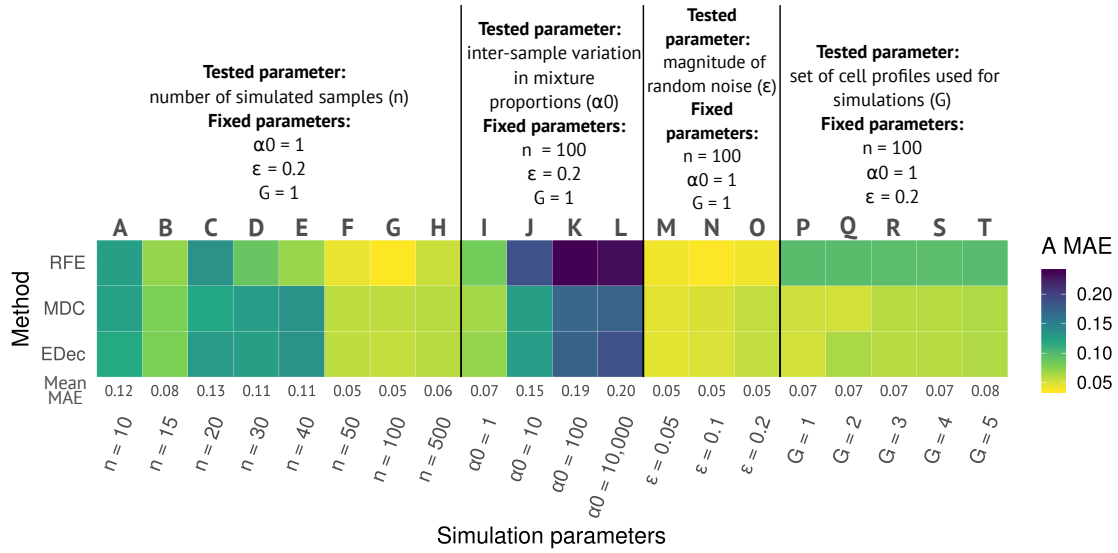


FIGURE 4.3 – **Résumé des résultats des trois méthodes de déconvolution sur les différentes simulations.** L'échelle de couleurs représente l'erreur absolue moyenne sur la matrice A. Les lignes correspondent aux trois méthodes : RFE pour RefFreeEwas, MDC pour MeDeCom et EDec pour EDec étape 1. Figure issue du papier.

Le nombre d'échantillons, n, était fixé par défaut à 100, et on a aussi testé les valeurs 10, 15, 20, 30, 40, 50 et 500. Le paramètre $\alpha 0$ qui représente la variabilité entre les proportions en type cellulaire des différents échantillons était fixé par défaut à 1, ce qui correspond à des échantillons très différents entre eux, nous avons donc testé des paramètres plus restrictifs regroupant les proportions autour de la valeur moyenne : 10, 100 et 10 000. Nous avons également testé l'effet de l'amplitude du bruit gaussien sur la matrice D, avec un écart-type ε fixé par défaut à 0,2 avec comme autres valeurs 0,05 et 0,1. Enfin, nous avons voulu nous assurer que les types cellulaires choisis pour la matrice T n'étaient

pas déterminants pour nos résultats, et nous avons donc essayé des simulations avec d'autres fonds génétiques (G), c'est-à-dire avec d'autres lignées cellulaires comme bases pour simuler T.

Pour chaque jeu de paramètres étudié, nous avons fait tourner les 3 méthodes de déconvolution sur dix réalisations de bruit différentes pour chacune des matrices D. Les résultats obtenus par chacune des méthodes sur les différentes simulations sont résumés dans la figure 4.3. Par la suite, nous allons commenter dans un premier temps l'influence de chaque paramètre puis comparer les trois méthodes de déconvolution entre elles.

4.2.2 Comparaison entre les paramètres

Nombre d'échantillons (n)

Le nombre d'échantillons présents dans le jeu de données est un paramètre important. En effet, un nombre plus important d'échantillons augmente l'hétérogénéité inter-individus et pourrait faciliter la déconvolution. C'est effectivement ce qu'on observe globalement (figure 4.4, panneau A). On remarque surtout un changement à partir du seuil $n = 50$, au-delà duquel l'erreur absolue moyenne sur la déconvolution de A diminue significativement pour toutes les méthodes. Les bons résultats observés pour $n = 15$ reflètent la sensibilité du pipeline d'analyse à la génération aléatoire de la matrice A ; ce point sera abordé dans partie Discussion.

Proportion des types cellulaires ($\alpha 0$)

Le paramètre $\alpha 0$, qui correspond à la variabilité des proportions en type cellulaires entre les échantillons, a lui aussi un fort impact sur la déconvolution. Plus $\alpha 0$ est petit, plus les proportions des différents types cellulaires varient autour de la valeur donnée, et donc plus les patients ont des compositions en type cellulaires différentes. Comme pour le nombre d'échantillons, cette augmentation de la diversité des observations facilite la déconvolution (voir figure 4.4, panneau B).

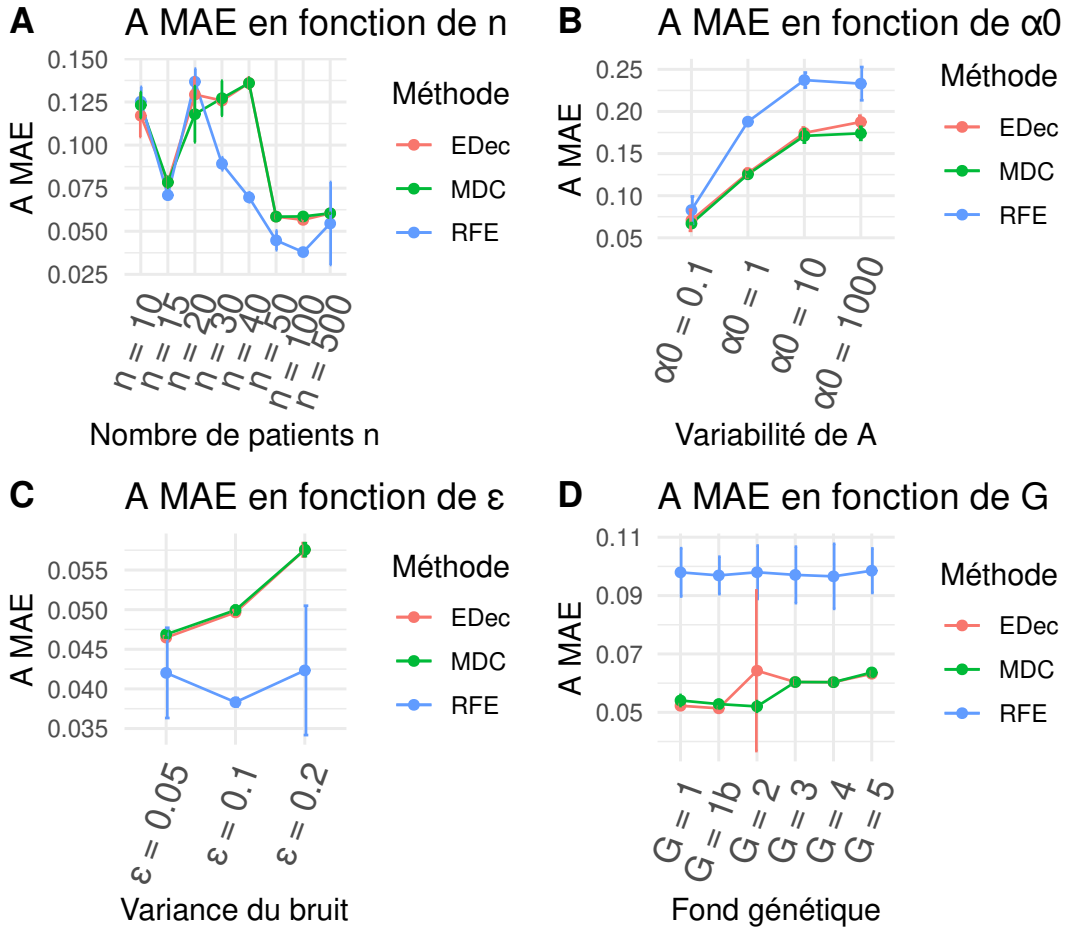


FIGURE 4.4 – **Variation de l'erreur absolue moyenne en fonction des différents paramètres de simulation pour les trois méthodes de déconvolution.** Les points correspondent à l'erreur absolue moyenne pour chaque paramètre (A : n, nombre d'échantillons, B : $\alpha 0$, variabilité de A, C : ϵ, écart type du bruit, D : G, composition de T), et pour chaque méthode de déconvolution : RFE pour RefFreeEwas, MDC pour MeDeCom et EDec pour EDec étape 1. Les barres d'erreur correspondent aux dix réalisations de bruit gaussien.

Écart-type du bruit (ε)

Notre analyse montre un faible impact du bruit sur les prédictions (figure 4.4, panneau C). Les trois méthodes de déconvolution sont robustes au bruit, mais ce faible effet peut aussi s'expliquer par notre façon de le simuler, un bruit gaussien étant assez peu réaliste.

Lignée du fond génétique (G)

	Épithélial cancer	Mésenchyme cancer	Épithélial sain	Fibroblaste sain	Lymphocyte
G1 (ref)	GSM1560930	GSM1560925	GSM2743808	GSM1354676	GSM1641099
G1bis	GSM1560930	GSM1560925	**GSM2743808**	GSM1354676	**GSM1641099**
G2	**GSM1560911**	**GSM1560931**	**GSM2743807**	**GSM1354675**	**GSM1641101**
G3	GSM1560930	GSM1560925	**GSM999346**	GSM1354676	GSM1641099
G4	GSM1560930	GSM1560925	**GSM999358**	GSM1354676	GSM1641099
G5	GSM1560930	GSM1560925	**GSM1337281**	GSM1354676	**GSM688851**

TABLEAU 4.1 – **Récapitulatif des lignées cellulaires utilisées pour simuler les matrices T alternatives.** Le fond génétique G1 correspond à la matrice T du data challenge. Les sondes en gras représentent les changements dans les autres matrices T, le orange représente les lignées séquencées en 450k et le rouge les lignées séquencées en 450k avec le type de sonde corrigé par BMIQ (voir texte). Dans la lignée G1bis, on corrige les sondes 450k. Dans la lignée G2, on change toutes les lignées de G1 par d'autres issues des mêmes jeux de données (voir tableau 3.2 pour le détail des jeux de données). Pour G3 et G4, on change uniquement les lignées 450k par d'autres issues des mêmes jeux de données. Pour la lignée G5, on change les lignées d'épithélial sain et de lymphocytes T par des lignées séquencées en 27k.

Lors du data challenge, les lignées utilisées pour simuler la matrice T avaient été choisies en mélangeant des technologies de sondes différentes (27k pour les lignées cancer et fibroblastes, 450k pour les lymphocytes T et l'épithélial contrôle, voir tableau 3.2), puis seules les sondes en commun avaient été conservées. Cependant les sondes de la technologie 450k ont deux types de conception : celles de type I sont identiques aux sondes 27k, mais celles de type II ont une distribution des valeurs de méthylation très différente, ce qui peut fausser les analyses [115]. Certaines sondes en commun avec la technologie 27k peuvent

avoir été remplacées par des sondes de type II en technologies 450k, ce qui introduirait un biais dans la déconvolution.

Pour une même matrice A, différentes matrices T ont été composées pour tester cet impact, leur composition est résumée dans le tableau 4.1. Nous avons testé d'autres lignées cellulaires issues des mêmes jeux de données (voir partie 3.2) que G1 (G2, G3, G4), pour tester l'effet de la lignée en elle-même. Nous avons aussi testé les lignées G1 en corrigeant l'effet des sondes de type II des lignées 450k avec le package BMIQ [115], en se basant sur une liste de références des sondes précédemment publiée [56] (G1bis). Enfin, nous avons remplacé les lignées issues de séquençage 450k des types cellulaires immunitaire et épithélial par des lignées séquencées en technologie 27k (G5).

Les résultats des méthodes de déconvolution (figure 4.4, panneau D) montrent que les lignées utilisées pour la matrice T n'ont pas un gros impact pour notre façon de simuler, toutes les matrices T permettent d'obtenir un score similaire, suggérant que la matrice A joue le rôle principal.

Variabilité en fonction de A

Dans cette partie, plusieurs matrices A différentes ont été simulées pour des jeux de paramètres communs, ainsi on retrouve 4 fois la combinaison $n = 100$, $\alpha 0 = 1$, $\epsilon = 0{,}2$ et G = 1 (figures 4.3 et 4.4), avec des comportements différents sur les résultats des méthodes de déconvolution. Ces résultats soulignent le rôle important de la réalisation de la matrice A (voir Discussion).

4.2.3 Comparaison entre les méthodes

Les résultats des trois méthodes de déconvolution, EDec, RefFreeEwas et MeDeCom sont visibles sur les figures 4.3 et 4.4. Nous n'observons pas de tendances claires pour une méthode ou une autre, la "meilleure" méthode semblant dépendre fortement de la simulation considérée.

4.3 Étude du nombre k de composantes

4.3.1 Comparaison des méthodes de sélection de k

Les quatre méthodes du choix de k ont été testées dans 4 conditions différentes : pour 3 ou 5 types cellulaires dans nos simulations, et avec ou sans retrait des sondes corrélées aux facteurs de confusion CF (figure 4.5 pour $k = 3$ et figure 4.6 pour $k = 5$). Le filtrage des facteurs de confusion sera abordé plus en détails dans la section suivante.

Sur les panneaux A des figures, on voit les résultats de la méthode par clustering hiérarchique. Pour $k = 3$ la méthode ne semble pas fonctionner, on retrouve $k = 4$ avant retrait des CF et $k = 5$ après. Pour $k = 5$ les résultats sont également étranges car la courbe ne se stabilise pas vraiment. Cette méthode est donc mise de côté car les résultats ne sont pas satisfaisants.

Pour la méthode de bootstrap de RefFreeEwas (panneau B), les résultats pour $k = 3$ sont satisfaisants avant ou après retrait des CF. Pour $k = 5$ le choix est compliqué, la visualisation par boîtes à moustaches rend difficile le choix optimal de k car les valeurs de déviance obtenues sont très proches. En revanche, elle reste robuste aux facteurs de confusion puisque dans la méthode les valeurs des sondes sont attribuées aléatoirement et ne suivent plus les valeurs biologiques.

Pour la méthode de validation croisée de MeDeCom, le nombre de types cellulaires choisi correspond à celui du coude. Pour $k = 3$ il apparaît clairement panneau C que les facteurs de confusion sont détectés comme un type cellulaire, la valeur de k est très bien déterminée après la sélection des sondes, alors qu'on trouve un type cellulaire de trop sans pré-filtrage. Pour $k = 5$ l'impact des facteurs de confusion semble moins fort, mais le choix est plus net après leur retrait.

Pour l'ACP, le nombre de composantes choisies par la règle de Cattel correspond à la valeur avant le coude sur la courbe ; le nombre de types cellulaires est ensuite égal au nombre de composantes + 1 puisque le premier axe distingue deux types cellulaires. Il apparaît clairement panneau D que les facteurs de confusion apparaissent dans l'ACP pour $k = 3$ et pour $k = 5$: un des axes disparaît après leur retrait, et le nombre de types cellulaires déduits devient

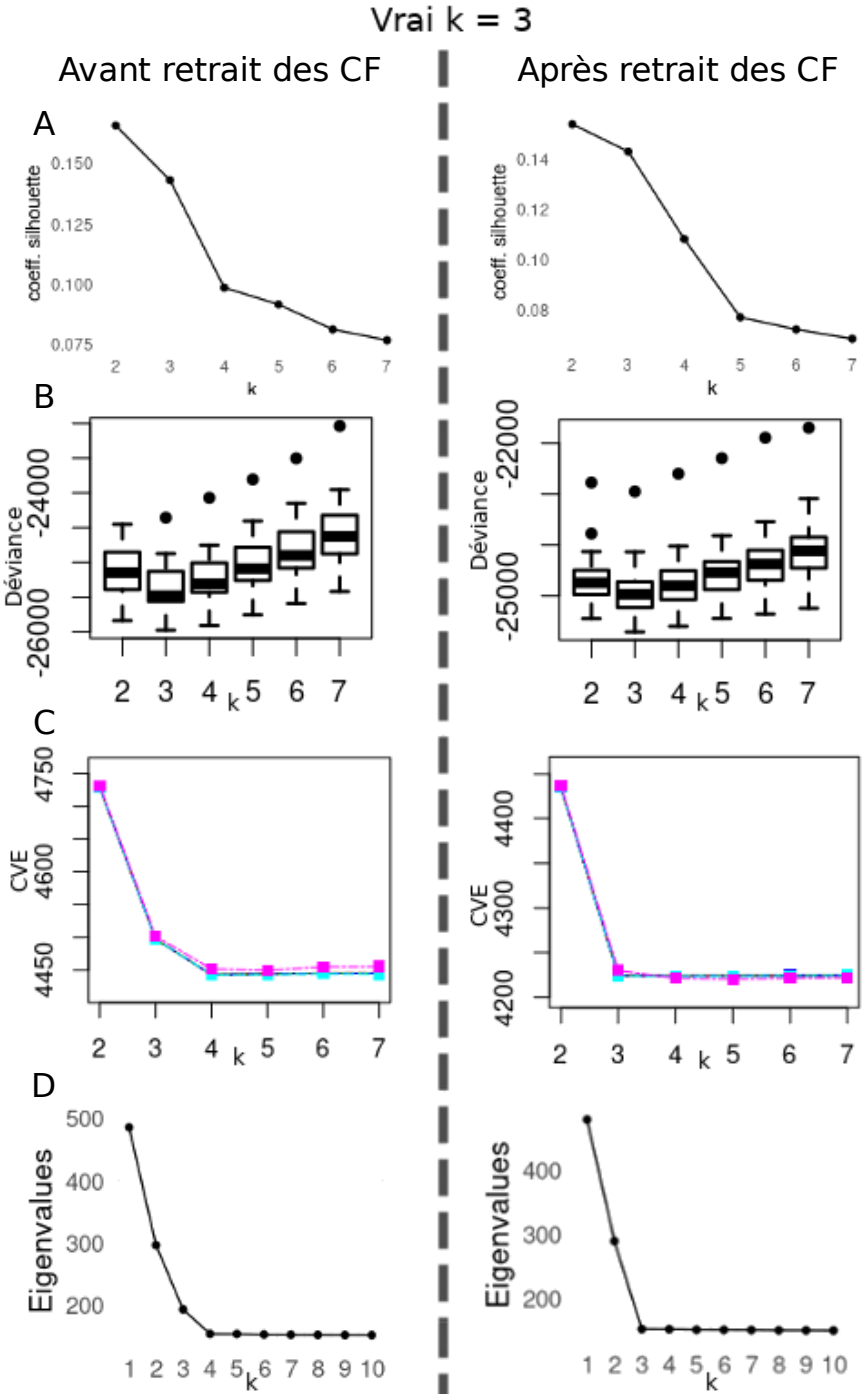


FIGURE 4.5 – **Résultats des différentes méthodes de choix de k pour k réel = 3.** À gauche les méthodes ont été lancées avant retrait des sondes corrélées aux facteurs de confusion, à droite après. Panneau A : méthode par clustering hiérarchique, le k choisi correspond au k maximal conservant un bon coefficient de silhouette. Panneau B : méthode par bootstrap, le k choisi correspond au k minimisant la déviance. Panneau C : méthode par cross-validation, le k choisi correspond au k minimisant la CVE. Panneau D : méthode par ACP, le k choisi correspond au nombre d'axes avant le coude +1. Voir partie 4.1.3 pour le détail des méthodes.

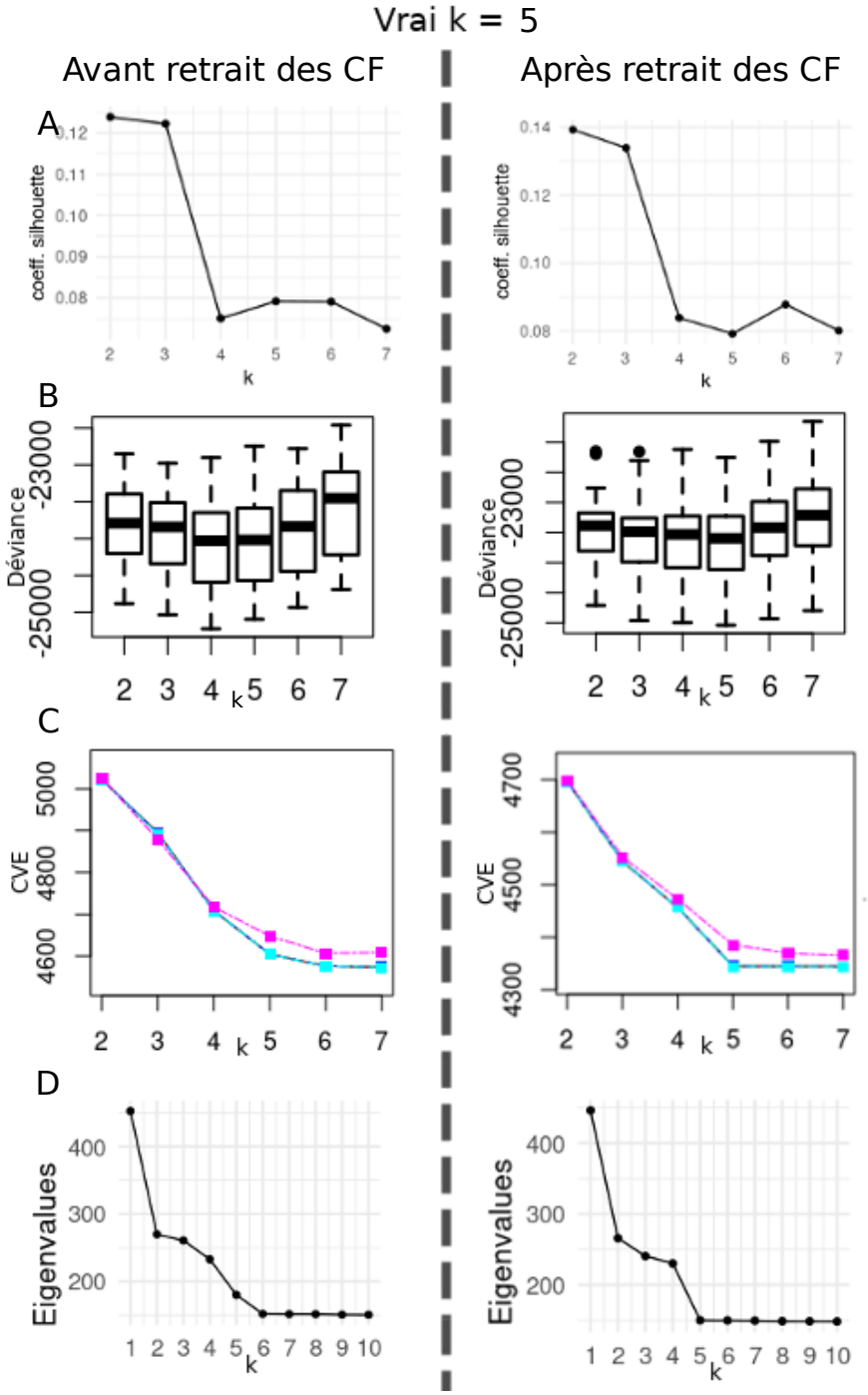


FIGURE 4.6 – **Résultats des différentes méthodes de choix de k pour k réel = 5.** À gauche les méthodes ont été lancées avant retrait des sondes corrélées aux facteurs de confusion, à droite après. Panneau A : méthode par clustering hiérarchique, le k choisi correspond au k maximal conservant un bon coefficient de silhouette. Panneau B : méthode par bootstrap, le k choisi correspond au k minimisant la déviance. Panneau C : méthode par cross-validation, le k choisi correspond au k minimisant la CVE. Panneau D : méthode par ACP, le k choisi correspond au nombre d'axes avant le coude +1. Voir partie 4.1.3 pour le détail des méthodes.

juste.

Globalement, les méthodes par validation croisée et bootstrap sont beaucoup plus complexes en ressources et en temps de calcul, et ne permettent pas d'améliorer la détection du bon k de manière significative par rapport à l'ACP. Pour la suite de nos analyses, nous utiliserons donc toujours la méthode par Analyse en Composantes Principales, qui donne de manière simple et rapide des résultats facilement interprétables. L'étape du choix de k doit se placer après le retrait des sondes corrélées aux facteurs de confusion, au risque sinon qu'ils soient détectés comme une des composantes.

4.3.2 Impact du choix de k

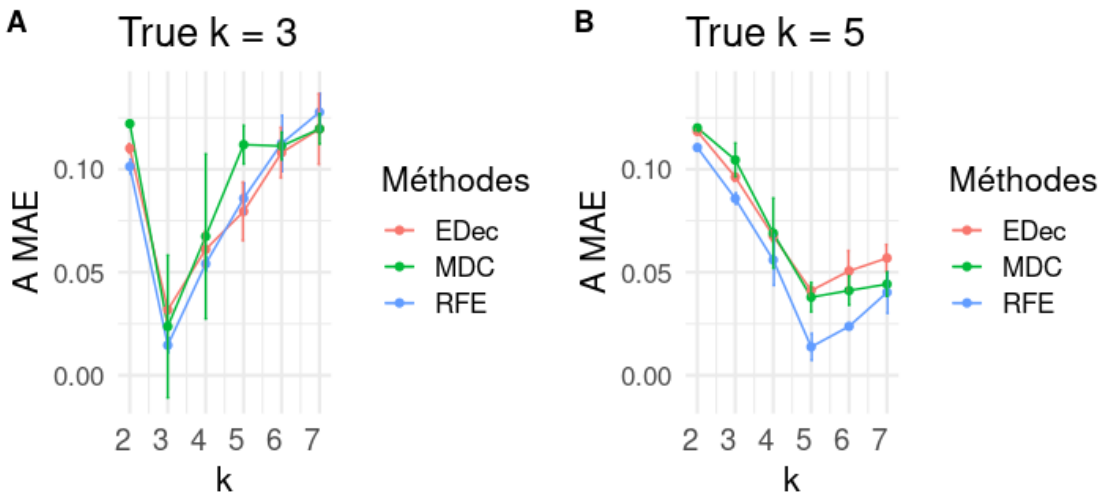


FIGURE 4.7 – **Impact du choix de k sur l'erreur absolue moyenne.** L'erreur absolue moyenne sur la déconvolution des trois méthodes est calculée pour une même matrice D et différents choix de k, pour k réel = 3 (A) et k réel = 5 (B). Les barres d'erreur correspondent à 10 réalisations de bruit différentes. Figure issue du papier.

Toutes les méthodes testées permettent d'entrer en paramètre le nombre de types cellulaires à déconvoluer, nous les avons donc faites tourner pour un k imposé variant entre 2 et 7, et un k réel de 3 ou de 5, puis nous avons calculé l'erreur absolue moyenne pour chacun des résultats (figure 4.7). Pour calculer

l'erreur absolue moyenne entre des matrices A de tailles différentes, notre algorithme rajoute des types cellulaires avec une composition nulle (ligne de 0 dans la matrice A) si le k utilisé est inférieur au k réel, ou à l'inverse sélectionne uniquement les types cellulaires déconvolués avec la meilleure corrélation si le k utilisé est supérieur au k réel.

Comme on peut s'y attendre, une mauvaise estimation de k a un gros impact sur les résultats, et à la fois le sous- ou le sur-estimer augmente considérablement l'erreur (par exemple, pour $k = 3$ et RefFreeEwas la MAE moyenne passe de 0,015 pour $k = 3$ à 0,101 pour $k = 2$ et 0,054 pour $k = 4$).

4.4 Impact de la pré-sélection des sondes

4.4.1 La détection des facteurs de confusion (CF)

Nous venons de voir que le retrait des sondes corrélées aux facteurs de confusion améliorait nettement la détection du bon nombre de types cellulaires dans les échantillons. C'est en fait aussi le cas pour le résultat de la déconvolution. En effet, en retirant les sondes comme décrit partie 4.1.2, c'est-à-dire 898 sondes sur 23381 pour $n = 100$ et 1894 pour $n = 20$, puis en faisant tourner les méthodes de déconvolution sur ces données, on peut voir des changements significatifs (figure 4.8). L'erreur est réduite d'environ 30 % en moyenne après le retrait des sondes corrélées. Les sondes corrélées aux données cliniques comme le sexe ou l'âge ont donc bien un impact négatif sur la qualité de la déconvolution.

4.4.2 La sélection des sondes informatives (FS)

Nous avons ensuite étudié l'effet de la sélection des sondes les plus informatives, et son comportement combiné avec le retrait de celles liées aux facteurs de confusion.

Sur la partie gauche des figures 4.9, on peut voir que les sélections basées sur la variance et sur l'ACP ne permettent pas de diminuer significativement l'erreur. En revanche, sélectionner les sondes en se basant sur la littérature avec infloci permet effectivement d'améliorer les résultats, avec en moyenne 47 % de

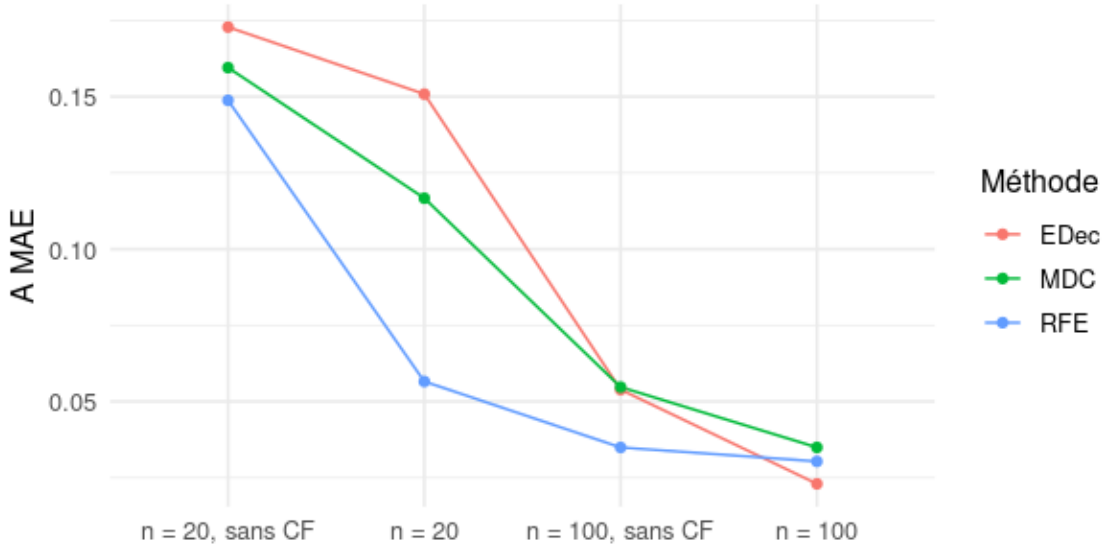


FIGURE 4.8 – **Impact du retrait des sondes reliées aux facteurs de confusion sur l'erreur absolue moyenne.** Les trois méthodes de déconvolution ont tourné avant ou après retrait des sondes reliées aux facteurs de confusion, pour $n = 20$ et $n = 100$. Le point correspond à la moyenne de 10 réalisations différentes du bruit.

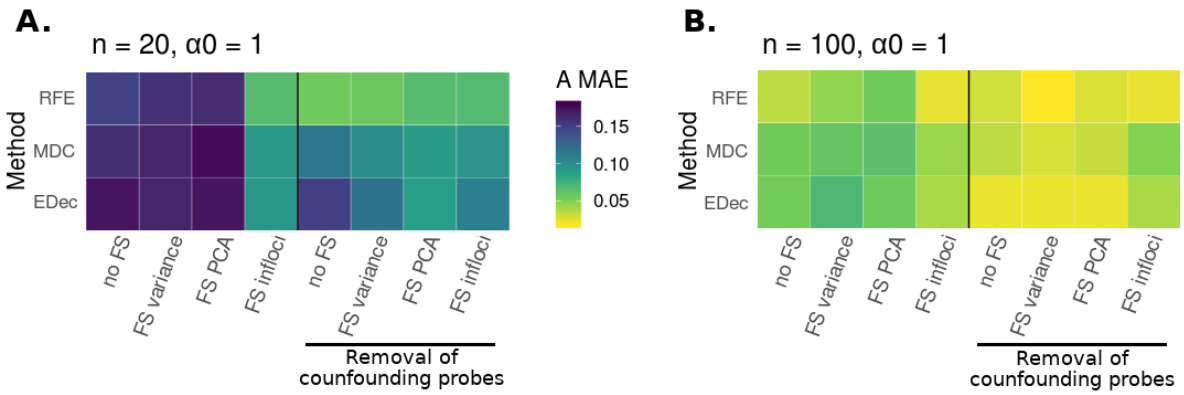


FIGURE 4.9 – **Impact de la sélection des sondes de méthylation sur les résultats de déconvolution.** Trois méthodes de sélection des sondes (FS) ont été testées (variance, ACP et infloci, voir partie 4.1.2), combinées ou non au retrait des sondes reliées aux facteurs de confusion (partie droite des figures). L'erreur absolue moyenne (sur 10 réalisations de bruit) est représentée pour les trois méthodes de déconvolution. Deux nombres de patients ont été testés : $n = 20$ (A) et $n = 100$ (B). Figure issue du papier.

réduction d'erreur pour $n = 20$ et 26 % pour $n = 100$.

Sur les parties droites des figures, on peut voir l'effet de l'étape de retrait des CF combiné aux différentes méthodes de FS. Cette fois, aucune des méthodes de FS ne permet de réduire significativement l'erreur. Ce qui est notable, c'est que les méthodes de FS variance et ACP laissent entre 6 000 et 10 000 sondes, alors que la méthode infloci en conserve uniquement dans les 600, on peut donc penser qu'elle inclue également le retrait de sondes liées aux facteurs de confusion, ce qui peut expliquer en partie ses bons résultats même quand elle est utilisée seule sans pré-filtrage des facteurs de confusion.

En revanche, même si diminuer le nombre de sondes ne permet pas d'améliorer les résultats de la déconvolution sur nos simulations, il y a un réel effet sur le temps de calcul. RefFreeEwas s'exécute toujours en moins d'une minute, mais les méthodes EDec et MeDeCom sont beaucoup plus longues, et on passe pour $n = 100$ de 25 minutes en moyenne sur toutes les sondes, à une dizaine de minutes après FS variance, et à moins d'une minute pour la FS infloci.

4.5 Conclusion sur les méthodes de déconvolution

4.5.1 Recommandations

À partir de toutes ces comparaisons, nous pouvons édicter certaines lignes de conduite quant au développement de pipeline de déconvolution (figure 4.10). La première étape doit être le retrait des sondes corrélées avec les facteurs de confusion, car elle a un impact significatif sur toutes les autres étapes. Ensuite, le choix de k peut être fait de différentes façons, mais nous avons montré qu'une simple ACP donnait des résultats concluants. L'étape suivante est la sélection des sondes, qui peut être faite par infloci si on a des à priori biologiques claires ou si on ne peut pas effectuer l'étape de retrait des sondes corrélées aux facteurs de confusion par manque de données cliniques sur les patients, ou par une autre méthode comme la sélection des sondes les plus variables si on veut gagner du temps de calcul. Enfin, les trois méthodes de déconvolution semblent donner des résultats équivalents, chacune étant légèrement meilleure dans des conditions particulières, sans qu'aucune des trois ne sorte particulièrement du lot.

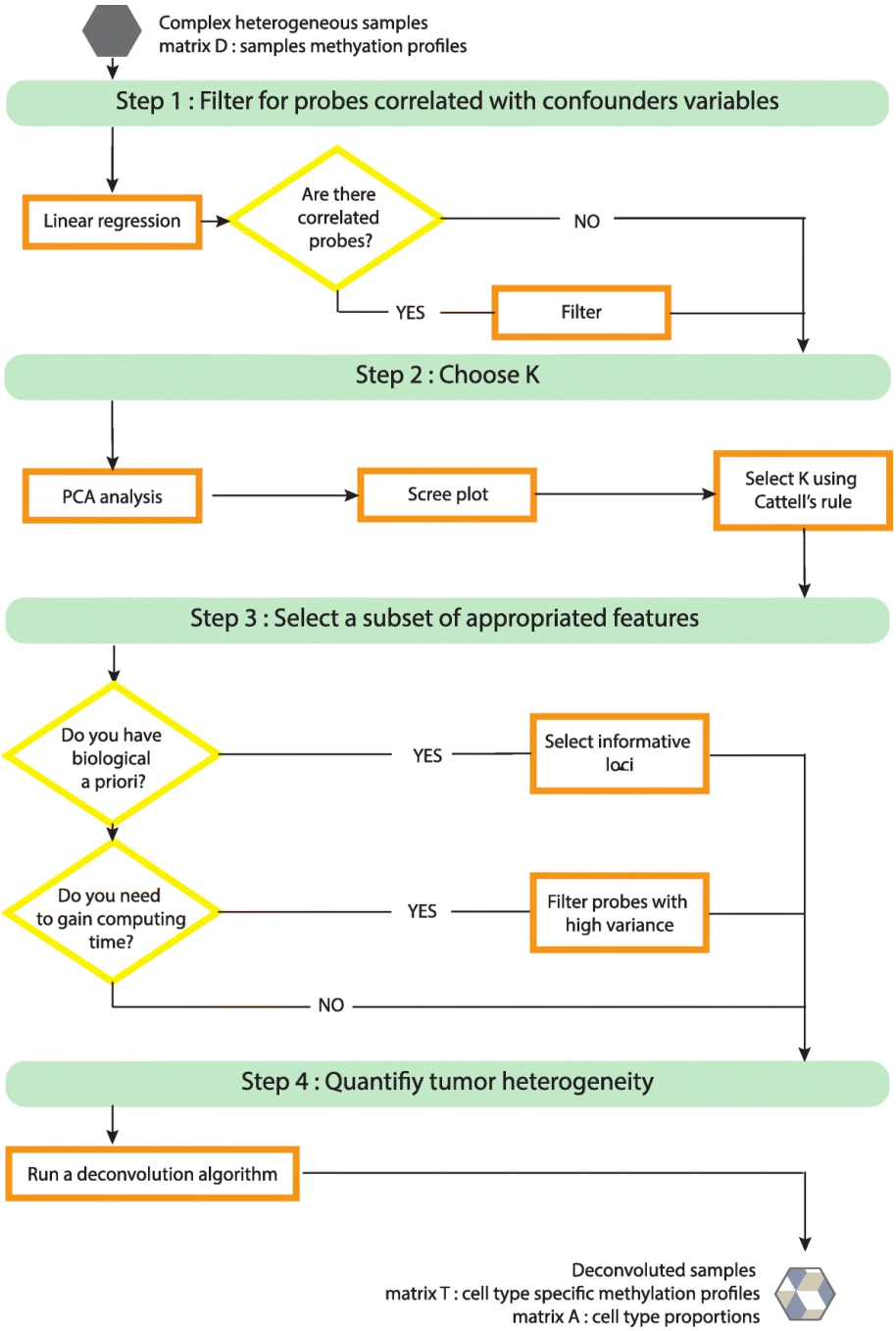


FIGURE 4.10 – **Résumé du pipeline de déconvolution.** La première étape consiste à filtrer les sondes corrélées à des facteurs de confusion, puis il faut établir le nombre de types cellulaires k. On peut faire une sélection dans les sondes avant de finalement lancer l'algorithme de déconvolution. Figure issue du papier.

4.5.2 Medepir

Pour effectuer simplement l'application de ces étapes, nous les avons implémentées dans R sous la forme d'un package appelé "Medepir" pour "MEthylation DEconvolution PIpeline in R" disponible sur Github. Le package permet de simuler la matrice de proportion A, puis la matrice D à partir des lignées souhaitées, et même de rajouter du bruit ou des sondes reliées aux facteurs de confusion, comme nous l'avons réalisé dans nos simulations (voir partie 3.2).

Ensuite pour la déconvolution, l'étape de sélection des CF par régression linéaire se fait avec la fonction *medepir : :CF_detection*, puis le choix de k par ACP avec *medepir : :plot_k*. Les sondes les plus variables sont sélectionnées par *medepir : :feature_selection*, et les méthodes de déconvolution sont lancées avec *medepir : :RFE*, *medepir : :EDec* et *medepir : :MDC*.

4.5.3 Limites et discussion

Une des limites de cette analyse comparative est d'être basée sur des simulations. Même si le pipeline Medepir a été testé sur un vrai jeu de données du cancer (voir partie 4.6), il faut être prudent quant à la généralisation de nos conclusions. Cependant, on peut être confiant sur l'intérêt du filtrage des sondes reliées aux facteurs de confusion et sur l'importance du choix de k.

Un problème important que nous avons observé au cours de notre analyse est l'extrême sensibilité des résultats données par les méthodes de déconvolution à la réalisation de la matrice A dans les simulations : pour des paramètres identiques de simulation, les résultats des méthodes de déconvolution peuvent beaucoup varier entre deux jeux de simulations (voir la figure 4.11). Ainsi, pour les résultats générés à partir de dix matrices A différentes, on observe une grande variabilité dans l'erreur absolue moyenne, avec parfois une méthode devenant significativement meilleure que les autres. Après retrait des sondes reliées aux facteurs de confusion (partie B de la figure), on voit une nette amélioration générale des résultats, mais il reste une variabilité entre les matrices. Il est important de prendre en compte cette sensibilité aux proportions de la matrice A pour les comparaisons ultérieures de méthodes de déconvolution.

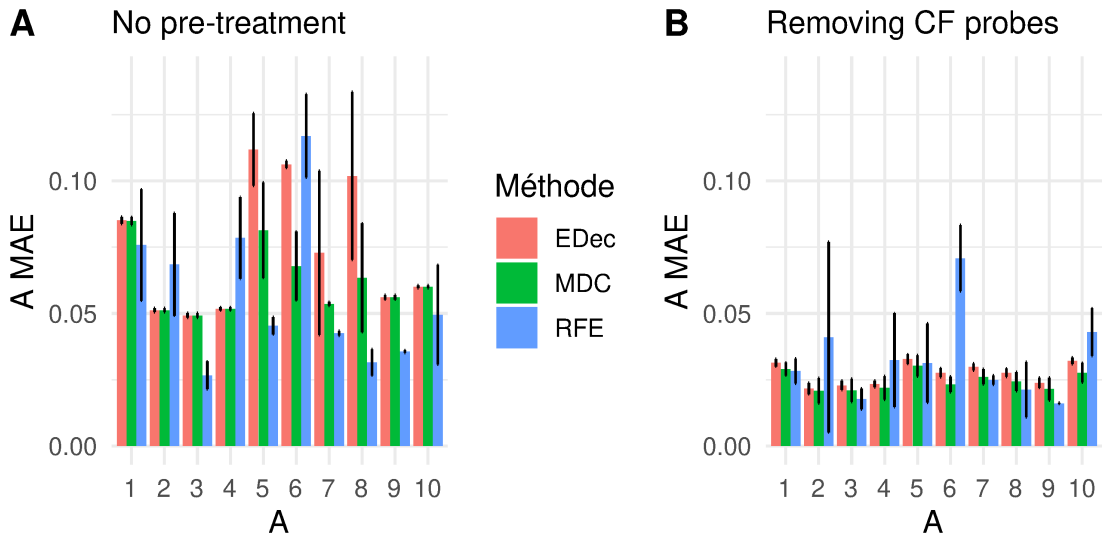


FIGURE 4.11 – **Variabilité des résultats en fonction de la matrice A.** Dix matrices D ont été générées à partir de dix matrices A différentes simulées avec les mêmes paramètres ($n = 100$, $\alpha 0 = 1$, $\epsilon = 0{,}2$). Les trois méthodes de déconvolution ont tourné avant (A) et après (B) retrait des sondes reliées aux facteurs de confusion. La barre d'erreur correspond à dix réalisations du bruit.

Un autre reproche qui peut être fait à notre travail est de nous concentrer sur l'analyse et l'optimisation de l'erreur sur la matrice A. Nous avons obtenus des résultats satisfaisants en comparant par corrélation de Pearson les différentes matrices T obtenues avec les types cellulaires utilisés dans les simulations (figure 4.12), cependant la distribution des valeurs de méthylation n'est pas toujours très réaliste (seule la méthode MeDeCom contraint la matrice T à avoir des valeurs proches de 0 ou de 1 pour une sonde donnée, ce qui est généralement le cas dans un tissu pur), et dans le cas de vraies données, il peut être difficile d'attribuer un sens biologique aux matrices T déconvoluées (voir partie III du manuscrit).

Enfin, on peut noter que si les méthodes de déconvolution se valent au niveau des résultats pour l'erreur sur A, elles ne sont pas toutes aussi simples à utiliser. Comme évoqué dans les résultats FS, RefFreeEwas a l'avantage de s'exécuter

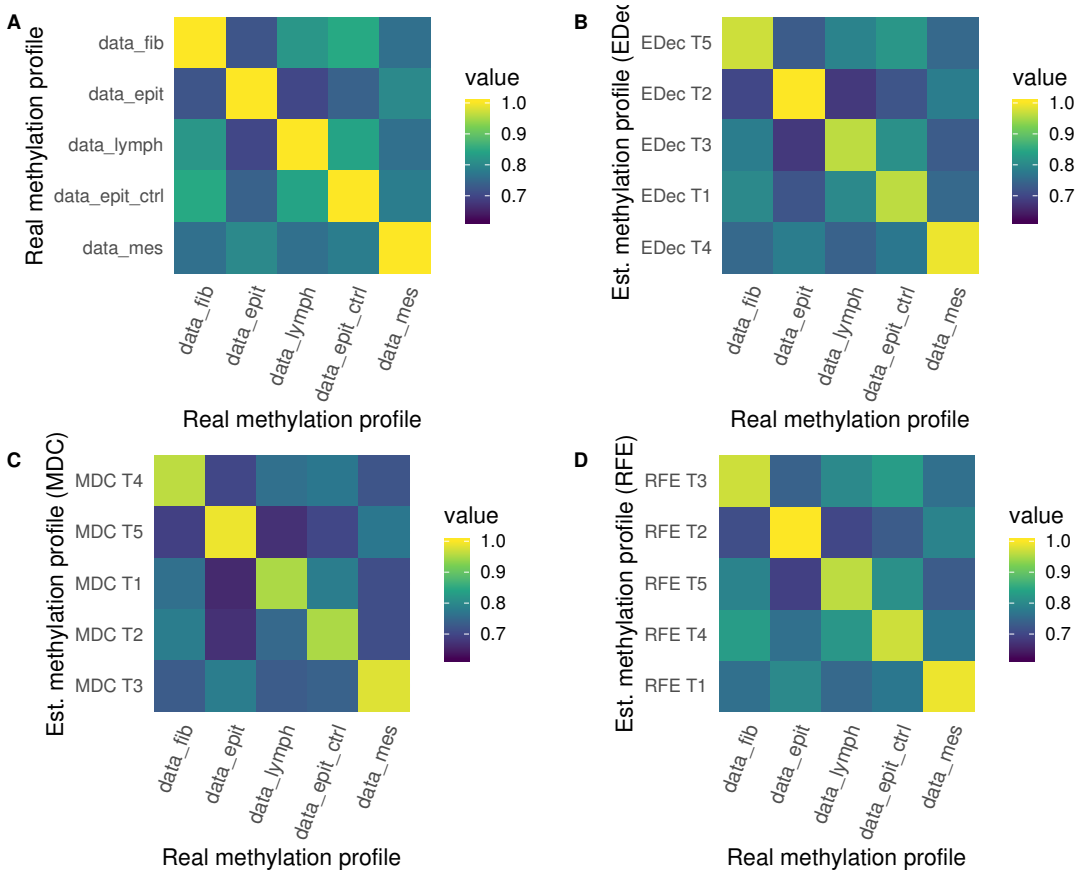


FIGURE 4.12 – **Corrélation entre les matrices T obtenues et la matrice T utilisée pour les simulations.** (A) Corrélation entre les différents types cellulaires utilisés pour la simulation, data_fib = fibroblastes, data_epith = épithélial cancéreux, data_lymph = lymphocytes T, data_epit_ctrl = épithélial sain et data_mes = mésenchyme cancéreux. (B,C,D) Les matrices T ont été obtenues par les différentes méthodes de déconvolution (B : EDec, C : MeDeCom et D : RefFreeEwas) après retrait des sondes corrélées aux facteurs de confusion. Figure issue du papier.

beaucoup plus rapidement que les deux autres. EDec, quant à elle, comporte tout un pipeline permettant d'inclure des sondes de la littérature et des données de RNA-seq, ce que ne proposent pas les autres méthodes, et ce qui n'a pas été évalué ici. Enfin, MeDeCom s'est révélée très difficile d'utilisation sans de puissantes ressources computationnelles (plus de 8 Go de mémoire vive) et un système d'exploitation adapté (le package ne s'installe que sous Linux, même si les développeurs travaillent à des versions compatibles pour Mac et Windows).

4.6 Application aux données LUAD et LUSC de TCGA

4.6.1 Pipeline

Pour tester notre pipeline sur un jeu de données réel, nous l'avons appliqué aux cohortes TCGA de l'adénocarcinome pulmonaire (LUAD) et du carcinome épidermoïde pulmonaire (LUSC) (voir partie 4 de l'introduction). Les données de méthylation de l'ADN ont été séquencées en 450k et comportent 370 échantillons pour LUSC et 456 pour LUAD. Avant d'appliquer le pipeline, les sondes avec des valeurs non détectées (NA) ou négatives dans tous les patients ont été retirées, et la variabilité entre les sondes de type I et II a été normalisée avec le package BMIQ [115] et la référence [56].

Le pipeline résumé en figure 4.10 a été appliqué. Pour la première étape, la détection des sondes associées aux facteurs de confusion, 18 826 sondes pour LUSC et 98 580 pour LUAD ont été retirées. Ensuite, l'ACP nous a permis de déterminer $k = 5$ dans LUAD et $k = 4$ dans LUSC (figure 4.13).

L'étape de sélection des sondes par la variance a permis de finalement conserver 29 053 sondes dans LUAD et 97 600 dans LUSC. Enfin, les 3 méthodes de déconvolution EDec, MeDeCom et RefFreeEwas ont été appliquées à la matrice D finale.

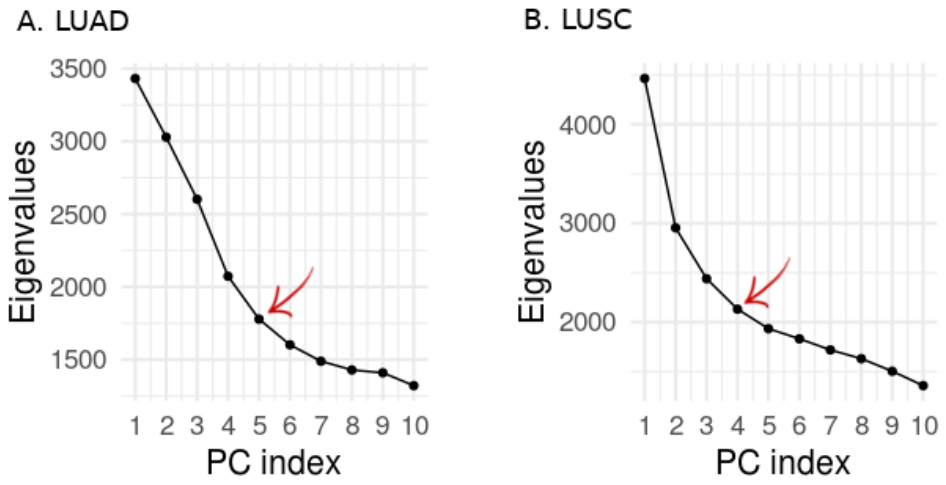


FIGURE 4.13 – **Résultats de l'ACP pour déterminer k dans les cancers TCGA LUAD et LUSC.** Les flèches rouges correspondent au coude choisi, la règle de Cattel indique de prendre le nombre de composants (PC) précédent. Le nombre de types cellulaires correspond au nombre de composant + 1 car le premier axe discrimine deux types cellulaires. On retrouve donc $k = 5$ pour LUAD (A) et $k = 4$ pour LUSC (B).

4.6.2 Comparaison avec les méthodes existantes

Nous avons voulu comparer les résultats de notre pipeline avec deux autres approches courantes pour inférer la composition du micro-environnement tumoral, EpiDISH et Estimate.

EpiDISH [103] est une méthode de déconvolution supervisée de l'ADNm se basant sur des profils de types cellulaires de référence pour inférer les proportions cellulaires. La méthode EpiDISH a été appliquée aux données LUAD et LUSC grâce à la fonction *epidish* et le jeu de référence *centEpiFibIC.m*.

Estimate [99] est quant à elle basée sur les données de RNA-seq (qui sont également disponibles pour les cohortes LUAD et LUSC), et utilise des signatures d'expression de gènes pour déterminer la pureté tumorale et la proportion en cellules immunitaires et stromales dans la tumeur. Les proportions de cellules immunitaires d'Estimate ont directement été récupérées à partir de leurs analyses sur leur site internet (estimate) pour les échantillons communs entre la méthylation 450k et le RNA-seq, soit 449 pour LUAD et 365 pour LUSC.

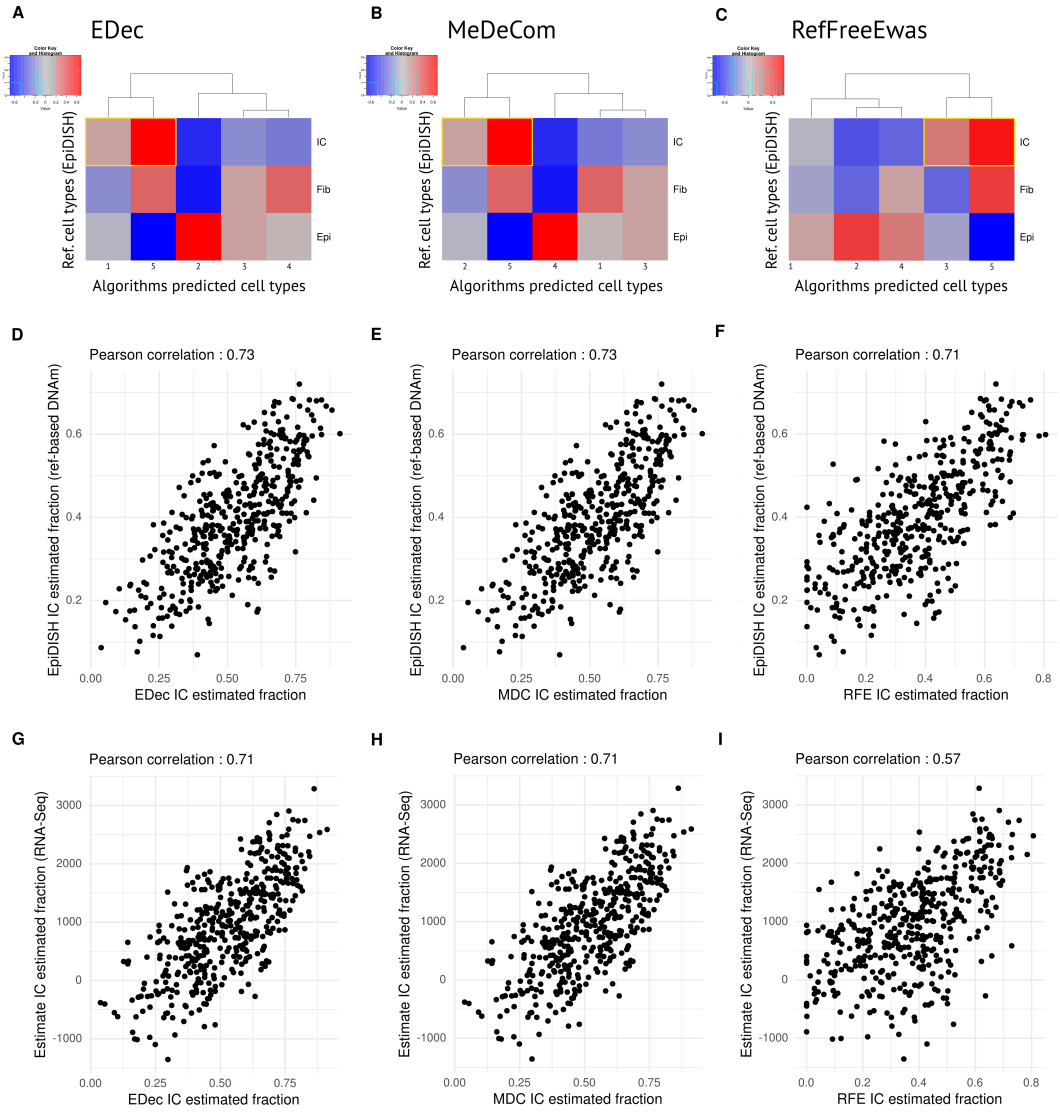


FIGURE 4.14 – **Application du pipeline sur les données TCGA LUAD.** Les matrices T obtenues par les méthodes reference-free sont corrélées (méthode de Pearson) aux références Epidish (A pour EDec, B pour MeDeCom, C pour RefFreeEwas). Le carré jaune correspond aux types cellulaires identifiés comme étant le type "cellule immunitaire". Les proportions des types cellulaires sont ensuite corrélées aux proportions obtenues par Epidish (D, E, F) et Estimate (G, H, I).

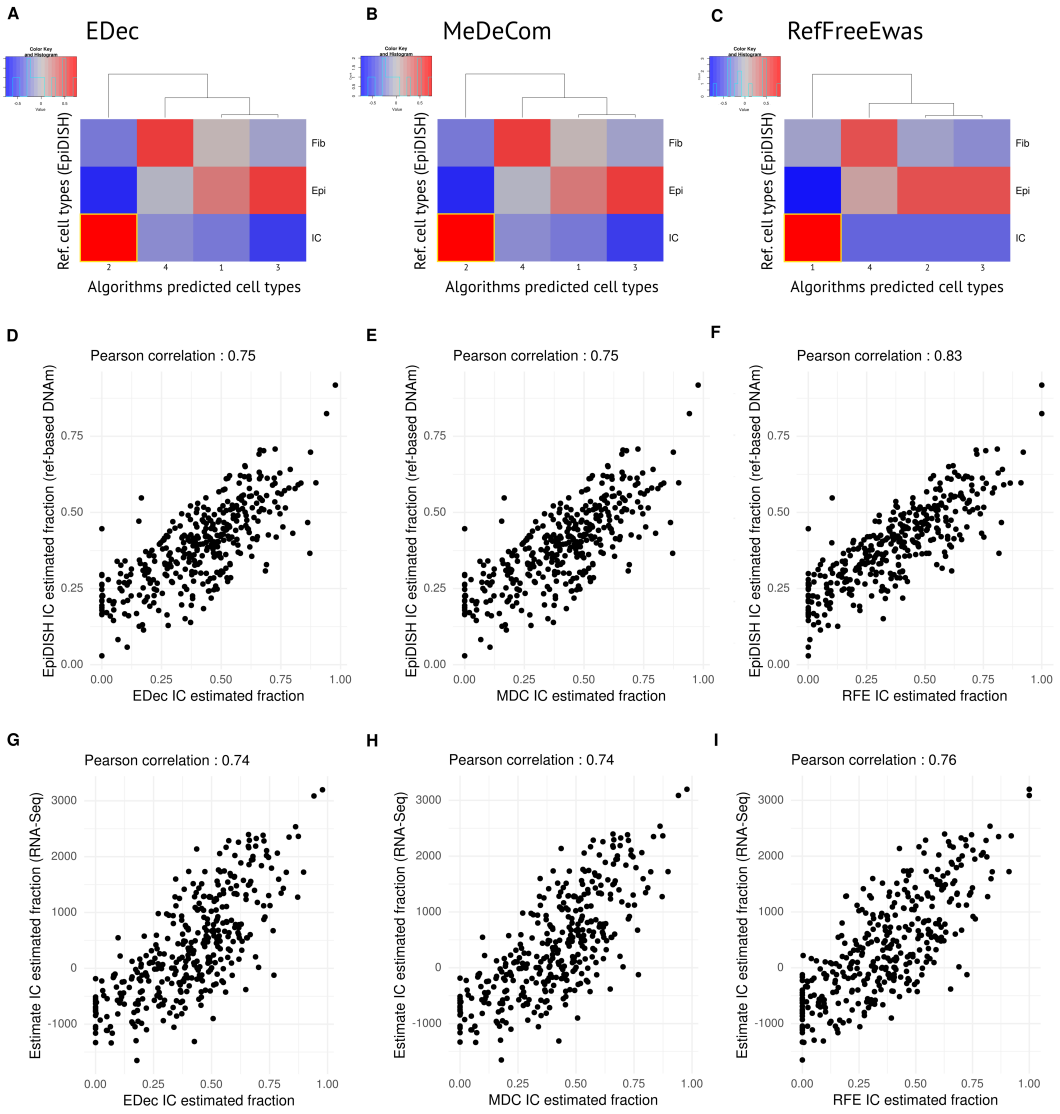


FIGURE 4.15 – **Application du pipeline sur les données TCGA LUSC.** Les matrices T obtenues par les méthodes reference-free sont corrélées (méthode de Pearson) aux références Epidish (A pour EDec, B pour MeDeCom, C pour RefFreeEwas). Le carré jaune correspond au type cellulaire identifié comme étant le type "cellule immunitaire". Les proportions des types cellulaires sont ensuite corrélées aux proportions obtenues par Epidish (D, E, F) et Estimate (G, H, I).

Dans un premier temps, nous avons utilisé la matrice de référence d'EpiDISH pour identifier les types cellulaires inférés par les méthodes reference-free en la corrélant aux matrices T obtenues avec le pipeline Medepir (figures 4.14 et 4.15, panneaux A, B, C). Par la suite, nous nous sommes particulièrement intéressés aux proportions des cellules immunitaires pour comparer les méthodes. Pour LUAD (figure 4.14), elles sont identifiées dans deux types déconvolués, et c'est donc la somme de ces proportions qui sont utilisées. Pour LUSC (figure 4.15), un seul type déconvolué est relié au type immunitaire.

Dans un second temps, nous avons comparé les proportions de cellules immunitaires déconvoluées par EDec, MeDeCom et RefFreeEwas en suivant notre pipeline avec celles d'Estimate et d'Epidish (figures 4.14 et 4.15, panneaux D à H). Pour les deux cancers, on retrouve une bonne corrélation, à la fois avec la méthode se basant sur des références et avec la méthode sur du RNA-seq. Ces résultats suggèrent que le pipeline Medepir est applicable à de vrais jeux de données hors simulations, et que les matrices déconvoluées ainsi peuvent avoir un sens biologique.

Chapitre 5

Élargissement de la question de la déconvolution

Les résultats de la déconvolution sans référence sur les données de méthylation étaient prometteurs et nous ont poussé à nous tourner vers la déconvolution à partir de données de RNA-seq. En effet, même si l'ADNm nous semblait à priori la meilleure source d'information en raison du nombre élevé de sondes et de la stabilité entre les types cellulaires, beaucoup de méthodes existantes de déconvolution se basent sur du RNA-seq et l'avantage d'un type de données ou de l'autre sur l'efficacité de la déconvolution n'est pas établi. De plus, l'intégration des deux types de données pourrait améliorer les résultats, et cette approche multi-omique nous semblait sous-exploitée.

Dans un premier temps, nous nous sommes donc intéressés à cette question intégrative à travers un deuxième data challenge. Ensuite, la question de la maintenabilité dans le temps de ces comparaisons entre méthodes sera abordée. Une partie de ces résultats est contenue dans un article tout juste accepté dans BMC Bioinformatics et donné en annexe 7.2.4.

5.1 Data challenge sur l'intégration multi-omique

5.1.1 Objectifs et organisation

Un deuxième data challenge a été organisé en 2019 sur le même modèle que le premier, mais cette fois sur la possibilité d'améliorer les résultats de la déconvolution en intégrant différents types de données omiques. À priori, combiner deux sources d'informations pour un même échantillon devrait faciliter la déconvolution, mais les données de RNA-seq et d'ADNm sont très différentes et le lien entre les deux n'est pas trivial, un travail exploratoire sur le sujet est donc nécessaire.

Ce projet a été mené en collaboration avec une équipe de la ligue contre le cancer travaillant sur des données de RNA-seq (programme carte d'identité tumorale), et Jérôme Cros, un collaborateur anatomo-pathologiste des hôpitaux de Paris spécialiste du cancer du pancréas. Dans le cadre d'une collaboration européenne, une équipe de l'Université d'Uppsala travaillant sur l'intérêt pédagogique du data challenge était également présente pour recueillir les avis des participants.

Le data challenge était de nouveau organisé en deux parties. Dans la première, les participants étaient en équipe travaillant soit uniquement sur des données de RNA-seq, soit uniquement d'ADNm. Dans la deuxième partie du challenge, les équipes étaient mélangées pour rassembler des personnes ayant déjà vu les deux types de données, et devaient désormais travailler à développer une méthode les intégrant toutes les deux.

5.1.2 Simulations

Les simulations sont dans l'ensemble proches de celles du 1er challenge (partie 3.2), mais ont été améliorées sur certains aspects. La vue d'ensemble se trouve en figure 5.1, je vais détailler chaque partie dans ce paragraphe.

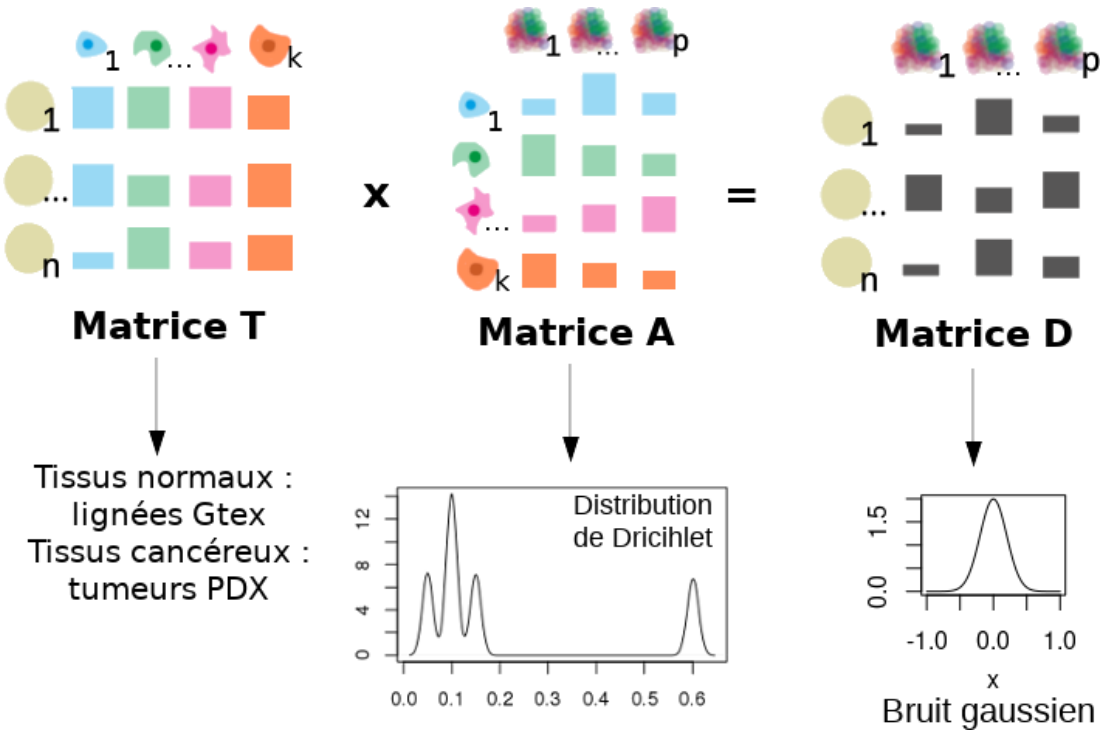


FIGURE 5.1 – **Résumé de la procédure de simulations du deuxième data challenge.** Les matrices T sont générées à partir de lignées cellulaires, provenant de la bases de données Gtex pour les types normaux et provenant de tumeurs PDX pour les types cancéreux. La matrice A des proportions en types cellulaire est calculée par une distribution de Dirichlet. Le produit de ces deux matrices donne la matrice D, sur laquelle un bruit gaussien suivant une loi $\mathrm{N}(0, 0, 2^2)$ est appliqué.

Matrices A

Les matrices A sont toujours simulées par une fonction de Dirichlet. On fixe $n = 30$ et $\alpha 0 = 10$ afin de ne pas obtenir des matrices trop simples à déconvoluer. Pour la première partie du data challenge, trois types cellulaires sont fixés : 45 % fibroblaste, 10 % immunitaire, 45 % cancer. Pour la deuxième partie, nous avons choisi 15 % normal, 45 % fibroblaste, 10 % immunitaire et 30 % de cancer, mais une nouvelle hétérogénéité au sein du type cellulaire cancer est introduite. Les 30 % sont partagés entre les deux sous-types cancéreux qui constituent les tumeurs du pancréas : le type basal et le type classique, soit à part égale, soit à trois quarts classique, soit à 90 % classique. Cette simulation se veut plus réaliste que

celle du premier data challenge, d'une part par le choix des proportions réalistes déterminées avec l'aide de J. Cros, et d'autre part par le choix de paramètres de Dirichlet plus complexes (part des deux types de cancer variables, $\alpha 0$ plus grand).

Il a été abordé dans la partie précédente, lors de notre analyse comparative des méthodes dans le pipeline Medepir, que la réalisation de A pouvait grandement influencer les résultats des méthodes. Pour éviter d'avantager par hasard une méthode en particulier, nous avons donc réalisé dix matrices A, puis nous avons fait tourner EDec, MeDeCom, RefFreeEwas et Epidish sur chacune d'entre elles et choisi des réalisations de A "neutres" ne favorisant aucune d'entre elles.

Matrices T

Pour les matrices T, nous bénéficions de meilleures données grâce à l'aide de l'équipe de la ligue contre le cancer. Pour les types cellulaires immunitaire, normal et fibroblaste, nous avons pris les données RNA-seq et ADNm sur la base de données publiques GTEX [116]. Pour les deux types de tumeurs (basal et classique), ce sont des lignées PDX de nos collaborateurs. Les lignées PDX proviennent d'une méthode où les cellules issues de tumeurs primaires sont retirées par chirurgie, triées par cytométrie en flux puis cultivées par xénogreffe dans des souris, ce qui permet d'obtenir des lignées cellulaires tumorales pures [117].

Toutes les données de méthylation (15 lignées cancer classique, 3 cancer basal, 3 immune, 3 normal et 8 fibroblaste) ont été séquencées en 850k. Les données de RNA-seq (15 cancer classique, 3 cancer basal, 16 immune, 29 normal et 33 fibroblaste) ont quant à elles été normalisées par edgeR [42] avec la méthode TMM [118] pour normaliser les échantillons entre eux, puis sont converties en CPM (count per million, en fonction de la taille des gènes).

Afin de tester la robustesse des méthodes, deux variations des matrices T sont effectuées à partir de ces lignées pour chacun des deux types de données omiques (figure 5.2). La première option est d'utiliser la médiane des différentes

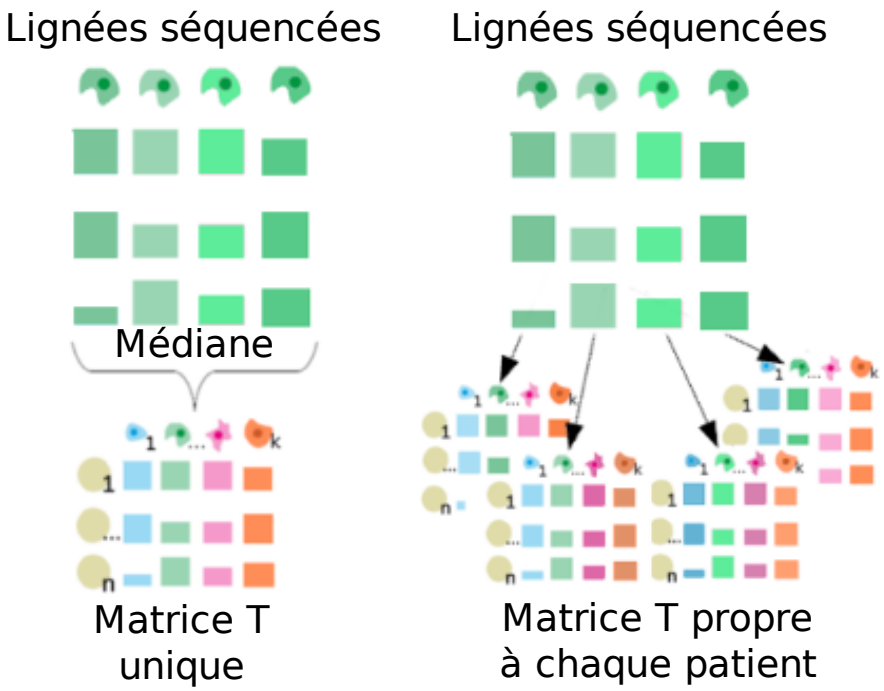


FIGURE 5.2 – **Les deux types de matrices T pour le deuxième data challenge.** Les matrices T sont générées à partir de lignées cellulaires. À gauche, on utilise la médiane des différentes lignées comme type de référence, et on obtient une matrice T unique pour tous les individus. À droite, on utilise une lignée différente à chaque fois, et on obtient une matrice T par patient.

lignées GTEX correspondant au type considéré, comme pour le premier data challenge, et d'obtenir une matrice T unique commune à tous les échantillons. La seconde est d'utiliser des lignées différentes dans chaque matrice, et donc d'obtenir une matrice T différente pour chaque patient. La deuxième option est à priori plus difficile à déconvoluer, mais est plus réaliste biologiquement. Dans la première partie du data challenge, le score des participants sera calculé sur ces deux options afin d'évaluer l'impact de l'hétérogénéité de la matrice T sur la déconvolution. Dans la seconde partie, on retourne à une matrice T unique afin de se concentrer sur l'intégration multi-omiques.

Notons que dans ce data challenge, il n'est plus question de la gestion des facteurs de confusion, là encore afin de se concentrer sur l'intégration des différents types de données.

Matrices D

On génère dix matrices D (RNA-seq et ADNm) en rajoutant un bruit gaussien aux matrices D issues de la multiplication de A et T, suivant une loi normale $N(0, 0, 1^2)$.

5.1.3 Méthodes obtenues

La première phase du data challenge consistait seulement à tester les méthodes de déconvolution existantes. En revanche, à l'issue de la deuxième phase, huit méthodes ont été développées par les participants. Trois utilisent uniquement la méthylation de l'ADN, deux les données de RNA-seq et trois méthodes se veulent intégratives et utilisent les deux types de données. Ces méthodes sont résumées dans le tableau 5.1, et comme dans la partie 3.3.1, elles sont décomposées en une étape de pré-traitement et de sélection des sondes/gènes (FS) et une étape de déconvolution. Pour les méthodes intégratives, une étape supplémentaire pour croiser les résultats des différents types de données est souvent nécessaire. Ces trois étapes vont être détaillées l'une à la suite de l'autre.

Présélection des sondes

Pour le RNA-seq, la présélection des gènes à analyser se fait dans toutes les méthodes obtenues par une analyse en composantes indépendantes (ICA) (voir tableau 5.1, lignes FS). Les gènes sélectionnés sont les plus importants (contribuant avec un FDR $> 0{,}2$) des composantes les plus stables (avec une stabilité moyenne $> 0{,}8$). La méthode RNA_wICA trie ensuite les gènes pour retirer les doublons, c'est-à-dire les gènes détectés pour plusieurs composantes. En revanche, les autres méthodes basées sur RNA_wNMF ne font pas ce tri et utilisent plusieurs fois les gènes importants dans plusieurs composantes, ce qui permet d'augmenter leur poids lors de la déconvolution.

Pour filtrer les sondes de méthylation, c'est la méthode du package Medepir avec les valeurs par défaut qui a été systématiquement choisie : les 5 000 sondes les plus variables sont conservées.

	RNA_wICA	RNA_wNMF	DNAm_Edec	DNAm_MeDeCom	DNAm_wICA	both_wICA	both_wNMFMeDeCom	both_meanwNMFMeDeCom
Surnom	*r_WIC*	*r_WNM*	*m_EDC*	*m_MDC*	*m_WIC*	*b_WIC*	*b_COM*	*b_MEA*
Data type	RNA	RNA	DNAm	DNAm	DNAm	both	both	both
FS DNAm	/	/	5000 most var	5000 most var	/	/	5000 most var	5000 most var
FS RNA	ICA, most important genes of most stable components, removing of duplicated genes	ICA, most important genes of most stable components, not removing of duplicated genes	/	/	/	/	ICA, most important genes of most stable components, not removing of duplicated genes	ICA, most important genes of most stable components, not removing of duplicated gene
Deconvolution DNAm	/	/	EDec	MeDeCom	ICA weighted on 30 most important genes	ICA weighted on 30 most important genes	MeDeCom with the A matrix computed on RNA as startA parameter	MeDeCom
Deconvolution RNA	ICA weighted on 30 most important genes	NMF with snmf/r method	/	/		ICA weighted on 30 most important genes	NMF with snmf/r method	NMF with snmf/r method
Time 10 A	~10mn	~20mn	~3h	~17h	~10mn	~10mn	~17h	~17h30
Time 1 A	~1mn	~2mn	~20mn	~1h40	~1mn	~1mn	~1h40	~1h45

TABLEAU 5.1 – **Description des méthodes obtenues à l'issue du deuxième data challenge.** Chaque colonne est une méthode. FS = pré-sélection des sondes, DNAm = sur données de méthylation de l'ADN, RNA = sur données d'expression des gènes. Les lignes "time" indiquent le temps d'exécution sur respectivement 10 (10 A) et une (1 A) matrices à déconvoluer.

Déconvolution

Pour la déconvolution des données de RNA-seq, deux méthodes ont été utilisées. L'ICA pondérée, abrégée wICA commence par exécuter une ICA rapide [96] sur D en utilisant la fonction *run_fastica* avec le package R deconica [119] en imposant le nombre de composantes correspondant au nombre de types cellulaires. Les 30 gènes les plus importants de chaque composante, correspondant donc supposément à des marqueurs de chaque type cellulaires, sont ensuite extraits avec la fonction *deconica : :generate_markers*. Ces gènes sont utilisés pour calculer le score d'abondance dans les différentes composantes avec la fonction *deconica : :get_scores*. Enfin, les proportions sont extraites à partir des scores pondérés sur les marqueurs avec la fonction *deconica : :stacked_proportions_plot*.

La NMF, quant à elle, est utilisée avec le package NMF [120] et la méthode "snmf/r" qui implémente un algorithme de sparse NMF estimant les matrices A et T alternativement ; en commençant par la matrice A. Cet algorithme de NMF favorisant la présence de 0 dans les matrices déconvoluées a déjà fait ses preuves

pour des données d'expression des gènes [121].

Pour la déconvolution de la méthylation, trois méthodes ont été employées : EDec et MeDeCom à travers le package Medepir, comme décrit précédemment, et l'ICA pondérée appliquée de la même façon que pour les données RNA-seq (voir ci-dessus).

Intégration

L'intégration des deux types de données (méthodes "both" dans le tableau) est réalisée de deux façons. Pour les méthodes b_WIC et b_MEA, on calcule simplement la moyenne des matrices A obtenues respectivement par les méthodes r_WIC et m_WIC, et par les méthodes r_WNM et d_MDC. La correspondance entre les types cellulaires déconvolués est choisie pour maximiser la corrélation entre les deux matrices A.

La méthode b_COM est la seule intégrant vraiment les deux types de données, la matrice A calculée avec la méthode r_WNM sert ensuite d'initialisation à MeDeCom qui calcule la matrice A finale sur les données d'ADNm.

Résultats

Les résultats obtenus par ces méthodes sur les 10 matrices A différentes simulées pour le data challenge sont visibles figure 5.3.

On peut remarquer que la méthode d'ICA pondérée (wICA) est plus efficace sur les données de RNA-seq (r_WIC) que sur les données de méthylation (m_WIC), et que si moyenner les deux matrices A obtenues n'améliore pas l'erreur, son amplitude et donc sa sensibilité à la variabilité des matrices en entrée diminue significativement.

Quatre méthodes se démarquent par leur faible erreur : m_MDC pour les données de méthylation de l'ADN, r_NMF pour le RNA-seq, et les intégratives mélangeant celles-ci : b_MEA et b_COM.

Cependant, ces quatre méthodes ont des résultats similaires, et ici l'intégration des différents types de données ne permet pas vraiment d'améliorer significativement l'efficacité de la déconvolution par rapport aux méthodes se basant sur un seul type. D'autres pistes devront donc être explorées pour permettre

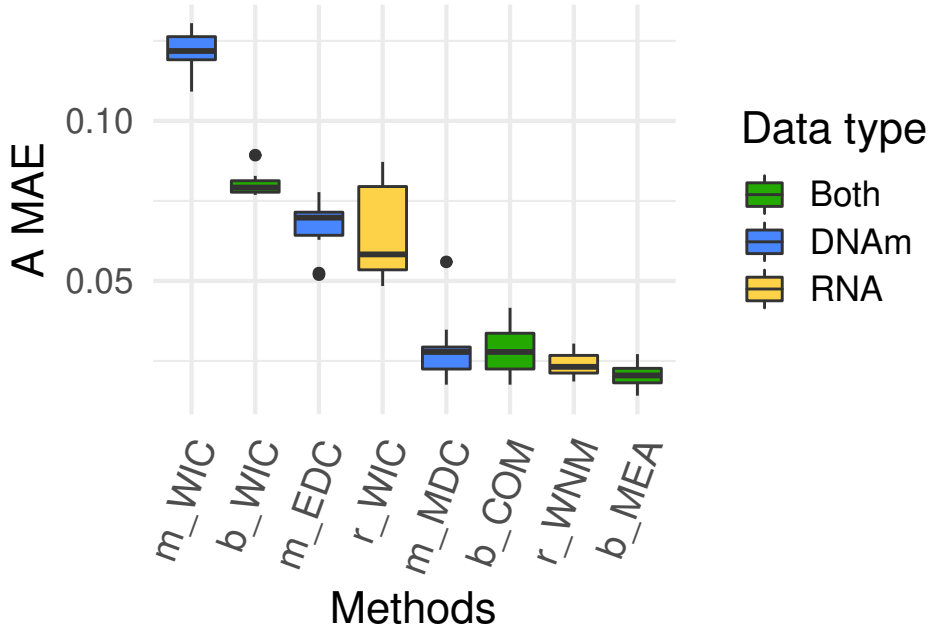


FIGURE 5.3 – **Résultats des méthodes du data challenge.** Les méthodes sont décrites en tableau 5.1, le score affiché ici correspond à l'erreur absolue moyenne de la déconvolution réalisées sur 10 matrices A simulées selon les mêmes paramètres avec chacune 10 réalisations de bruit.

de développer une méthode se démarquant par son efficacité (voir discussion).

5.2 Pérenniser le projet via une plateforme permanente

Malgré des résultats encourageants, la tenue de ce deuxième data challenge nous a montré les limites de ce format court ; bien qu'il y ait de nombreux avantages au travail collaboratif qu'il permet, le temps est trop limité pour à la fois aborder une problématique complexe et développer des méthodes vraiment innovantes. Il nous a donc semblé nécessaire d'aller plus loin et de proposer un moyen de pérenniser l'évaluation systématique des méthodes entre elles : c'est le projet Deconbench.

Le but de Deconbench est de proposer une plateforme permanente, utilisant dans un premier temps le même support que nous avions utilisé durant le data challenge (Codalab), pour permettre de d'évaluer des méthodes de déconvolution, mais aussi de soumettre des jeux de données, sur lesquels pourront tourner toutes les méthodes soumises. L'intérêt est donc double : pour les développeurs de méthodes de déconvolution, elle offre la possibilité d'évaluer facilement la leur et de la comparer aux autres, et pour les biologistes ou cliniciens la possibilité de faire tourner facilement plusieurs méthodes sur leurs jeux de données. Ce projet a été décrit dans un article en cours de publication (annexe 7.2.4), l'interface est visualisable ici : lien de la plateforme. L'accès à la plateforme Deconbench devrait ouvrir très prochainement (fin 2021, dès la publication de l'article).

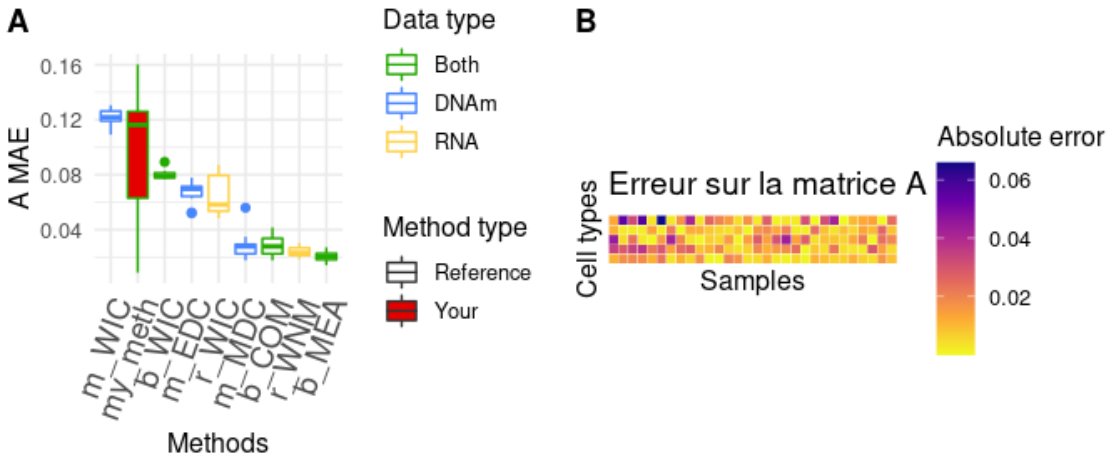


FIGURE 5.4 – **Exemple des figures obtenues sur Deconbench**. En A, l'erreur absolue moyenne obtenue par la méthode soumise (my_meth) est positionnée en fonction des 8 méthodes de référence. En B, on peut regarder le détail de chaque matrice déconvoluée pour évaluer si la méthode est robuste selon les simulations, et si l'erreur se concentre sur un type cellulaire ou un échantillon en particulier.

Par défaut, les huit méthodes décrites dans le tableau 5.1 servent de méthodes de référence, et les participants obtiennent leur positionnement en fonction de celles-ci. Différentes figures sont générées pour chaque soumission (figure 5.4). La principale figure correspond aux scores obtenus par les méthodes de réfé-

rences sur les matrices déconvoluées, c'est-à-dire à la figure 5.3, mais avec l'ajout de la méthode soumise (figure 5.4, panneau A). Les méthodes sont classées par erreur moyenne, et la position de la méthode soumise est donc facilement visualisable. D'autres figures, comme celle du panneau B, permettent de visualiser les erreurs pour chaque matrice A, et donc d'identifier la source d'erreur majoritaire de notre méthode (entre les matrices A, plutôt sur certains échantillons, plutôt sur certains types cellulaires, etc.).

Discussion et conclusion

Dans cette deuxième partie de ma thèse, nous avons vu que la déconvolution sans référence pouvait être un outil puissant pour inférer la composition des échantillons broyés dont le signal était mélangé. Différents axes ont été explorés, en particulier, nous avons montré que le pré-traitement des données et la prise en compte des facteurs de confusion (comme les données cliniques) étaient essentiels pour améliorer les analyses. Le pipeline d'analyse Medepir constitue une première étape vers la reproductibilité des résultats, mais des efforts sont encore nécessaires pour permettre une comparaison robuste et égalitaire entre les différentes méthodes de déconvolution développées.

Limites

Choix des méthodes utilisées

De très nombreuses méthodes de déconvolution existent, utilisant différents types de données comme le méthylome ou le transcriptome, avec ou sans utilisation de données de référence. Des benchmarks existent déjà pour certains types de méthodes, notamment celles basées sur les références, par exemple celles spécifiques aux données de RNA-seq [122] ou celles spécialisées dans la détection des cellules immunitaires [123]. Cependant, ici nous n'avons pas souhaité faire une comparaison exhaustive des méthodes de déconvolution existantes. Ce choix a plusieurs explications, premièrement les questions que nous nous posions étaient plus larges que le choix de la meilleure méthode, comme l'intérêt du pré-traitement des données, ou la possibilité de faire de l'intégration de données multi-omiques. Par ailleurs, le choix de travailler par exploration

collective lors de data challenges nous a fait réaliser que chaque méthode pouvait être optimisée pour un jeu de données particulier, et qu'il n'y avait pas encore de manière robuste de comparer objectivement l'ensemble des méthodes existantes. Ces constats nous ont poussé vers la création de la plateforme Deconbench, qui devrait à terme permettre de tester plusieurs méthodes de déconvolution sur de nombreux jeux de données, à la fois simulé et obtenus expérimentalement.

Enfin, les méthodes de déconvolution existantes en elles-mêmes montrent des limites assez importantes. Ainsi, une étude parue récemment compare les résultats de pureté tumorale du cancer de la prostate obtenus par 10 méthodes informatique sur des données moléculaires (ADNm, RNA-seq, génome et micro-ARN) aux résultats de quantification obtenus par plusieurs cliniciens [124]. On observe une forte différence entre les résultats obtenus cliniquement, considérés comme justes, et les résultats obtenus *in silico*. Il est donc nécessaire d'améliorer encore les méthodes de déconvolution. Affiner l'initialisation des matrices A et T pourrait par exemple être une piste à exploiter puisque on a montré l'importance de cette étape.

Intégration multi-omiques

Les méthodes intégratives développées lors du 2ème data challenge ont donné des résultats plutôt décevants puisqu'elles ne permettent pas de surpasser les méthodes utilisant un seul type de données. En effet, on pourrait à priori penser que combiner deux types d'information (transcriptome et méthylome) améliorerait forcément la quantité de signal, faciliterait le lissage du bruit dû à chacune des méthodes de séquençage, et donc augmenterait naturellement l'efficacité de la déconvolution. En réalité, nous avons vu que l'intégration n'est pas si triviale, en effet le lien entre méthylome et transcriptome n'est pas direct et une "vraie" intégration demande des recherches supplémentaires. Un travail d'interprétation biologique sur les données devrait permettre de relier de manière plus efficace les deux informations, et amener à terme une intégration réussie. Globalement, l'intégration multi-omiques est un sujet d'étude en pleine expansion qui promet une nette amélioration des analyses dans le futur [125].

Choix du score et de l'évaluation

Dans la partie 3.2, j'expliquais que nous évaluions les méthodes des différents data challenges sur la matrice de proportion A. Ce choix était motivé par la variabilité structurale de la matrice T. En effet, en fonction du nombre de sondes ou de gènes sélectionnés en pré-traitement, sa taille peut fortement varier et il était donc difficile de comparer différentes méthodes sur la matrice T. Ainsi, la validité de la matrice T déconvoluée n'a pas vraiment été évaluée et étudiée dans cette partie, bien qu'un des objectifs de l'analyse d'hétérogénéité soit aussi d'inférer les types cellulaires présents dans la tumeur. Il n'est donc pas évident que les valeurs de la matrice T inférée soient des valeurs d'expressions de gène ou de méthylation crédibles biologiquement, d'autres analyses sur ce sujet restent nécessaires. Cet enjeu sera abordé plus longuement dans la partie suivante, notamment dans la section 7.1.2.

En plus du choix d'évaluer uniquement A, nous avons également, au cours des différentes analyses, choisi comme métrique l'erreur absolue moyenne MAE. Cependant, ce score a ses limites car il n'est pas forcément représentatif de tous les types d'erreur obtenus par les méthodes. Par exemple, si parmi deux méthodes l'une d'elles déconvolue toujours moyennement, et l'autre déconvolue très mal certaines tumeurs et très bien d'autres, elles auront un score moyen identique. Il serait donc intéressant de considérer aussi des scores qui évaluent l'amplitude moyenne et la localisation des erreurs. Dans ce sens, dans le cadre du "cometh course" (voir ci-dessous), nous avons ainsi choisi un score combinant MAE, corrélation sur les lignes (types cellulaires) et corrélation sur les colonnes (échantillons).

Travail sur des simulations

Même si l'efficacité du pipeline Medepir a été validé sur le cancer du poumon TCGA en comparant les résultats obtenus avec ceux de précédentes analyses, la majorité du travail a été réalisé sur des données simulées. Un des axes d'amélioration rejoint la partie 5.2 : Deconbench devrait permettre de tester les méthodes sur de nombreux jeux de données simulés ou non. Pour l'instant, des jeux de données expérimentaux, fiables contenant le signal d'un mélange contrôlé d'échan-

tillons dont la composition (profils individuels et proportions dans le mélange) est connue sont difficiles à obtenir. Des essais sont en cours dans notre équipe pour obtenir ce type d'échantillons expérimentalement en collaboration avec Jérome Cros. D'autres pistes sont la quantification par immuno-histochimie des types cellulaires présents dans la tumeur avant broyage, ainsi que la construction de faux échantillons mélangés à partir de données séquencées sur des cellules uniques. Nous pourrons ensuite tester les différentes méthodes sur ces jeux de données, mais aussi améliorer les simulations pour se rapprocher des données moléculaires obtenues expérimentalement.

Enfin, nous pourrions également développer des simulations plus complexes, par exemple en cherchant à simuler un bruit plus réaliste sur D. Cependant, les simulations actuelles permettent déjà d'analyser beaucoup de paramètres, et leur complexification multiplie vite les combinaisons de paramètres à tester, et rend aussi l'analyse des résultats plus difficile.

Perspectives

D'autres données omiques

Dans cette partie, nous avons vu que la déconvolution pouvait être un outil puissant, qu'il est intéressant de continuer à développer et améliorer. L'équipe a plusieurs projets en cours à ce sujet. Gedepir est un package en cours de développement pour implémenter un pipeline d'analyse semblable à Medepir mais pour la déconvolution non-supervisée de données transcriptomiques. Par ailleurs, l'équipe travaille également sur le single cell : l'objectif de ce projet est de réussir à identifier des listes de biomarqueurs robustes et spécifiques à chaque lignée, pour pouvoir ensuite les utiliser comme référence lors de déconvolution semi-supervisée d'échantillons mélangés. Cependant, l'identification précise des types cellulaires, même en cellule unique, reste un défi quand on vise un degré fin de caractérisation (par exemple, séparer différents sous-types de fibroblastes).

Plateformes

Un enjeu actuel du benchmarking, dans un domaine émergeant comme celui de la déconvolution, est de rendre les outils accessibles à la communauté de chercheurs, mais également de permettre l'analyse dynamique des nouvelles méthodes publiées. Nous avons différents projets en cours pour suivre ces objectifs.

Deconbench d'une part, devrait permettre à terme d'avoir une plateforme de benchmark dynamique et ouverte, intégrant au fur et à mesure les nouvelles méthodes et les nouveaux jeux de données disponibles.

En parallèle, une application web interactive réalisée grâce au package R shiny appelée Decomics est en cours de développement, pour permettre une application intuitive et sans programmation du package Gedepir, puis à terme de Medepir. Le but est d'ouvrir ces outils à une communauté plus large que les bio-informaticiens et ne maîtrisant pas forcément l'utilisation de R.

Enfin, dans cette même lignée, un premier "Cometh course" a été réalisé en Février 2021. Ce workshop en ligne autour de la quantification de l'hétérogénéité tumorale était particulièrement adressé aux cliniciens, pour leur apprendre à analyser rapidement et facilement leurs données, à travers une application web facile d'utilisation (une sorte de prototype pour Decomics). Le workshop a finalement été adapté pour convenir également aux profils plus bio-informatiques qui s'étaient inscrits, avec une session parallèle sous forme de mini data challenge sur la déconvolution non supervisée.

Application à d'autres jeux de données

Actuellement, nos efforts se sont concentrés sur des apports méthodologiques à la déconvolution, sans explorer des données et des champs d'application en détail. En effet, l'intérêt de la caractérisation de l'hétérogénéité cellulaire ou de la pureté tumorale est déjà bien documenté, et nous ne nous sommes pas attardés sur ce point. Cependant, il serait intéressant d'appliquer nos pipelines (medepir et gedepir) à d'autres données que le cancer du poumon, que ce soit des échantillons dont la pureté est déjà bien caractérisée (TCGA par exemple), ou des nouveaux échantillons, pour voir si les résultats sont consistants avec les précédentes méthodes et si on obtient des résultats prédictifs, pour la survie

notamment. Dans l'immédiat, nos efforts vont se concentrer sur une approche innovante détaillée dans la partie suivante.

Conclusion

La déconvolution nous permet donc d'obtenir des informations fiables sur la composition du micro-environnement tumoral. Or, on sait que cette composition joue un rôle important sur des paramètres cliniques comme la prolifération ou la résistance aux traitements. Ainsi, sachant caractériser cette hétérogénéité, un enjeu important est de comprendre comment elle est régulée et d'en caractériser les facteurs. Dans la suite de ma thèse, je me suis donc intéressée à relier l'expression des gènes dans les tumeurs, et en particulier la dérégulation détectée par Penda, avec la composition du micro-environnement.

Troisième partie

Lien entre composition du micro-environnement tumoral et dérégulation des gènes

Introduction

Contexte, problématique et objectifs

Dans les parties précédentes, nous avons développé un outil innovant pour détecter les gènes dérégulés dans une tumeur de manière individuelle et personnalisée, puis nous avons exploré les outils permettant de retrouver la composition en types cellulaires d'un échantillon mélangé. Dans cette partie, nous allons essayer de combiner ces deux informations, afin de détecter les gènes dont la dérégulation est en lien avec la composition du micro-environnement tumoral.

Le micro-environnement tumoral regroupe l'ensemble des types cellulaires autour et dans la tumeur, il est composé entre autres de cellules immunitaires, de tissus conjonctifs (comme les fibroblastes) et de tissus sains (comme des cellules épithéliales). Il joue un rôle très important dans plusieurs mécanismes tumoraux comme la métastase [126], ou la réponse aux traitements [127]. Auparavant estimée visuellement par le clinicien, la pureté tumorale est désormais quantifiée et caractérisée *in silico* grâce aux données issues de séquençage haut débit par diverses méthodes, comme celles décrites dans la partie précédente (déconvolution avec ou sans référence, estimations par biomarqueurs pour les cellules immunitaires, etc.).

Des études récentes ont montré que cette hétérogénéité en types cellulaires pouvait représenter un biais important pour l'analyse différentielle et la classification des tumeurs en fonction de la dérégulation des gènes, et qu'il était donc crucial de le corriger [128]. Malgré tout, les mécanismes régulant la composition

du micro-environnement ou les biomarqueurs associés à cette composition ne sont encore que peu caractérisés.

Certains travaux ont toutefois déjà abordé cette question. Ainsi, plusieurs études ont déjà identifié des gènes dont la dérégulation est reliée à la pureté tumorale et à la survie en corrélant l'expression des gènes et les proportions du micro-environnement données par la méthode ESTIMATE basée sur références [129, 130, 131].

D'autres méthodes ont également été développées pour répondre à cette question en allant plus loin que la corrélation expression-pureté tumorale. Par exemple, un article actuellement pré-publié sur bioRxiv propose un outil appelé BayesPrism [132], utilisant un modèle bayésien pour inférer les matrices A et T à partir d'échantillons mélangés et de profils en single-cell. Leur modèle permet d'obtenir une matrice T différente pour chaque échantillon, et également d'effectuer une analyse différentielle pour relier la composition du micro-environnement et l'expression dans chaque tumeur. Cependant, leur étude intègre 37 000 échantillons en single-cell comme référence, et de telles données ne sont pas disponibles pour tous les types de cancers. D'autres études s'intéressent également au lien entre la dérégulation d'un gène et la composition du micro-environnement mais toujours à l'aide de données single-cell, comme le lien entre les sous-types de gliomes et la proportion de cellules immunitaires [133] et même dans le cas du COVID-19, entre un récepteur du virus et le micro-environnement cellulaire de celui-ci [134].

Notre objectif est de développer une méthode permettant de partir des échantillons mélangés ("bulk") existants, et de retrouver les gènes dérégulés en lien avec la composition cellulaire. L'utilisation de Penda pour l'analyse différentielle personnalisée nous permettra d'utiliser une information supplémentaire, la dérégulation binaire pour chaque échantillon, par rapport aux méthodes basées sur l'expression brute.

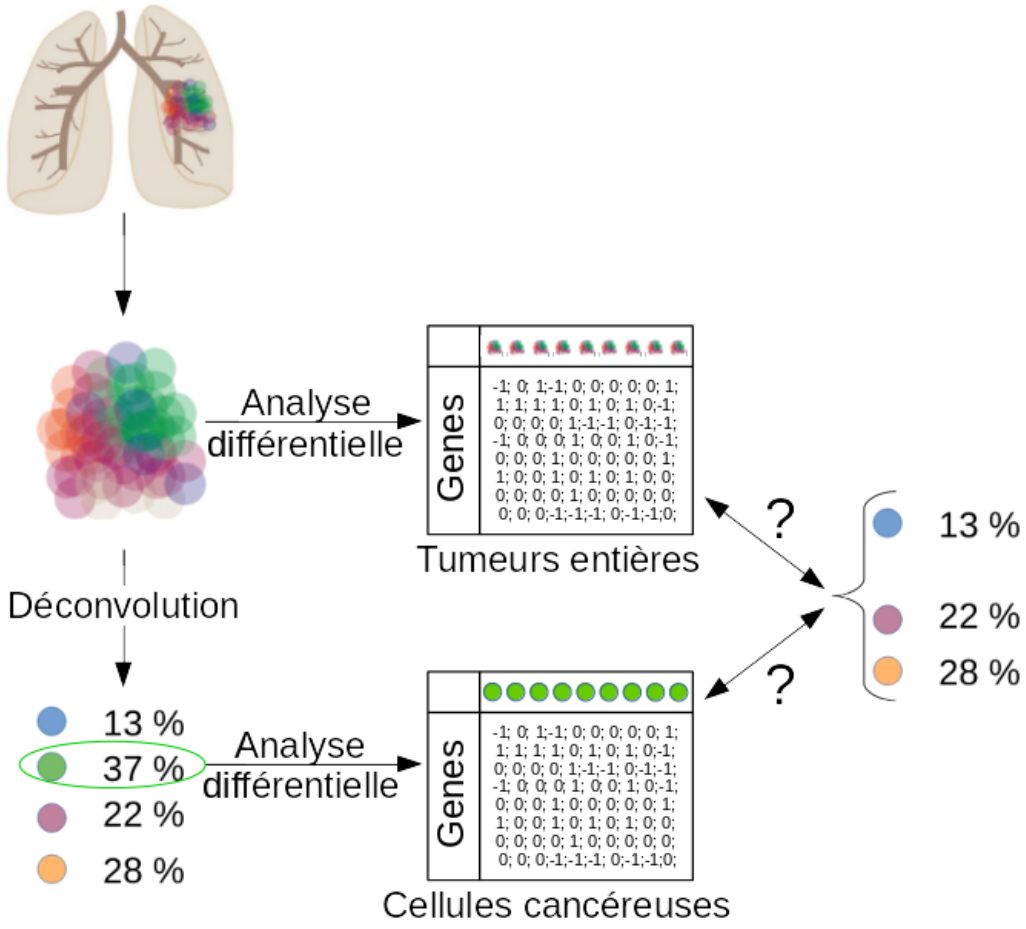


FIGURE 5.5 – **Représentation schématique des analyses différentielles effectuées dans cette partie.** L'analyse différentielle est effectuée avec la méthode Penda, dans un premier temps sur le signal de la tumeur entière (chapitre 6), puis sur le signal du type cellulaire tumoral isolé (chapitre 7). Dans chacun des cas, la matrice d'expression différentielle obtenue est reliée aux proportions des types cellulaires du micro-environnement grâce à différentes métriques.

Dans cette partie, nous allons explorer la question du lien entre composition du micro-environnement et dérégulation des gènes à travers deux aspects (voir figure 5.5). L'analyse différentielle va d'abord être effectuée sur la tumeur entière, et les gènes dérégulés vont être reliés aux proportions des différents types cellulaires déconvolués (partie haute de la figure). Ici, l'objectif est de voir le lien entre les gènes dérégulés à l'échelle de la tumeur et la composition du micro-environnement. Dans un second temps, l'analyse différentielle sera appliquée uniquement aux cellules correspondant au type "cancer purifié" obtenu à l'issue de la déconvolution, et les dérégulations de ces cellules pourront à leur tour être reliées aux proportions du micro-environnement (partie basse de la figure). Cette fois, on observe directement le lien entre la dérégulation des gènes dans les cellules cancéreuses et la composition du micro-environnement. À chaque fois, le lien entre dérégulation et composition tumorale sera quantifié grâce à différentes métriques détaillées ci-dessous.

Métriques utilisées

Plusieurs métriques ont été définies pour établir le lien entre proportions de dérégulation et dérégulation des gènes. En appliquant sur les données d'expression une première étape d'analyse différentielle individuelle (effectuée avec la méthode Penda), on peut obtenir pour chaque gène un statut binaire : dérégulé ou non-dérégulé. La plupart des métriques se basent sur ce principe, mais la corrélation directe entre l'expression des gènes et la composition du micro-environnement sera également utilisée comme référence. La corrélation permet de détecter des liens très forts et linéaires entre la proportion d'un type cellulaire et l'expression d'un gène. En revanche, elle ne permet pas de détecter tous les types de lien, comme les effets de seuils qui ne s'activent qu'à partir d'une certaine proportion cellulaire.

Les autres métriques se basant sur la catégorisation binaire de dérégulation (oui/non) se décomposent en deux groupes : les mesures sur un type cellulaire en particulier, et les mesures sur la composition globale.

1. Mesures sur l'abondance de chaque type cellulaire

Les résultats de l'analyse différentielle permettent, pour chaque gène, de séparer les échantillons en deux populations, et de calculer des métriques pour comparer leurs distributions en proportions cellulaires. Différentes métriques sont testées pour évaluer la distance entre les deux distributions, elles sont illustrées dans la figure 5.6 et détaillées dans les parties suivantes.

a. Distance de Kantorovitch.

La première des métriques est couramment utilisée en théorie du transport : la distance de Kantorovitch (aussi appelée distance de Wasserstein) représente la distance minimale pour passer d'une distribution à l'autre [135]. Elle est calculée avec la fonction *kantorovich* du package R ptlmapper [136]. Concrètement, on calcule la somme cumulée de la différence entre les deux distributions f_1 et f_2, la distance de Kantorovitch correspond à l'aire sous la valeur absolue de cette courbe.

$$Dist.Kantorovitch(f_1, f_2) = \int_{-\infty}^{+\infty} |\int_{-\infty}^{x} f_1(t) - f_2(t)\ dt|dx$$

b. Test de Student.

La deuxième métrique correspond au résultat d'un test de Student d'égalité des moyennes sur les deux échantillons. Il est réalisé par la fonction R de base *t.test*, la p-valeur est ensuite transformée sur une échelle logarithmique afin de faciliter sa lecture : $Val.Student = -log10(p - valeur)$. Au final, une grande valeur indique des moyennes très significativement différentes.

c. Test de Kolmogorov-Smirnov.

On utilise également le test non-paramétrique de Kolmogorov-Smirnov sur deux échantillons, qui permet de déterminer si les deux distributions suivent la même loi avec la fonction *ks.test* par défaut dans R [137]. On récupère la statistique du test qui correspond à la différence maximale entre les distributions

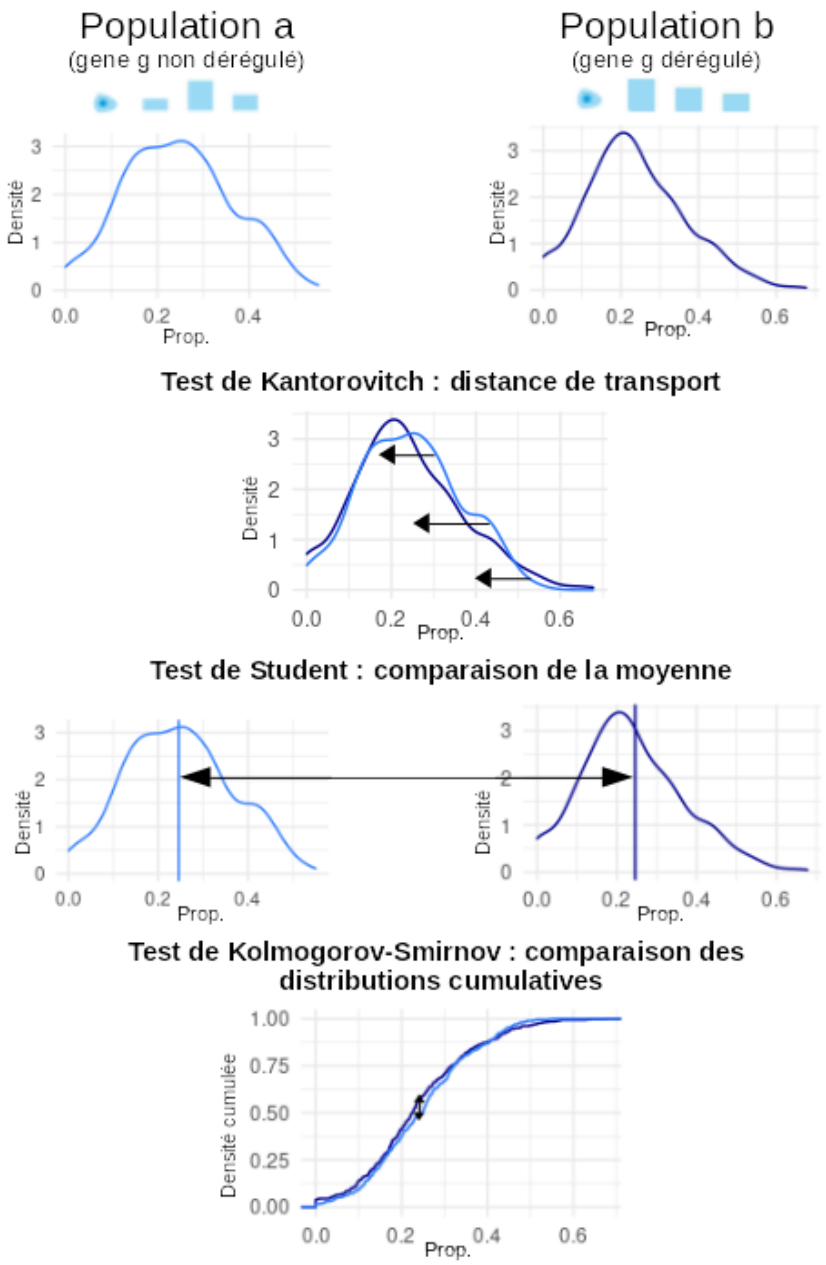


FIGURE 5.6 – **Représentation schématique des mesures à l'échelle d'un type cellulaire.** Les populations a (bleu clair) et b (bleu foncé) de tumeurs sont séparées en fonction de la dérégulation d'un gène g fixé. On compare ensuite les distributions des proportions d'un type cellulaire donné à l'aide de trois métriques : distance de Kantorovitch, test de Student et test de Kolmogorov-Smirnov.

cumulatives :

$$Dist.Ks(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$$

une valeur élevée indique donc des distributions plus éloignées.

2. Mesures sur la composition globale du micro-environnement

Les mesures de la section précédente permettent de relier la composition de chaque type cellulaire à l'expression des gènes, cependant on perd l'information de la structure tumorale globale. Ainsi, chaque type cellulaire est regardé indépendamment, sans prendre en compte toutes les autres proportions du reste du micro-environnement tumoral. Pour mesurer l'influence de l'ensemble des proportions, nous nous sommes inspirés de la méthode ptlmapper [136].

Cette méthode a été développée dans le cadre de la détection de variants génétiques associés à une probabilité de distribution phénotypique chez la levure. Pour chaque génotype (ou individu) considéré, le trait phénotypique d'intérêt est mesuré dans chaque cellule d'une population clonale, permettant ainsi d'obtenir la distribution du trait pour un génotype donné. La méthode ptlmapper permet ensuite de déterminer l'effet du génotype sur la probabilité d'expression du trait. ptlmapper procède en trois étapes (voir figure 5.7). Premièrement, la distance de Kantorovitch (entre deux distributions du trait phénotypique) est calculée pour toutes les paires d'individus, et permet de placer les individus dans un espace multidimensionnel où les points proches ont des phénotypes comparables (panneau A). Les individus sont ensuite annotés en fonction de leur génotype pour un marqueur génétique donné (panneau B). Enfin, une analyse discriminante permet de mesurer si le marqueur génétique sépare significativement les individus dans l'espace (panneau C).

Dans notre cas, on veut adapter cette méthode pour détecter les dérégulations de gènes associés à la distribution en types cellulaires entre les échantillons. Ainsi, les différents échantillons seront donc placés dans un espace multidimensionnel en fonction de leurs proportions en types cellulaires, et la dérégulation des gènes sera utilisée comme facteur discriminant à la place des marqueurs

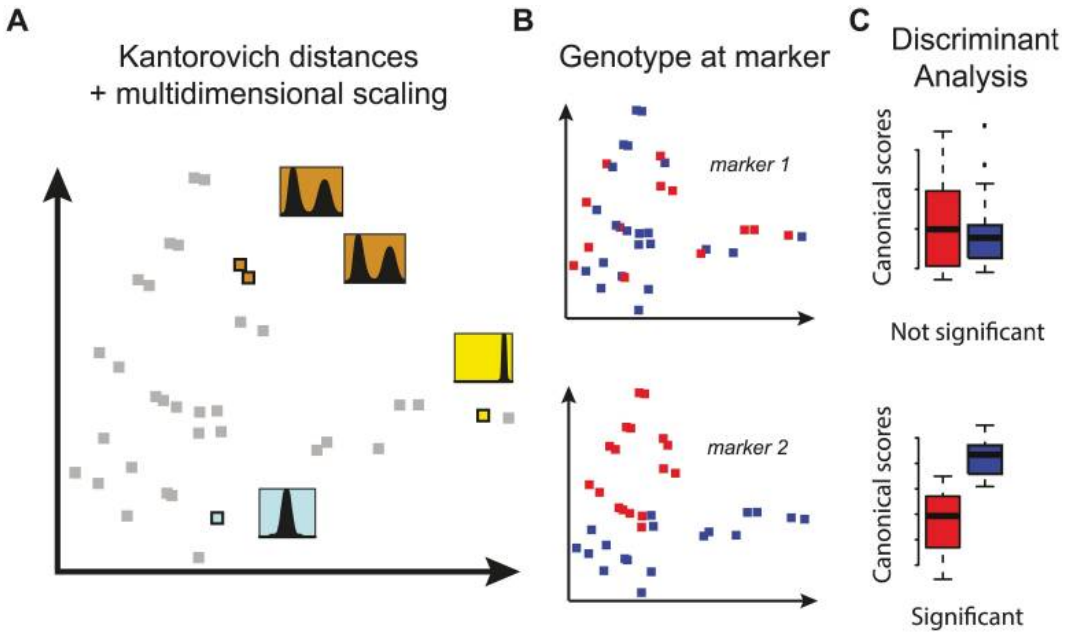


FIGURE 5.7 – **Principe de la méthode ptlmapper.** Pour une population donnée, on étudie la proportion du trait X (phénotype) dans chaque individu. A) Les distances de Kantorovitch sont calculées pour toutes les paires d'individus, ce qui permet de placer les individus dans un espace multidimensionnel. La distance entre les individus (carrés) dans l'espace reflète des phénotypes comparables (ex : carrés oranges). B) Les individus sont annotés (bleu ou rouge) en fonction de leur génotype pour un marqueur génétique donné (marker 1 en haut, marker 2 en bas). C) Une analyse discriminante canonique est réalisée pour tester si le génotype du marqueur discrimine les individus dans l'espace phénotypique. Dans cet exemple, le lien génétique est significatif pour le marqueur 2 mais pas pour le marqueur 1. Figure issue de l'article "Exploiting Single-Cell Quantitative Data to Map Genetic Variants Having Probabilistic Effects" de F. Chuffart et al., [136].

génétiques.

a. Placement des points dans l'espace

Pour la première étape de la méthode ptlmapper, nous testons deux façons de placer les échantillons dans l'espace en fonction de leurs proportions en type cellulaire.

Comme nous n'avons en général que 4 ou 5 types cellulaires, nous allons dans un premier temps utiliser directement la matrice de proportion telle quelle en entrée de la fonction. Les échantillons seront donc placés dans un espace à 4 ou 5 dimensions. Chaque dimension correspond alors à un type cellulaire. Le désavantage est que cet espace est difficile à visualiser, mais l'avantage est que chaque axe a un sens biologique facile à interpréter.

Dans un second temps, on va également tester la réduction des dimensions de la matrice de proportion. Pour cela, on utilise la méthode implémentée dans ptlmapper. On commence par calculer la distance (ici euclidienne) basée sur les proportions en type cellulaire entre chaque paire d'échantillons : on obtient ainsi une matrice de distance entre les tumeurs. Grâce à cette matrice, on peut ensuite appliquer un algorithme de positionnement multidimensionnel, qui permet de réduire les dimensions en gardant proches les points similaires. En ne conservant que les deux première dimensions principales, qui contribuent le plus aux différences inter-individus, on réduit alors le nombre de coordonnées pour chaque tumeur.

b. Analyse discriminante

Pour chaque gène, les échantillons sont ensuite annotés en fonction de leur statut de dérégulation obtenu avec la méthode Penda. Une analyse discriminante est alors effectuée avec la fonction *get_wilks_score* du package ptlmapper [136]. La métrique obtenue avec cette méthode est le score de Wilks, qui est calculé grâce à une analyse discriminante généralisée basée sur un modèle linéaire. Cela revient à tester si les proportions des différents types cellulaires sont expliquées par le statut de dérégulation du gène. Finalement, on obtient pour chaque gène le score de Wilks correspondant à la proportion de la variance totale

de l'analyse discriminante n'étant pas expliquée par les différences entre les groupes de dérégulation. Concrètement, plus elle est petite, plus la distribution des proportions est expliquée par la dérégulation du gène.

Dans cette partie de ma thèse, je vais décrire comment nous avons utilisé ces métriques. Dans un premier chapitre, nous avons effectué l'analyse différentielle à l'échelle de la tumeur mélangée (figure 5.5 partie haute. Puis, dans le second chapitre, en effectuant l'analyse différentielle uniquement sur l'expression type cancer isolé (partie basse de la figure 5.5). Ce travail impliquant de nombreuses étapes (quantification de l'hétérogénéité tumorale, identification de la composante cancéreuse, puis "purification" du signal et analyse différentielle), elles seront implémentées à travers un package R : Ritmic pour 'RegulatIon of Tumor MICro-environment'. Cette partie III de ma thèse est beaucoup plus exploratoire que les parties précédentes, et le travail n'est pas terminé au moment de l'écriture de ce manuscrit.

Chapitre 6

Analyse de la régulation génétique du micro-environnement à partir des échantillons mélangés

Dans un premier temps, nous avons donc cherché à quantifier le lien entre la dérégulation des gènes à l'échelle tumorale globale et les proportions des différents types cellulaires du micro-environnement.

6.1 Données

6.1.1 Données biologiques

Comme dans les parties précédentes, nous avons choisi le cancer du poumon pour nos premières explorations. L'avantage principal est que nous possédons plusieurs cohortes : celles du TCGA et celles du CHU de Grenoble du Pr. Brambilla qui nous permettra de valider nos résultats. Ces jeux de données permettent aussi de croiser les valeurs d'expression des gènes (RNA-seq) avec celles de méthylation de l'ADN (ADNm) pour un même patient. Les cohortes du TCGA comportent en plus un nombre d'échantillons contrôles satisfaisant pour l'utilisation de Penda. Ainsi, dans un premier temps, nous nous sommes concentrés sur l'analyse des données de l'adénocarcinome (LUAD) TCGA, composées de 59

échantillons contrôles et de 451 échantillons tumoraux.

6.1.2 Analyse différentielle

L'analyse différentielle est effectuée avec la méthode Penda, qui est appliquée en suivant les vignettes du package, comme décrit dans la partie 2.2. Les paramètres de la méthode du centile sélectionnés sur des simulations réalistes sont donc un quantile de 0,1 et un facteur de 1,2, le seuil du test Penda sélectionné est quant à lui de 0,7. 5 286 gènes sont retirés car leur expression est très faible. À la fin de cette étape, on obtient donc pour chacun des 20 074 gènes des 451 tumeurs, un statut de dérégulation : sur-exprimé, sous-exprimé ou non dérégulé. Globalement, on trouve en moyenne 17% de dérégulation dans les données.

6.1.3 Déconvolution

La déconvolution est réalisée en suivant le pipeline Medepir, comme décrit dans la partie 4.6. Les données séquencées en 450k sont donc normalisées en fonction du type de sonde, puis on filtre les 98 580 sondes reliées aux facteurs de confusion. On détermine ensuite qu'il y a 5 types cellulaires par ACP, et la déconvolution est effectuées par EDec après sélection des 29 053 sondes les plus variables. Dans cette partie, on s'intéresse uniquement à la matrice de proportions A, on obtient donc pour chaque tumeur la proportion des 5 types cellulaires la composant. Les proportions varient entre 0 et 0,72, la somme est de 1 dans chaque tumeur.

6.1.4 Assignation des types déconvolués

Les types cellulaires déconvolués sont assignés comme dans la figure 4.14 de la partie 4.6. Les matrices T d'expression de chaque type cellulaire sont corrélées aux références de Epidish, puis séparées par regroupement hiérarchique. On retrouve deux types cellulaires associés à la référence immunitaire ($IC1$ et $IC2$), un à l'épithélial (Epi) et deux aux fibroblastes ($Fib1$ et $Fib2$). Ici, le type "cancer" est probablement le type épithélial.

6.1.5 Pré-traitement des résultats de Penda

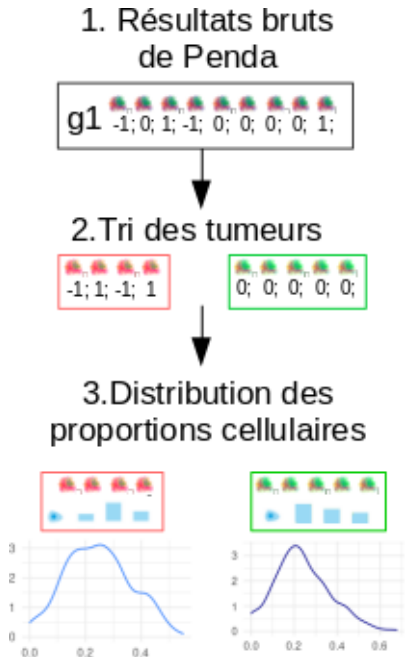


FIGURE 6.1 – **Représentation schématique du pré-traitement des résultats de Penda.** Étape 1 : à l'issue de l'analyse différentielle par Penda, on obtient pour le gène $g1$ un vecteur contenant une valeur par échantillon : -1 si le gène est sous-exprimé, 0 si le gène n'est pas dérégulé et 1 si le gène est sur-exprimé. Étape 2 : les échantillons sont regroupés en fonction de la dérégulation ou non de $g1$. Étape 3 : ces groupes serviront à comparer les distributions des différents types cellulaires, ils doivent donc être d'une taille suffisante pour que les distributions soient bien définies et éviter des effets statistiques de faible échantillonnage.

Les étapes du pré-traitement des résultats de Penda sont résumées dans la figure 6.1. Pour cette analyse, nous choisissons de classifier les gènes en deux groupes : dérégulés ou non, sans prendre en compte le sens de la dérégulation. En effet, comme nous l'avons vu dans la partie I, pour un même gène le sens est souvent le même, et pour avoir des échantillons suffisamment grands, il est plus simple de travailler avec deux classes. Pour chaque gène, les patients sont regroupés en fonction de leur dérégulation, et ce sont les distributions des proportions des différents types cellulaires qui seront comparées. Afin d'obtenir des distributions statistiquement bien définies, nous fixons un seuil arbitraire

sur la taille des groupes : seuls les gènes avec plus de 100 tumeurs dans chaque groupe de dérégulation sont conservés. Nous obtenons donc 5 119 gènes qui sont dérégulé dans au moins 100 patients, et non-dérégulés dans au moins 100 autres.

6.2 Analyse de l'abondance de chaque type cellulaire

La distance entre les deux distributions de proportions pour chaque gène et chaque type cellulaire est calculée par les différentes métriques détaillées dans l'introduction, partie III. Pour l'ensemble des couples gènes-type cellulaire étudiés, on obtient une distance de Kantorovitch comprise entre 0 et 0,2 avec une médiane à 0,042, une -log10(p-valeur) de Student entre 0 et 41 avec une médiane à 3 et une valeur de Kolmogorov-Smirnov variant entre 0 et 0,55 avec une médiane à 0,15. Les distributions des métriques sont visibles figure 6.2.

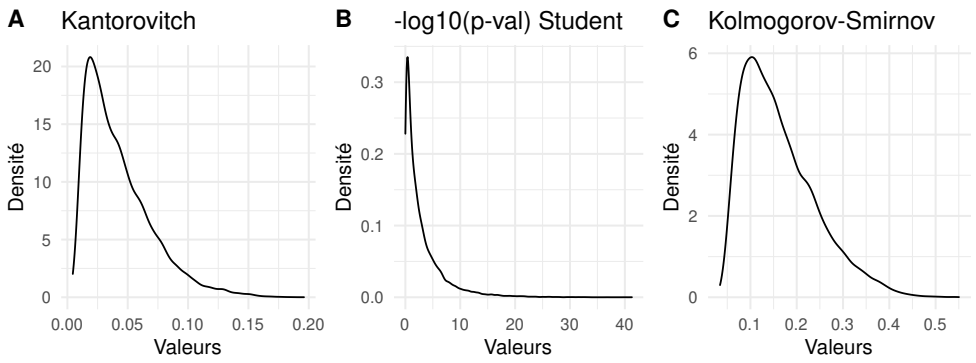


FIGURE 6.2 – **Distribution des différentes métriques obtenues.** En A, la distribution de la distance de Kantorovitch, en B le -log10 de la p-valeur de Student et en C la valeur de Kolmogorov-Smirnov.

Si on regarde plus en détail les valeurs de la distance de Kantorovitch pour chaque type déconvolué, on obtient la figure 6.3, avec chaque point correspondant à un gène et un type cellulaire.

La somme d'informations à traiter étant importante, dans un premier temps, on adopte une stratégie "top-gènes". Pour une métrique donnée, on sélectionne les 5 gènes qui maximisent la distance entre les deux groupes. Les 5 top-gènes

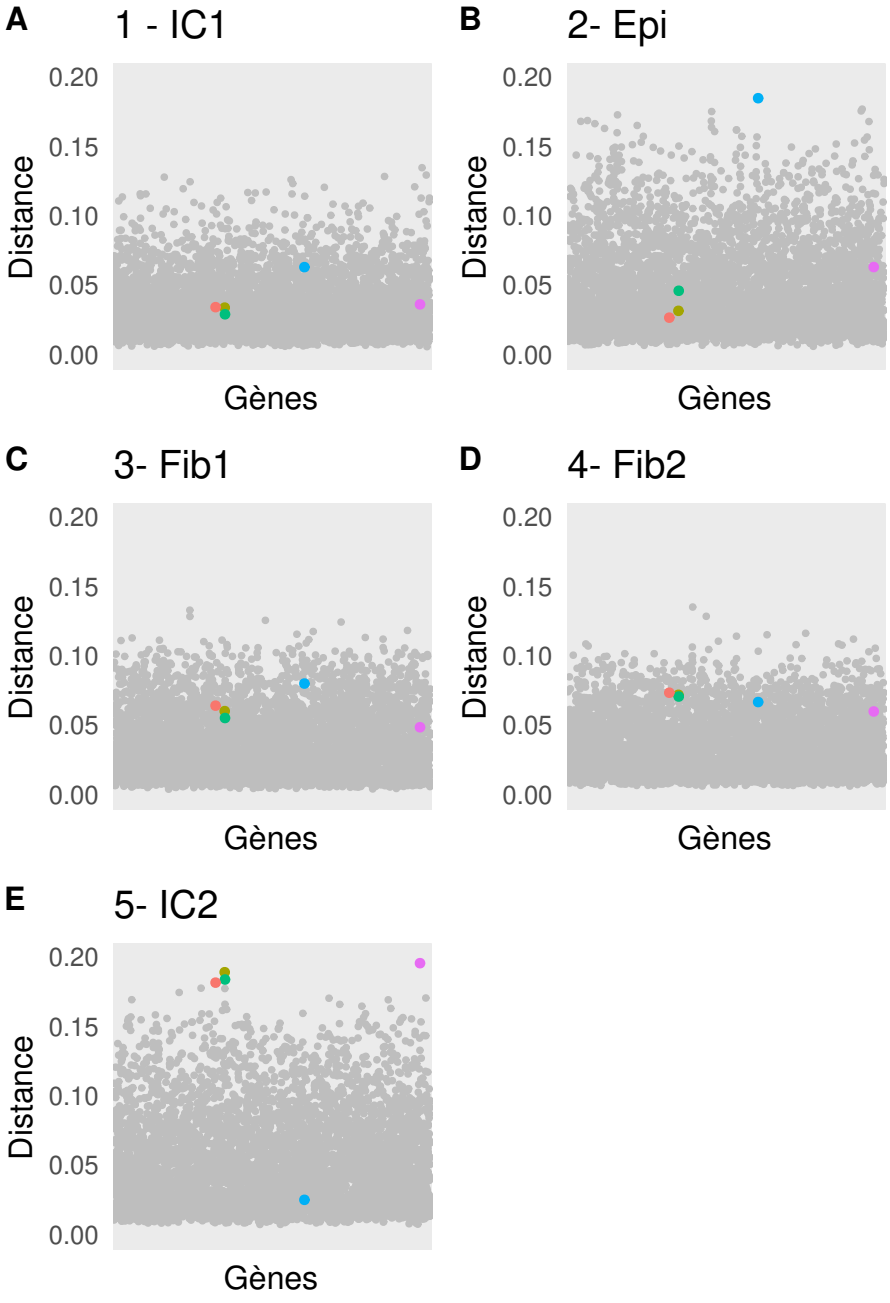


FIGURE 6.3 – **Représentation graphique de la distance de Kantorovitch entre le groupe avec et sans dérégulation, pour tous les gènes et pour les différents types cellulaires.** Chaque point est un gène et chaque panneau un type cellulaire (IC1/2 = Cellules immunitaires, Epi = Epithélial, Fib1/2 = Fibroblastes). Les points plus gros et colorés sont les 5 gènes ayant la plus grande distance, tout types cellulaires confondus. On en retrouve un pour Epi : Nek2 (bleu) et quatre pour IC2 : VSIR (violet), GIMAP1 (jaune), GIMAP7 (vert) et FLI1 (rouge).

de chaque métrique sont résumés dans le tableau 6.1.

	Dist. kanto.	Pval. Student	Dist. KS	Wilks 5 dim.	Wilks réduit
Top1	VSIR - IC2	MAGEA3 - Epi	GIMAP1 - IC2	NEK2	CEP55
Top2	GIMAP1 - IC2	MAGEA6 - Epi	VSIR - IC2	TPX2	NEK2
Top3	NEK2 - Epi	VSIR - IC2	MAGEA3 - Epi	KIF2C	TPX2
Top4	GIMAP7 - IC2	TMEM125 - Epi	MAGEA6 - Epi	MAGEA3	BIRC5
Top5	FLI1 - IC2	CIP2A - Epi	NEK2 - Epi	KIF23	KIF4A

TABLEAU 6.1 – **Top-gènes pour chacune des métriques testées.** Pour chaque métrique on extrait les 5 gènes les plus significatifs (maximisant la distance entre les groupes). Les trois premières colonnes correspondent à la partie 6.2 : "Dist. kanto." correspond à la distance de Kantorovitch, "Pval Student" au -log10 de la p-valeur du test de Student et "Dist. KS" à la distance Kolmogorov-Smirnov. Les deux autres colonnes correspondent à la partie 6.3.1 : "Wilks 5 dim." correspond à l'analyse discriminante sur les 5 types cellulaires et "Wilks réduit" correspond à l'analyse après réduction des axes par positionnement multidimensionnel.
Les gènes identiques entre deux métriques sont de la même couleur.

Pour la distance de Kantorovitch, on obtient les 5 gènes dont la dérégulation maximise la différence entre les proportions d'un type cellulaire avec un seuil de distance supérieure à 0,18. Parmi ces gènes, 4 sont associés au type immunitaire ($IC2$), et un associé au tissu épithélial (Epi). Ce dernier est NEK2, codant une kinase du cycle cellulaire trouvée sur-exprimée dans de nombreux cancers et impliquée dans la tumorigenèse ou la résistance au traitement [138]. Concernant les quatre gènes "immunitaires", on retrouve GIMAP1 et GIMAP7 : la dérégulation des gènes de la famille GIMAP (GTPase des protéines associées à l'immunité) dans le cancer du poumon a déjà été associée à la réponse immunitaire dans une étude préliminaire, avec notamment un lien dans la régulation des lymphocytes T [139]. FLI1 est un facteur de transcription dont le rôle dans la dérégulation du cycle cellulaire a été observé dans des cellules cancéreuses [140], et VSIR (anciennement appelé VISTA) est un récepteur immunitaire inhibant la réponse des lymphocytes T [141]. Les gènes retrouvés ont donc une cohérence biologique, mais on ne sait pas vraiment si le gène sort dérégulé dans la tumeur car le type cellulaire concerné est sur-représenté par rapport aux échantillons contrôles, ou si il y a réellement un lien avec les cellules cancéreuses. La figure 6.3, peu informative, ne sera pas reproduite pour les deux autres métriques.

Les cinq top-gènes retrouvés par le test de Student sont majoritairement associés au type épithélial, MAGEA6 et MAGEA3 codent des antigènes MAGE-A qui sont exprimés à la surface de nombreuses cellules cancéreuses de manière assez spécifique et qui sont des cibles thérapeutiques [142], CIP2A est une oncoprotéine bien caractérisée [143] et TMEM125 une protéine transmembranaire. On retrouve également VSIR associé au type immunitaire, comme pour la distance de Kantorovitch.

Enfin, avec la statistique de Kolmogorov-Smirnov on retrouve des gènes des deux premières métriques : MAGEA6, MAGEA3, NEK2, GIMAP et VSIR.

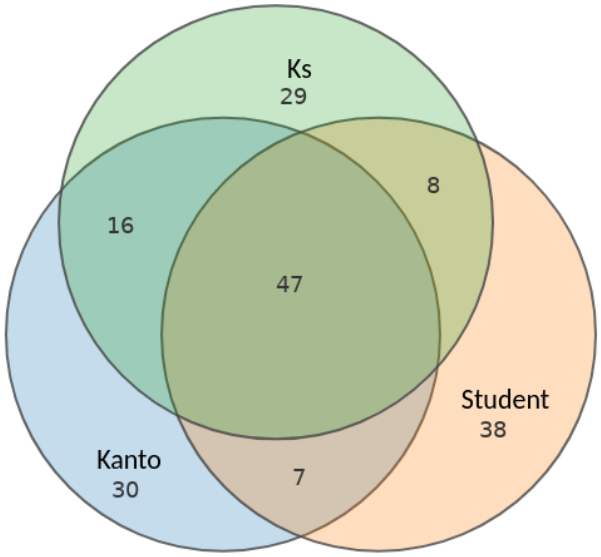


FIGURE 6.4 – **Répartition des 100 meilleures combinaisons gène-type cellulaire pour les trois métriques.** Pour chaque métrique (Ks = statistique du test de Kolmogorov-Smirnov, Kanto = distance de Kantorovitch, Student = -log10(p-valeur) du test de Student), on regarde la répartition des 100 meilleures associations gène-type cellulaire.

Même si le classement des meilleures combinaisons gène-type cellulaire n'est pas strictement le même entre les trois métriques, on retrouve une grande partie de gènes en commun (environ 70% des gènes sont communs à au moins 2 métriques, voir le diagramme de Venn figure 6.4). Ainsi, quand on regarde les

100 top-gènes pour chaque métrique, 47 sont communs aux 3 métriques, et entre 29 et 38 sont spécifiques à une seule (et ne se retrouvent pas non plus forcément dans les 200 top-gènes des autres métriques). Les 47 gènes communs sont associés au type cellulaire épithélial (27 gènes) et au type cellulaire immunitaire $IC2$ (20 gènes). Une rapide analyse d'enrichissement Gene Ontology [144] sur ces gènes montre une très forte sur-représentation du cycle cellulaire (microtubules, arrangements chromosomiques et mitose) pour les gènes associés au type épithélial, mais aucun résultat significatif n'est à noter pour les gènes associés à $IC2$.

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW! ?)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
positive regulation of exit from mitosis	6	2	.01	> 100	+	1.41E-04	4.18E-02
cellular response to interleukin-3	6	2	.01	> 100	+	1.41E-04	4.10E-02
mitotic spindle assembly checkpoint signaling	25	3	.06	52.58	+	3.48E-05	1.44E-02
mitotic spindle assembly	47	4	.11	37.29	+	5.45E-06	2.76E-03
mitotic sister chromatid segregation	115	7	.26	26.67	+	1.09E-08	1.15E-05
mitotic cytokinesis	68	4	.16	25.78	+	2.17E-05	9.73E-03
establishment of chromosome localization	75	4	.17	23.37	+	3.13E-05	1.37E-02

FIGURE 6.5 – **Résultats de l'analyse d'enrichissement sur les 27 gènes associés à l'épithélial.** Extrait des résultats de l'analyse d'enrichissement, seuls les premiers niveaux de processus biologiques sont affichés.

6.3 Analyse de la composition globale

Les résultats préliminaires de la partie précédente sont intéressants car les gènes retrouvés ont un sens biologique. Cependant, on relie un type cellulaire donné à l'expression d'un gène, mais on perd l'information de la structure tumorale et de la composition globale du micro-environnement de chaque tumeur. Une deuxième voie est donc explorée via l'analyse discriminante sur la structure tumorale. Le but de cette analyse, détaillée dans l'introduction partie III, est donc de tester directement si le statut de dérégulation d'un gène discrimine la distribution des proportions des différents types cellulaires.

6.3.1 Multidimensionnelle

Dans un premier temps, on utilise directement la matrice de proportion telle quelle en entrée de la fonction ptlmapper. On place donc les échantillons dans un espace à 5 dimensions (une par type cellulaire), et on regarde quel gène discrimine les distributions en fonction de sa dérégulation ou non.

Le top-5 des gènes les plus significatifs est visible dans le tableau 6.1, colonne "Wilks 5 dim". Le gène retrouvé avec la plus petite p-valeur est NEK2, qu'on retrouvait déjà dans la partie 6.2, ce résultat est rassurant car il montre une consistance avec la partie précédente (NEK2 sera détaillé dans la partie 6.4). On retrouve également TPX2, un facteur relié aux microtubules dont l'expression différentielle dans le cancer du poumon a déjà été remarquée [145] et deux gènes de la famille des kinésines (KIF), des protéines se déplaçant sur les microtubules. Ces résultats recoupent les indications de l'enrichissement Gene Ontology de l'analyse précédente, ou rejoignent les gènes détectés par les autres métriques.

6.3.2 Avec réduction des dimensions

Dans un second temps, on essaie l'analyse discriminante avec réduction des axes avant le calcul du score de Wilks. Réduire les dimensions peut permettre de concentrer le signal, de filtrer le bruit, et permet également une visualisation graphique des résultats.

Cette fois, c'est CEP55 le gène le plus significatif, il code une protéine du centrosome, et sa sur-expression induit la tumorigénèse chez la souris [146]. NEK2 a la deuxième plus petite p-valeur. On retrouve également TPX2 et un autre gène de la famille KIF. BIRC5 quant à lui est un gène régulateur de l'apoptose. L'avantage de cette analyse est de permettre une visualisation graphique en deux dimensions : chaque tumeur est placée en fonction de ses coordonnées, puis colorée en fonction du statut de dérégulation d'un gène (figure 6.6).

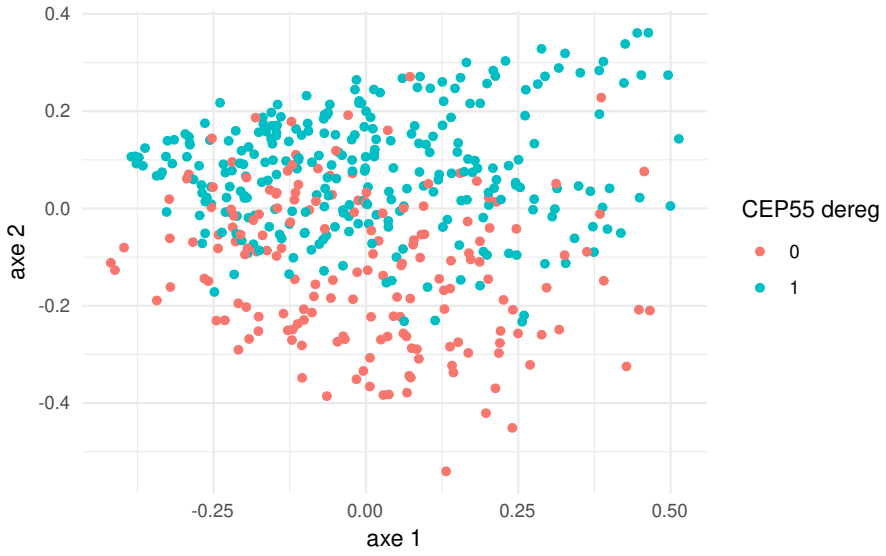


FIGURE 6.6 – **Visualisation graphique de l'analyse discriminante sur deux dimensions.** Les axes 1 et 2 sont calculés pour réduire les dimensions de la matrice de proportions en type cellulaires. Chaque point correspond à une tumeur, coloré en fonction du résultat de la méthode Penda sur le gène CEP55 (0 : non dérégulé, 1 : dérégulé). Visuellement, on voit que l'axe 2 sépare bien les deux groupes.

6.4 Exemple d'un gène : NEK2

Les analyses précédentes nous permettent de trouver des gènes dont la dérégulation est associée à la composition tumorale, mais des analyses plus poussées sont nécessaires pour évaluer la pertinence de ces gènes. Nous prenons donc comme exemple le gène NEK2 qui ressort clairement dans les analyses discriminantes, pour étudier plus en détail le lien entre son expression et les proportions des différents types cellulaires.

NEK2 est détecté sur-exprimé par Penda dans 315 tumeurs sur les 451, cette sur-expression est connue dans la littérature et est associée à un mauvais pronostic et une augmentation de la résistance aux traitements [138, 147].

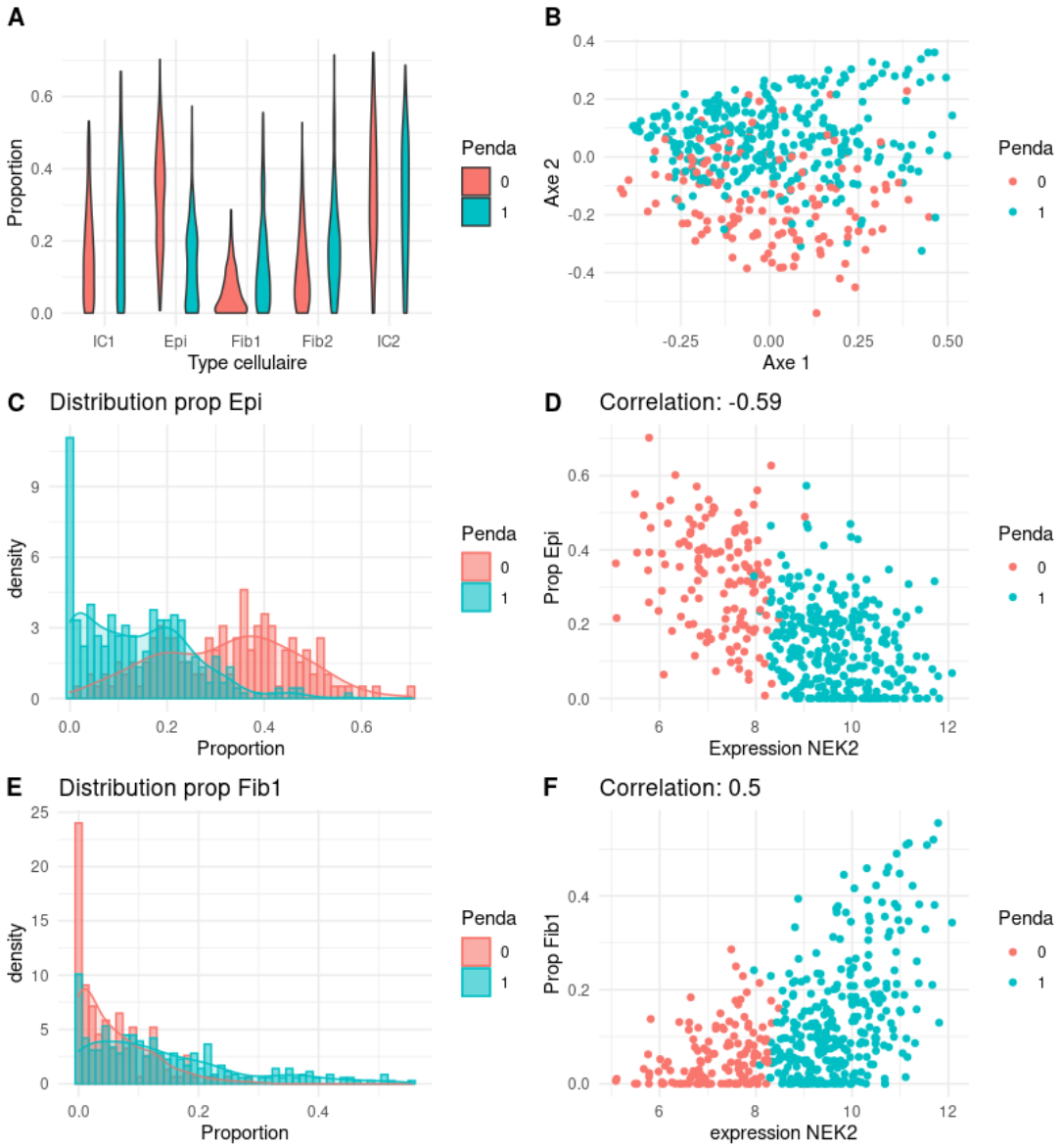


FIGURE 6.7 – **Exploration du gène NEK2.** En A, distribution des proportions des différents types cellulaire en fonction de la dérégulation Penda de NEK2 (0 : non dérégulé, 1 : dérégulé). En B, représentation des tumeurs en fonction des axes réduits, chaque tumeur est colorée en fonction de la dérégulation de NEK2. En C et E, distribution de la proportion du type cellulaire en fonction de la dérégulation de NEK2, pour respectivement les types "Epi" et "Fib1". En D et F, corrélation entre l'expression normalisée de NEK2 et la proportion déconvoluée du type cellulaire, chaque tumeur est colorée en fonction de la dérégulation Penda.

Si on regarde graphiquement la distribution des proportions des 5 types cellulaires déconvolués en fonction de la dérégulation Penda (figure 6.7 panneau A), on remarque que le type "Fibroblaste 1" semble plus présent dans les tumeurs où NEK2 est dérégulé. En revanche, le type "Épithélial" est sous-représenté dans ces cas-là. Ces résultats sont cohérents avec les métriques de distances calculées dans la partie 6.2. De même, quand on place graphiquement les tumeurs en fonction des coordonnées réduites sur deux dimensions, la dérégulation de NEK2 sépare visuellement les tumeurs en fonction de l'axe 2 (figure 6.7, panneau B).

On s'intéresse donc plus en détail aux types cellulaires déconvolués "Epi" et "Fib1". Quand on regarde leur distribution (panneaux C et E), on retrouve la même chose que dans la figure A. La différence est particulièrement nette pour le type épithélial, qui est sur-représenté quand NEK2 n'est pas dérégulé. La sur-expression de NEK2 serait donc plutôt associée aux tumeurs avec plus de tissus conjonctifs (fibroblastes) que de tissus cancéreux (dérivés de l'épithélial), or les fonctions des fibroblastes associés au cancer sont très large : activateur de croissance cellulaire et de métastase, mais aussi régulation du métabolisme cancéreux et recrutement de cellules immunitaires, leur présence peut à la fois être associée à un effet promoteur ou répresseur [86].

On regarde ensuite le lien entre l'expression brute normalisée de NEK2 dans les tumeurs (avant l'analyse différentielle), et la proportion des types cellulaires déconvolués (panneaux D et F). Premièrement, le résultat du test Penda de dérégulation (couleur des points) est très nettement relié avec l'expression de NEK2 dans la tumeur. On obtient une corrélation de -0,6 entre la proportion du type épithélial et l'expression de NEK2, et une corrélation de 0,5 avec la proportion de Fib1. Ces valeurs de corrélations sont relativement fortes, NEK2 aurait donc probablement été retenu sur un test de corrélation classique également.

6.5 Conclusion

Tous ces résultats indiquent un lien entre la dérégulation de certains gènes et la proportion des types cellulaires. De plus, ces gènes ont une pertinence biologique, ce qui est très encourageant. Cependant, l'interprétation des résultats de Penda est compliquée car on ne sait pas si c'est la dérégulation du gène qui influe sur la composition du micro-environnement tumoral, ou si le gène est juste un marqueur d'un type cellulaire sur - ou sous - représenté. Ainsi, la composition du micro-environnement est un facteur de confusion lors de l'application de Penda sur les profils mélangés.

Par ailleurs, la tumorigénèse implique également des dérégulations dans les cellules du micro-environnement, comme on a par exemple déjà pu l'aborder à propos des cellules immunitaires et des fibroblastes associés au cancer. Ce mécanisme amplifie l'incertitude sur l'origine de la dérégulation détectée par Penda quand on analyse l'expression à l'échelle tumorale. Il faudrait donc pouvoir appliquer l'analyse différentielle non pas sur la tumeur en entier, mais uniquement à l'échelle du type "cancer" purifié.

Pour extraire la part d'expression des gènes imputable au type cancer, et analyser uniquement la dérégulation dans celle-ci, il est nécessaire de retravailler l'étape de déconvolution et de caractérisation des types cellulaires obtenus. Nous souhaitons également utiliser l'information des échantillons contrôles en les incluant dans les étapes de déconvolution et d'analyse de dérégulation avec Penda. Toutes ces étapes vont être implémentées dans un pipeline d'analyse décrit dans la partie suivante.

Chapitre 7

Analyse de la régulation génétique du micro-environnement à partir de la composante tumorale purifiée

Dans la partie précédente, nous avons vu qu'utiliser l'échantillon mélangé pour l'analyse différentielle pouvait introduire des biais car la dérégulation détectée pouvait être directement liée à une variation de la composition du micro-environnement et pas forcément à une dérégulation du niveau d'expression dans un type cellulaire.

Nous décidons donc de développer un pipeline d'analyse nous permettant d'établir un lien entre gènes dérégulés spécifiquement dans les cellules cancéreuses, et la composition du micro-environnement. Ce pipeline est construit à partir des données d'expression et de méthylation de l'ADN brutes.

Ce chapitre se compose de deux parties : la première est consacrée à la construction d'un pipeline d'analyse robuste, implémenté sous la forme d'un package R appelé Ritmic pour *RegulatIon of Tumor MIcroenvironment Composition*. La construction du pipeline est couplée à une phase exploratoire sur deux jeux de données (un jeu de simulations simples et un jeu de données réelles) permettant de guider notre choix des méthodes implémentées. La deuxième partie décrit l'application du pipeline Ritmic entier à des simulations plus réalistes.

7.1 Explorations pour le développement d'un pipeline

La phase de développement du pipeline Ritmic est réalisée à l'aide de deux types de données : les données réelles LUAD du TCGA (voir partie précédente) et les simulations sur le cancer du pancréas issues du 2ème data challenge sur la déconvolution (voir partie 5.1.2).

Le pipeline d'analyse en lui-même est composé de quatre grandes étapes : (1) la déconvolution de la matrice obtenue sur des échantillons mélangés D en entrée pour retrouver les différents types cellulaires T et leurs proportions A, (2) l'identification de la part cancéreuse et l'isolement de son signal par rapport à la part du micro-environnement, (3) l'inférence des gènes dérégulés dans la part cancéreuse, puis enfin (4) l'analyse de la dérégulation de ces gènes en lien avec la composition du micro-environnement. Ces étapes sont résumées dans la figure 7.1.

7.1.1 Déconvolution

Dans nos précédents travaux sur la déconvolution (Partie II), celle-ci était réalisée principalement sur la méthylation de l'ADN et seule la matrice de proportion A était conservée pour nos analyses. Ici, on veut également obtenir une matrice T fiable pour avoir accès aux profils d'expression des gènes pour chacun des types cellulaires présents dans la tumeur. Pour rappel, la déconvolution se formule ainsi :

$$\begin{aligned} D_{RNAseq} &= T_{RNAseq} * A_{RNAseq} \\ D_{ADNm} &= T_{ADNm} * A_{ADNm} \end{aligned}$$

avec D_t la matrice obtenue par séquençage de l'échantillon mélangé, A_t la matrice de proportion des différents types cellulaires dans chaque tumeur et T_t le profil moléculaire des différents types cellulaires pour les données de type $t \in \{RNAseq, ADNm\}$. Théoriquement, lorsque ce sont les mêmes échantillons qui ont été séquencés avec les deux technologies différentes (RNA-seq,

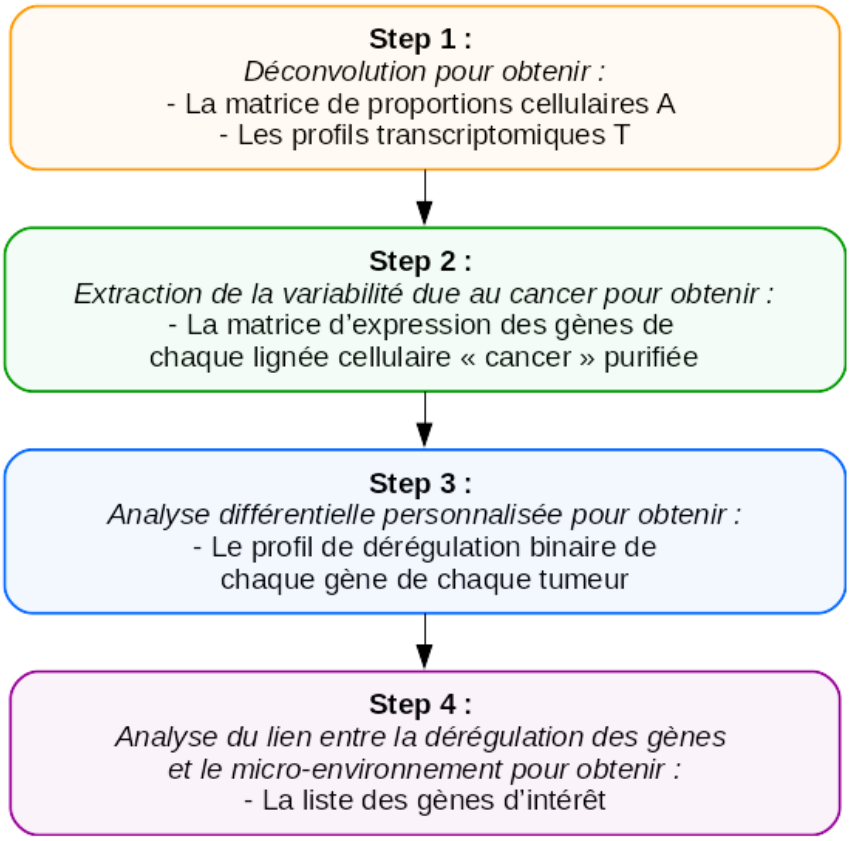


FIGURE 7.1 – **Représentation schématique du pipeline recherché.** Étape 1 (jaune) : déconvolution des matrices D. Étape 2 (vert) : isolement de l'expression des cellules cancéreuses de chaque échantillon. Étape 3 (bleu) : analyse des gènes dérégulés dans chaque échantillon par application de la méthode Penda. Étape 4 (violet) : identification des gènes d'intérêt, dérégulés en lien avec la composition du micro-environnement.

ADNm), A_{ADNm} et A_{RNAseq} peuvent être considérées comme étants identiques ($A_{RNAseq} = A_{ADNm} \equiv A$).

Pour avoir accès à cette matrice de proportion A et aux profils d'expression de gènes (T_{RNAseq}), trois stratégies ont été testées.

Détail des trois approches

Déconvolution par wNMF. Pendant le second data challenge sur la déconvolution (voir partie 5.1.3), wNMF a été la meilleure méthode pour inférer la matrice de proportion A en se basant uniquement sur les données RNA-seq. La méthode est en deux parties : une pré-sélection des sondes sur leur contribution aux principaux axes d'une Analyse en Composantes Indépendantes (ICA), puis la déconvolution par factorisation non-négative de la matrice (NMF). On obtient donc directement A et T_{RNAseq} en partant de D_{RNAseq}.

Reconstitution par inversion. Pour reconstituer T_{RNAseq}, nous avons aussi développé une méthode simple à partir de la déconvolution sur la méthylation. Dans un premier temps, la matrice D_{ADNm} est déconvoluée en suivant le pipeline Medepir (voir partie 4.5.2). On calcule ensuite la pseudo-inverse de la matrice A obtenue, afin de pouvoir obtenir T_{RNAseq} directement à partir de D_{RNAseq}.

$$D_{RNAseq} * inv(A_{ADNm}) = T_{RNAseq}$$

EDec. Jusqu'ici, EDec avait été utilisé dans nos analyses uniquement à travers son étape 1, la déconvolution des données de méthylation de l'ADN. Cependant, la méthode possède aussi une fonction pour retrouver le profil de RNA-seq à partir de la déconvolution sur l'ADNm : $run\_edec\_stage\_2$. Comme pour l'inversion ci-dessus, cette fonction estime T_{RNAseq} à partir de la matrice A calculée sur la méthylation et de D_{RNAseq}. Le principe mathématique appliqué pour cela est la méthode des moindres carrés appliquée pour minimiser la distance entre $A*T$ et D, et contrainte pour que les valeurs d'expression obtenues dans T_{RNAseq} soient supérieures ou égales à 0 [102].

Résultats

Les trois méthodes de déconvolution sont exécutées sur les données LUAD du TCGA et sur les simulations du 2ème data challenge de déconvolution. Nous avons ensuite comparé les matrices T_{RNAseq} obtenues pour les gènes en commun entre toutes les données (après pré-sélection des sondes par chacune des méthodes sur chacun des jeux de données, c'est-à-dire 10 015 gènes pour LUAD et 11 668 gènes pour les simulations).

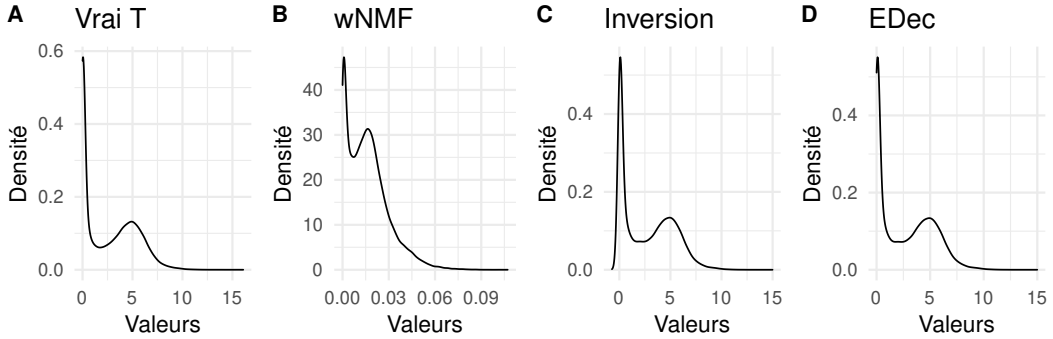


FIGURE 7.2 – **Distribution des matrices T pour les données simulées.** A) T utilisée pour les simulations (T théorique), B) T déconvoluée obtenue par wNMF, C) T déconvoluée obtenue par inversion, D) T déconvoluée obtenue par EDec. Voir partie 7.1.1 pour le détail des méthodes.

Les valeurs des matrices T et A obtenues diffèrent beaucoup entre les méthodes de déconvolution. La distribution des données pour chacune des méthodes est visible figure 7.2 pour les données simulées et figure 7.3 pour les données LUAD. La méthode NMF nous permet d'obtenir des "expressions" comprises entre 0 et 0,5 avec une médiane à 0.0004 pour LUAD et entre 0 et 0,1 avec une médiane à 0.015 pour les simulations. Ces valeurs sont vraiment faibles. Avec la méthode d'inversion, on retrouve des valeurs comprises entre -530 000 et +870 000 avec une médiane à 600 pour LUAD, et entre -0,7 et +15 avec une médiane à 2,3 pour les simulations. La présence de valeurs négatives (12% pour les données réelles et 4% pour les simulations) n'a bien évidemment aucun sens biologique et pourrait éventuellement poser problème par la suite si certaines méthodes d'analyse ne les prennent pas en charge. Enfin, EDec nous donne des expressions entre 0 et 620 000 avec une médiane à 577 pour LUAD (déconvo-

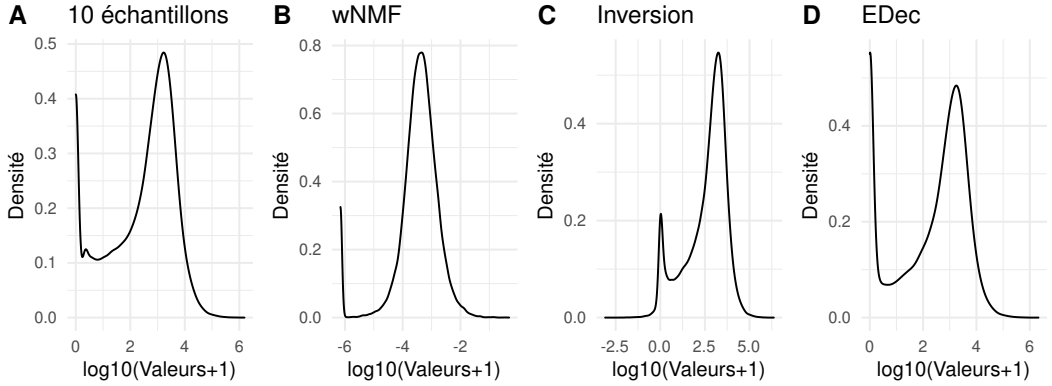


FIGURE 7.3 – **Distribution des données LUAD.** A) 10 premiers échantillons de la matrice D d'origine, B) T déconvoluée obtenue par wNMF, C) T déconvoluée obtenue par inversion, D) T déconvoluée obtenue par EDec. Les données ont été log-transformées pour faciliter la lecture des distributions (log10(valeur+1) pour A, C et D, log10(valeur+x) avec x la médiane des valeurs de wNMF divisée par la médiane des 10 échantillons de la matrice D d'origine pour le panneau B). Voir partie 7.1.1 pour le détail des méthodes.

luées sur des "reads" non log-transformés), et entre 0 et 15 avec une médiane à 2,3 pour les simulations (fabriquées à partir de données log-transformées). Les valeurs et les distributions obtenues par EDec sont les plus cohérentes avec les données biologiques.

Pour comparer les méthodes deux à deux, on calcule la corrélation de Pearson entre les vecteurs T obtenus. Les résultats sont visibles sur la figure 7.4. Les trois graphiques du haut représentent les résultats pour les données LUAD. On observe une assez bonne corrélation entre les types obtenus par les deux méthodes partant de la matrice A_{ADNm} (inversion et EDec) (voir panneau C). Ainsi, chaque type inféré par EDec peut être associé visuellement avec un type issu de l'inversion ($T1_{inv}$ avec $T1_{Edec}$, $T2_{inv}$ avec $T2_{Edec}$, etc.). En revanche, quand on regarde la corrélation de ces méthodes avec wNMF, l'association est moins nette pour certains types cellulaires (panneaux A et B). Par exemple, on peut considérer que $T2_{nmf}$, $T3_{nmf}$, $T4_{nmf}$ et $T5_{nmf}$ correspondent respectivement à $T3_{inv/EDec}$, $T5_{inv/EDec}$, $T1_{inv/EDec}$ et $T2_{inv/EDec}$ mais l'association restante, c'est à dire $T1_{nmf}$ avec $T4_{inv/EDec}$, est beaucoup moins claire.

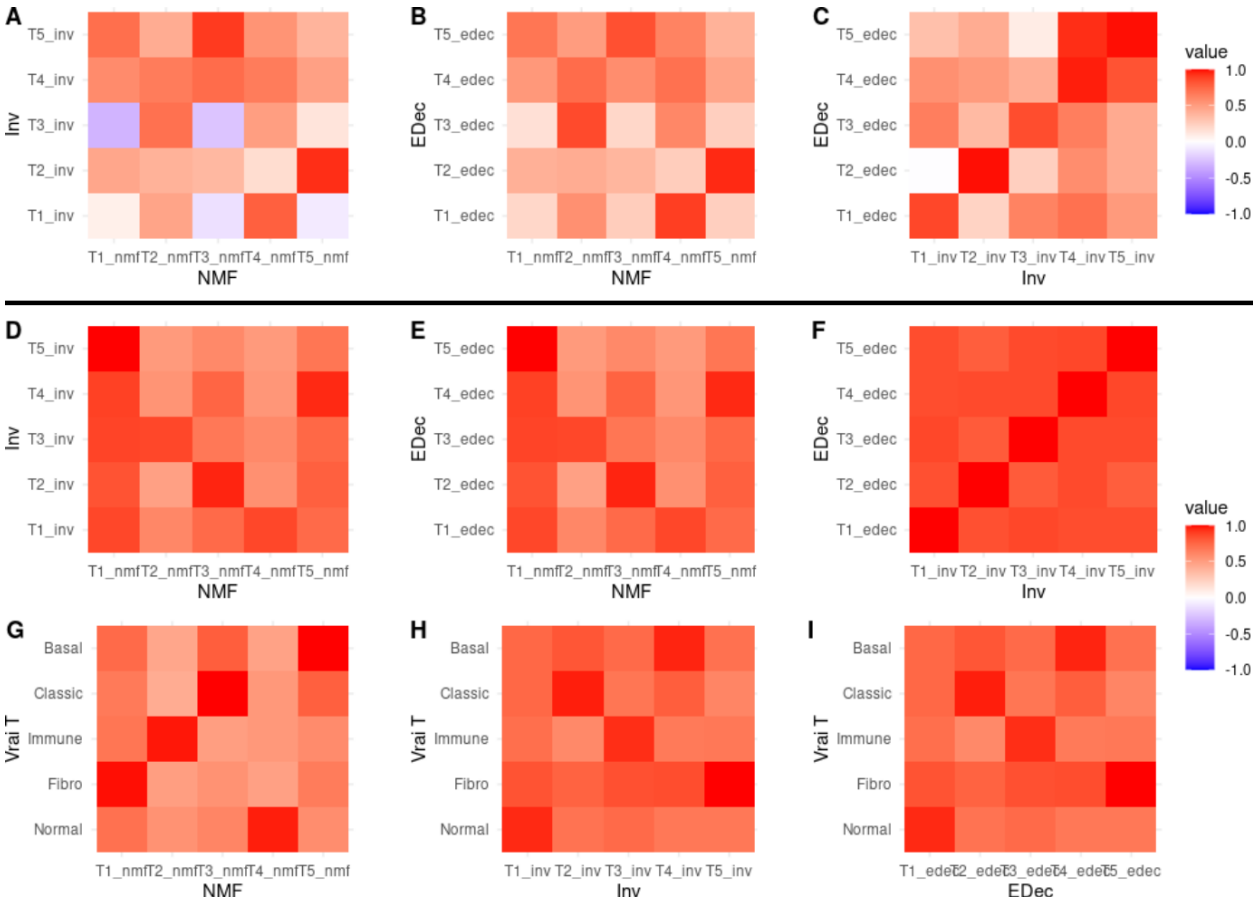


FIGURE 7.4 – **Corrélations entre les matrices T déconvoluées.** En A-C, les données LUAD, en D-I les données du pancréas simulées. En A-F les types cellulaires déconvolués par les trois méthodes (Inv = Inversion, NMF = wNMF et EDec) sont comparés deux à deux. En G-I, les types déconvolués sont comparés à la "réalité terrain" ayant servi aux simulations (Vrai T).

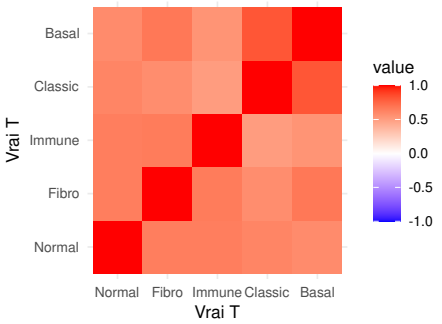


FIGURE 7.5 – **Corrélation entre les types cellulaires utilisés pour les simulations.**

Sur les trois figures de la deuxième ligne, on observe la même tendance pour les types déconvolués à partir des matrices D simulées. Encore une fois, les résultats sont très fortement corrélés entre EDec et Inv (panneau F). La corrélation avec les types cellulaires déconvolués par wNMF est moins claire (panneaux D et E).

Enfin, sur les trois graphiques du bas, on voit la corrélation entre les types déconvolués et les types cellulaires ayant servi à faire la simulation. Pour les trois méthodes, le résultat est satisfaisant et on retrouve clairement une correspondance entre les types cellulaires d'origine et ceux déconvolués. Il est à noter que globalement toutes les corrélations sont très fortes, y compris au sein d'une même matrice T, on peut le voir également entre les types servant aux simulations sur la figure 7.5.

Méthode retenue pour Ritmic

Pour l'implémentation du pipeline, nous avons finalement choisi d'utiliser la méthode EDec pour plusieurs raisons. Premièrement, la méthode a déjà été publiée et est robuste, ce qui est un avantage pour pouvoir se concentrer sur les autres aspects innovants de Ritmic. De plus, EDec intègre des étapes pour contraindre biologiquement la pertinence des résultats (par exemple en calculant des profils d'expression et de méthylation positifs, et en contraignant les proportions à être comprises entre 0 et 1 avec une somme égale à 1 pour chaque tumeur). Un autre argument en faveur d'EDec est que l'approche multi-omique de la déconvolution permet théoriquement d'obtenir des résultats plus robustes aux bruits expérimentaux de chacune des méthodes de séquençage. Enfin, les distributions des valeurs de T obtenues par EDec sont beaucoup plus proches des valeurs réelles que celles des autres méthodes de déconvolution.

Ainsi, l'approche retenue est dans un premier temps d'appliquer les méthodes de pré-traitement développées dans le pipeline Medepir : les données de méthylation de l'ADN sont donc filtrées pour retirer les sondes dont la valeur est corrélée aux données cliniques. À cette étape, on peut utiliser une Analyse en Composantes Principales pour estimer le nombre de types cellulaires composant l'échantillon s'il est inconnu. Ensuite, les 2000 sondes les plus variables sont

sélectionnées, ce qui permet de se concentrer sur les sondes les plus informatives et d'accélérer le temps de calcul des étapes suivantes. Cette étape de sélection des sondes est facultative, et le nombre de sondes conservées est au choix de l'utilisateur.

La fonction *run_edec_step*1 permet dans un second temps de déconvoluer les données d'ADNm pour obtenir la proportion des différents types cellulaires dans chaque échantillon. Cette matrice de proportion est utilisée dans la fonction *run_edec_step*2 pour obtenir finalement les profils d'expression de chaque type cellulaire déconvolué.

7.1.2 Isoler la part "cancer"

L'étape suivante du pipeline est d'isoler la part "cancer" qui correspondrait aux cellules cancéreuses purifiées, ce qui implique dans un premier temps de séparer les types cellulaires déconvolués en deux groupes : ceux qui correspondent au type cancéreux et ceux qui correspondent au micro-environnement. Pour cette étape, nous explorons trois stratégies.

Exploration de différentes approches

1. Profils de référence. Jusqu'ici (dans la partie II et dans le chapitre précédent d'exploration), la stratégie de caractérisation des types cellulaire était de corréler les profils déconvolués (sur la méthylation) et les profils de références de la méthode Epidish.

Les résultats obtenus sont visibles sur la figure 7.6, ils permettent d'identifier assez bien le type 5 déconvolué comme des cellules immunitaires, le type 2 déconvolué comme de l'épithélial, et les types 3 et 4 comme des fibroblastes. L'identification du type 1 reste plus difficile, de plus il y a malgré tout une corrélation assez forte (0,3) entre le type 5 et le profil fibroblaste.

Globalement, cette méthode convient quand des références ont déjà été développées pour les types cellulaires étudiés, mais elle n'est pas applicable à tous les types de données. Par ailleurs, dans le cas de LUAD, cette approche ne

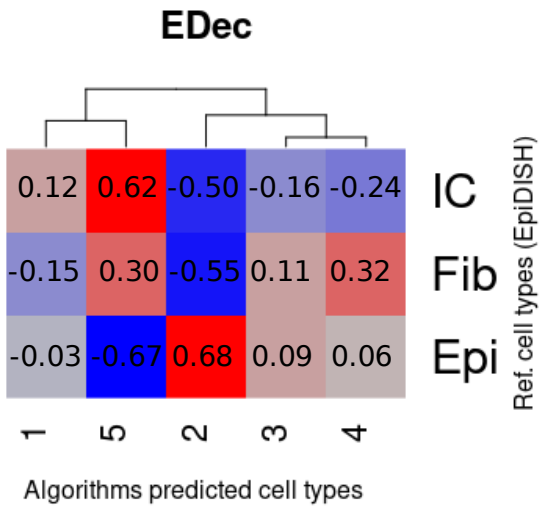


FIGURE 7.6 – **Corrélation entre les profils déconvolués et les profils d'Epidish.** Heatmap des corrélations entre les types déconvolués par EDec (en colonnes) et les types de référence d'Epidish (en lignes). Epi = épithélial, Fib = fibroblastes, IC = cellules immunitaires. Les colonnes sont ordonnées par clustering hiérarchique.

permet pas vraiment une identification précise (on avait seulement trois classes : fibroblaste, épithélial ou cellule immunitaire pour 5 types déconvolués). Nous avons donc exploré deux nouvelles stratégies pour identifier la part cancéreuse dans les données LUAD.

2. Marqueurs. La seconde stratégie est de chercher dans les profils déconvolués des biomarqueurs des différents types cellulaires qu'on peut retrouver dans le cancer pulmonaire.

Dans un premier temps, une liste de biomarqueurs provenant d'une étude en single cell sur l'adénocarcinome et regroupant les marqueurs canoniques des types cellulaires à partir de nombreuses références est sélectionnée [148]. Les marqueurs de l'étude Kim sont regroupés en cinq grands types cellulaires : épithélial avec 4 gènes (EPCAM, KRT19, CDH1, KRT18), immunitaire avec 16 gènes regroupant différents types cellulaires (Lymphocytes T : CD3D, CD3E,

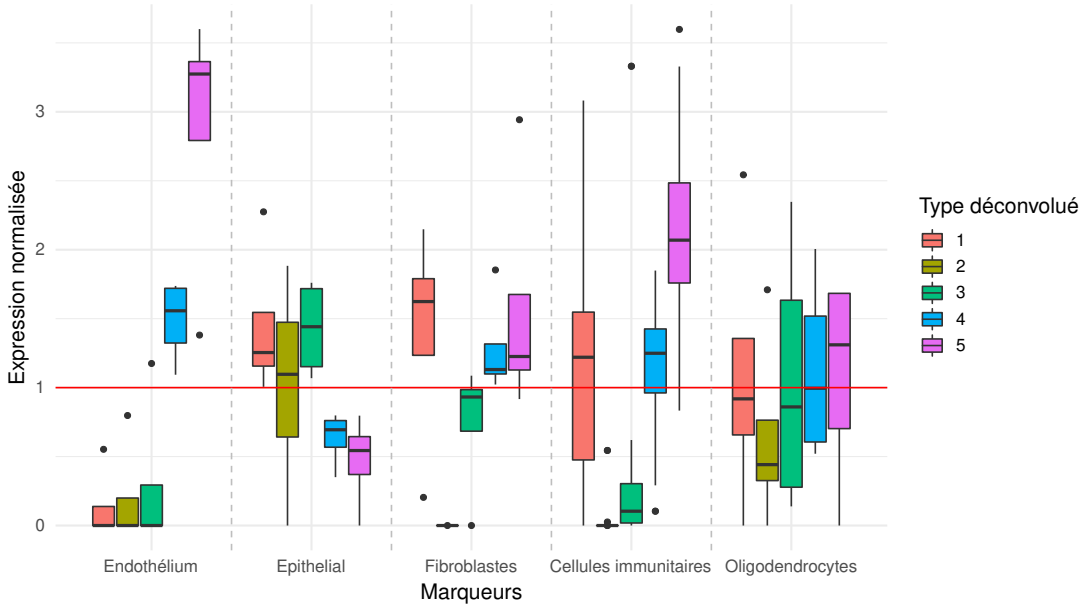


FIGURE 7.7 – **Niveau d'expression des biomarqueurs Kim [148] dans chaque type cellulaire déconvolué.** L'expression de chaque gène est divisée par son expression moyenne dans les différents types cellulaires. La barre rouge indique un score de 1, donc une expression égale à la moyenne.

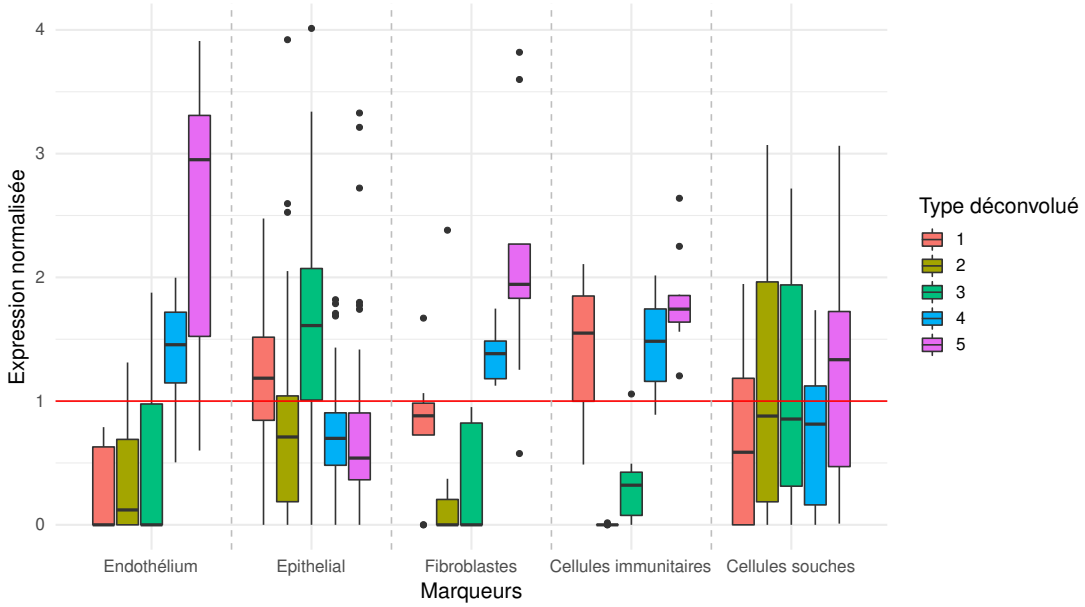


FIGURE 7.8 – **Niveau d'expression des biomarqueurs CellMarker dans chaque type cellulaire déconvolué.** L'expression de chaque gène est divisée par son expression moyenne dans les différents types cellulaires. La barre rouge indique un score de 1, donc une expression égale à la moyenne.

CD3G, lymphocytes B : CD79A, cellules myéloïdes : CD68, MARCO, FCGR3A, LYZ, lymphocytes natural killer : NCAM1, NKG7, GNLY, KLRD1 et mastocytes : NCAM1, NKG7, GNLY, KLRD1), fibroblastes avec 4 gènes (DCN, COL1A1, COL1A2, THY1), endothélium (vaisseaux sanguins) avec 4 gènes (PECAM1, CLDN5, FLT1, RAMP2) et enfin oligodendrocytes (nerfs) avec également 4 gènes (OLIG1, OLIG2, MOG, CLDN11).

Pour étudier les profils déconvolués, on divise ensuite l'expression de chaque gène par sa moyenne dans tous les types cellulaires afin de normaliser les niveaux d'expression entre les biomarqueurs. Un score de 1 signifie donc que le gène est pile sur la moyenne, supérieur à 1 il est plus exprimé dans ce type cellulaire, inférieur à 1 il est moins exprimé. Graphiquement, on peut ensuite visualiser l'expression des différents marqueurs dans chaque type cellulaire de chaque matrice de profils cellulaires déconvolués.

Les résultats obtenus sont visibles dans la figure 7.7. Globalement, ils donnent des idées d'associations entre tissus déconvolués et types cellulaires, mais il n'y a pas de consensus clair. Ainsi, certains types déconvolués sont associés à plusieurs biomarqueurs (par exemple le type 1 qui sur-exprime à la fois les biomarqueurs épithélial, fibroblastes et immunitaire) et d'autres à aucun (par exemple le type 2).

Le manque de robustesse pouvant provenir des listes de biomarqueurs utilisées qui sont très petites, nous avons donc essayé la même technique sur des listes plus larges. Ces listes sont obtenues à partir de CellMarker, une base de données regroupant des marqueurs de types cellulaires décrits dans la littérature [149]. Les résultats obtenus dans cette analyse recoupent les premiers, mais ne sont pas plus concluants, avec toujours des types déconvolués non définis et d'autres qui expriment plusieurs marqueurs différents (figure 7.8).

Les types cellulaires déconvolués que nous obtenons ne semblent donc pas interprétables comme les profils de single cell utilisés pour définir les biomarqueurs. Dans ce cas, d'autres approches doivent être testées (comme dans la prochaine section), ou les listes de marqueurs utilisés affinées.

3. Répartition contrôles/tumeurs. Puisque les types cellulaires ont été déconvolués à la fois dans des échantillons sains et tumoraux, on choisi de tester une autre approche plus pragmatique : regarder directement la composition des types déconvolués dans chaque condition clinique.

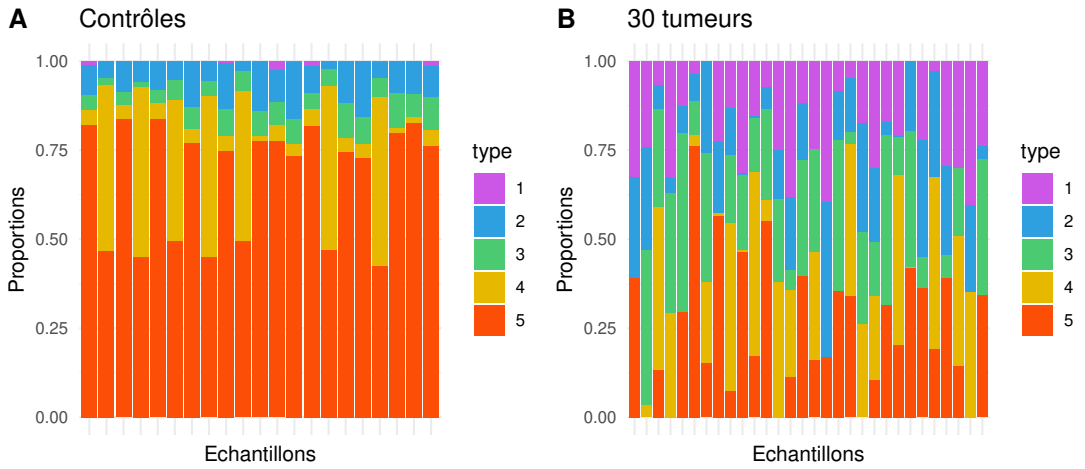


FIGURE 7.9 – **Visualisation des matrices de proportions dans les échantillons sains et tumoraux déconvolués.** Les matrices ont été déconvoluées par EDec étape 1 sur les données de méthylation de l'ADN. En A, les échantillons contrôles, en B, 30 échantillons tumoraux tirés au hasard.

Comme on peut voir sur la figure 7.9, les proportions des différents types cellulaires sont très différentes entre les échantillons sains et tumoraux. Premièrement, on voit que les profils des contrôles sont beaucoup plus stables, avec majoritairement le type 5 et une proportion variable du type 4. Le type cellulaire 1 est particulièrement minoritaire dans tous les échantillons contrôles, on retrouve également peu les types 2 et 3. Les profils des tumeurs sont beaucoup plus variés, on observe les 5 types cellulaires en proportions variables entre les échantillons.

Comparaison des méthodes d'assignation. Les résultats des 3 méthodes d'assignation des types cellulaires déconvolués (corrélation avec des types cellulaires de référence pour l'ADNm déconvolué, expression des biomarqueurs dans les types RNA-seq déconvolués et comparaison de la répartition des types cellulaires

entre contrôles et tumeurs) sont regroupés dans le tableau 7.1. Pour Epidish, on assigne les types déconvolués qui ont une corrélation supérieure ou égale à 0,3. Pour les biomarqueurs, il faut qu'au moins 75% des marqueurs soient sur-exprimés dans le type déconvolué.

	T1	**T2**	**T3**	**T4**	**T5**
Corrélation Epidish	/	Epi	Fib	Fib	IC, Fib
Marqueurs Kim	Epi, Fib	/	Epi	Endo, Fib, IC	Endo, Fib, IC
Cell Marker	Epi, IC	/	Epi	Endo, Fib, IC	Endo, Fib, IC
Proportion	Cancer	/	/	Sain	Sain

TABLEAU 7.1 – **Résumé de l'interprétation des types cellulaires pour les différentes méthodes d'assignation** On cherche à relier les types déconvolués (T1, T2, T3, T4 et T5) avec un des types cellulaires réels (Fib = fibroblastes, Epi = épithélial, IC = cellules immunitaires, Endo = endothélial). "Corrélation Epidish" consiste à corréler la matrice T déconvoluée sur la méthylation avec les types de référence de la méthode Epidish. "Marqueurs Kim" et "Cell marker" se basent sur la sur-expression des biomarqueurs des différents types cellulaires dans la matrice T_{RNAseq} déconvoluée. Enfin, "Proportion" consiste à regarder la répartition des types déconvolués de la matrice A entre les échantillons contrôles et les échantillons tumoraux.

Les types 4 et 5, majoritairement présents dans les échantillons contrôles, étaient assignés respectivement aux fibroblastes et aux cellules immunitaires par la corrélation sur la méthylation, et à la fois à l'endothélium, aux fibroblastes et aux cellules immunitaires par les biomarqueurs. Ces résultats sont assez cohérents avec la composition possible du micro-environnement tumoral.

Le type 1 qui est seulement présent dans les échantillons tumoraux déconvolués était faiblement corrélé pour Epidish, mais est fortement associé aux biomarqueurs des cellules épithéliales pour les deux types de marqueurs. Les cellules cancéreuses étant dérivées de l'épithélial, ces résultats semblent plutôt consistants. Enfin, les types 2 et 3 variant entre les échantillons cancer et

contrôles semblent également être des cellules plutôt épithéliales.

Regarder la répartition des composantes entre tissus sains et tumoraux peut donc venir compléter la méthode des biomarqueurs ou de la corrélation à des types cellulaires de référence, même si cela ne permet pas forcément de différencier les cellules "cancer" des autres tissus propres à la tumeur (fibroblastes associés au cancer par exemple).

Méthode retenue pour Ritmic

Dans Ritmic, nous avons choisi de ne pas imposer à l'utilisateur de méthode pour l'identification des types déconvolués car, comme on l'a vu précédemment, le choix dépend beaucoup des données analysées et des données disponibles.

Quand une bonne liste de biomarqueurs ou des profils de référence pour les types cellulaires composant l'échantillon sont disponibles, on peut directement les utiliser pour identifier les profils déconvolués à partir de T.

En l'absence de biomarqueurs et de référence, il peut être plus difficile d'identifier les types déconvolués. Si des contrôles ont été intégrés à l'étude, un bon indicateur peut être de regarder les types cellulaires qui les composent. On peut également utiliser la matrice de proportion pour retrouver des corrélations entre la présence de certains types cellulaires et le score de pureté tumorale calculé par des outils comme Infinium purity [150].

7.1.3 Expression différentielle dans le type cellulaire cancer

Une fois le type "cancer" identifié, l'objectif est d'inférer la dérégulation des gènes dans ce type pour chaque échantillon. Cependant les méthodes de déconvolution donnent des profils T constants : ces profils représentent l'expression moyenne des gènes dans les types cellulaires en question. Pour avoir une expression des gènes unique à chaque échantillon, on fait l'hypothèse que la variabilité résiduelle de l'expression, non-expliquée par la composition différentielle en type cellulaire, se concentre essentiellement dans le type cellulaire cancéreux.

De ce fait, on définit :

$$T_{cancer^i} = \frac{D_i - \sum_{j=1}^{K_{ME}} T_k * A_{i,k}}{A_{cancer}} \tag{7.1}$$

, avec T_{cancer^i} le vecteur moléculaire de la composante cancéreuse dans chaque échantillon i, D_i la matrice d'expression convoluée de i, T_k les profils d'expression moyens des K_{ME} types cellulaires non-cancéreux, $A_{i,k}$ la proportion du type k dans l'échantillon i et $A_{i,cancer}$ la proportion totale des types cellulaires cancéreux dans i.

On souhaite maintenant identifier les gènes dérégulés dans chaque échantillon de cellules cancéreuses purifiées T_{cancer^i}, en utilisant la méthode Penda décrite dans la première partie de ma thèse. Cette méthode permet d'obtenir une information sur la dérégulation de chaque gène dans chaque échantillon étudié. Plusieurs hypothèses sont envisagées pour définir les échantillons contrôles qui serviront pour le test.

Choix des contrôles

Le choix des échantillons contrôles utilisés pour la méthode Penda est une étape très importante, car ils servent à la fois de références pour les rangs relatifs, mais aussi pour la méthode des centiles quand le test Penda ne peut pas être appliqué. Comme on l'a déjà abordé dans la partie 2.3.3, de mauvais contrôles peuvent compromettre fortement la pertinence des résultats. Plusieurs pistes sont envisagées.

Lignée non cancéreuse déconvoluée. La première possibilité est d'utiliser le type cellulaire sain dont dérivent les cellules cancéreuses comme référence. Cependant, adapter Penda pour utiliser uniquement un échantillon contrôle nous fait perdre beaucoup de robustesse puisqu'on a normalement besoin d'un jeu de contrôles de taille suffisamment importante pour établir les listes de références de rangs stables entre les gènes. Par ailleurs, l'identification précise de cette lignée parmi les types inférés peut s'avérer très délicate. Cette solution

a donc été écartée pour nos analyses.

Contrôles purifiés. La seconde piste est d'utiliser le signal des échantillons contrôles déconvolués après retrait des types identifiés comme non-cancéreux. Cependant, retirer l'expression des types constituant le micro-environnement des contrôles revient à se concentrer sur une expression résiduelle qui dans notre modèle correspond au bruit, cette solution ne nous semble donc pas pertinente non plus.

Profils cellulaires de référence. Finalement, nous décidons de comparer l'expression du cancer à des profils d'expression de cellules saines dont le type de cancer étudié est dérivé. Dans nos analyses, nous avons utilisé la base de données publiques GTEX qui permet d'avoir accès à un grand nombre de tissus sains avec un nombre important d'échantillons par tissu [116]. Penda est ensuite appliqué en suivant les vignettes pour estimer les meilleurs paramètres (Partie 1.3.4).

Résultats sur LUAD

Les simulations de Penda pour choisir les meilleurs paramètres sur la lignée purifiée LUAD (où l'expression des types 4 et 5 identifiés comme "non-cancéreux" est retirée) donnent de très mauvais résultats, avec des courbes ROC diagonales.

Une hypothèse pour expliquer ces résultats est peut-être qu'à l'issue de l'étape précédente, on se retrouve avec près de 25% des données T_{cancer^i} qui sont négatives. Le problème vient peut-être également des contrôles, ou d'une des étapes précédentes comme la déconvolution et l'identification des types cellulaires. Comme l'étape suivante nécessite des simulations adaptées afin d'évaluer les performances du pipeline, et qu'il semble difficile d'identifier simplement le problème rencontré sur le jeu de données LUAD à ce stade, nous décidons d'arrêter l'exploration ici pour la reprendre sur des simulations adaptées plus tard.

7.1.4 Lien entre gène et micro-environnement

L'objectif de la dernière étape du pipeline est de retrouver les gènes dont la dérégulation est reliée à la proportion du micro-environnement.

La première étape consiste à pré-traiter les résultats de la méthode Penda. Premièrement, on veut trier les gènes pour conserver uniquement ceux dont le statut de dérégulation varie entre les tumeurs. Comme on va utiliser des métriques comparant les distributions entre les deux groupes, il est nécessaire d'avoir assez d'échantillons dans chaque catégorie pour conserver une robustesse statistique. Le nombre d'échantillon minimum est fixé par l'utilisateur à travers le paramètre de la fonction $ritmic :: pre\_treat$. Le format des données est également modifié, pour passer de la liste Penda (sur-exprimé : vrai ou faux, sous-exprimé : vrai ou faux) à une matrice 0/1, avec 0 un gène non dérégulé et 1 un gène dérégulé.

La seconde étape permet de calculer les métriques : pour chaque gène, on sépare les tumeurs en fonction du statut de dérégulation. On mesure ensuite l'écart entre les proportions de chaque type cellulaire de chacun des deux groupes, grâce aux métriques décrites dans l'introduction partie III. Pour chaque gène et chaque type cellulaire, on obtient donc la distance de Kantorovitch, la p-valeur du test de Student, et la distance du test de Kolmogorov-Smirnov. Si la dérégulation d'un gène est corrélée à la présence d'un type cellulaire, la distance entre les deux groupes sera plus grande. Pour estimer l'intérêt de l'étape Penda, on calcule également la corrélation directe entre l'expression dans les cellules cancéreuses et la proportion de chaque type cellulaire. Il est à noter qu'en plus de ces mesures par type cellulaire, il est possible d'effectuer une analyse discriminante pour voir si la dérégulation d'un gène est reliée à la structure tumorale globale (l'ensemble des proportions).

7.1.5 Conclusion

La version finale du pipeline est implémentée sous la forme d'un package R (Ritmic) englobant toutes les étapes, il est schématisé dans la figure 7.10.

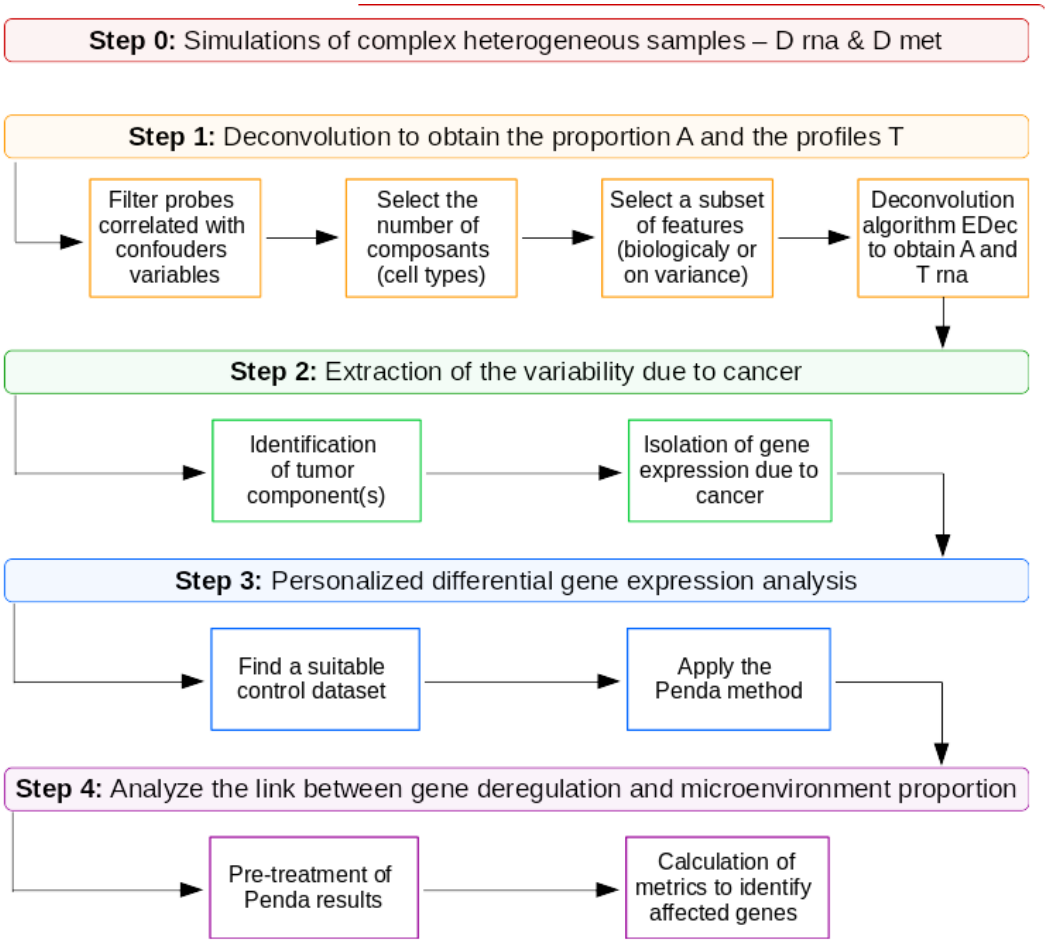


FIGURE 7.10 – **Représentation schématique du pipeline Ritmic.** Étape 0 (rouge) : différentes fonctions du package permettent de simuler des matrices D de transcriptome et de méthylome, elles peuvent aussi être obtenues expérimentalement. Étape 1 (jaune) : les matrices D sont déconvoluées pour obtenir la matrice de proportion A et le profil d'expression de chaque type cellulaire T_{rna}, ces étapes sont effectuées avec les packages Medepir et EDec. Étape 2 (vert) : l'expression des cellules cancéreuses de chaque échantillon est extraite grâce à la méthode implémentée dans Ritmic. Étape 3 (bleu) : les gènes dérégulés dans chaque échantillon sont identifiés grâce à l'application de la méthode Penda sur la composante "cellules cancéreuses". Étape 4 (violet) : les gènes dérégulés en lien avec la composition du micro-environnement sont identifiés grâce à différentes métriques implémentées dans Ritmic.

Il permet donc de déconvoluer des tumeurs mélangées pour en retrouver la composition, puis ensuite d'extraire du mélange l'expression identifiée comme étant propre au type cancer, avant d'appliquer une méthode d'analyse différentielle pour obtenir la dérégulation des gènes dans chaque échantillon et ainsi de pouvoir évaluer si ces dérégulations sont associées à la composition du micro-environnement tumoral.

Afin d'évaluer ce pipeline, nous allons développer dans la prochaine partie des simulations complexes permettant de représenter une tumeur mélangée, et tester l'influence des différents paramètres sur les résultats.

7.2 Application du pipeline à des simulations

L'application du pipeline Ritmic sur des simulations a deux objectifs : évaluer l'impact de tous les paramètres de simulation, et tester la pertinence de notre méthode avant une application future aux données réelles, pour lesquelles on a vu précédemment qu'il pouvait exister certaines difficultés. Dans un premier temps, je vais présenter les simulations, puis ensuite les différents paramètres ayant varié, et les résultats de chaque étape du pipeline. Dans cette partie, nous choisissons de tester la variation de l'expression de certains gènes tirés au hasard avec la proportion du type immunitaire (IC), car l'infiltration immunitaire est une donnée couramment étudiée [151].

7.2.1 Principe des simulations

Le principe général des simulations est le même que celui décrit dans la partie 5.1.2 pour le 2ème data challenge. On simule donc des échantillons du pancréas. Le profil complexe D_i de chaque échantillon i est simulé de la manière suivante :

$$D_i = \sum_{j=1}^{k} T_j * A_{i,j}$$

Avec T_j un vecteur correspondant au profil moléculaire de référence de chaque type cellulaire, et $A_{i,j}$ la proportion du type j dans l'échantillon du

patient i.

Matrices A et T

La même matrice de proportions simulées A est utilisée pour les jeux de données transcriptome et méthylome d'un même jeu d'échantillons, elle est générée par une distribution de Dirichlet comme dans la partie 5.1.2.

De même, la matrice T du méthylome est générée avec les mêmes profils médians que pour le data challenge, c'est-à-dire à partir des lignées GTEx pour les types cellulaires sains et des tumeurs PDX pour le type cancéreux. On conserve quatre types cellulaires : cancer de type classique, cellules immunitaires, épithélial sain et fibroblastes.

La réalisation de la matrice T du transcriptome diffère un peu de la partie 5.1.2. Alors que les profils des types cellulaires sains sont considérés comme identiques entre chaque patient, une approche plus réaliste est choisie pour le type cellulaire cancer, avec le profil moléculaire qui varie désormais entre chaque échantillon.

Cette stratégie a plusieurs objectifs, d'une part s'approcher de la réalité où chaque processus oncogénique est unique et chaque tumeur très différente d'une autre, mais aussi permettre de déréguler les gènes de manière individuelle dans les cellules cancéreuses, et donc d'obtenir à la fin de la déconvolution des profils hétérogènes entre les individus. Afin de simplifier les simulations, nous n'appliquons pas ce processus aux données de méthylome qui ne sont utilisées qu'à l'étape de l'obtention de A.

Concrètement, le principe de la simulation est d'utiliser un profil cellulaire différent pour le type cancer de chaque échantillon. Nos données PDX contiennent les profils cellulaires de seulement 15 lignées, il peut donc être nécessaire d'en simuler des supplémentaires pour obtenir un nombre d'échantillons supérieur.

Dans ce cas, on utilise la fonction R normalmixEM [57] pour inférer pour chaque gène une distribution bimodale à partir de son expression dans les 15 lignées. On tire ensuite l'expression des profils manquants dans cette distribu-

tion. Si la distribution de l'expression du gène dans les 15 lignées ne suit pas une loi bimodale, ce qui est par exemple le cas quand on a une valeur extrême ou beaucoup de 0, alors on tire l'expression dans la tumeur simulée parmi celles existantes.

On obtient alors la formule suivante :

$$D_i = T_{cancer^i} * A_{i,cancer} + \sum_{j=1}^{k_{ME}} T_j * A_{i,j} \tag{7.2}$$

Avec k_{ME} le nombre total de types cellulaires présents dans le micro-environnement, à l'exception de la composante du type cellulaire cancer T_{cancer}, et T_{cancer^i} la composante cancer du patient i.

Lien entre l'expression et le micro-environnement

L'enjeu du pipeline est de permettre de retrouver des gènes dont la dérégulation est reliée à la composition du micro-environnement : il faut donc simuler ce phénomène. Deux types de modèles ont été implémentés.

Simulation par seuil (t). La simulation par seuil consiste à augmenter l'expression du gène dans les cellules cancéreuses si la proportion d'un type cellulaire du micro-environnement donné j_0 (par la suite le type IC) dépasse un seuil t_ME fixé par l'utilisateur. Cela équivaut à modéliser un gène dont la sur-expression serait reliée à une plus forte représentation de ce type cellulaire dans le micro-environnement, avec un effet on/off sur l'expression du gène déclenché à partir d'une certaine proportion.

Concrètement, on choisit un set de gènes qui sera associé au type j_0. Pour un gène g de ce set, dans chaque patient i, on regarde :

$$\text{Si } A_{i,j_0} > t\_ME \Rightarrow T^n_{cancer^i,g} = T^o_{cancer^i,g} * f_t$$

, avec A_{i,j_0} la proportion du type cellulaire j_0 dans la tumeur i, $T^{n,o}_{cancer^i,g}$ l'expression de g dans la composante cancer du patient i après (n) ou avant (o) correction et f_t un facteur de dérégulation fixé par l'utilisateur.

Simulation par facteur (f). La simulation par facteur modélise la dérégulation de l'expression des gènes liés au micro-environnement de manière linéaire en fonction de la proportion d'un type cellulaire j_0 : plus le type cellulaire est présent, plus le gène est sur-exprimé.

Ainsi, comme précédemment, pour un gène g fixé, dans chaque patient i, on a :

$$T^n_{cancer^i,g} = T^o_{cancer^i,g} * (1 + f_f * A_{i,j_0})$$

, avec f_f un autre facteur de dérégulation fixé par l'utilisateur.

7.2.2 Variation des paramètres

Les simulations sont composées de différents types de paramètres : des paramètres génériques comme le nombre d'échantillons p et la présence ou non de contrôles ; des paramètres spécifiques à la matrice A simulée comme la variabilité entre les patients $\alpha 0$ ou la proportion moyenne du type immunitaire α_{IC} ; des paramètres associés à la dérégulation des gènes du type cellulaire cancer (T_{cancer}) en fonction de la proportion immunitaire ; et des paramètres liés à D comme le bruit gaussien rajouté à la fin. Les paramètres variés dans cette partie sont résumés dans la figure 7.11.

Paramètres génériques (p, $ctrl$)

Les paramètres génériques concernent toute la simulation. Le nombre de tumeurs p est fixé à 60 par défaut, on teste également 30 et 120. D'après notre expérience sur la déconvolution (voir partie 4.2.2), on s'attend à ce qu'un plus grand nombre d'échantillons améliore les résultats.

Par défaut, on ajoute 15 échantillons contrôles, ils sont générés avec des proportions moyennes de 45% d'épithélial sain, 45% de fibroblastes et 10% de cellules immunitaires. La présence d'échantillons contrôles devrait augmenter l'hétérogénéité entre les échantillons et donc également faciliter la déconvolution. En outre, c'est une bonne référence pour l'étape d'assignation des types cellulaires, et pour vérifier la pertinence des types déconvolués.

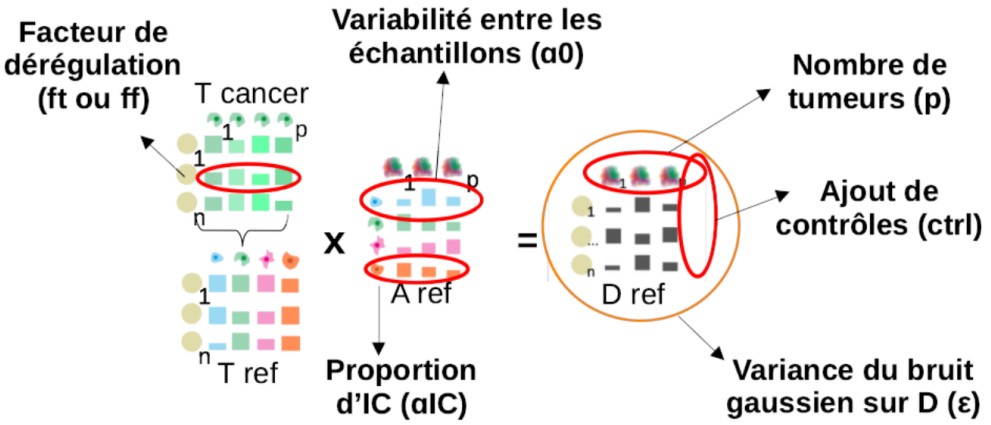


FIGURE 7.11 – **Représentation schématique des différents paramètres variants lors des simulations.** La matrice T_{cancer} représente les différents profils "cancer" de chaque échantillon simulé. La matrice T_{ref} comprend tous les profils cellulaires : les types épithélial, immunitaire et fibroblaste qui sont les mêmes pour tous les échantillons, et le type cancéreux (T_{cancer}) qui varie. La matrice A_{ref} représente les proportions des différents types cellulaires entre les échantillons. La multiplication de T_{ref} par A_{ref} donne la matrice D_{ref} qui représente l'expression de l'échantillon complexe sur lequel on peut rajouter un bruit gaussien.

Paramètres sur A ($\alpha 0$, α_{IC})

La matrice des proportions A joue un rôle important dans ces simulations puisque les gènes qu'on cherche à retrouver à l'issue du pipeline sont dérégulés en fonction de la proportion en type immunitaire.

Le paramètre $\alpha 0$ est fixé par défaut à 10, ce qui correspond à une distribution de Dirichlet plutôt stable entre les échantillons. On teste également $\alpha 0 = 1$, ce qui correspond à une forte variabilité, et $\alpha 0 = 100$ qui implique une grande stabilité des proportions en types cellulaires. D'après nos analyses sur Medepir, une matrice A plus variable est plus simple à déconvoluer (voir partie 4.2.2).

Toujours lors de nos tests sur la déconvolution, nous avons constaté que les valeurs α des différents types cellulaires n'était pas un paramètre important. En revanche, ici la proportion de type immunitaire α_{IC} est utilisée pour impacter la dérégulation du type cancéreux. Par défaut, nous avons choisi de fixer α_{IC} pour obtenir 10% de type immunitaire (et 15% normal, 45% fibroblastes, 30% cancer) et nous testons également α_{IC} tel que la proportion de cellules immunitaires soit

de 25%, avec 10% de normal, 40% de fibroblastes et 25% de cancer.

Paramètres sur T (f_t, f_f)

Pour les données d'ADNm, il n'y a aucune variation dans la matrice T. Pour celles du transcriptome, les profils des types cellulaires épithélial sain, fibroblastes et cellules immunitaires sont également identiques entre tous les échantillons simulés.

A l'inverse, la matrice T_{cancer} qui sert à simuler un profil de cancer différent pour chaque échantillon du transcriptome est amenée à varier. La base de T_{cancer} est la même pour toutes les simulations avec $p = 60$, mais change pour $p = 30$ et $p = 120$. La dérégulation des gènes change quant à elle à chaque fois, puisqu'elle dépend de la matrice A, mais les gènes choisis pour être dérégulés (100 gènes tirés au hasard) sont identiques pour toutes les simulations.

Pour chaque jeu de paramètres, les deux façons de simuler le lien entre expression du gène et proportion du type immunitaire (seuil et facteur, voir partie 7.2.1) sont appliquées indépendamment. Les facteurs f_t et f_f sont définis par défaut par v, la moyenne sur tous les gènes de l'écart-type de l'expression de chaque gène entre les 60 lignées cancer, c'est-à-dire :

$$f_t = f_f = v \equiv \frac{1}{n_g} \sum_{g=1}^{n_g} \sigma_g \tag{7.3}$$

Avec n_g le nombre de gènes et σ_g l'écart-type du vecteur d'expression $\{T_{cancer^i,g}\}$ du gène g dans les 60 lignées.

Pour nos simulations, $v = 11,46$. Quatre autres valeurs de facteurs sont testées : $v/10$, $v/2$, $v * 2$ et $v * 10$ afin de voir la sensibilité de notre méthode à l'amplitude de dérégulation.

Paramètres sur D (ε)

Par défaut, on ne rajoute pas de bruit sur la matrice D simulée. Cependant, on teste l'ajout de deux amplitudes de bruit gaussien pour simuler le bruit expérimental, d'écart-type ε. Pour la matrice de méthylation, le bruit faible suit

	p	$\alpha 0$	α**IC**	**ctrl**	f	ε_{RNA}	ε_{ADNm}
Référence	60	10	0,1	Oui	v	0	0
p = **30**	**30**	10	0,1	Oui	v	0	0
p = **120**	**120**	10	0,1	Oui	v	0	0
$\alpha 0$ = **1**	60	**1**	0,1	Oui	v	0	0
$\alpha 0$ = **100**	60	**100**	0,1	Oui	v	0	0
α_{IC} +	60	10	**0,25**	Oui	v	0	0
ctrl-	60	10	0,1	**Non**	v	0	0
f–	60	10	0,1	Oui	v**/10**	0	0
f-	60	10	0,1	Oui	v**/2**	0	0
f+	60	10	0,1	Oui	v***2**	0	0
f++	60	10	0,1	Oui	v***10**	0	0
ε+	60	10	0,1	Oui	v	$0,3^2$	300^2
ε-	60	10	0,1	Oui	v	$0,15^2$	150^2

TABLEAU 7.2 – **Résumé des différentes combinaisons de paramètres.** Valeurs des différents paramètres (colonnes) pour les différentes simulations (lignes). p correspond au nombre d'échantillons, α_0 à la variabilité des proportions entre les échantillons, α_{IC} à la proportion de cellules immunitaires, ctrl à la présence ou non de contrôles, f à la valeur du facteur de dérégulation en lien avec le micro-environnement (ici, $v = 11{,}46$) et ε à la variance du bruit gaussien rajouté à la fin de la simulation, respectivement pour les valeurs de transcriptome (ε_{RNA}) et de méthylome (ε_{ADNm}).

une loi N(0, $0,15^2$) et le bruit fort suit une loi N(0, $0,3^2$). Pour l'expression des gènes, ce sont des lois N(0, 150^2) et N(0, 300^2). Les variances choisies pour le bruit fort correspondent environ à la variance observée entre toutes les valeurs de chaque matrice D, pour le bruit faible, c'est la moitié.

Conclusion sur les paramètres

Pour cette analyse, on simule donc 13 conditions expérimentales différentes, résumées dans le tableau 7.2 : 30 ou 120 échantillons au lieu de 60, une matrice A plus ou moins variables, plus de cellules immunitaires, quatre autres amplitudes de dérégulation des gènes en lien avec la proportion de cellules immunitaires, un jeu de données sans contrôles, et deux réalisations de bruit. Pour chacune de ces conditions expérimentales, on teste deux façons de simuler le lien expression dans le cancer - proportion de cellules immunitaires : par seuil (t) ou par facteur (f). On obtient donc 26 matrices D pour le méthylome et 26 matrices D associées pour le transcriptome à analyser.

7.2.3 Résultats

Dans cette partie, je vais détailler les résultats de chacune des étapes du pipeline décrite dans la partie 7.1 et dans la figure 7.10.

Déconvolution et isolement de la part "cancer"

La déconvolution est en deux étapes, la première est d'obtenir la matrice A à partir des données de méthylation. À ce stade, on a seulement 9 conditions expérimentales puisqu'il n'y a pas de variation f pour la méthylation (voir tableau 7.2). On obtient globalement de bons résultats en regardant l'erreur absolue moyenne sur la matrice A (voir la figure 7.12 panneau A), avec des valeurs variant de 0,01 à 0,14. Si on compare les variations de paramètres à la simulation de référence (colonne gauche) on obtient les résultats attendus : l'ajout d'un bruit gaussien sur la matrice D complique la déconvolution de A, plus il y a de patients plus l'erreur est faible, et c'est le même comportement quand l'hétérogénéité entre échantillons des proportions en type cellulaires augmente. La condition

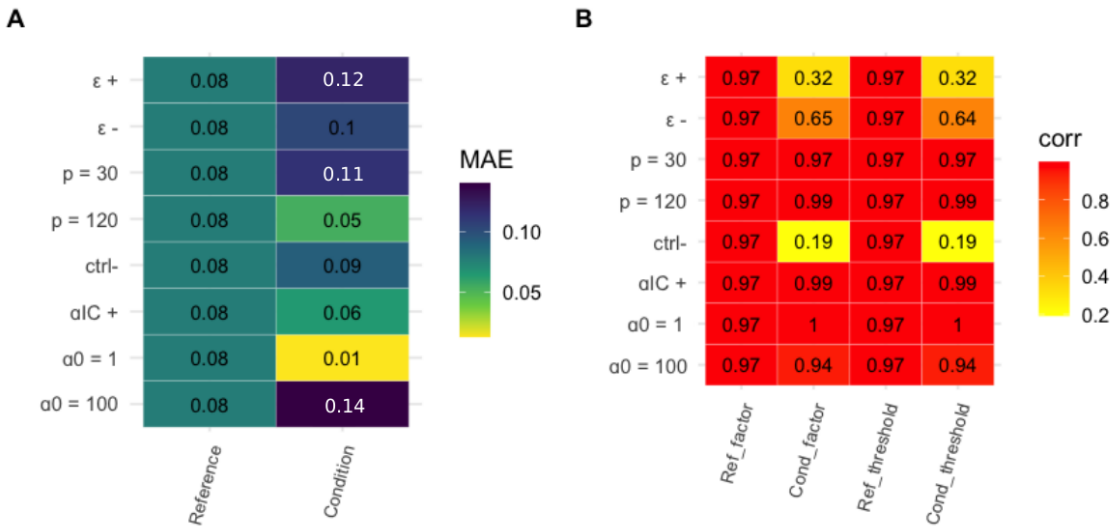


FIGURE 7.12 – **Résultats de la déconvolution en fonction des conditions de simulations.** Les différentes variations de paramètres de simulation sont détaillées dans le tableau 7.2.
A. L'erreur absolue moyenne (MAE) est calculée entre la matrice A déconvoluée et la matrice A utilisée pour les simulations.
B. La corrélation de Pearson moyenne est calculée entre tous les types non-cancéreux déconvolués et ceux utilisées pour les simulation. Les deux types de simulation du lien proportion IC - expression cancer sont étudiées (voir partie 7.2.2) : facteur (deux premières colonnes) et seuil (deux colonnes suivantes).
La simu de référence correspond aux paramètres $p = 60$, $\alpha 0 = 10$, $\alpha_{IC} = 0,1$, présence de contrôles, pas de bruit, et dérégulation égale à v. Les autres paramètres sont visibles dans le tableau 7.2.

αIC + améliore légèrement le score de déconvolution sur la matrice A, mais il est possible que ce soit seulement dû à la réalisation aléatoire d'une matrice A différente.

La seconde étape de déconvolution permet d'inférer T_{RNA} à partir de A et de D_{RNA}. Avant de comparer les résultats de l'inférence, nous devons identifier les différents types cellulaires, en particulier le type cancer. Pour cela, j'utilise un jeu de 455 marqueurs identifiés dans l'équipe par Yasmina Kermezli grâce à un important travail de recherche bibliographique. Ces marqueurs sont ciblés sur différents types cellulaires constituant les tumeurs du pancréas : trois sous-types épithéliaux : les cellules acinaires avec 20 marqueurs, les cellules ductales avec 13 marqueurs et les cellules endocrines avec 32 marqueurs, 8 marqueurs pour les fibroblastes, 320 marqueurs pour les cellules de cancer classiques, et 62 pour les cellules immunitaires. Pour faire correspondre les types cellulaires déconvolués et les marqueurs, on commence par calculer le z-score (ou variable centrée réduite) de chaque marqueur g sur T dans chaque type j, tel que :

$$\text{z-score} = \frac{T_{j,g} - \mu}{\sigma}$$

avec μ la moyenne de l'expression de g dans les types cellulaires et σ son écart-type. À partir des heatmaps, on peut ensuite visuellement retrouver quels marqueurs sont sur-exprimés dans les différents types déconvolués. Dans l'exemple figure 7.13, à gauche, on peut voir le résultat d'une déconvolution qui permet de bien associer chaque composante à une série de biomarqueurs : le type 1 correspond à de l'épithélial, le type 2 à des cellules immunitaires, le type 3 aux fibroblastes et le type 4 au cancer. À droite, on peut voir une matrice T dont la déconvolution était moins réussie à cause du bruit rajouté sur D : l'assignation est la même que pour la matrice de gauche, mais les types cellulaires déconvolués semblent visiblement être moins biologiquement pertinents. Une fois cette identification faite, on peut étudier plus en détail l'efficacité de la déconvolution sur T_{RNA}.

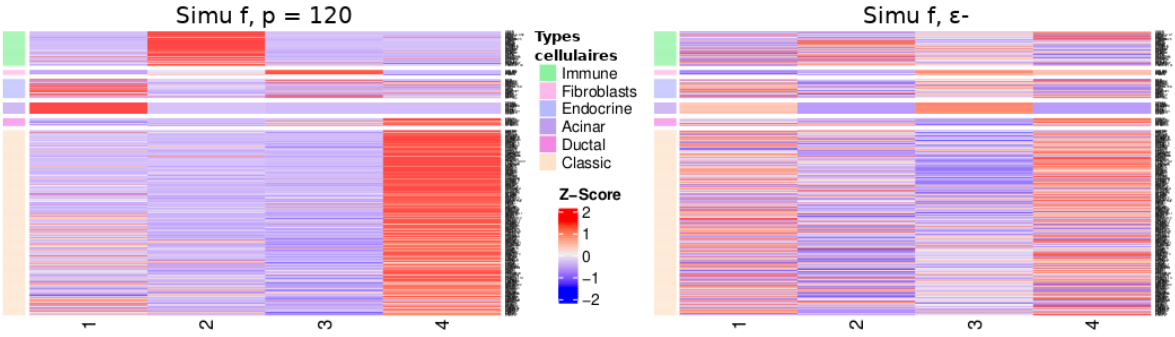


FIGURE 7.13 – **Identification des types cellulaires par z-score.** Pour chacun des 455 marqueurs, le z-score est calculé pour la matrice T déconvoluée. Visuellement, on peut identifier les types cellulaires déconvolués. À gauche : exemple de la simulation par facteur, avec un grand nombre d'échantillons $p = 120$. À droite : exemple de la simulation par facteur avec ajout d'un faible bruit gaussien ε-.

La figure 7.12, panneau B, donne la corrélation de Pearson moyenne entre la matrice T déconvoluée et celle utilisée pour faire les simulations, en réorganisant les colonnes suivant l'identification faite précédemment, et en excluant le type cancer. On regarde de nouveau seulement les 9 conditions de A, pour les deux sortes de simulation : facteur et seuil. Les variations des facteur de dérégulation f_t et f_f ne sont pas étudiées à cette étape car leur effet est négligeable. On retrouve un très bon score ($> 0{,}94$) pour les variations de $\alpha 0$ et α_{IC} et pour celles du nombre d'échantillons. En revanche, le bruit affecte beaucoup la déconvolution de la matrice T comme on pouvait déjà l'observer à l'étape d'identification (7.13, à droite). Un effet aussi fort du bruit sur D est assez surprenant, et nécessitera des investigations supplémentaires dans le futur. Le bruit rajouté sur D_{RNAseq} étant fort, il serait sans doute nécessaire de l'ajuster en fonction de l'expression d'origine du gène. Enfin, le score de corrélation le plus bas (0,19) est atteint pour la déconvolution sans contrôle, ce qui est cette fois assez logique car les contrôles constituent une sorte de référence interne dans l'échantillon et facilitent donc la déconvolution des types du micro-environnement. Par ailleurs, on peut noter que les scores de corrélation obtenus sont stables entre la simulation de dérégulation liée au micro-environnement par seuil et celle par facteur.

Une fois l'identification faite, nous conservons la part d'expression "cancer" de chaque échantillon en retirant l'expression des autres types cellulaires identifiés comme appartenant au micro-environnement, comme décrit partie 7.1.2. A cette étape, on se retrouve avec des valeurs d'expression négatives pour certains gènes, ce point sera abordé dans la discussion.

Analyse différentielle

À partir de la matrice D_{cancer} obtenue en isolant l'expression des cellules cancéreuses de chaque échantillon, on cherche maintenant à identifier les gènes dérégulés de manière personnalisée en appliquant la méthode Penda. On choisit d'utiliser comme contrôles 200 lignées GTEX [116] d'épithélial sain qui n'ont pas servi aux simulations, et comme seuil Penda 0,8 car, d'après notre expérience (voir la partie 2.3), il faut un seuil élevé dans le cas de l'utilisation de contrôles dont le profil d'expression est éloigné des échantillons analysés. Les paramètres de la méthode du centile sont fixés par simulations à c = 0,07 et f = 1. Les résultats de la méthode Penda se présentent sous la forme d'une information oui/non de sur-expression ou de sous-expression, sans valeur numérique d'amplitude de dérégulation associée.

Lien entre la dérégulation et la proportion du micro-environnement

Une fois l'analyse différentielle effectuée, on cherche à identifier les gènes dont la dérégulation est reliée à la proportion des types cellulaires du micro-environnement. Dans un premier temps, on trie les données pour conserver uniquement les gènes qui sont respectivement dérégulés et non-dérégulés dans au moins 10 échantillons pour chaque statut. On choisit ce seuil de 10 au lieu de 100 dans le chapitre précédent pour éviter de retirer trop de gènes d'intérêt. Le but est d'avoir des distributions permettant de calculer correctement les métriques : le nombre de gènes conservés est récapitulé dans le tableau 7.3. On peut constater que le nombre de gènes dérégulés conservé après cette étape de tri varie beaucoup, avec un minimum de 9 pour $\varepsilon+$ et un maximum de 83 pour $p = 120$. Dans les simulations avec l'ajout d'un bruit, énormément de gènes sont détectés comme étant dérégulés (92% pour la simulation par facteur $\varepsilon+$,

	ref	p **= 30**	p **= 120**	$\alpha 0$ **= 1**	$\alpha 0$ **= 100**	α**IC+**	**ctrl-**
Genes f	16 009	7 072	18 329	14 196	18 831	15 911	15 569
dont dereg	78	39	70	83	81	76	83
Genes t	16 010	7 103	18 328	18 195	14 782	15 895	15 569
dont dereg	79	50	82	82	22	57	78

	$f--$	$f-$	$f+$	$f++$	ε**+**	ε**-**
Genes f	16 007	16 007	16 012	15 998	2 392	5 395
dont dereg	69	74	80	66	10	28
Genes t	16 006	16 014	16 017	15 963	2 505	5 063
dont dereg	68	76	77	26	9	19

TABLEAU 7.3 – **Nombre de gènes conservés pour chaque condition après le tri initial.** Lors de l'analyse Penda, les gènes dont l'expression est systématiquement égale à 0 sont retirés. Lors du tri initial de cette étape du pipeline, ce sont les gènes qui ont un effectif inférieur à 10 dans l'une des deux conditions de dérégulation (dérégulé ou non-dérégulé) qui sont retirés. Pour chaque condition de simulation (voir 7.2), le nombre de gènes conservés est indiqué dans ce tableau. Les lignes "dont dereg" correspondent au nombre de gènes conservés parmi les 100 gènes dérégulés en lien avec la proportion d'IC lors de la simulation.

contre 31% pour la simulation par facteur de référence), ce qui fausse cette étape. Pour $p = 30$, on retrouve moins de gènes car il est plus rare d'avoir 10 échantillons pour chaque condition que pour $p = 60$ ou $p = 120$. Il est à noter que ce tri sur l'effectif à un sens pour les métriques de distribution, basées sur une séparation binaire des données, mais qu'il est toujours possible d'utiliser des métriques entre l'expression et les proportions du micro-environnement, comme la corrélation directe, sur ces gènes retirés ici.

A partir de ces données, on peut calculer les différentes métriques pour les groupes Penda (distance de Kantorovitch, p-valeur du test de Student et distance de Kolmogorov-Smirnov), ainsi que la corrélation directe avec l'expression des gènes, puis obtenir des courbes ROC de détection en faisant varier les seuils de détection. Dans cette analyse, nous n'utilisons pas l'analyse discriminante avec le test de Wilks décrit en introduction car nos simulations sont basées sur le lien direct entre un type cellulaire et l'expression d'un gène, et pas sur un effet global de la structure tumorale.

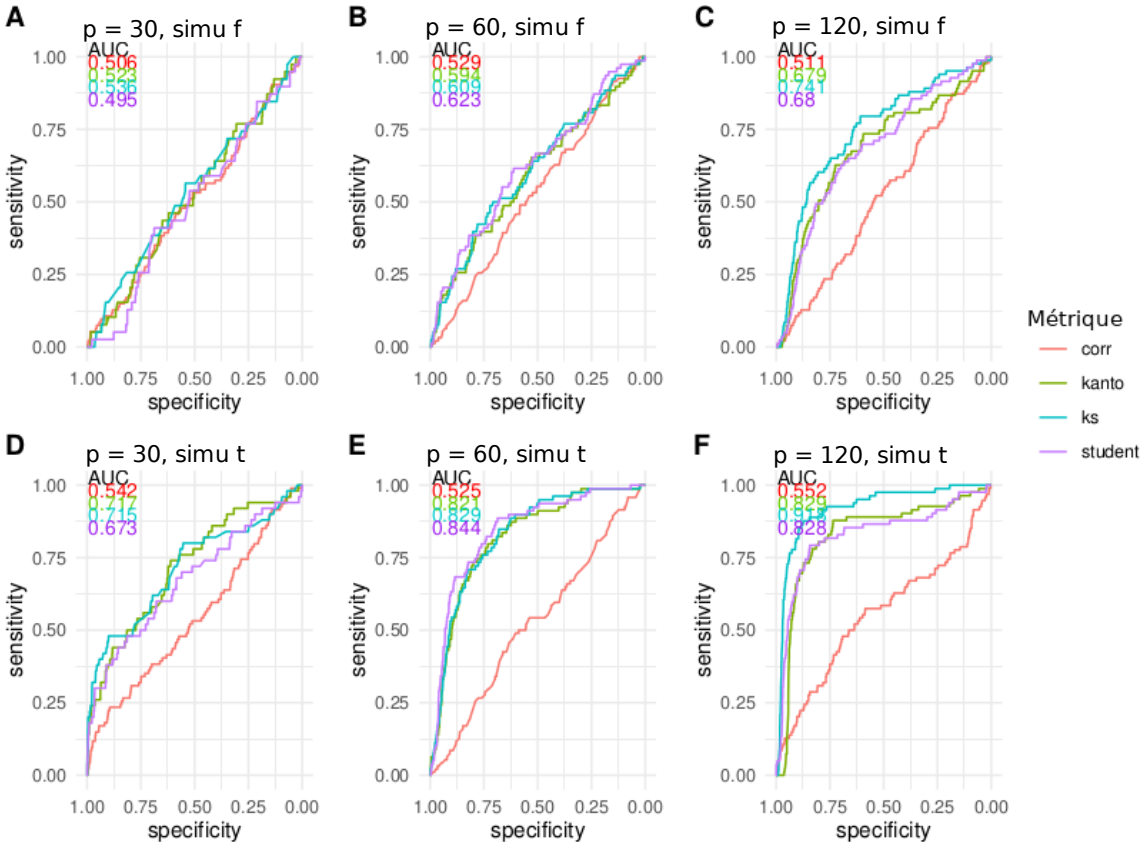


FIGURE 7.14 – **Résultat de la détection des gènes dérégulés en lien avec la proportion des cellules immunitaires pour différentes tailles d'échantillons.** Pour les trois simulations en fonction du nombre d'échantillons, $p = 30$, $p = 60$ et $p = 120$, on fait varier les seuils des différentes métriques. A, B et C : simu par facteur, D, E, et F : simu par seuil. Corr = corrélation de Pearson avec l'expression, Kanto = distance de Kantorovitch entre les groupes Penda, ks = distance de Kolmogorov-Smirnov entre les groupes Penda, Student = p-valeur du test de Student entre les groupes Penda.

Par exemple, dans la figure 7.14 on peut voir les courbes obtenues pour les variations du nombre d'échantillons p. Globalement, la simulation par facteur (en haut) donne des résultats beaucoup moins bons que la simulation par seuil (en bas). Cette différence est assez logique, car dans la simulation par facteur on multiplie l'expression du gène par v*proportion(IC), alors que dans la simulation par seuil on multiplie l'expression du gène directement par v, la dérégulation est donc plus forte. La courbe rouge correspond à la corrélation entre l'expression de T_{cancer} et la proportion de cellules immunitaire, alors que les 3 autres courbes sont calculées à partir de la matrice binaire Penda. Pour toutes les conditions, les résultats à partir de Penda sont meilleurs que ceux sur les données d'expression brute, même si la différence est moins visible pour la simulation par facteur et $p = 30$ car aucune méthode n'offre une bonne détection. Par ailleurs, les trois métriques calculées sur Penda (p-valeur de Student, distance de Kantorovitch et distance de Kolmogorov-Smirnov) semblent équivalentes. Enfin, les résultats s'améliorent lorsque le nombre d'échantillons augmente, ce qui est probablement dû à la meilleure déconvolution (voir la partie précédente 7.2.3).

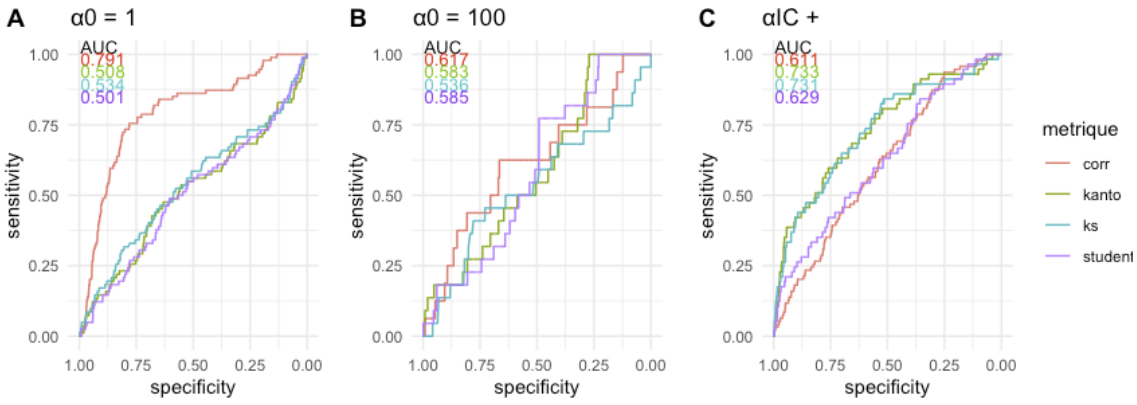


FIGURE 7.15 – **Résultat de la détection des gènes dérégulés en lien avec la proportion des cellules immunitaires pour les variations sur A.** La dérégulation a été simulée par seuil. A : matrice A plus variable. B : matrice A moins variable. C : plus de cellules immunitaires. Corr = corrélation de Pearson avec l'expression, Kanto = distance de Kantorovitch entre les groupes Penda, ks = distance de Kolmogorov-Smirnov entre les groupes Penda, Student = p-valeur du test de Student entre les groupes Penda.

Les résultats des simulations par facteur pour les paramètres sur A donnent le même type de courbes diagonales que $p = 60$, on se concentre donc sur les simulations par seuil qui présentent plus de variabilité. Les résultats sur les paramètres de la matrice A sont surprenants (voir figure 7.15). Comme attendu, une matrice A moins variable complique la déconvolution et donc la détection, mais une matrice A très variable avantage nettement la corrélation directe et fait perdre beaucoup d'efficacité aux résultats sur Penda. En regardant plus en détail la proportion de cellules immunitaires dans chacune des conditions $\alpha 0$ (figure 7.16), on s'aperçoit que pour $\alpha 0 = 1$ seulement 12 échantillons dépassent le seuil de 0,1 qui marque l'activation de la dérégulation des différents gènes. La réalisation aléatoire de cette matrice A a donc été très défavorable au type immunitaire, le gène est ainsi très rarement dérégulé.

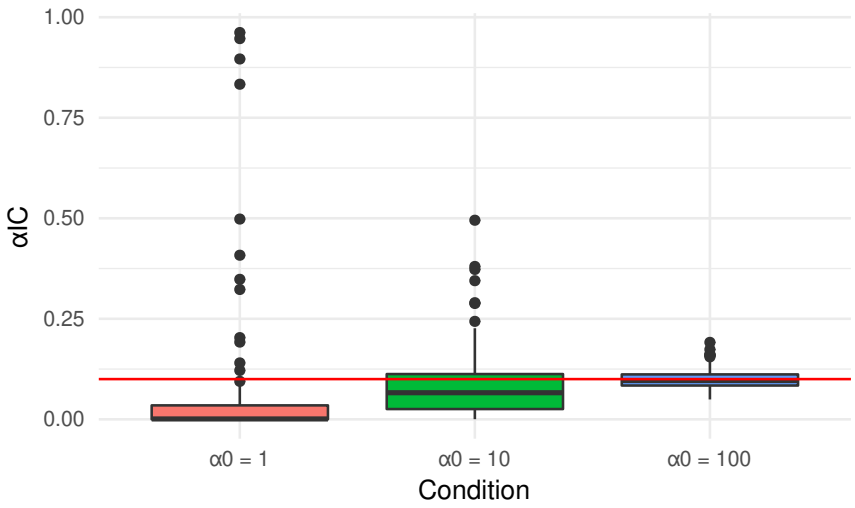


FIGURE 7.16 – **Répartition des valeurs de α_{IC} pour les trois valeurs de $\alpha 0$.** Répartition de la proportion de cellules immunitaires dans les 60 échantillons de chacune des condition $\alpha 0$. La ligne rouge marque la limite $\alpha_{IC} = 0{,}1$.

La méthode Penda perd également en efficacité quand la proportion de cellules immunitaire augmente (figure 7.15, panneau C), cela est dû au fait que si la proportion dépasse quasiment systématiquement le seuil, le gène se retrouve toujours dérégulé de la même façon, et le lien entre proportion et expression est perdu.

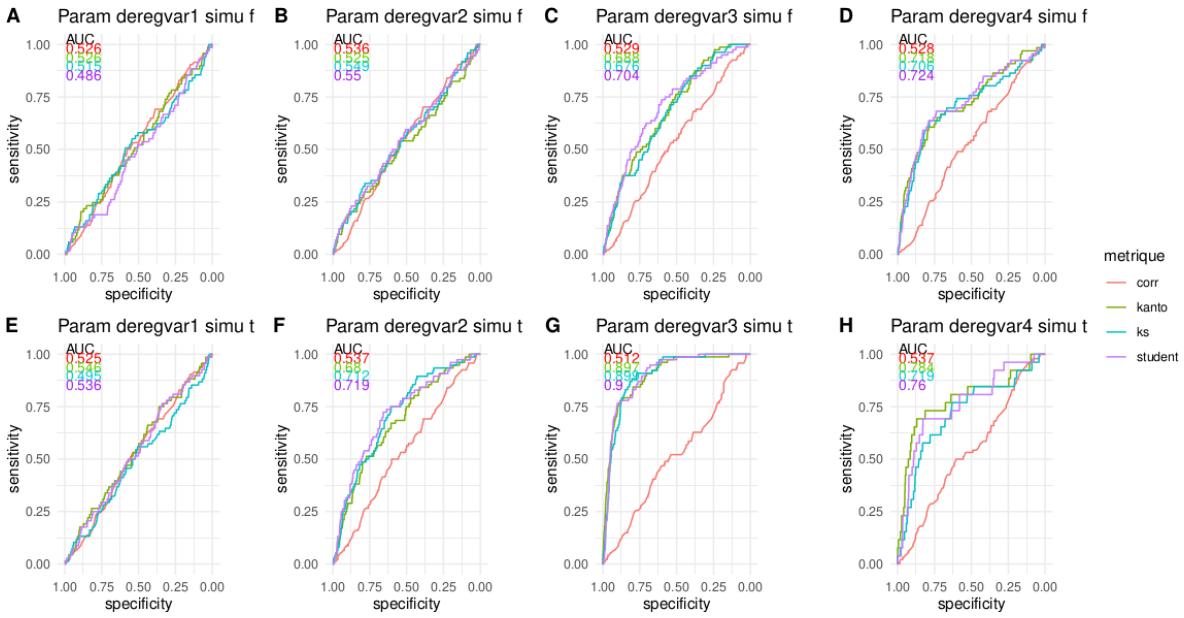


FIGURE 7.17 – **Résultat de la détection des gènes dérégulés en lien avec la proportion des cellules immunitaires pour différents facteurs.** Des simulations ont été effectuées avec différents facteurs de dérégulation. A, B, C, et D : simulation par facteur. E, F, G, et H : simulation par seuil. deregvar1 correspond au facteur classique divisé par 10, deregvar2 divisé par 2, deregvar3 multiplié par 2, et deregvar4 multiplié par 10. Corr = corrélation de Pearson avec l'expression, Kanto = distance de Kantorovitch entre les groupes Penda, ks = distance de Kolmogorov-Smirnov entre les groupes Penda, Student = p-valeur du test de Student entre les groupes Penda.

Enfin, le dernier paramètre important est le facteur de dérégulation. Pour rappel, dans la simulation par facteur f, on multiplie l'expression du gène dérégulé par $1 + f_f * A_{IC,i}$. Dans la simulation par seuil t, on multiplie l'expression du gène dérégulé par f_t si la proportion d'IC dépasse le seuil. La valeur v des facteurs est fixée à 11,46, et on teste quatre autres valeurs : $v/10$ (deregvar1), $v/2$ (deregvar2), $v * 2$ (deregvar3) et $v * 10$ (deregvar4). Les résultats sont visibles sur la figure 7.17. On observe un décalage entre les deux types de simulations, la courbe obtenue pour $f_f = v * 10$ ressemble à celle de $f_t = v/2$, ce décalage est assez logique car f_f est multiplié par la proportion en cellules immunitaires qui est en moyenne 0,1. Pour la simulation par facteur, plus la valeur du facteur augmente, plus les gènes sont détectés via la méthode Penda, sans que les valeurs testées ne permettent de dépasser une aire sous la courbe de 0,72. Pour la simulation par seuil, le même comportement est observé jusqu'à $f_t = v * 2$, en revanche les résultats diminuent pour $f_t = v * 10$, mais seulement 26 gènes ont été conservés contre 66-80 pour les autres simulations ce qui peut causer ce décalage. Globalement, plus la dérégulation est forte plus le gène est facilement retrouvé, mais il faudrait tester d'autres valeurs pour f_t et f_f afin d'affiner le résultat.

7.2.4 Conclusion

Pour conclure, l'effet de tous les paramètres n'est pas encore finement étudié, mais on observe déjà des tendances semblables aux résultats obtenus sur la déconvolution. La plupart des simulations permettent d'obtenir des résultats de détection satisfaisants, avec une aire sous la courbe ROC de plus de 0,8 pour la simulation par seuil pour les paramètres de référence. Globalement, l'utilisation de Penda et des métriques de distribution permet d'améliorer grandement les résultats par apport à la corrélation sur l'expression brute, même si les écarts entre les différentes métriques méritent une exploration supplémentaire. Enfin, il semble qu'il faille un effet de dérégulation assez fort pour que la détection soit efficace.

Ces résultats sont encore exploratoires, les simulations méritent d'être affinées puis appliquées sur des données réelles, mais ces premières analyses sont très encourageantes pour la suite du travail.

Discussion et conclusion

Dans cette troisième partie, nous avons exploré une partie du lien entre dérégulation des gènes et composition du micro-environnement. Nos analyses préliminaires nous ont conduits à la construction d'un pipeline d'analyse regroupant différentes étapes : déconvolution d'échantillons mélangés, identification et extraction de la composante cancéreuse purifiée, analyse différentielle sur ses données d'expression puis inférence des gènes dérégulés en lien avec la composition tumorale. L'application du pipeline sur des données d'essai simulées est convaincante, et ouvre la voie à une application sur des données biologiques afin de réussir à mieux caractériser le lien entre dérégulation et micro-environnement tumoral dans les tumeurs.

Limites

Choix des méthodes utilisées

Déconvolution

Pour l'étape de déconvolution, nous avons choisi d'utiliser EDec pour inférer les proportions à partir des données de méthylation, puis les profils des types cellulaires à partir des données transcriptomiques.

Le choix de cette méthode présente cependant plusieurs contraintes. Premièrement, c'est une méthode non-supervisée, et les profils déconvolués peuvent donc être compliqués à interpréter. De plus, il est nécessaire d'avoir en entrée des données multi-omiques car il faut le transcriptome et le méthylome de chaque patient.

Ce choix a été fait malgré tout car, comme on l'a déjà vu, l'approche multi-omiques et non supervisée a également des avantages : en particulier la réduction du bruit de chacune des méthodes de séquençage, et l'application à des échantillons variés sans connaissance à priori des composantes cellulaires. Par ailleurs, comme on a pu le voir au moment de l'exploration des méthodes du pipeline (partie 7.1), tester trop de méthodes à chaque étape complique l'analyse de la qualité des résultats et la compréhension des erreurs. La construction du pipeline Ritmic permettant de garder chaque bloc indépendant, l'utilisation d'autres méthodes de déconvolution peut cependant se faire très simplement.

Pour l'application sur des données réelles, il pourrait par exemple être judicieux d'appliquer une méthode semi-supervisée sur un échantillon bien caractérisé afin de tester la méthode sans le biais de la déconvolution. Une autre méthode publiée récemment, Rodeo, cherche spécifiquement à inférer T à partir de A [152] et mériterait d'être testée en remplacement de EDec step 2.

Extraction du type "cancer"

Dans notre méthode, nous partons du principe que toute l'hétérogénéité d'expression est due aux cellules cancéreuses et que les cellules du micro-environnement ont un profil stable : c'est un à priori très fort. Cette hypothèse simplificatrice reste néanmoins nécessaire pour rendre l'extraction du profil "cancer" possible sans informations supplémentaires sur les échantillons bulk, même si elle peut biaiser l'interprétation des résultats.

Un autre point délicat de cette étape est l'apparition de valeurs négatives quand l'expression des différents types cellulaires du micro-environnement est extraite de D pour obtenir T_{tum}. En effet, il arrive que le produit $T * A$ déconvolué soit supérieur au D d'origine. Il pourrait être intéressant de rajouter une contrainte à ce sujet dans l'algorithme de déconvolution, ou d'instaurer une normalisation au moment de l'extraction de T_{tum} pour éviter d'obtenir des expressions de gène négatives.

Métriques de distance

Dans la dernière étape du pipeline Ritmic, plusieurs méthodes sont appliquées pour mesurer la distance entre les proportions des deux groupes d'expression. Sur les simulations, les trois métriques utilisées (p-valeur de Student, distance de Kantorovitch et distance de Kolmogorov-Smirnov) donnent des résultats globalement similaires, même si elles évaluent des paramètres différents des distributions.

Lors de cette étape, on peut également appliquer une approche plus globale grâce à l'analyse discriminante décrite dans la partie exploratoire. Sur nos simulations, la structure tumorale globale n'est pas modifiée car seule la composante immunitaire est affectée donc l'application de cette méthode n'a pas d'intérêt. Cependant, il serait intéressant de faire des tests sur des données réelles pour évaluer sa pertinence.

D'autres métriques peuvent facilement être implémentées au sein du pipeline Ritmic. La principale contrainte sur ces tests reste la nécessité de comparer des populations de taille suffisante pour avoir une bonne quantification de la statistique et donc d'avoir un nombre d'échantillons suffisamment important.

Simulations

Les simulations utilisées dans cette partie présentent plusieurs limitations.

Dérégulation

Dans nos analyses, la dérégulation des gènes en lien avec le micro-environnement est implémentée de manière un peu naïve, avec un facteur de multiplication fixé arbitrairement. De plus, elle ne s'effectue que dans une direction : la sur-expression. Dans la réalité biologique l'autre cas de figure est tout à fait envisageable : par exemple, la sur-représentation de cellules immunitaires pourrait inhiber l'expression de gènes dans les cellules cancéreuses. Nos simulations permettent à priori de générer une sous-expression : pour la simulation par facteur il faut fixer f_f tel que $f_f \in]-10;0[$, pour la simulation par seuil fixer f_t tel que $f_f \in]0;1[$, cependant cette possibilité reste limitée, par exemple elle ne

permet pas encore de sous-exprimer un gène sous une certaine proportion d'un type cellulaire. Il pourrait notamment être intéressant de faire un jeu de données combinant plusieurs types de dérégulation sur des gènes différents, avec à la fois de la sur- et de la sous-expression, implémentées par facteur ou par seuil.

Aléatoires des simulations

Dans la partie d'application du pipeline, nous avons testé tous les paramètres sur une seule réalisation des matrices A, alors que nous avons montré dans la partie 2 que l'aléatoire de sa génération pouvait fortement influencer les résultats. Il est prévu de relancer l'analyse sur d'autres jeux de données simulées afin de tester la robustesse du pipeline.

Analyse différentielle

L'analyse différentielle avec la méthode Penda nécessite un jeu de référence. Or, ici, le pipeline Ritmic fait qu'il est impossible d'obtenir des contrôles directement à partir des données déconvoluées de tissus sains. En effet, on purifie la tumeur en partant du principe que toute la variabilité d'expression entre deux échantillons peut avoir deux causes : soit les proportions du microenvironnement, soit la dérégulation dans les cellules tumorales en elles-mêmes. On n'a donc pas accès à la variabilité de la composante "contrôle" entre les échantillons qui nous permettrait de construire avec fiabilité les listes L et H de PenDA.

Notre choix a donc été d'utiliser des lignées épithéliales publiques, dont dérivent les cellules tumorales qu'on étudie, et de fixer le paramètre de seuil du test assez haut pour ne pas être trop sensible aux différences entre les échantillons. Cette stratégie semble fonctionner d'après les premiers résultats, mais peut-être que des simulations Penda plus fines permettraient de définir des paramètres plus optimaux.

Dans tous les cas, obtenir une information binaire et individuelle grâce à Penda (dérégulé / non dérégulé) plutôt qu'une information continue d'expression permet d'utiliser des métriques plus spécifiques que la corrélation, et dans nos tests semble permettre de détecter plus efficacement les gènes d'intérêt.

Perspectives

Cette partie étant encore au stade de développement, les perspectives sont nombreuses. Dans le cadre des simulations, nous pourrions tester d'autres paramètres notamment de plus grands facteurs pour la simulation du lien environnement-dérégulation. Il faudrait également rapidement tester la robustesse de nos premiers résultats en variant les matrices A utilisées pour les simulations.

L'étape actuelle du développement de la méthode est l'application à des données réelles, qui est en cours sur des jeux du cancer du poumon. Sur les données réelles, on se confronte pour l'instant à des difficultés pour interpréter les matrices T déconvoluées. Plusieurs stratégies sont en cours de développement, comme utiliser une méthode de déconvolution semi-supervisée pour affiner les profils T obtenus, ou rechercher des marqueurs de types cellulaires plus robustes. On pourrait également renforcer l'identification des types T_{RNA} en intégrant l'information de A et de T_{ADNm}.

La priorité actuelle est donc d'améliorer les méthodes d'interprétation de la matrice T ainsi que l'étape d'extraction de la partie tumorale.

Enfin, dans le futur, nous allons comparer Ritmic aux autres méthodes existantes proposant de détecter un lien entre expression des gènes et micro-environnement, comme les méthodes décrites dans l'introduction ou Demix [153] qui a été développée pour les données de micro-puces à ADN en utilisant comme référence des gènes exprimés dans les contrôles et pas dans les tumeurs. A terme, Ritmic devrait permettre de trouver de nouveaux biomarqueurs reliés à la pureté tumorale et à la composition du micro-environnement. Ces biomarqueurs seront peut-être également des indicateurs de survie ou de réponse aux traitements.

Conclusion

Bien qu'étant toujours en cours de développement, le projet du pipeline Ritmic offre déjà des résultats intéressants. Un article de recherche est en cours de rédaction à ce sujet, le projet devrait se prolonger sur mes derniers mois de thèse et être conclu peu après.

Conclusion générale et perspectives

Conclusion

L'objectif de ma thèse était de développer de nouveaux outils méthodologiques permettant d'améliorer la compréhension et la caractérisation de l'hétérogénéité tumorale. L'hétérogénéité dans le cancer est présente à de multiples niveaux, nous l'avons traitée à travers deux aspects. Au niveau inter-tumoral, nous avons choisi de nous focaliser sur la caractérisation des différences d'expression entre deux tumeurs indépendantes. Au niveau intra-tumoral, sur la composition en types cellulaires d'un échantillon donné.

Penda est une méthode très prometteuse pour l'analyse différentielle personnalisée. Développée pour permettre d'inférer les gènes dérégulés à l'échelle d'une tumeur, elle a des nombreuses autres applications possibles : n'importe quel type de données quantitatives, à condition d'avoir un jeu de données "contrôle" de taille suffisante et des échantillons "cas" à comparer individuellement. Penda a été implémentée sous la forme d'un package R, puis appliquée à différents jeux de données. Dans le cancer du poumon, nous avons pu isoler des gènes d'intérêts, systématiquement dérégulés dans les tumeurs ou au contraire très conservés. Nous avons également identifié des gènes biomarqueurs dont la dérégulation était associée à la survie. Ces résultats n'auraient pas pu être obtenus avec les méthodes traditionnelles analysant les dérégulations à l'échelle populationnelle ou avec d'autres méthodes "individuelles" induisant un très haut taux de faux positifs.

Plusieurs travaux ont été menés sur la caractérisation de la composition en types cellulaires des tumeurs, qui contiennent à la fois des cellules cancéreuses et des cellules du micro-environnement (cellules immunitaires, tissus sains, stroma, etc.). Le pipeline Medepir permet d'inférer cette composition à partir de données de méthylation de l'ADN. Il intègre des étapes de pré-traitement des données (retrait des sondes corrélées aux facteurs de confusion, réduction du nombre de sondes) et l'application des méthodes existantes de déconvolution sans référence (EDec, RefFreeEwas, MeDecom). Des tests ont été effectués pour la recherche d'une méthode efficace combinant les données RNA-seq et ADNm, et

ont conduit à la construction d'une plateforme de "benchmarking", Deconbench. Ce développement a eu lieu dans un cadre collaboratif grâce à l'organisation de différents "datas challenges" réunissant des chercheurs d'horizons variés, aussi bien du côté mathématique et informatique que du côté biologique et clinique. Si aucune méthode de déconvolution ne se détache clairement aujourd'hui, les pistes ouvertes sont nombreuses et Deconbench devrait faciliter ce travail.

Enfin, une partie plus exploratoire de ma thèse a été consacrée au lien entre les deux premières méthodes. L'objectif de Ritmic est de détecter des gènes dérégulés dans les cellules cancéreuses purifiées *in silico*, et d'associer cette dérégulation à la composition du micro-environnement tumoral. Si une première version du pipeline est déjà construite et appliquée à des simulations, des tests sont encore nécessaire pour appliquer la méthode à des données cliniques et pouvoir ainsi inférer de nouveaux biomarqueurs. Ce travail sera réalisé dans les prochains mois.

Toutes ces contributions répondent bien aux problématiques initiales : développer des méthodes adaptées aux enjeux de l'analyse individuelle de l'hétérogénéité tumorale.

Discussion et perspectives

Une discussion sur les méthodes développées et leurs perspectives se trouve déjà au sein de chaque partie des résultats, cette section sera donc plutôt axée sur les discussions et les perspectives à l'échelle globale de ma thèse.

Choix des données

Dans ma thèse, nous nous sommes intéressés à deux types de données, méthylome et transcriptome, car ces deux types de données sont largement disponibles et déjà bien caractérisés, nous permettant ainsi de nous concentrer sur le développement des méthodes en elles-mêmes. Cependant, d'autres types de données existent et pourraient nous apporter des informations supplémentaires, comme

le protéome ou les variations du nombre de copies des gènes au niveau du génome (cnv). En revanche, comme on l'a déjà abordé dans le chapitre de la déconvolution, l'intégration multi-omique qui permettrait de réunir toutes ces informations et de réduire le bruit de chacune des méthodes reste encore un défi aujourd'hui. L'intégration multi-omique peut se faire de deux principales façons [154]. La première est de combiner les résultats obtenus sur différents types d'omiques, dans ce cas nos méthodes sont compatibles. La deuxième est de vraiment regrouper les données et de réduire les dimensions de celles-ci [125], dans ce cas nos méthodes pourraient éventuellement apporter une information supplémentaire via leurs matrices de résultats (proportions des types cellulaires pour la déconvolution, dérégulation binaire pour Penda) qui seraient alors traitées comme un autre type de données à intégrer. Une perspective intéressante sera donc de travailler sur ces deux axes en intégrant plus de données.

D'autres choix restrictifs ont été faits, comme celui d'utiliser principalement des méthodes sans références, ce qui complique parfois l'interprétation des résultats mais permet d'appliquer la déconvolution à des tissus solides broyés sans à priori des types cellulaires le composant [155]. Par ailleurs, nous n'avons pas utilisé de données en cellules uniques ou "single cell" car cette technologie n'est encore que peu utilisée à l'échelle clinique. Cependant, il est certain que dans un futur proche, ce type de donnée se démocratisera [156]. Cela permettra de caractériser plus simplement le micro-environnement tumoral, et l'utilisation de Penda et de Ritmic à l'échelle de la cellule unique pourrait permettre une caractérisation encore plus précise de l'hétérogénéité. Concernant Medepir, la déconvolution ne sera plus utile si le single-cell devient la routine clinique. En revanche, des profils single-cell précis pourraient servir de référence pour les types cellulaires le temps de la transition. Des travaux sont en cours en ce sens dans l'équipe, mais actuellement assigner avec certitude l'identité des cellules séquencées n'est pas trivial et cette solution est donc difficilement applicable.

Ainsi, dans l'avenir, il serait intéressant de repenser les méthodes pour intégrer différentes technologies. Pour Penda, l'application à n'importe quel type de données quantitatives devrait être assez simple car la conception de la méthode,

basée sur les rangs, n'est pas spécifique au RNA-seq. Pour la déconvolution, des recherches sont encore nécessaires. Une vraie intégration multi-omique demandera forcément un travail important, mais une piste pourrait par exemple être de chercher un lien entre les données génomiques (variations du nombre de copies des gènes par exemple) et les gènes détectés sur-exprimés par Penda dans le transcriptome.

Différents types d'hétérogénéité

L'hétérogénéité dans le cancer est un sujet très vaste, à de nombreux niveaux. Avec Penda, nous avons regardé l'échelle entre les tumeurs, sans chercher à caractériser l'origine des variations d'expression, qui pourraient entre autres provenir de la composition cellulaire de la tumeur ou des mutations génétiques. Dans la partie Ritmic, on cherche justement à améliorer cette analyse différentielle en prenant en compte la composition de la tumeur et la présence du micro-environnement.

En analysant des données "bulk", il faut garder à l'esprit que l'échantillon représente un petit morceau de la tumeur, qui peut également présenter une forte hétérogénéité spatiale, et des clones de cellules cancéreuses très différents dans un autre morceau. Ces cellules cancéreuses peuvent même se différencier suffisamment pour devenir des types cellulaires différents, comme les types "classique" et "basal" du cancer du pancréas qui co-existent dans certaines tumeurs. Penda pourrait alors être appliquée sur des échantillons de "spatial single cell", où l'on récolte les profils en cellules uniques à différentes positions de la tumeur. Ce type d'analyse pourrait permettre de mieux comprendre la dérégulation liée à l'hétérogénéité spatiale.

Par ailleurs, il faut également noter que nos échantillons "contrôles" sont en réalité des tissus adjacents aux tumeurs, qui ont potentiellement subi des traitements lourds et qui peuvent être impactés par la proximité des cellules cancéreuses et présenter une hétérogénéité propre. Enfin, il y a aussi un niveau temporel à l'hétérogénéité tumorale, qui n'est pas du tout traitée ici. Un autre aspect qu'il pourrait être intéressant d'aborder est l'intégration des données cliniques en notre possession. Le stade tumoral par exemple pourrait être un fac-

teur de confusion dans les différentes analyses, même si lors de l'application de Penda aux cancers du poumon nous n'avions pas trouvé de corrélation évidente.

Reproductibilité, maintenabilité et diffusion

Lors du développement d'outils et de méthodes (bio)informatiques dans le cadre d'un projet "à court terme" comme une thèse, des questions sur leur avenir se posent. Dans ma thèse, une grande attention a été portée à la reproductibilité et à la maintenabilité des outils. Les méthodes ont toutes été développées sous la forme de fonctions bien documentées, regroupées dans des packages et disponibles librement via Github. Des exemples y figurent, ainsi que des manuels d'utilisations (nommés "vignettes"), qui dans le cas de Penda génèrent automatiquement une section "matériel et méthodes". Pour la déconvolution, nous travaillons à la reproductibilité à travers la plateforme Deconbench, qui devrait permettre d'intégrer au fur et à mesure toutes les méthodes existantes. Cependant, il est possible que dans l'avenir certaines fonctions ne soient plus utilisables malgré tout, par exemple si elles deviennent obsolètes dans une nouvelle version de R. Dans ce cas, il est difficile de prédire ce que les méthodes deviendront, même si on peut très bien imaginer leurs implémentations dans un autre langage de programmation. Par ailleurs, notre équipe étant petite et sans ingénieur permanent, il est difficile de savoir qui va maintenir les packages si je ne suis plus disponible.

La diffusion de nos méthodes me parait en revanche un challenge plus important. De nouveaux outils sont développés continuellement, et dans le flot quotidien de publications, il est parfois difficile de sortir du lot. Nous avons essayé de valoriser nos outils à travers des diffusions dans différentes conférences internationales (posters et présentations orales), mais je n'ai pas l'impression que ce soit suffisant pour se démarquer sans avoir de collaborations avec de "gros" groupes expérimentaux/cliniques. Cependant, Penda, publiée depuis un peu plus d'un an a actuellement été citée dans un article paru et deux soumis dont nous ne connaissons pas les auteurs. Pour Medepir, les conditions sont un peu différentes puisque le pipeline est directement issu du travail du data chal-

lenge, et implique donc un gros consortium (les participants au data challenge) pouvant aider à sa reconnaissance. La difficulté à faire connaître des méthodes actuellement tient aussi sans doute de cette période où les conférences ne se passent pas en présentiel, ce qui rend plus difficiles les discussions informelles. Malgré tout, l'écart entre le développement d'un nouvel outil et son application réelle, à l'échelle clinique par exemple, me paraît assez complexe à franchir. Pour la déconvolution, nous avons essayé d'organiser un "cometh course" avec des cliniciens et une application web facile d'utilisation, mais encore une fois l'obligation de distanciel instaurée par la pandémie du virus SARS-CoV-2 a à mon sens largement amoindri l'impact en compliquant les interactions.

L'organisation des data challenges de la partie II sur une semaine chacun, qui se sont déroulés à Aussois, constituent en revanche une bonne manière de réunir des chercheurs autour d'une thématique, en regroupant conférences et exercices pratiques. Les data challenge permettent de rencontrer d'autres chercheurs et de nouer des collaborations, mais c'est aussi et surtout une méthode de pédagogie innovante, basée sur l'expérimentation, l'entraide et la compétition, qui gagne à se démocratiser. Dans notre équipe, l'organisation de ce type d'événements devrait pouvoir reprendre à partir de l'année prochaine.

Bibliographie

[1] Alexis Battle et al. « Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals ». In : *Genome Research* 24.1 (jan. 2014), p. 14-24. doi : 10.1101/gr.155192.113.

[2] Yi-Fan Lu et al. « Personalized Medicine and Human Genetic Diversity ». In : *Cold Spring Harbor Perspectives in Medicine* 4.9 (sept. 2014), a008581. doi : 10.1101/cshperspect.a008581.

[3] Cristian Tomasetti, Lu Li et Bert Vogelstein. « Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention ». In : *Science (New York, N.Y.)* 355.6331 (24 mar. 2017), p. 1330-1334. doi : 10.1126/science.aaf9011.

[4] Michael S. Lawrence et al. « Discovery and saturation analysis of cancer genes across 21 tumor types ». In : *Nature* 505.7484 (23 jan. 2014), p. 495-501. doi : 10.1038/nature12912.

[5] Katherine A. Hoadley et al. « Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin ». In : *Cell* 158.4 (14 août 2014), p. 929-944. doi : 10.1016/j.cell.2014.06.049.

[6] Thomas Hensing et al. « A personalized treatment for lung cancer : molecular pathways, targeted therapies, and genomic characterization ». In : *Advances in Experimental Medicine and Biology* 799 (2014), p. 85-117. doi : 10.1007/978-1-4614-8778-4_5.

[7] The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM). « A Genomics-Based Classification of Human Lung Tumors ». In : *Science Translational Medicine* 5.209 (30 oct. 2013), 209ra153-209ra153. doi : 10.1126/scitranslmed.3006802.

[8] Koji Tsuta et al. « The utility of the proposed IASLC/ATS/ERS lung adenocarcinoma subtypes for disease prognosis and correlation of driver gene alterations ». In : *Lung Cancer* 81.3 (1er sept. 2013), p. 371-376. doi : 10.1016/j.lungcan.2013.06.012.

[9] Prudence A. Russell et al. « Does Lung Adenocarcinoma Subtype Predict Patient Survival ? : A Clinicopathologic Study Based on the New International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Lung Adenocarcinoma Classification ». In : *Journal of Thoracic Oncology* 6.9 (1er sept. 2011), p. 1496-1504. DOI : 10.1097/JTO.0b013e318221f701.

[10] Fuyan Hu et al. « Gene Expression Classification of Lung Adenocarcinoma into Molecular Subtypes ». In : *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.4 (juil. 2020). Conference Name : IEEE/ACM Transactions on Computational Biology and Bioinformatics, p. 1187-1197. DOI : 10.1109/TCBB.2019.2905553.

[11] Fengju Chen et al. « Multiplatform-based Molecular Subtypes of Non-Small Cell Lung Cancer ». In : *Oncogene* 36.10 (mar. 2017), p. 1384-1393. DOI : 10.1038/onc.2016.303.

[12] Federica Zito Marino et al. « Molecular heterogeneity in lung cancer : from mechanisms of origin to clinical implications ». In : *International Journal of Medical Sciences* 16.7 (10 juin 2019), p. 981-989. DOI : 10.7150/ijms.34739.

[13] Franziska Michor et Kornelia Polyak. « The Origins and Implications of Intratumor Heterogeneity ». In : *Cancer prevention research (Philadelphia, Pa.)* 3.11 (nov. 2010), p. 1361-1364. DOI : 10.1158/1940-6207.CAPR-10-0234.

[14] Olivier Trédan et al. « Drug resistance and the solid tumor microenvironment ». In : *Journal of the National Cancer Institute* 99.19 (3 oct. 2007), p. 1441-1454. DOI : 10.1093/jnci/djm135.

[15] Elza C. de Bruin et al. « Spatial and temporal diversity in genomic instability processes defines lung cancer evolution ». In : *Science (New York, N.Y.)* 346.6206 (10 oct. 2014), p. 251-256. DOI : 10.1126/science.1253462.

[16] Claudia Calabrese et al. « Genomic basis for RNA alterations in cancer ». In : *Nature* 578.7793 (2020), p. 129-136. DOI : 10.1038/s41586-020-1970-0.

[17] Borros Arneth. « Tumor Microenvironment ». In : *Medicina (Kaunas, Lithuania)* 56.1 (30 déc. 2019), E15. DOI : 10.3390/medicina56010015.

[18] Rebecca L. Siegel, Kimberly D. Miller et Ahmedin Jemal. « Cancer statistics, 2020 ». In : *CA : A Cancer Journal for Clinicians* 70.1 (2020). _eprint : https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21590, p. 7-30. DOI : 10.3322/caac.21590.

[19] Eric A. Collisson et al. « Comprehensive molecular profiling of lung adenocarcinoma ». In : *Nature* 511.7511 (9 juil. 2014), p. 543-550. doi : 10.1038/nature13385.

[20] Peter S. Hammerman et al. « Comprehensive genomic characterization of squamous cell lung cancers ». In : *Nature* 489.7417 (sept. 2012), p. 519-525. doi : 10.1038/nature11404.

[21] Joshua D. Campbell et al. « Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas ». In : *Nature genetics* 48.6 (juin 2016), p. 607-616. doi : 10.1038/ng.3564.

[22] Sophie Rousseaux et al. « Ectopic Activation of Germline and Placental Genes Identifies Aggressive Metastasis-Prone Lung Cancers ». In : *Science Translational Medicine* 5.186 (22 mai 2013). Publisher : American Association for the Advancement of Science Section : Research Article, 186ra66-186ra66. doi : 10.1126/scitranslmed.3005723.

[23] Michael Friberg Bruun Nielsen, Michael Bau Mortensen et Sönke Detlefsen. « Key players in pancreatic cancer-stroma interaction : Cancer-associated fibroblasts, endothelial and inflammatory cells ». In : *World Journal of Gastroenterology* 22.9 (7 mar. 2016), p. 2678-2700. doi : 10.3748/wjg.v22.i9.2678.

[24] Edwin Southern, Kalim Mir et Mikhail Shchepinov. « Molecular interactions on microarrays ». In : *Nature Genetics* 21.1 (jan. 1999). Bandiera_abtest : a Cg_type : Nature Research Journals Number : 1 Primary_atype : Reviews Publisher : Nature Publishing Group, p. 5-9. doi : 10.1038/4429.

[25] F. S. Collins. « The Human Genome Project : Lessons from Large-Scale Biology ». In : *Science* 300.5617 (11 avr. 2003), p. 286-290. doi : 10.1126/science.1084564.

[26] Eric S. Lander et al. « Initial sequencing and analysis of the human genome ». In : *Nature* 409.6822 (fév. 2001), p. 860-921. doi : 10.1038/35057062.

[27] Brian T. Wilhelm et al. « Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution ». In : *Nature* 453.7199 (juin 2008), p. 1239-1243. doi : 10.1038/nature07002.

[28] Ana Conesa et al. « A survey of best practices for RNA-seq data analysis ». In : *Genome Biology* 17 (2016), p. 13. doi : 10.1186/s13059-016-0881-8.

[29] Daniel Hebenstreit. « Methods, Challenges and Potentials of Single Cell RNA-seq ». In : *Biology* 1.3 (16 nov. 2012), p. 658-667. doi : 10.3390/biology1030658.

[30] Alessandra Dal Molin et Barbara Di Camillo. « How to design a single-cell RNA-sequencing experiment : pitfalls, challenges and perspectives ». In : *Briefings in Bioinformatics* 20.4 (19 juil. 2019), p. 1384-1394. doi : 10.1093/bib/bby007.

[31] C. David Allis et Thomas Jenuwein. « The molecular hallmarks of epigenetic control ». In : *Nature Reviews Genetics* 17.8 (août 2016), p. 487-500. doi : 10.1038/nrg.2016.59.

[32] Miho M. Suzuki et Adrian Bird. « DNA methylation landscapes : provocative insights from epigenomics ». In : *Nature Reviews Genetics* 9.6 (juin 2008). Bandiera_abtest : a Cg_type : Nature Research Journals Number : 6 Primary_atype : Reviews Publisher : Nature Publishing Group, p. 465-476. doi : 10.1038/nrg2341.

[33] Sergio Villicaña et Jordana T. Bell. « Genetic impacts on DNA methylation : research findings and future perspectives ». In : *Genome Biology* 22 (30 avr. 2021), p. 127. doi : 10.1186/s13059-021-02347-6.

[34] Julian Broche et al. « Genome-wide investigation of the dynamic changes of epigenome modifications after global DNA methylation editing ». In : *Nucleic Acids Research* 49.1 (9 déc. 2020), p. 158-176. doi : 10.1093/nar/gkaa1169.

[35] Marina Bibikova et al. « Genome-wide DNA methylation profiling using Infinium® assay ». In : *Epigenomics* 1.1 (oct. 2009), p. 177-200. doi : 10.2217/epi.09.14.

[36] Juan Sandoval et al. « Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome ». In : *Epigenetics* 6.6 (juin 2011), p. 692-702. doi : 10.4161/epi.6.6.16196.

[37] Sebastian Moran, Carles Arribas et Manel Esteller. « Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences ». In : *Epigenomics* 8.3 (mar. 2016), p. 389-399. doi : 10.2217/epi.15.114.

[38] Ryan Lister et al. « Human DNA methylomes at base resolution show widespread epigenomic differences ». In : *Nature* 462.7271 (19 nov. 2009), p. 315-322. doi : 10.1038/nature08514.

[39] Sébastien A. Smallwood et al. « Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity ». In : *Nature Methods* 11.8 (août 2014), p. 817-820. doi : 10.1038/nmeth.3035.

[40] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018.

[41] Michael I. Love, Wolfgang Huber et Simon Anders. « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 ». In : *Genome Biology* 15.12 (5 déc. 2014), p. 550. doi : 10.1186/s13059-014-0550-8.

[42] Mark D. Robinson, Davis J. McCarthy et Gordon K. Smyth. « edgeR : a Bioconductor package for differential expression analysis of digital gene expression data ». In : *Bioinformatics* 26.1 (1er jan. 2010), p. 139-140. doi : 10.1093/bioinformatics/btp616.

[43] Matthew E. Ritchie et al. « limma powers differential expression analyses for RNA-sequencing and microarray studies ». In : *Nucleic Acids Research* 43.7 (20 avr. 2015), e47. doi : 10.1093/nar/gkv007.

[44] David M Mutch et al. « The limit fold change model : A practical approach for selecting differentially expressed genes from microarray data ». In : *BMC Bioinformatics* 3 (21 juin 2002), p. 17. doi : 10.1186/1471-2105-3-17.

[45] Wilson Wen Bin Goh, Wei Wang et Limsoon Wong. « Why Batch Effects Matter in Omics Data, and How to Avoid Them ». In : *Trends in Biotechnology* 35.6 (juin 2017), p. 498-507. doi : 10.1016/j.tibtech.2017.02.012.

[46] Peipei Li et al. « Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data ». In : *BMC Bioinformatics* 16 (28 oct. 2015). doi : 10.1186/s12859-015-0778-7.

[47] Ciaran Evans, Johanna Hardin et Daniel M Stoebel. « Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions ». In : *Briefings in Bioinformatics* 19.5 (27 fév. 2017), p. 776-792. doi : 10.1093/bib/bbx008.

[48] Francesca Vitali et al. « Developing a 'personalome' for precision medicine : emerging methods that compute interpretable effect sizes from single-subject transcriptomes ». In : *Briefings in Bioinformatics* (18 déc. 2017). doi : 10.1093/bib/bbx149.

[49] Likun Wang et al. « DEGseq : an R package for identifying differentially expressed genes from RNA-seq data ». In : *Bioinformatics* 26.1 (1er jan. 2010), p. 136-138. doi : 10.1093/bioinformatics/btp612.

[50] Sonia Tarazona et al. « Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package ». In : *Nucleic Acids Research* 43.21 (2 déc. 2015), e140. doi : 10.1093/nar/gkv711.

[51] Jianxing Feng et al. « GFOLD : a generalized fold change for ranking differentially expressed genes from RNA-seq data ». In : *Bioinformatics* 28.21 (1er nov. 2012), p. 2782-2788. doi : 10.1093/bioinformatics/bts515.

[52] Hongwei Wang et al. « Individual-level analysis of differential expression of genes and pathways for personalized medicine ». In : *Bioinformatics* 31.1 (1er jan. 2015), p. 62-68. doi : 10.1093/bioinformatics/btu522.

[53] Xiangyu Li et al. « A rank-based algorithm of differential expression analysis for small cell line data with statistical control ». In : *Briefings in Bioinformatics* (2018). doi : 10.1093/bib/bbx135.

[54] Lishuang Qi et al. « Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels : a case study for resected stage I non-small-cell lung cancer ». In : *Briefings in Bioinformatics* 17.2 (1er mar. 2016), p. 233-242. doi : 10.1093/bib/bbv064.

[55] Qingzhou Guan et al. « Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms ». In : *Oncotarget* 7.42 (18 oct. 2016). doi : 10.18632/oncotarget.11996.

[56] Wanding Zhou, Peter W. Laird et Hui Shen. « Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes ». In : *Nucleic Acids Research* 45.4 (28 fév. 2017), e22-e22. doi : 10.1093/nar/gkw967.

[57] Tatiana Benaglia et al. « mixtools : An R Package for Analyzing Mixture Models ». In : *Journal of Statistical Software* 32.1 (21 oct. 2009). Number : 1, p. 1-29. doi : 10.18637/jss.v032.i06.

[58] Dirk Eddelbuettel et Romain Francois. « Rcpp : Seamless R and C++ Integration ». In : *Journal of Statistical Software* 40.1 (13 avr. 2011). Number : 1, p. 1-18. doi : 10.18637/jss.v040.i08.

[59] Simon Anders et Wolfgang Huber. « Differential expression analysis for sequence count data ». In : *Genome Biology* 11.10 (27 oct. 2010), R106. doi : 10.1186/gb-2010-11-10-r106.

[60] The Cancer Genome Atlas Research Network. « Comprehensive genomic characterization of squamous cell lung cancers ». In : *Nature* 489.7417 (sept. 2012), p. 519-525. doi : 10.1038/nature11404.

[61] Julie George et al. « Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors ». In : *Nature Communications* 9.1 (13 mar. 2018). Number : 1 Publisher : Nature Publishing Group, p. 1048. DOI : 10.1038/s41467-018-03099-x.

[62] Shu Zhang et al. « Landscape of transcriptional deregulation in lung cancer ». In : *BMC Genomics* 19.1 (5 juin 2018), p. 435. DOI : 10.1186/s12864-018-4828-1.

[63] Adam J. Bass et al. « SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas ». In : *Nature Genetics* 41.11 (nov. 2009). Number : 11 Publisher : Nature Publishing Group, p. 1238-1242. DOI : 10.1038/ng.465.

[64] Pierre P. Massion et al. « Significance of p63 amplification and overexpression in lung cancer development and prognosis ». In : *Cancer Research* 63.21 (1er nov. 2003), p. 7113-7121.

[65] Nataliya Mar, James J. Vredenburgh et Jeffrey S. Wasser. « Targeting HER2 in the treatment of non-small cell lung cancer ». In : *Lung Cancer* 87.3 (1er mar. 2015). Publisher : Elsevier, p. 220-225. DOI : 10.1016/j.lungcan.2014.12.018.

[66] H. M. Jeong et al. « ESRP1 is overexpressed in ovarian cancer and promotes switching from mesenchymal to epithelial phenotype in ovarian cancer cells ». In : *Oncogenesis* 8.9 (29 août 2019). Number : 9 Publisher : Nature Publishing Group, p. 1-3. DOI : 10.1038/s41389-019-0155-x.

[67] Hyung Ho Lee et al. « Epithelial Splicing Regulatory Protein (ESPR1) Expression in an Unfavorable Prognostic Factor in Prostate Cancer Patients ». In : *Frontiers in Oncology* 10 (2020). Publisher : Frontiers. DOI : 10.3389/fonc.2020.556650.

[68] Guanglei Chen et al. « RILPL2 regulates breast cancer proliferation, metastasis, and chemoresistance via the TUBB3/PTEN pathway ». In : *American Journal of Cancer Research* 9.8 (2019), p. 1583-1606.

[69] Terry M. Therneau et Patricia M. Grambsch. « The Cox Model ». In : *Modeling Survival Data : Extending the Cox Model*. Sous la dir. de Terry M. Therneau et Patricia M. Grambsch. Statistics for Biology and Health. New York, NY : Springer, 2000, p. 39-77. DOI : 10.1007/978-1-4757-3294-8_3.

[70] John N. Weinstein et al. « Comprehensive molecular characterization of urothelial bladder carcinoma ». In : *Nature* 507.7492 (mar. 2014). Number : 7492 Publisher : Nature Publishing Group, p. 315-322. doi : 10.1038/nature12965.

[71] Daniel C. Koboldt et al. « Comprehensive molecular portraits of human breast tumours ». In : *Nature* 490.7418 (oct. 2012). Number : 7418 Publisher : Nature Publishing Group, p. 61-70. doi : 10.1038/nature11412.

[72] Donna M. Muzny et al. « Comprehensive molecular characterization of human colon and rectal cancer ». In : *Nature* 487.7407 (juil. 2012). Number : 7407 Publisher : Nature Publishing Group, p. 330-337. doi : 10.1038/nature11252.

[73] Jihun Kim et al. « Integrated genomic characterization of oesophageal carcinoma ». In : *Nature* 541.7636 (jan. 2017). Number : 7636 Publisher : Nature Publishing Group, p. 169-175. doi : 10.1038/nature20805.

[74] Michael S. Lawrence et al. « Comprehensive genomic characterization of head and neck squamous cell carcinomas ». In : *Nature* 517.7536 (jan. 2015). Number : 7536 Publisher : Nature Publishing Group, p. 576-582. doi : 10.1038/nature14129.

[75] Christopher J. Ricketts et al. « The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma ». In : *Cell Reports* 23.1 (3 avr. 2018). Publisher : Elsevier, 313-326.e5. doi : 10.1016/j.celrep.2018.03.075.

[76] Adam Abeshouse et al. « The Molecular Taxonomy of Primary Prostate Cancer ». In : *Cell* 163.4 (5 nov. 2015). Publisher : Elsevier, p. 1011-1025. doi : 10.1016/j.cell.2015.10.025.

[77] Adam J. Bass et al. « Comprehensive molecular characterization of gastric adenocarcinoma ». In : *Nature* 513.7517 (sept. 2014). Number : 7517 Publisher : Nature Publishing Group, p. 202-209. doi : 10.1038/nature13480.

[78] Douglas A. Levine. « Integrated genomic characterization of endometrial carcinoma ». In : *Nature* 497.7447 (mai 2013). Number : 7447 Publisher : Nature Publishing Group, p. 67-73. doi : 10.1038/nature12113.

[79] Arabel Vollmann-Zwerenz et al. « Tumor Cell Invasion in Glioblastoma ». In : *International Journal of Molecular Sciences* 21.6 (12 mar. 2020). doi : 10.3390/ijms21061932.

[80] Michele Ceccarelli et al. « Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma ». In : *Cell* 164.3 (28 jan. 2016), p. 550-563. doi : 10.1016/j.cell.2015.12.028.

[81] Cameron W. Brennan et al. « The Somatic Genomic Landscape of Glioblastoma ». In : *Cell* 155.2 (10 oct. 2013). Publisher : Elsevier, p. 462-477. doi : 10.1016/j.cell.2013.09.034.

[82] Brad A. Friedman et al. « Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models ». In : *Cell Reports* 22.3 (16 jan. 2018), p. 832-847. doi : 10.1016/j.celrep.2017.12.066.

[83] Anders Lundin et al. « Human iPS-Derived Astroglia from a Stable Neural Precursor State Show Improved Functionality Compared with Conventional Astrocytic Models ». In : *Stem Cell Reports* 10.3 (13 mar. 2018), p. 1030-1045. doi : 10.1016/j.stemcr.2018.01.021.

[84] Michael J. Gerdes et al. « Emerging Understanding of Multiscale Tumor Heterogeneity ». In : *Frontiers in Oncology* 4 (18 déc. 2014). doi : 10.3389/fonc.2014.00366.

[85] Ibiayi Dagogo-Jack et Alice T. Shaw. « Tumour heterogeneity and resistance to cancer therapies ». In : *Nature Reviews Clinical Oncology* 15.2 (fév. 2018), p. 81-94. doi : 10.1038/nrclinonc.2017.166.

[86] Raghu Kalluri. « The biology and function of fibroblasts in cancer ». In : *Nature Reviews Cancer* 16.9 (sept. 2016). Number : 9 Publisher : Nature Publishing Group, p. 582-598. doi : 10.1038/nrc.2016.73.

[87] Yan-gao Man et al. « Tumor-Infiltrating Immune Cells Promoting Tumor Invasion and Metastasis : Existing Theories ». In : *Journal of Cancer* 4.1 (5 jan. 2013), p. 84-95. doi : 10.7150/jca.5482.

[88] Nasser K. Altorki et al. « The lung microenvironment : an important regulator of tumour growth and metastasis ». In : *Nature reviews. Cancer* 19.1 (jan. 2019), p. 9-31. doi : 10.1038/s41568-018-0081-9.

[89] Vladimir Yu Kiselev, Tallulah S. Andrews et Martin Hemberg. « Challenges in unsupervised clustering of single-cell RNA-seq data ». In : *Nature Reviews. Genetics* 20.5 (mai 2019), p. 273-282. doi : 10.1038/s41576-018-0088-9.

[90] Christoph Ziegenhain et al. « Comparative Analysis of Single-Cell RNA Sequencing Methods ». In : *Molecular Cell* 65.4 (16 fév. 2017). Publisher : Elsevier, 631-643.e4. doi : 10.1016/j.molcel.2017.01.023.

[91] Romina E. Araya et Romina S. Goldszmid. « Characterization of the tumor immune infiltrate by multiparametric flow cytometry and unbiased high-dimensional data analysis ». In : *Methods in Enzymology* 632 (2020), p. 309-337. doi : 10.1016/bs.mie.2019.11.012.

[92] Wei Chang Colin Tan et al. « Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy ». In : *Cancer Communications (London, England)* 40.4 (avr. 2020), p. 135-153. doi : 10.1002/cac2.12023.

[93] Francisco Avila Cobos et al. « Computational deconvolution of transcriptomics data from mixed cell populations ». In : *Bioinformatics* 34.11 (1er juin 2018), p. 1969-1979. doi : 10.1093/bioinformatics/bty019.

[94] Laura Cantini et al. « Assessing reproducibility of matrix factorization methods in independent transcriptomes ». In : *Bioinformatics* 35.21 (1er nov. 2019), p. 4307-4313. doi : 10.1093/bioinformatics/btz225.

[95] Urszula Czerwinska. *deconica : Deconvolution of transcriptome through Immune Component Analysis*. R package version 0.1.1. 2019.

[96] A. Hyvärinen et E. Oja. « Independent component analysis : algorithms and applications ». In : *Neural Networks* 13.4 (juin 2000), p. 411-430. doi : 10.1016/S0893-6080(00)00026-5.

[97] Xianlu Laura Peng et al. « De novo compartment deconvolution and weight estimation of tumor samples using DECODER ». In : *Nature Communications* 10.1 (18 oct. 2019), p. 4729. doi : 10.1038/s41467-019-12517-7.

[98] Yuna Blum et al. « Dissecting heterogeneity in malignant pleural mesothelioma through histo-molecular gradients for clinical applications ». In : *Nature Communications* 10.1 (22 mar. 2019), p. 1333. doi : 10.1038/s41467-019-09307-6.

[99] Kosuke Yoshihara et al. « Inferring tumour purity and stromal and immune cell admixture from expression data ». In : *Nature Communications* 4.1 (11 oct. 2013). Number : 1 Publisher : Nature Publishing Group, p. 2612. doi : 10.1038/ncomms3612.

[100] E. Andres Houseman et al. « Reference-free deconvolution of DNA methylation data and mediation by cell composition effects ». In : *BMC Bioinformatics* 17 (29 juin 2016). doi : 10.1186/s12859-016-1140-4.

[101] Pavlo Lutsik et al. « MeDeCom : discovery and quantification of latent components of heterogeneous methylomes ». In : *Genome Biology* 18.1 (24 mar. 2017), p. 55. doi : 10.1186/s13059-017-1182-6.

[102] Vitor Onuchic et al. « Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types ». In : *Cell reports* 17.8 (15 nov. 2016), p. 2075-2086. doi : 10.1016/j.celrep.2016.10.057.

[103] Andrew E. Teschendorff et al. « A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies ». In : *BMC Bioinformatics* 18.1 (13 fév. 2017), p. 105. doi : 10.1186/s12859-017-1511-5.

[104] Alexander J. Titus et al. « Cell-type deconvolution from DNA methylation : a review of recent applications ». In : *Human Molecular Genetics* 26 (R2 1er oct. 2017), R216-R224. doi : 10.1093/hmg/ddx275.

[105] Peter A. Jones. « Functions of DNA methylation : islands, start sites, gene bodies and beyond ». In : *Nature Reviews Genetics* 13.7 (juil. 2012). Number : 7 Publisher : Nature Publishing Group, p. 484-492. doi : 10.1038/nrg3230.

[106] Dirk Schübeler. « Function and information content of DNA methylation ». In : *Nature* 517.7534 (15 jan. 2015), p. 321-326. doi : 10.1038/nature14192.

[107] Eilis Hannon et al. « Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins ». In : *PLoS Genetics* 14.8 (9 août 2018). doi : 10.1371/journal.pgen.1007544.

[108] Tanya Barrett et al. « NCBI GEO : archive for functional genomics data sets—update ». In : *Nucleic Acids Research* 41 (Database issue jan. 2013), p. D991-D995. doi : 10.1093/nar/gks1193.

[109] Reza Mohammadi. *bmixture : Bayesian Estimation for Finite Mixture of Distributions*. Version Rpackage version 1.5. 2019.

[110] Scott Chasalow. *combinat : combinatorics utilities*. Version Rpackage version 0.0-8. 2012.

[111] J. A. Ferreira et A. H. Zwinderman. « On the Benjamini–Hochberg method ». In : *The Annals of Statistics* 34.4 (août 2006). Publisher : Institute of Mathematical Statistics, p. 1827-1849. doi : 10.1214/009053606000000425.

[112] Florian Privé et al. « Efficient analysis of large-scale genome-wide data with two R packages : bigstatsr and bigsnpr ». In : *Bioinformatics* 34.16 (15 août 2018), p. 2781-2787. doi : 10.1093/bioinformatics/bty185.

[113] Raymond B. Cattell. « The Scree Test For The Number Of Factors ». In : *Multivariate Behavioral Research* 1.2 (1er avr. 1966). Publisher : Routledge _eprint : https ://doi.org/10.1207/s15327906mbr0102_10, p. 245-276. doi : 10.1207/s15327906mbr0102_10.

[114] Peter J. Rousseeuw. « Silhouettes : A graphical aid to the interpretation and validation of cluster analysis ». In : *Journal of Computational and Applied Mathematics* 20 (1er nov. 1987), p. 53-65. doi : 10.1016/0377-0427(87)90125-7.

[115] Andrew E. Teschendorff et al. « A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data ». In : *Bioinformatics* 29.2 (15 jan. 2013), p. 189-196. doi : 10.1093/bioinformatics/bts680.

[116] John Lonsdale et al. « The Genotype-Tissue Expression (GTEx) project ». In : *Nature Genetics* 45.6 (juin 2013). Number : 6 Publisher : Nature Publishing Group, p. 580-585. doi : 10.1038/ng.2653.

[117] Manuel Hidalgo et al. « Patient Derived Xenograft Models : An Emerging Platform for Translational Cancer Research ». In : *Cancer discovery* 4.9 (sept. 2014), p. 998-1013. doi : 10.1158/2159-8290.CD-14-0001.

[118] Mark D. Robinson et Alicia Oshlack. « A scaling normalization method for differential expression analysis of RNA-seq data ». In : *Genome Biology* 11.3 (2 mar. 2010), R25. doi : 10.1186/gb-2010-11-3-r25.

[119] Urszula Czerwinska. *deconica : Deconvolution of transcriptome through Immune Component Analysis*. Version R package version 0.1.1. 2019.

[120] Renaud Gaujoux et Cathal Seoighe. « A flexible R package for nonnegative matrix factorization ». In : *BMC Bioinformatics* 11.1 (2 juil. 2010), p. 367. doi : 10.1186/1471-2105-11-367.

[121] Hyunsoo Kim et Haesun Park. « Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis ». In : *Bioinformatics* 23.12 (15 juin 2007), p. 1495-1502. doi : 10.1093/bioinformatics/btm134.

[122] Francisco Avila Cobos et al. « Benchmarking of cell type deconvolution pipelines for transcriptomics data ». In : *Nature Communications* 11.1 (6 nov. 2020), p. 5650. doi : 10.1038/s41467-020-19015-1.

[123] Gregor Sturm et al. « Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology ». In : *Bioinformatics* 35.14 (15 juil. 2019), p. i436-i445. doi : 10.1093/bioinformatics/btz363.

[124] Syed Haider et al. « Systematic Assessment of Tumor Purity and Its Clinical Implications ». In : *JCO Precision Oncology* 4 (1er nov. 2020). Publisher : Wolters Kluwer, p. 995-1005. doi : 10.1200/PO.20.00016.

[125] Laura Cantini et al. « Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer ». In : *Nature Communications* 12 (5 jan. 2021), p. 124. doi : 10.1038/s41467-020-20430-7.

[126] Johanna A. Joyce et Jeffrey W. Pollard. « Microenvironmental regulation of metastasis ». In : *Nature Reviews Cancer* 9.4 (avr. 2009), p. 239-252. doi : 10.1038/nrc2618.

[127] Melissa R. Junttila et Frederic J. de Sauvage. « Influence of tumour micro-environment heterogeneity on therapeutic response ». In : *Nature* 501.7467 (19 sept. 2013), p. 346-354. doi : 10.1038/nature12626.

[128] Dvir Aran, Marina Sirota et Atul J. Butte. « Systematic pan-cancer analysis of tumour purity ». In : *Nature Communications* 6 (4 déc. 2015), p. 8971. doi : 10.1038/ncomms9971.

[129] Yi Liu et al. « Identification of tumor microenvironment-related prognostic genes in colorectal cancer based on bioinformatic methods ». In : *Scientific Reports* 11.1 (22 juil. 2021). Bandiera_abtest : a Cc_license_type : cc_by Cg_type : Nature Research Journals Number : 1 Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Cancer microenvironment ;Computational biology and bioinformatics Subject_term_id : cancer-microenvironment ;computational-biology-and-bioinformatics, p. 15040. doi : 10.1038/s41598-021-94541-6.

[130] Qian Chen et al. « Identification of a tumor microenvironment-related gene signature to improve the prediction of cervical cancer prognosis ». In : *Cancer Cell International* 21 (25 mar. 2021), p. 182. doi : 10.1186/s12935-021-01867-2.

[131] Yong Li et al. « Weighted gene correlation network analysis identifies microenvironment-related genes signature as prognostic candidate for Grade II/III glioma ». In : *Aging (Albany NY)* 12.21 (7 nov. 2020), p. 22122-22138. doi : 10.18632/aging.104075.

[132] Tinyi Chu et Charles G. Danko. *Bayesian cell-type deconvolution and gene expression inference reveals tumor-microenvironment interactions*. Company : Cold Spring Harbor Laboratory Distributor : Cold Spring Harbor Laboratory Label : Cold Spring Harbor Laboratory Section : New Results Type : article. 19 août 2020, p. 2020.01.07.897900. doi : 10.1101/2020.01.07.897900.

[133] Qianghu Wang et al. « Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment ». In : *Cancer Cell* 32.1 (10 juil. 2017), 42-56.e6. doi : 10.1016/j.ccell.2017.06.003.

[134] Jinyu Cheng et al. « Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19 ». In : *Briefings in Bioinformatics* 22.2 (1er mar. 2021), p. 988-1005. doi : 10.1093/bib/bbaa327.

[135] L. Kantorovitch. « On the Translocation of Masses ». In : *Management Science* 5.1 (1958), p. 1-4.

[136] Florent Chuffart et al. « Exploiting Single-Cell Quantitative Data to Map Genetic Variants Having Probabilistic Effects ». In : *PLoS genetics* 12.8 (août 2016), e1006213. doi : 10.1371/journal.pgen.1006213.

[137] George Marsaglia, Wai Wan Tsang et Jingbo Wang. « Evaluating Kolmogorov's Distribution ». In : *Journal of Statistical Software* 8.1 (10 nov. 2003). Number : 1, p. 1-4. doi : 10.18637/jss.v008.i18.

[138] Yanfen Fang et Xiongwen Zhang. « Targeting NEK2 as a promising therapeutic approach for cancer treatment ». In : *Cell Cycle* 15.7 (28 mar. 2016), p. 895-907. doi : 10.1080/15384101.2016.1152430.

[139] Yu-Ming Shiao et al. « Dysregulation of GIMAP genes in non-small cell lung cancer ». In : *Lung Cancer* 62.3 (1er déc. 2008). Publisher : Elsevier, p. 287-294. doi : 10.1016/j.lungcan.2008.03.021.

[140] Beiping Miao et al. « The transcription factor FLI1 promotes cancer progression by affecting cell cycle regulation ». In : *International Journal of Cancer* 147.1 (1er juil. 2020), p. 189-201. doi : 10.1002/ijc.32831.

[141] J. Louise Lines et al. « VISTA is an immune checkpoint molecule for human T cells ». In : *Cancer Research* 74.7 (1er avr. 2014), p. 1924-1932. doi : 10.1158/0008-5472.CAN-13-1504.

[142] Paul Zajac et al. « MAGE-A Antigens and Cancer Immunotherapy ». In : *Frontiers in Medicine* 4 (8 mar. 2017), p. 18. doi : 10.3389/fmed.2017.00018.

[143] Saiedeh Razi Soofiyani, Mohammad Saeid Hejazi et Behzad Baradaran. « The role of CIP2A in cancer : A review and update ». In : *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 96 (déc. 2017), p. 626-633. doi : 10.1016/j.biopha.2017.08.146.

[144] Huaiyu Mi et al. « PANTHER version 14 : more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D419-D426. doi : 10.1093/nar/gky1038.

[145] Ke Wang et al. « Identification of differentially expressed genes in non-small cell lung cancer ». In : *Aging (Albany NY)* 11.23 (9 déc. 2019), p. 11170-11185. doi : 10.18632/aging.102521.

[146] Debottam Sinha et al. « Cep55 overexpression promotes genomic instability and tumorigenesis in mice ». In : *Communications Biology* 3.1 (21 oct. 2020), p. 1-16. doi : 10.1038/s42003-020-01304-6.

[147] Rui Bai et al. « NEK2 plays an active role in Tumorigenesis and Tumor Microenvironment in Non-Small Cell Lung Cancer ». In : *International Journal of Biological Sciences* 17.8 (11 mai 2021), p. 1995-2008. doi : 10.7150/ijbs.59019.

[148] Nayoung Kim et al. « Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma ». In : *Nature Communications* 11.1 (8 mai 2020), p. 2285. doi : 10.1038/s41467-020-16164-1.

[149] Xinxin Zhang et al. « CellMarker : a manually curated resource of cell markers in human and mouse ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D721-D728. doi : 10.1093/nar/gky900.

[150] Yufang Qin et al. « InfiniumPurify : An R package for estimating and accounting for tumor purity in cancer methylation research ». In : *Genes & Diseases* 5.1 (21 fév. 2018), p. 43-45. doi : 10.1016/j.gendis.2018.02.003.

[151] Ziyi Chen et al. « Inference of immune cell composition on the expression profiles of mouse tissue ». In : *Scientific Reports* 7 (13 jan. 2017), p. 40508. doi : 10.1038/srep40508.

[152] Maria K Jaakkola et Laura L Elo. « Computational deconvolution to estimate cell type-specific gene expression from bulk data ». In : *NAR Genomics and Bioinformatics* 3.1 (1er mar. 2021). doi : 10.1093/nargab/lqaa110.

[153] Jaeil Ahn et al. « DeMix : deconvolution for mixed cancer transcriptomes using raw measured data ». In : *Bioinformatics* 29.15 (1er août 2013), p. 1865-1871. doi : 10.1093/bioinformatics/btt301.

[154] Vessela N. Kristensen et al. « Principles and methods of integrative genomic analyses in cancer ». In : *Nature Reviews Cancer* 14.5 (mai 2014), p. 299-313. doi : 10.1038/nrc3721.

[155] Andrew E. Teschendorff et Shijie C. Zheng. « Cell-type deconvolution in epigenome-wide association studies : a review and recommendations ». In : *http ://dx.doi.org/10.2217/epi-2016-0153* (14 mar. 2017). Archive Location : London, UK Publisher : Future Medicine Ltd London, UK. doi : 10.2217/epi-2016-0153.

[156] Mario L. Suvà et Itay Tirosh. « Single-Cell RNA Sequencing in Cancer : Lessons Learned and Emerging Challenges ». In : *Molecular Cell* 75.1 (11 juil. 2019), p. 7-12. doi : 10.1016/j.molcel.2019.05.003.

Annexes

Annexe 1

PenDA, a rank-based method for personalized differential analysis : Application to lung cancer

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

PenDA, a rank-based method for personalized differential analysis: Application to lung cancer

Magali Richard[1]*, Clémentine Decamps[1], Florent Chuffart[2], Elisabeth Brambilla[3], Sophie Rousseaux[2], Saadi Khochbin[2], Daniel Jost[1,4]*

1 Univ Grenoble Alpes, CNRS, Grenoble INP, TIMC-IMAG, Grenoble, France, **2** CNRS UMR 5309, Inserm U1209, Univ Grenoble Alpes, Institute for Advanced Biosciences, Grenoble, France, **3** CHUGA, Inserm U1209, Univ Grenoble Alpes, Institute for Advanced Biosciences, Grenoble, France, **4** University of Lyon, ENS de Lyon, Univ Claude Bernard, CNRS, Laboratory of Biology and Modelling of the Cell, Lyon, France

* magali.richard@univ-grenoble-alpes.fr (MR); daniel.jost@ens-lyon.fr (DJ)

Citation: Richard M, Decamps C, Chuffart F, Brambilla E, Rousseaux S, Khochbin S, et al. (2020) PenDA, a rank-based method for personalized differential analysis: Application to lung cancer. PLoS Comput Biol 16(5): e1007869. https://doi.org/10.1371/journal.pcbi.1007869

Editor: Amin Emad, McGill University, CANADA

Received: January 22, 2020

Accepted: April 11, 2020

Published: May 11, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007869

Data Availability Statement: LUAD and LUSC expression data were downloaded from The Cancer Genome Atlas program (https://portal.gdc.cancer.gov/). Grenoble hospital expression data are accessible in the study GSE30219. Rmarkdown

Abstract

The hopes of precision medicine rely on our capacity to measure various high-throughput genomic information of a patient and to integrate them for personalized diagnosis and adapted treatment. Reaching these ambitious objectives will require the development of efficient tools for the detection of molecular defects at the individual level. Here, we propose a novel method, PenDA, to perform Personalized Differential Analysis at the scale of a single sample. PenDA is based on the local ordering of gene expressions within individual cases and infers the deregulation status of genes in a sample of interest compared to a reference dataset. Based on realistic simulations of RNA-seq data of tumors, we showed that PenDA outcompetes existing approaches with very high specificity and sensitivity and is robust to normalization effects. Applying the method to lung cancer cohorts, we observed that deregulated genes in tumors exhibit a cancer-type-specific commitment towards up- or down-regulation. Based on the individual information of deregulation given by PenDA, we were able to define two new molecular histologies for lung adenocarcinoma cancers strongly correlated to survival. In particular, we identified 37 biomarkers whose up-regulation lead to bad prognosis and that we validated on two independent cohorts. PenDA provides a robust, generic tool to extract personalized deregulation patterns that can then be used for the discovery of therapeutic targets and for personalized diagnosis. An open-access, user-friendly R package is available at https://github.com/bcm-uga/penda.

Author summary

The hopes of precision medicine rely on our capacity to measure individual molecular information for personalized diagnosis and treatment. These challenging perspectives will be only possible with the development of efficient methodological tools to identify patient-specific molecular defects from the many precise molecular information that one can access at the single-individual, single tissue or even single-cell levels. Such methods

vignettes (S1 & S2 Texts) for reproducing all figures and tables in this paper can be found in a R package named penda (https://github.com/bcm-uga/penda). Preprocessed data used to generate the figures can be found at http://membres-timc.imag.fr/Magali.Richard/publication.html: tcga_data_ctrl.rds, tcga_data_case.rds, tcga_exp_grp.rds, grenoble_data_ctrl.rds, grenoble_data_case.rds, grenoble_exp_grp.rds.

Funding: The research leading to these results was supported by ITMO Cancer (Plan Cancer 2014-2019, Biologie des Systèmes n°BIO2015-08) [EB, SK, DJ] and University Grenoble-Alpes via the Grenoble Alpes Data Institute [MR] and the SYMER program [SK, DJ] (which are funded by the French National Research Agency under the "Investissements d'Avenir" program ANR-15-IDEX-02). SK acknowledges additional funding from Plan Cancer ASC16079CSA, Pitcher, LIFE program of University Grenoble Alpes (ANR-15-IDEX-02), Fondation ARC "Canc'air" (RAC16042CLA) and project PGA1 RF20190208471. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

will provide a better understanding of disease-specific biological mechanisms and will promote the development of personalized therapeutic strategies. Here we describe a novel method, named PenDA, to perform differential analysis of gene expression at the individual level. Based on a realistic benchmark of simulated tumors, we demonstrated that PenDA reaches very high efficiency in detecting sample-specific deregulated genes. We then applied the method to two large cohorts associated with lung cancer. A detailed statistical analysis of the results allowed to isolate genes with specific deregulation patterns, like genes that are up-regulated in all tumors or genes that are expressed but never deregulated in any tumors. Given their specificities, these genes are likely to be of interest in therapeutic research. In particular, we were able to identified 37 new biomarkers associated to bad prognosis that we validated on two independent cohorts.

Introduction

General medicine still largely relies on detecting diseases after the apparition of symptoms and on curing them with generic treatments. However, many studies have highlighted how the natural genetic or genomic diversities observed in a population, as well as patient history, or environment exposure, may strongly affect diseases risks, prognoses and responses to treatments [1,2]. This is particularly critical for cancer, where each individual tumor may be viewed as an independent disease, with specific and variable responses to generic therapeutic treatments [3]. Recently, thanks to the development of cheap and robust next-generation sequencing techniques, getting better insights into inter-individual heterogeneities was made possible by the analyses of large cohorts of patients. This led to the identification of individual molecular signatures or biomarkers associated with better prognosis, or better response to targeted treatment [4–6]. This new knowledge paves the way to precision and personalized medicine where the genetic, genomic, and molecular information of each patient will be integrated to develop personalized diagnosis and treatment [2,3]. However, such challenging perspectives will be only possible with the concomitant development of efficient and robust methodological tools that allow the identifications of molecular defects or deregulation patterns at the individual level.

Many statistical or bioinformatic methods do already exist to identify deregulated genes at the population level. For example, in the context of gene expression, standard methods like DESeq2 [7], edgeR [8] or limma [9] are designed and routinely used to identify genes that are differentially expressed *in average* between two groups of patients [10]. These methods are usually based on modelling of the data distribution and statistical testing for differential expression (fold change analysis). While valuable to detect consistent *typical* deregulation patterns, such analyses do not provide precise information at the individual level. In addition, these global methods are usually very sensitive to batch effects that, without corrections, may lead to false discoveries or to confound important subpopulation effects [11]. Prior application of normalization routines to the investigated samples are used to mitigate such technical biases, but improper normalization may still perturb the biological signal [12,13].

Novel methods, robust to technical interference, are therefore needed to capture specific, individual data. Few promising techniques already allow to extract interpretable information from personalized omics data (see [14] for a review). Rankcomp [15,16] uses pairs of genes with a stable, relative order in a reference dataset to infer deregulated genes in individual samples [17–19]. This method, based on ranking, avoids the problem of normalization between samples, but results in very high false discovery rates (above 20%, see Methods). Alternative

methods, like DEGseq [20], NOISeq [21] or Gfold [22], exploit paired samples from the same patient (one control versus one malignant) to perform differential analysis. However, such matched samples are usually rare (for example, in the case of cancer, a single sample from the tumorous biopsy is usually available for one patient). Above all, it is not clear if the variabilities observed between paired samples are due to actual deregulation, to intrinsic inter-sample heterogeneities, or to technical biases. For example, in lung cancer, correlations between paired tumorous and normal samples are similar than between tumors of two different patients, and are only slightly higher than between a tumorous sample and an unmatched normal tissue (Fig 1A).

To overcome all these limitations, we developed PenDA, for Personalized Differential Analysis, a rank-based method, robust to batch or normalization effects, that uses information extracted from a reference dataset to infer the deregulation status of genes in individual samples of interest.

For illustrating the power of the method, we focused on lung cancer, which is the first cause of cancer-related death world-wide [23] and represents a major public health issue. In particular, we studied two datasets provided by The Cancer Genome Atlas (TCGA) for two of the most common histologies of non-small-cell lung cancers (NSCLCs): adenocarcinoma (ADC [24]) and squamous cell carcinoma (SQCC [25]). Clinical implications, gene expression patterns and DNA mutation landscapes are largely distinct between both histologies even if some pathways are similarly altered [26]. Their mutation rates are unusually high compared to other lung cancers and molecular heterogeneity is important [24,25,27]. This molecular heterogeneity translates into a complex landscape of deregulation of gene expression [28,29]. Previous analyses of molecular abnormalities occurring in a large proportion of patients have already led to the development of biomarkers for target therapy [30] and for prognostic signatures [31] but it still remains an important biomedical priority [32]. More generally, observations of morphological, histological or molecular defects led to the classifications of ADC and SQC into various subtypes [24,25,28,33–36]. For example, ADC is generally classified into three subtypes according to transcriptional and histopathological data: terminal respiratory unit (TRU or bronchoid), proximal inflammatory (PI or squamoid) and proximal proliferative (PP or magnoid). These subtypes differ by gene expression but also by clinical behaviors like the stage-specific survival [24,36,37]. Recently, Chen et al [27] combined various molecular information (DNA methylation, copy number alteration, mRNA, miRNA and protein expression) to define 6 molecular subtypes of ADC that partially overlap with the standard classification and that show correlation with survival rate, immune profiles or cigarette exposure. By using our method PenDA on the TCGA datasets for ADC and SQC, we illustrated how personalized differential analysis can bring additional information compared to previous studies about inter-individual heterogeneity, can help to find gene classifiers for molecular subtypes and may be used to infer biomarkers related to prognosis.

Results

A robust algorithm to infer if genes are differentially regulated in individual samples

Description of the method. PenDA is a rank-based method that allows to infer if the expression of any gene in a given sample of interest is deregulated compared to a set of reference samples (see Methods for details). The fundamental assumption behind the algorithm is that a gene is seen as deregulated in an individual sample if its local ordering compared to other genes with similar expressions is perturbed, as similarly stated by the RankComp method [15]. Briefly, PenDA starts by inferring a reference of relative ordering in control samples: for

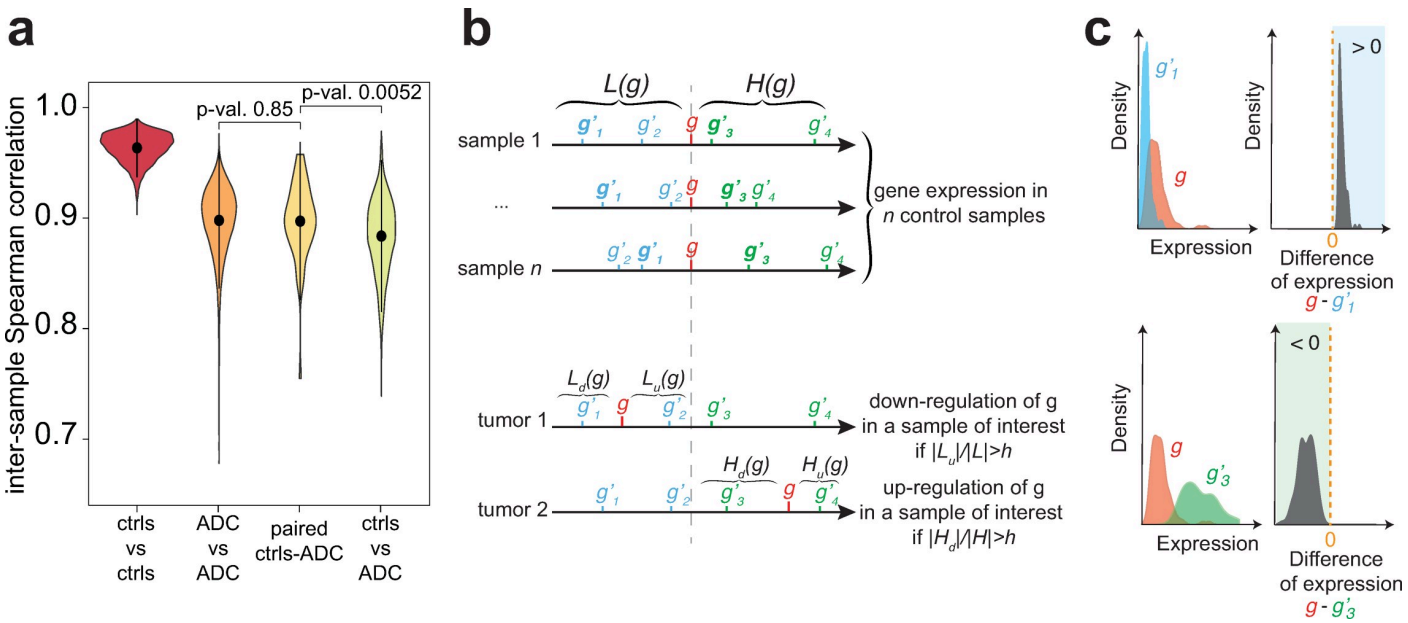


Fig 1. The PenDA method. (a) Violin-plots for the distributions of Spearman correlation between two samples taken from the TCGA database on lung adenocarcinoma: between two non-tumorous samples (ctrls vs ctrls, n = 4,656 pairs), between two tumorous samples (ADC vs ADC, n = 103,285), between paired normal and tumorous samples (paired ctrls-ADC, n = 48), and between unpaired controls and tumors (ctrls vs ADC, n = 44,135). Shown p-values correspond to Wilcoxon tests. (b) Basic scheme depicting the PenDA method. (Top) For each gene *g*, the algorithm infers sets of genes whose expressions are always lower (*L(g)*) or higher (*H(g)*) than that of *g* in a pool of control, reference samples. (Bottom) In a given individual (tumor) sample, *g* is viewed as deregulated if its relative ordering with genes in the *L(g)* and *H(g)* lists is modified. (c) Examples of genes in the *L* (g'_1, top) or *H* (g'_3, bottom) lists of a gene *g*. While the individual distributions of gene expression in the control samples may overlap (left), the distribution of the difference in gene expression in controls (right) is always positive or negative for genes in *L* and *H* lists respectively.

https://doi.org/10.1371/journal.pcbi.1007869.g001

every gene *g*, it constructs two lists *L(g)* and *H(g)* of genes whose expression is lower and higher respectively than that of *g* in almost all the samples of a given reference dataset (Fig 1B top and 1C). To avoid comparison with genes having very different expression levels and to increase sensitivity of the method, lists *L(g)* and *H(g)* are then limited to the subset of *l* genes whose expression in control samples are closest to *g*. Finally, for a given sample of interest, PenDA scans every gene *g* to determine if it might be up- or down-regulated in that sample. This step is performed by considering the number of genes $L_u(g)$ (respectively $H_d(g)$) in *L(g)* (resp. *H(g)*) in the studied case whose relative ordering to *g* has changed compared to controls (Fig 1B bottom). If the proportion of such genes with a modified order ($|L_u(g)|/|L(g)|$ or $|H_d(g)|/|H(g)|$) exceeds a given threshold *h*, the gene *g* is detected as deregulated. It has to be noted that a change of ordering between *g* and a gene *g'* of *L(g)* and *H(g)* might be caused by the deregulation of *g'* and not necessary by that of *g*. To limit the consequences of this effect on the detection of deregulation, PenDA iteratively applies the previous scheme until convergence by excluding at each iteration the current set of deregulated genes from every *L* and *H* lists (S1 Fig). In the cases where the *L(g)* or *H(g)* lists are empty, we used the percentile method (see Methods for details) to evaluate the deregulation of *g* (S2 Fig).

Impact of method parameters and of the dataset properties on performance. To test and validate our method, we generated a realistic simulated dataset where we controlled the identity of deregulated genes and the direction (up or down) of deregulation. Based on the RNA-seq profiles of 18,000 genes in normal and tumorous samples of two lung cancer cohorts (adenocarcinoma: ADC, squamous cells: SQCC) of the TCGA database [24,25], we simulated 10 tumorous samples each having on average 30% of deregulated genes (see Methods for details). Note that to avoid any bias in the analysis, simulations were not based on the same

principle that governed the PenDA method, ie, the relative order of gene expressions. Rather, each *in silico* tumor was generated by randomly choosing a normal sample and a list of deregulated genes was randomly assigned. Then, the perturbed gene expressions of these genes were obtained by adding to the normal levels random values typical of the differences in gene expression between tumorous and normal samples as observed in the actual dataset.

We first aimed at testing the method on this dataset by varying the two parameters of the algorithm: l the restricted size of the $L(g)$ and $H(g)$ lists, and h the detection threshold based on the $|L_u(g)|/|L(g)|$ and $|H_d(g)|/|H(g)|$ ratios (see above). We used the 97 non-tumorous lung samples of the TCGA dataset to determine $L(g)$ and $H(g)$ and then, apply the PenDA method to the 10 simulations. By varying h from 0 to 1, we built a ROC curve (true positive rate TPR vs false positive rate FPR) for different l values (Fig 2A). We observed that all the curves are well above the line of no-discrimination (dashed grey line), reaching simultaneously high sensitivity and high specificity. Using the maximal value of informedness (TPR-FPR) as a summary statistic of the ROC curve, we observed that the method reached an optimal prediction efficiency for *l~10–100*. For too short lists, finite size effects dominate and decrease the predictive power. For very large lists, $L(g)$ and $H(g)$ contain many genes whose expressions are very far

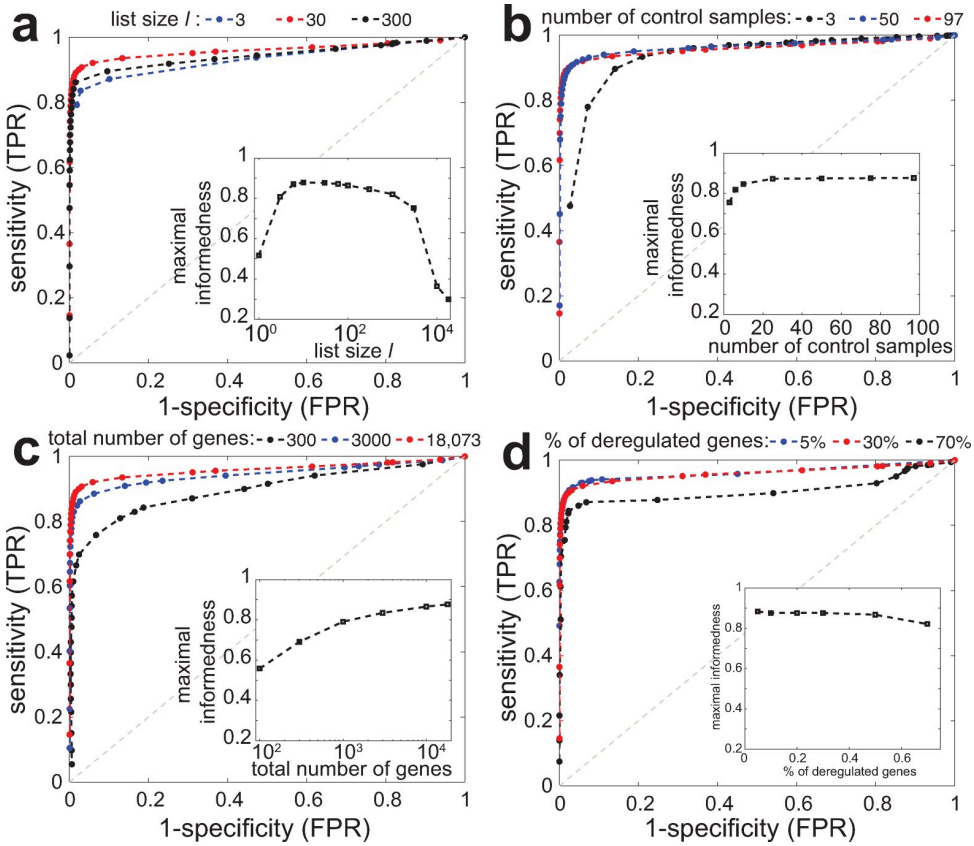


Fig 2. Parameter analysis and predictive power. ROC curves (true positive rate TPR vs false positive rate FPR) of the PenDA method on simulated datasets. The curves were obtained by varying the proportion threshold h for various values of other method parameters or of properties of the investigated dataset. Insets show the maximal informedness that represent the maximal value of the difference TPR-FPR computed for each ROC curve. (a) Effect of the maximal size l of L and H lists. (b) Impact of the number of control samples used to infer the L and H lists. (c) Effect of the total number of genes in the dataset. (d) Impact of the proportion of deregulated genes in the tumorous samples.

https://doi.org/10.1371/journal.pcbi.1007869.g002

from *g*. Thus, if *g* is weakly or mildly deregulated, these genes will keep their relative position compared to *g*, leading to a loss in sensitivity. In the next, we imposed $l = 30$.

We then evaluated how PenDA performance depends on the intrinsic properties of the investigated datasets. We determined $L(g)$ and $H(g)$ using different numbers of non-tumorous samples and run PenDA on the same set of 10 simulations. We observed that the method is very robust regarding the size of the reference datasets, achieving very high efficiency even for a limited number of control samples (Fig 2B). Next, we kept the reference pool fixed but varied the number of investigated genes from 100 to 18,000 and applied PenDA to the simulated dataset restricted to the corresponding limited set of genes (Fig 2C). We remarked that the reliability of the method is an increasing function of the number of genes, achieving very good performance for numbers higher than ~3,000. Indeed, a large number of genes augments the capacity of $L(g)$ and $H(g)$ lists to integrate genes that may be sensitive to changes in relative ordering. Finally, we tested the effect of the percentage of deregulated genes in the simulated datasets that may affect the current sizes of L and H lists during the iterations of the method. Fig 2D showed that the predictive power of PenDA is relatively insensitive to this quantity, performance slightly declining for very high percentage.

All these quantitative analyses illustrate that the method is very robust regarding parameters and dataset properties fine-tuning. In particular, PenDA remains performant even for a small number of reference datasets.

Comparison with other individual-based methods. We next sought to compare PenDA with other existing methods that also allow personalized diagnosis of gene deregulation. Using the same set of 10 simulations introduced before, we generated ROC curves (see Methods) for 4 alternative methods (Fig 3): 2 versions of the rank-based method RankComp [15,16], a simple percentile method based on outlier detection and DESeq2 [38], the popular algorithm for detecting differential expression at the population level but used here on an individual basis. We observed that PenDA outperforms these methods, in particular in the limit of high specificity (FPR< = 5%) where PenDA could reach very high sensitivity (TPR> = 90%) even for a limited number of control samples (Fig 3B). Surprisingly, outcomes of the RankComp methods were very dependent to the number of control samples and even lead to better results for smaller control datasets. Note that basing our definition of deregulation on relative rankings

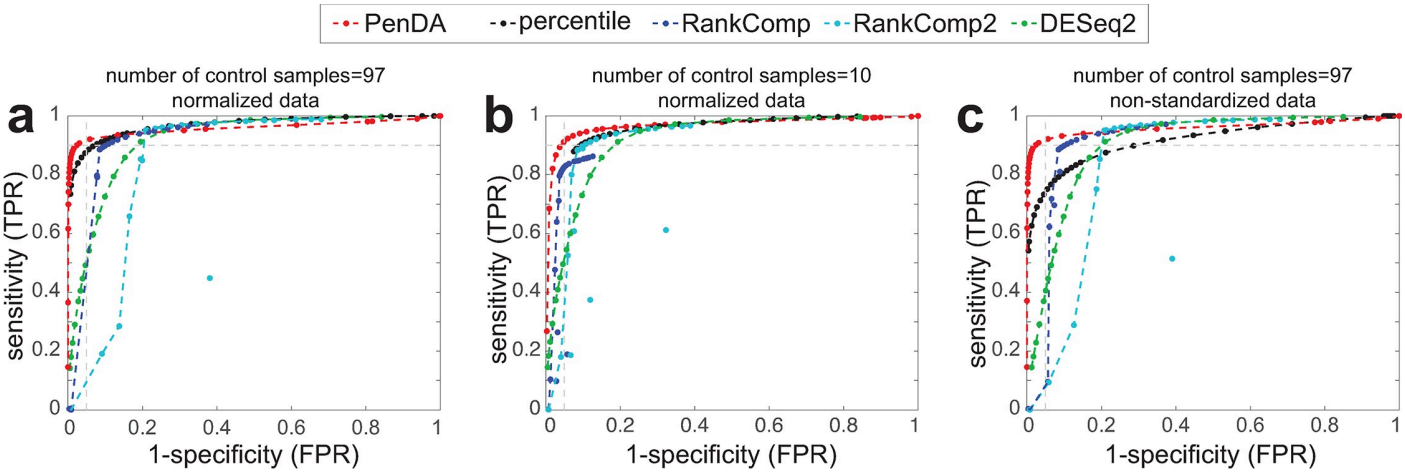


Fig 3. Comparison with other methods. (a) ROC curves on the same simulated dataset (normalized data, 97 control samples) as used in Fig 2 for PenDA, a simple percentile-based method, 2 versions of RankComp and DESeq2. (b) As in (a) but reference pool was composed by only 10 control samples. (c) As in (a) but data were not normalized.

https://doi.org/10.1371/journal.pcbi.1007869.g003

limits the sensitivity of PenDA (and RankComp) to batch or normalization effects compared to the percentile method (Fig 3C), DESeq2, thanks to its internal normalization routine, being also robust (S11C Fig).

The PenDA package. The PenDA method is available as a R package at https://github.com/bcm-uga/penda. The *penda* vignette (vignette_penda, S1 Text) runs the PenDA pipeline (S3 Fig) on the samples of interest. It takes as an input two dataframes corresponding to the reference dataset of control samples and the dataset to investigate. It first filters for genes whose expressions are very low in every samples. Then, it computes the *L* and *H* lists from control samples for a given list size *l*. Finally, in every sample, it run the iterative process to infer gene deregulation based on a user-defined threshold *h*. Optionally, the package offers the possibility to find the optimal set of parameters (in particular *h*) best adapted to: (i) the input data and (ii) a user-defined specific maximal false-discovery rate (vignette_simulation, S2 Text). It is based on realistic simulations built on the input dataframes and a ROC analysis, as described in the previous sections. Typically, on a standard personal computer (1 core of 3.6 GHz CPU), construction of *L* and *H* lists takes ~10 sec CPU time for 18,000 genes and 98 controls. Downstream analysis of gene deregulation is slower and requires ~2 min CPU time per analyzed sample.

Application of the PenDA method to personalized analysis of genetic deregulation in lung cancer

Overview of gene deregulation in adenocarcinoma and squamous cell carcinoma. We evaluated the performances of PenDA on two large cohorts of patients from The Cancer Genome Atlas (TCGA) project representing two of the most common types of non-small-cell lung cancers: lung adenocarcinoma (ADC, ~50%) and lung squamous cell carcinoma (SQCC, ~40%) [39]. Personalized differential analysis was performed on the normalized gene expression data (RNA-seq) of 455 ADC cases and 473 SQCC cases (S1 Table).

We observed that the proportion of deregulated genes per tumor is very variable (Fig 4A), ranging from 3% to 61% of deregulated genes in ADCs (with a mean of 33%, corresponding to 5960 genes) and from 0.4% to 55% of genes deregulated in SQCCs (with a mean of 42%, corresponding to 7659 genes). Analysis of variance revealed a slight effect of tumor stages on the total number of gene deregulations in both ADC and SQCC patients (S4A and S4B Fig). Multiple-comparisons with the Tukey method indicated a significant increase in the number of deregulated genes between an early stage of cancer (stage Ia) and the later stages (stage Ib to stage IV). We consistently observed a higher number of gene down-regulations compared to gene up-regulations in each patient (median ratio down/up of 1.25 in ADCs and of 1.31 in SQCCs). These ratios were invariant across tumor stages (one-way ANOVA non-significant, S4C and S4D Fig).

To test the accuracy of our method, we compared the gene deregulation behavior between ADC and SQCC disease groups. We examined, for each gene, the proportion of tumors where the gene was detected as deregulated within each cohort (Fig 4B and 4C). A two-proportion Z-test was used to compare, for each gene, the observed proportion of deregulation (S2 Table). We identified 5346 genes with a significant variation in down-regulation proportion between ADCs and SQCCs (Fig 4B) and 5616 genes with a significant variation in up-regulation proportion between ADCs and SQCCs (Fig 4C). Gene functional annotation indicated an enrichment in cell division, epidermis development and keratinocyte differentiation in genes specifically up-regulated in SQCCs (S5D Fig). In contrast, genes specifically up-regulated in ADCs display a significant enrichment in glycan processing (S5B Fig). Genes specifically down-regulated in either SQCC or ADC do not display significant enrichment (GO term

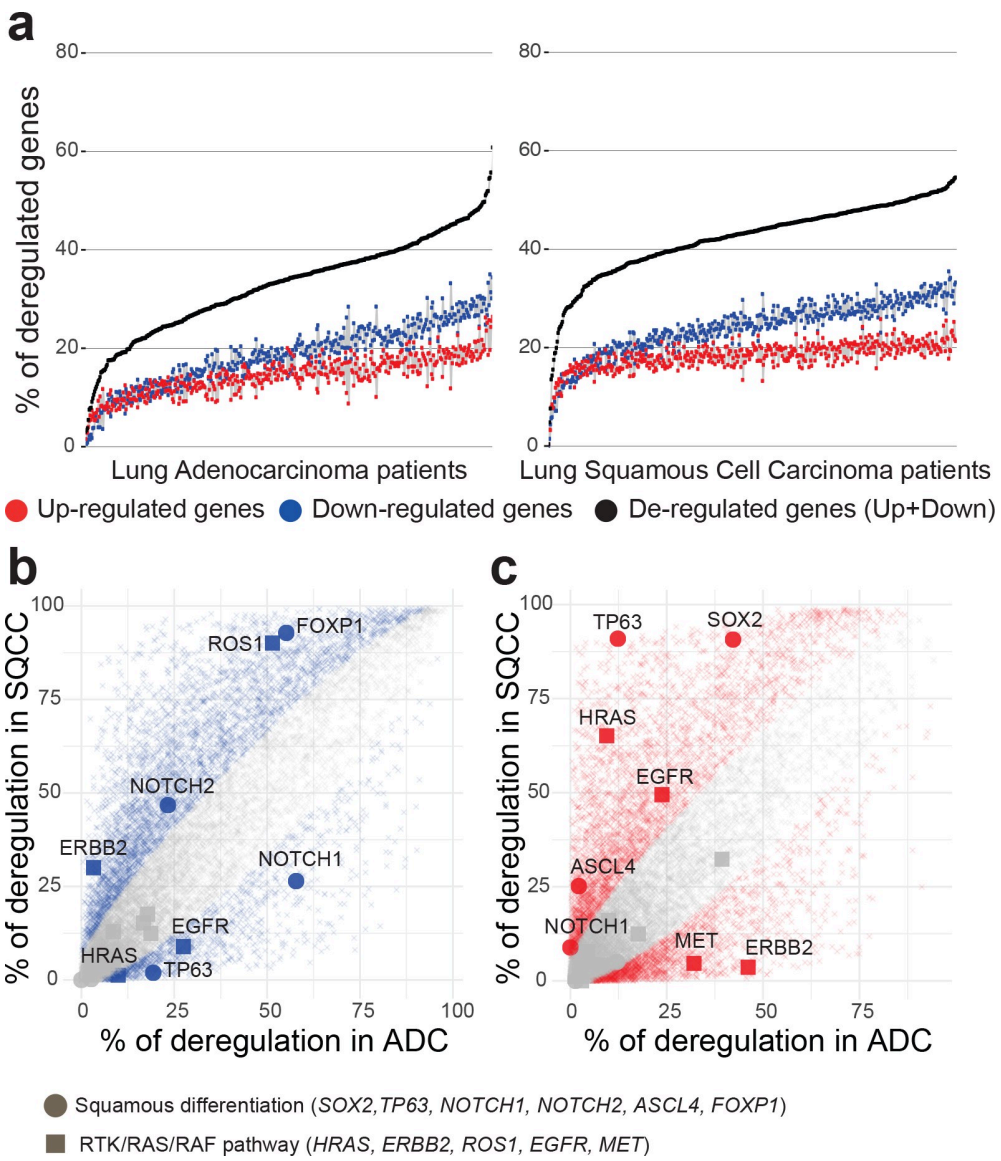


Fig 4. Overview of genetic deregulation in adenocarcinoma and squamous cell carcinoma. (a) The percentage of deregulated genes in ADC (left panel) and SQCC (right panel) patients. % of up-regulated genes is indicated in red, % of down-regulated genes is indicated in blue, total % of deregulated genes (up + down) is indicated in black. Patients are ordered by increasing total number of deregulated genes. (b,c) Scatterplot of the percentage of deregulated patients for each gene in the ADC cohort (x-axis) versus deregulated patients percentage in the SQCC cohort (y-axis). Left panel (b) represents downregulation events and right panel (c) represents upregulation events. Colored points represent significant differences between ADC and SQCC cohorts (two-sided two-proportion z-test, p-value < 0.05 after Bonferroni correction for 18143 multiple testing).

https://doi.org/10.1371/journal.pcbi.1007869.g004

significance score < 2). Thus, our method successfully managed to identify biological pathways differentially activated between ADCs and SQCCs.

To illustrate such differential behaviors, we specifically depicted genes belonging to two known pathways involved in cancer progression: the squamous differentiation, that often display somatic alterations in SQCC cancers [25], and the receptor tyrosine kinase (RTK)/RAS/RAF pathway, frequently mutated in ADC cancers [24] (Fig 4B and 4C). In agreement with

previous studies based on population level analysis [40,41], we observed a specific high proportion of up-regulation of SOX 2 and TP63 in SQCCs and of ERBB2 in ADCs. SOX2 is a transcription factor involved in normal squamous cell differentiation, which is frequently amplified in SQCCs [42]. TP63 belongs to the p53 tumor suppressor family, an overexpression of an altered TP63 isoform has been frequently associated with cancer squamous histology [43]. ERRB2 is a member of the epidermal growth factor (EGF) receptor family and is often overexpressed or mutated in ADC [44]. Interestingly, many genes frequently affected by somatic alterations, such as KRAS and EGFR in ADCs[45], exhibit a weaker gene deregulation. In contrast, some genes with a low occurrence of somatic alterations present a strong deregulation frequency in SQCCs, such as FOXP1 or NOTCH1 [25].

Taken together, these results suggest that personalized analysis of both genetic mutations and gene expression variations are required for a full understanding of regulation pathways involved in tumorigenesis.

Most deregulated genes are committed to specific deregulation patterns. Recurrent gene deregulations are considered as characteristic features of cancer initiation and progression. To explore the deregulation pattern of each gene, we analyzed their proportion of down-regulation and up-regulation in each cohort (Fig 5A and 5B). Most of the genes that are deregulated in more than ~30% of the patients exhibited a commitment toward up-regulation or down-regulation. For genes deregulated in less that ~30% of the patients, up-regulation and down-regulation are less constrained. Interestingly, ~ 5% of the genes that are either down or up-regulated in more than 30% of both SQCCs and ADCs display antagonistic commitment (S6 Fig). Thus, while the orientation of the deregulation commitment (towards up or down regulation) is generally conserved between ADC and SQCC, in some cases, it may be inverted.

We then decided to quantify extreme single gene deregulation frequencies using a one sample t-test in which we compared the mean deregulation of each gene to the mean deregulation of all genes. Using this approach, we were able to identify genes with specific deregulation patterns, that we defined as super-conserved (SC, genes almost never deregulated), super-up-regulated (SU, genes almost systematically up-regulated) and super-down-regulated (SD, genes almost systematically down-regulated) (S3 Table). While some of the genes with a 'super' regulation pattern are common to ADCs and SQCCs cancers, we observed that a significant proportion of them are specific to a given histology (Fig 5C). Functional profiling indicated that SQCCs SU genes are enriched in cell cycle processes, DNA replication and keratinocyte differentiation. Interestingly, a significant proportion of SQCCs and ADCs SD genes are related to angiogenesis and signal transduction processes (S7 Fig).

As an illustration of the 'super' regulation patterns, we examined more closely three characteristic genes: the SC gene *CAPS*, the SU gene *ESRP1* and the SD gene *RILPL2*. *CAPS* encodes for a calcium binding protein, ESPR1 is an epithelial cell-type-specific splicing regulator and RILPL2 is a rab-interacting lysosomal protein. In Fig 5D–5F, we plotted for these three genes the distribution of gene expression (normalized RNA-seq counts) within the control dataset, the ADC and the SQCC cohorts. Interestingly, for the *CAPS* gene, we do observe a difference in mean expression of the gene between cancer tissues and control whereas no differential expression was detected at individual level. Similarly, expression distributions of *ESRP1* and *RILPL2* genes in ADC and SQCC cohorts partially overlap with their respective distributions in control samples. However, our method identified deregulation in almost all patients of both ADC and SQCC cohorts, indicating that these two genes are committed to specific deregulation pattern during tumorigenesis.

These examples illustrate the power of individual-based approaches compared to population based-approaches. Indeed, extreme single gene deregulation frequencies detection is only possible when individual variations are considered. Those results clearly indicate that a small

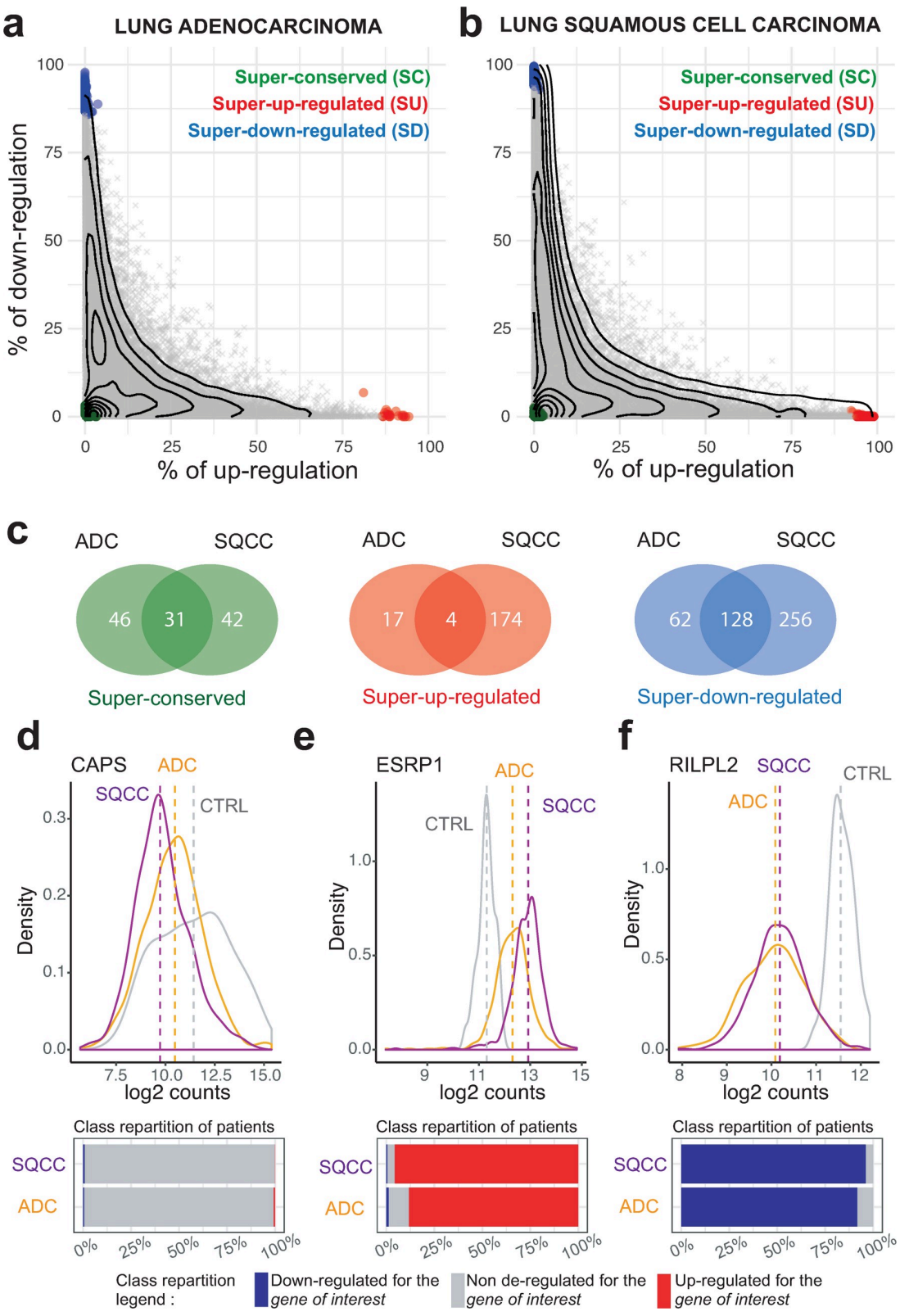

a
LUNG ADENOCARCINOMA
b
LUNG SQUAMOUS CELL CARCINOMA
Super-conserved (SC)
Super-up-regulated (SU)
Super-down-regulated (SD)
% of down-regulation
% of up-regulation
c
ADC
SQCC
46
31
42
Super-conserved
17
4
174
Super-up-regulated
62
128
256
Super-down-regulated
d
CAPS
e
ESRP1
f
RILPL2
CTRL
Density
log2 counts
Class repartition of patients
Class repartition legend :
Down-regulated for the gene of interest
Non de-regulated for the gene of interest
Up-regulated for the gene of interest

Fig 5. The gene deregulation pattern. (a-b) Scatterplots of the percentage of up-regulated versus down-regulated patients in the ADC (left panel) and SQCC (right panel) cohorts. Each dot corresponds to one gene. The x-axis indicates the percentage of up-regulation within the cohort, the y-axis indicates the percentage of down-regulation within the cohort. The contour lines correspond to the density of genes. Genes that are significantly differentially expressed at the individual level (t-statistic, q-value < 0.05) are represented using the following color code: green genes are super-conserved (SC), blue genes are super-down-regulated (SD), red genes are super-up-regulated (SU), other genes are depicted in gray. (C) Venn diagrams indicating the total number of SC, SU and SD genes in ADC and SQCC cohorts. (d-e-f) (Top panels) Distributions of gene expression levels (normalized counts) for three representative genes (the SC gene CAPS in (d), the SU gene ESPRP1 in (e), the SD gene RILPL2 in (f)) in the ADC cohort (yellow), in the SQCC cohort (purple), and for the control patients (gray). The dashed lines represent the mean expressions. (Bottom panels) The corresponding percentages of patients deregulated for each shown gene in ADC and SQCC cohorts are represented by bar plots: gray for non-deregulated patients, blue for down-regulated patients and red for up-regulation patients.

https://doi.org/10.1371/journal.pcbi.1007869.g005

proportion of genes are committed to specific deregulation patterns that occur in all patient of a given cohort. Given their specificities, the 'super' genes will likely be of interest in therapeutic research.

Individual genetic deregulations efficiently classify cancer histology and identify novel adenocarcinoma molecular subtype. ADC and SQCC histologies differ in gene expression. To assess the power of PenDA method compared to traditional analyses on normalized expression counts, we applied principal component analysis (PCA) on both PenDA differential expression matrix (values equal to -1 if a gene in a given tumor is down-regulated, 0 if a gene is not deregulated or 1 if a gene is up-regulated) and normalized count matrix (normalized RNA-seq counts with values between 0 and $3{,}7.10^6$ counts). In both cases, we observed a separate clustering of ADC and SQCC cohorts mainly driven by the first principal component (Fig 6A and 6B). We used a supervised learning algorithm (SVM, see Methods) to compare classification properties of *normalized count* versus *differential expression* inputs. Both approaches succeed to properly classify patients between ADC and SQCC histologies, though we observed that classification based on PenDA inputs performed slightly better (Fig 6C). We then applied hierarchical clustering to classify the 455 ADC and 473 SQCC samples together, using a subset of 875 genes defined in a previous independent study (based on RNA-seq counts) as lung cancer subtypes classifiers (Classification to Nearest Centroid, [40]). We clustered samples with a distance based on inter-sample Pearson correlations computed from the PenDA differential expression matrix (Fig 6D). We observed a clear separation between ADCs and SQCCs groups, thereby validating our methodological approach. We could identify one main SQCC class and three ADC subclasses (S3 Table). The majority of ADC patients clustered into 2 subclasses (class II and III), that were not distinguishable in the clustering analysis performed by George et al on different lung cancers, using the same classifier genes [40]. We compared the three ADC subclasses obtained with our approach with the six ADC genomic subtypes previously identified by Chen et al, using a multiplatform-based approach on the TCGA-LUAD dataset [27]. Class II ADC patients are mainly associated with AD1, AD2 and AD3 subtypes, whereas the majority of class III ADC patients is distributed among AD4 and AD5 subtypes (Fig 6E). Similarly, class II and class III ADC patients did not directly relate to the integrated ADC molecular subtypes defined by the pioneer work of The Cancer Genome Atlas Research Network [24] (S8A Fig). Interestingly, the same hierarchical clustering analysis using the same genes but with normalized counts did not clearly highlight the three ADC subtypes identified with the PenDA differential expression matrix (S8B Fig). Thus, clustering ADC according to their individual deregulation profiles identified new ADC subclasses. This demonstrates that personalized analysis using PenDA method brings new insights into histology classification.

Systematic up-regulation of 37 genes in adenocarcinoma is a strong predictor of poor prognosis. We then wondered what defined these novel ADC subclasses. First, we asked whether this segmentation into three classes was specific to the classifier genes chosen to perform the hierarchical clustering. We performed a principal component analysis on ADC

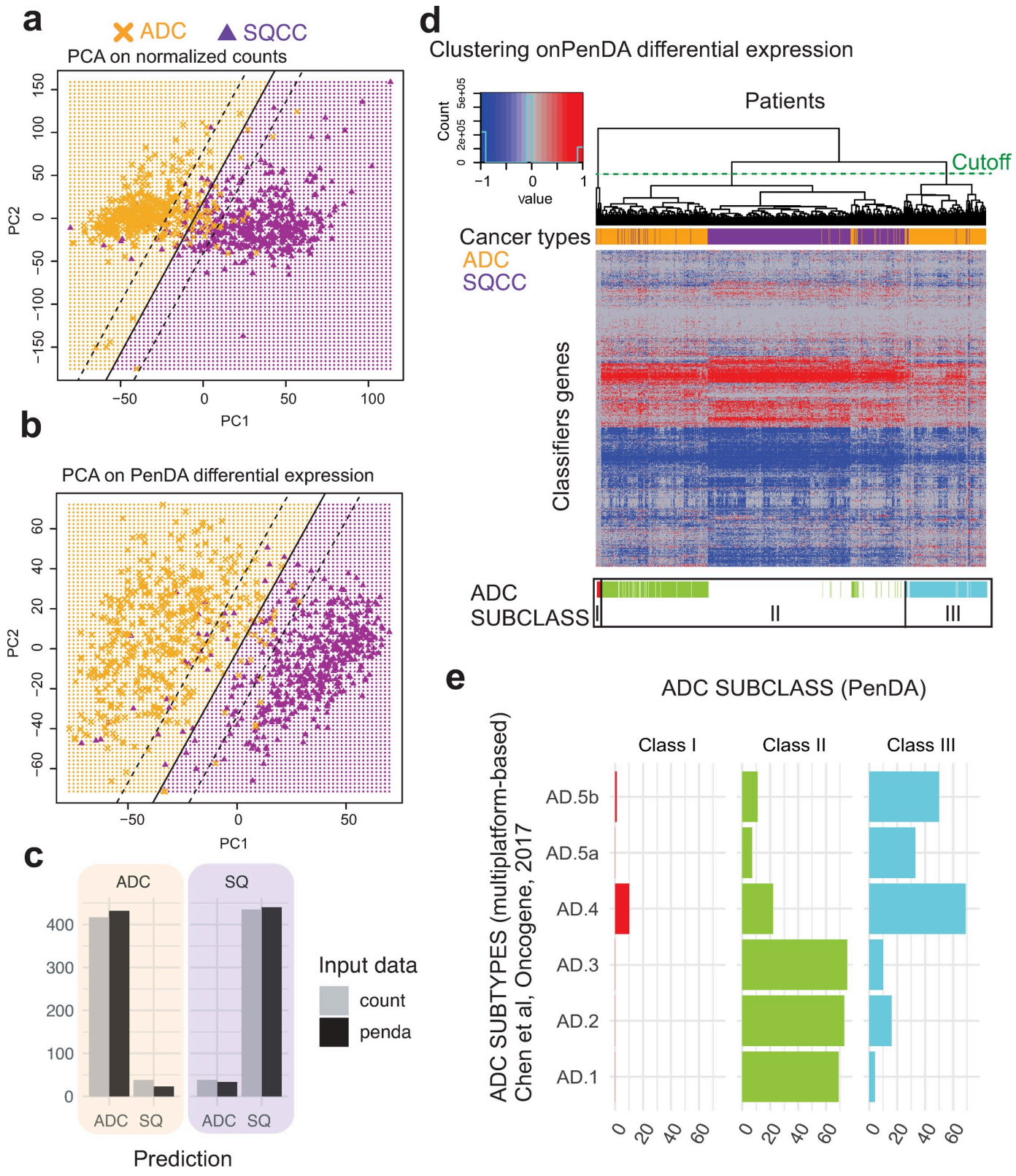

a
ADC
SQCC
PCA on normalized counts
PC2
PC1
b
PCA on PenDA differential expression
PC2
PC1
c
ADC
SQ
Input data
count
penda
Prediction
d
Clustering onPenDA differential expression
Count
value
Patients
Cutoff
Cancer types
ADC
SQCC
Classifiers genes
ADC
SUBCLASS
e
ADC SUBCLASS (PenDA)
Class I
Class II
Class III
ADC SUBTYPES (multiplatform-based)
Chen et al, Oncogene, 2017
AD.5b
AD.5a
AD.4
AD.3
AD.2
AD.1

Fig 6. Genetic deregulations efficiently classify cancer histologies. (a, b) Principal Component Analysis on TCGA non-small-cell lung cancers (ADC and SQCC cohorts) using normalized count matrix (a) or PenDA differential expression matrix (b) as input. Full lines represent the decision boundary between ADC and SQCC histologies (using a linear SVM classifier on the first two principal components). Dashed lines represent the upper and lower margins of the decision boundary. Each symbol represents an individual sample (orange crosses for ADC, purple triangles for SQCC). (c) At the bottom, the bar plot represents the histology predictions based on the SVM classifier. SVM on PenDA predicts correctly 95% of ADCs and 93% SQCCs. SVM on count predicts correctly 92% of ADCs and 92% SQCCs. (d) Heatmap of PenDA differential expression matrix applied to a specific set of classifier genes (n = 875) in TCGA non-small-cell lung cancers: ADC (orange) and SQCC (purple). Two hierarchical clustering analyses were performed: using Euclidean distance to sort genes and using Pearson correlation-based distance to classify patients, with a complete linkage function in both cases. ADC subclasses (color-coded, class I to III) are defined according to the dendrogram cutoff n = 3 groups (cutting section = green dashed line). (e) Graphical representation of the contingency table between ADC subtypes (Chen et al,) and ADC subclasses (PenDA analysis). Each bar plot represents the total number of patients in each cell of the table.

https://doi.org/10.1371/journal.pcbi.1007869.g006

cohort only using the corresponding PenDA differential expression matrix for all genes (Fig 7A). The first two principal components of the analysis nicely discriminated classes I, II and III. We then focused on the two major groups: class II and class III. We performed a Cox survival analysis on these two groups (Fig 7B) and observed that the class III patients have a better 5-year survival prognosis than class II patients (cox p-value = 0.00104). In order to better understand the molecular differences between class II and class III patients, we analyzed the pattern of deregulation of all genes in each class (Fig 7C). In class II, we observed a significant augmentation in the proportion of tumors where a given gene was detected as deregulated. In total, ~13% of the genes (n = 2432) were significantly more often deregulated in class II compared to class III patients (one-sided proportion test). We verified that the cancer stages, gender, and age were evenly distributed in class II and class III patients (chi square test p-value = 0.2133, p-value = 1, and p-value = 0.2133, respectively) and that the shift in genetic deregulation was detectable independently of stages, gender and age (S9 Fig). This indicated that this adenocarcinoma classification was not correlated with any of these putative confounding factors.

We decided to specifically study the 37 genes displaying the most extreme differences between the two classes, i.e. the genes deregulated in more than 75% of class II patients and in less than 25% of class III patients (red dots on Fig 7C, S4 Table). Since all these genes are committed toward up-regulation in class II patients, we tested if the up-regulation of these genes would be a good predictor of cancer survival. We added up the level of individual deregulation of the 37 genes (values equal to -1, 0 or 1, for each gene) to quantify the total deregulation score associated with those genes. Then we defined three groups using the 1st and the 3rd quantile of the score distribution. Analysis of the 5-years survival curve in the ADC LUAD-TCGA dataset showed a significant difference between groups, with a worst prognosis for patients that display up-regulation of most of the genes (score $\geq$ 34, Fig 7D). To validate our selected set of 37 genes as robust biomarkers, we applied the PenDA method on expression data (Affymetrix Human Genome U133 Plus 2.0 Array) of an independent adenocarcinoma cohort from the Grenoble Hospital (85 patients, GSE30219[4]) (see Methods). We then investigated the 5-years survival curve of the three groups predicted using 36 genes (all genes were analyzed in the Grenoble Hospital cohort, except FAM72D not measured by the array). Coherently with the results observed in TCGA-LUAD ADC cohort, patients up-regulated for many genes (score $\geq$ 15) have a worst prognosis (cox p-value = $5.2.10^{-4}$, Fig 7E). Thus, using the PenDA method, we identified 37 biomarkers predicting a bad outcome when they are all up-regulated. Altogether, these results suggest that PenDA method is a powerful approach to discover new biomarkers in cancer.

Discussion

The PenDA method provides a new rank-based approach to analyze personalized gene deregulation. The method outcompetes existing approaches to identified genetic deregulation at the

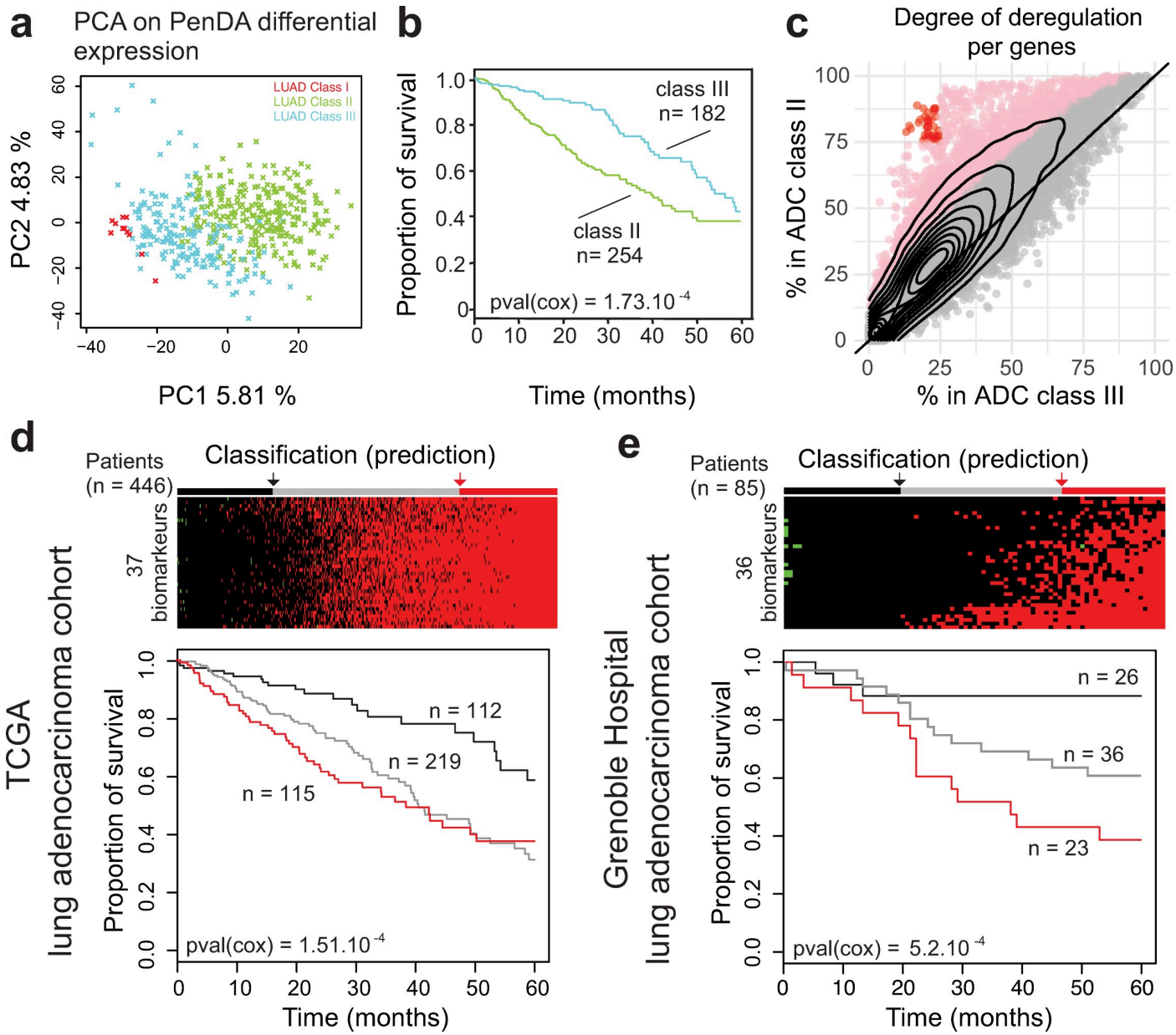


Fig 7. Upregulation of 37 genes in adenocarcinoma is a strong predictor of poor prognosis. (a) Principal Component Analysis on ADC cohort. Each cross represents an individual sample. The color of the dots represents the three subclasses defined in Fig 6. (b) Survival of ADC patients classified according to the 2 main subtypes (classes II and III). (c) The percentage of deregulated patients within the ADC class II (y-axis) or the ADC class III (x-axis). Each dot corresponds to one gene. The contour lines correspond to the density of genes. Pink dots indicate genes with a significant higher proportion of deregulation in the class II (proportion test, p-value < 0.05 after Bonferroni correction for multiple testing). Red dots define 37 genes highly deregulated (>75%) in the class II group and lowly deregulated (<25%) in the class III group. (d) (Top) Classification of ADC TCGA-LUAD built on the total number of up-regulated genes among the subset of 37 classifiers defined in (c). Patients are separated into 3 discrete groups: a group with a low upregulation (black, score < 4), a group with intermediate deregulation (gray, $4 \leq$ score < 34) and a group with most genes upregulated (red, $34 \leq$ score). (Bottom) Survival of patients according to these 3 groups. (e) As in (d) but for ADC Grenoble Hospital patients. Patients are separated into 3 discrete groups: a group with a low upregulation (black, score ≤ 0), a group with intermediate deregulation (gray, $0 <$ score < 15) and a group with most genes upregulated (red, $15 \leq$ score).

https://doi.org/10.1371/journal.pcbi.1007869.g007

individual level on simulated datasets. Applied to non-small-cell lung cancer expression data, our method showed that gene deregulation varies in a continuous manner between patients. When frequently deregulated, genes tend to commit to specific deregulation patterns (up or

down regulation). We observed that a small proportion of genes exhibits unusual 'super' deregulation pattern (always down, up or non-deregulated). Personalized differential analysis succeeds to properly cluster adenocarcinoma and squamous cells lung cancer histology. More specifically, clustering analysis leads to the identification of 37 biomarkers that efficiently predict 5-years survival in two independent adenocarcinoma cohorts. The method is available as an open source R package called *penda*. We provide user guidelines so that *penda* could be installed and run by users with limited computational experience. To ensure reproducibility of analysis, the *penda* vignette provides a summary of used parameters ready to be included in the method section of publications using PenDA.

PenDA is robust against different techniques of transcriptome analysis and against batch effects. Notably, the biomarkers that we identified on the ADCs TCGA cohorts based on an RNA-seq technology was validated on an independent ADC cohort where gene expressions were measured with microarrays. Another advantage of the method is that it is easily generalizable to other types of data like transcript expression, DNA methylation, proteomics, etc. For instance, several methods have been recently developed and benchmarked for the inference of isoform abundance from RNA-seq data [46]. However, classical differential expression analytical tools (on RNA-seq count data) are based on gene features and are not optimized for the estimate of transcripts abundance data. Thus, testing for individual differential isoform abundances with PenDA would be an interesting challenge. The PenDA approach could also be adapted for single cell analysis [47] to leverage the understanding of single cell expression and to quantify intra-sample heterogeneity at the single cell level.

The current PenDA method has however several limitations. First, though our method does not depend on replicates to identify individual deregulation, it relies on a control cohort that is supposed to reliably define a 'normal' ranking. Therefore, it is crucial to properly define suitable control datasets. Second, PenDA individual expression analysis requires the use of genome-wide transcriptomic data. In the future, we would like to explore the possibility to define a set of super conserved genes that could serve as internal reference for 'partial' PenDA analysis on sparse qPCR data. Third, our method is not suitable for genes with low expression levels in all samples, which are currently removed by filtering in the first step of the analysis.

The aim of population differential analysis is to detect consistently up or down regulated genes, *in average*. The PenDA method was based on the concept that individual level analyses are complementary of population approaches. Applying DESeq2, one of the most common DE analysis software, to the ADC and SQCC TCGA cohorts, highlighted similarities and differences for the genes with specific deregulation patterns identified by PenDA (super-conserved, super-up-regulated, super-down-regulated) (S10A and S10B Fig). For example, if all SU and SD genes were identified as differentially regulated by DESeq2 at the population level, many genes detected by DESeq2 as deregulated with a large fold-change and a low adjusted p-value are deregulated only in a limited subset of patients. Moreover, PenDA provides a unique way of identifying genes that are significantly never de-regulated (super-conserved), a category of genes hardly detectable by population methods. Similarly, compared to another meta-analysis of genetic deregulation at the population level in non-small-cell lung cancer based on microarray gene expression data [48], we observed that none of the three super-up genes common between SQCCs and ADCs (*PAFAH1B3*, *CBLC* and *ESRP1*) were identified as up-regulated by Tian et al, and only 28 of the 128 super-down-regulated genes common between SQCC and ADC were identified as down-regulated in the same study (S10C Fig). More surprisingly, *CD19* and *IL10*, two genes involved in the immune response and never deregulated in SQCCs and ADCs TCGA cohorts were identified as over-expressed by Tian *et al*. These comparisons suggest that applying the PenDA approach and identifying individual genetic deregulation patterns can bring new, complementary insights into the comprehensive analysis

of non-small-cell lung cancers or other types of cancers. In particular, genes displaying a 'super' profile can be considered as generic candidates for therapeutic strategies.

The PenDA method generates useful individual information that can be incorporated into further functional analysis. With PenDA, we provided generalized statistics at the level of a single individual/sample and at the level of a single gene (number of deregulated genes per tumor, number of tumors where a gene is deregulated, proportion of up-regulation if differentially regulated, etc.). At the gene level, this individual information can be combined to increase the power to detect significant association with phenotypic outcome, such as survival. As an illustration, we analyzed the synergic effect of gene deregulation of the GINS complex on survival, in the ADC cohort. GINS is a four-genes complex essential for initiation and elongation during DNA replication [49]. High expression of this complex has been related to tumorigenic properties [50]. ADC patients are heterogeneously deregulated for each of the GINS complex member, we classified them into three groups, based on PenDA differential analysis: (-): absence of gene deregulation for all the 4 constitutive genes; (+): 1 to 3 gene deregulations; and (+++): all genes are simultaneously deregulated (Fig 8A). Overall survival of ADC patients could be significantly discriminated using the synergic effect of GINS deregulation (Fig 8B), however, no significant effect of GINS deregulation could be identified using single gene Cox regression models (Fig 8C). This example demonstrates the interest of exploring possible synergic effects of single gene deregulations, in each individual. Besides survival analysis, single gene differential analysis could be profitably included into network analyses [51] to identify driver genes and functional communities. Moreover, a systematic exploration of the relationship between driver mutations [52] and individualized expression deregulations is a promising strategy to improve the accuracy of future pan-genomic studies.

Methods

Data and preprocessing

Two datasets of gene expression (HTSeq-Counts) were downloaded from The Cancer Genome Atlas program (https://portal.gdc.cancer.gov/). The datasets contain tumor ('01' barcoded

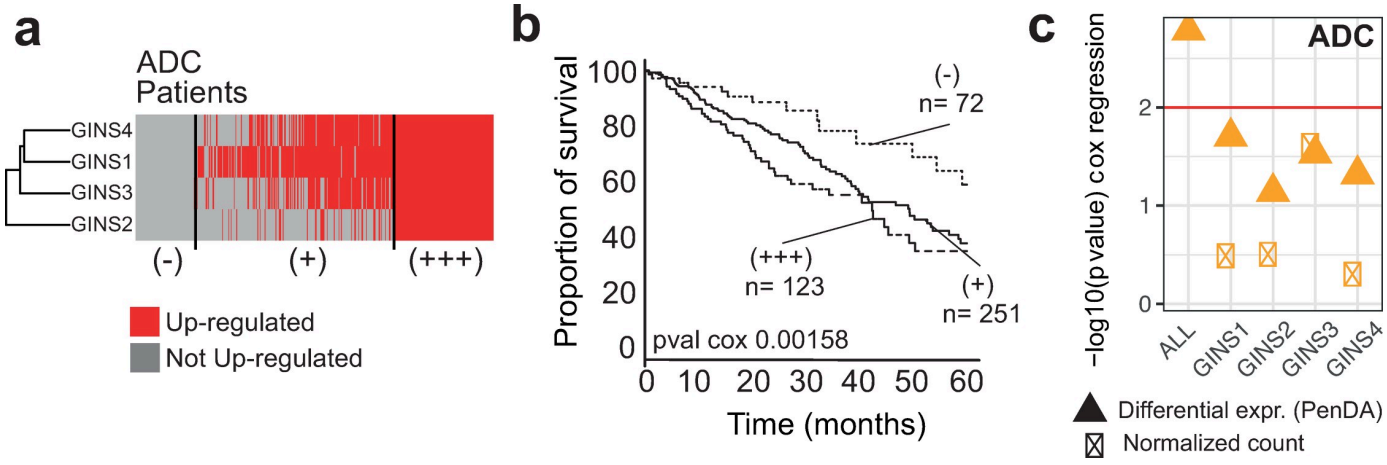


Fig 8. Synergic effects of gene deregulation within a protein complex. (a) Heatmap showing the distribution of gene deregulations of genes coding for the GINS complex in the ADC cohort. Patients are ordered from left to right according to an increasing number of gene deregulations within the GINS complex. The patients were separated into discrete deregulation groups of: 0 up-regulation (-), 1–3 up-regulations (+) and 4 up-regulations (+++). (b) Survival of ADC patients according to the deregulation groups defined in (a). (c) Cox regression p-values associated with different models (multivariate and univariate). Cox regression is applied on PenDA deregulation matrix (triangles) or expression matrix (ticked boxes, normalized count values). ALL corresponds to a multivariate cox model including the four genes of the GINS complex. The red line corresponds to the significance level of 0.01.

https://doi.org/10.1371/journal.pcbi.1007869.g008

samples) and control ('11' barcoded samples) tissues from two non-small cells lung cancers: lung adenocarcinoma (LUAD or ADC) and lung squamous cell (LUSC or SQCC). Patients with prior malignancies and replicated samples were removed from the analysis. We kept 1026 samples: 455 ADC tumors, 473 SQCC tumors and 98 control tissues consisting of normal adjacent lung tissue samples (50 from the ADC cohort, 48 from the SQCC cohort). For further analysis, we selected 19177 protein coding genes (hg38 reference genome). This corresponds to protein-coding genes of the base RefSeqGene (https://www.ncbi.nlm.nih.gov/refseq/rsg/). We then normalized the HTSeq-Counts using the *estimateSizeFactors* and the *count* functions of the DESeq2 package [38]. Finally, data were filtered to remove genes with null expression (counts = 0) in all samples (controls and tumors). At the end, we kept 18143 protein-coding genes.

The PenDA method

For each gene g, we first determined the lists $L(g)$ and $H(g)$ of other genes whose expressions are respectively lower or higher than that of g in at least 99% of the normal samples. These lists were next restricted to the subsets of l genes that have a median gene expression in normal samples closest to the corresponding median level of g, with l a user-defined parameter controlling the maximal size of L and H lists.

For a given tumor sample T, the personalized differential analysis was then performed iteratively:

a. For each gene g, we compared its expression $E(g,T)$ in the tumor T to the corresponding expression of genes present in the L and H lists. It allowed to defined four non-overlapping sets of genes:

$$L_d = \{g' \in L(g) \setminus E(g', T) < E(g, T)\}$$

$$L_u = \{g' \in L(g) \setminus E(g', T) > E(g, T)\}$$

$$H_d = \{g' \in H(g) \setminus E(g', T) < E(g, T)\}$$

$$H_u = \{g' \in H(g) \setminus E(g', T) > E(g, T)\}$$

$L_u \neq \emptyset$ or $H_d \neq \emptyset$ indicated than the relative ordering of g has changed in T compared to the control cases.

b. We considered that a gene g is deregulated in T if and only if

$$\left(\frac{|L_u|}{|L|} \geq h\right) \vee \left(\frac{|H_d|}{|H|} \geq h\right) \quad (1)$$

with $|X|$ the cardinality of ensemble X and h a user-defined parameter defining the minimal proportion of genes in L or H whose relative ordering with g has changed. If Eq (1) is satisfied then g is considered as down-regulated or up-regulated if $|L_d|+|H_d|<|L|$ or $|L_u|+|H_u|<|H|$ respectively. In the cases where the L or H lists are empty, we used the percentile method (see below) to take the decision on the status of g in T.

c. After having scanned all the genes, we aimed to minimize the potential bias that observed changes of ordering is actually due to the deregulation of genes in the L or H lists. Thus, we excluded in every L and H lists all the genes that had been diagnosed as deregulated in step

(b), and reiterated steps (a), (b) and (c) until convergence of the list of deregulated genes (S1 Fig, blue line), or until a user-specified number of iterations had been reached. It often happens that the final iterations oscillate between two lists (S1 Fig, red line). In this case, the union of both lists is considered as the predicted set of deregulated genes.

The percentile method

The percentile method consists in finding if the expression value of a gene in a test-sample is an outlier of the distribution of expression for the same gene within an ensemble of reference samples. More precisely, for each gene g, we determined p_l and p_u respectively the x and $(100\text{-}x)$ percentiles of the distribution of expression $E(g,S)$ for g within the ensemble of normal samples $\{S\}$, where x, given in %, is a user-tunable parameter. Then, a gene g in tumor sample T with an expression $E(g,T)$ was considered as differentially expressed in that sample if $E(g,T)<p_l/f$ (down-regulation) or $E(g,T)>p_u{}^*f$ (up-regulation), with $f\geq 1$ a user-defined factor allowing to expand the window of normal expression. A ROC curve analysis obtained by varying x and based on the simulated datasets (see below) suggested that using a factor $f\sim 1.2$ leads to an optimized diagnosis with this method (S2 Fig).

Simulated datasets

We generated realistic simulated datasets from the ensembles of normal and tumorous samples of the LUAD and LUSC TGCA studies. We first ranked all the gene expression values in normal samples and pooled them into consecutive packets. Each packet k contained 100 values of similar range $\{E(g_{k,1},S_{k,1}), E(g_{k,2},S_{k,2}),\ldots, E(g_{k,100},S_{k,100})\}$ with $E(g_{k,i},S_{k,i})$ the expression of gene $g_{k,i}$ in normal sample $S_{k,i}$. Then for each group, we computed the ensemble of expression differences in normal samples defined as $\Delta_n(k) = \{E(g_{k,i},S')-E(g_{k,i},S_{k,i}), 1\leq i\leq 100 \text{ and } \forall\ S'\neq S_{k,i}\}$. Similarly, we defined the ensemble of expression differences between tumorous and normal samples as $\Delta_c(k) = \{E(g_{k,i},T)-E(g_{k,i},S_{k,i}), 1\leq i\leq 100 \text{ and } \forall \text{ tumor } T\}$. From the 5% and 95% percentiles of $\Delta_n(k)$, noted $p_5(k)$ and $p_{95}(k)$ respectively, we isolated the subset $\Delta_c'(k)$ of values in $\Delta_c(k)$ that are smaller than $p_5(k)$ or greater than $p_{95}(k)$. We assumed that $\Delta_c'(k)$ represents typical abnormal expression differences observed in cancer for the packet k and that the ratio $r(k)$ between the number of elements in $\Delta_c'(k)$ and in $\Delta_c(k)$ is representative of the probability for a gene in this group to be deregulated.

Finally, to generate a simulated tumorous sample, we chose randomly one normal sample S. For each gene g, we determined the packet k containing $E(g,S)$ and its expression was modified with a probability $r(k)$ by adding a randomly-chosen element of $\Delta_c'(k)$. In average 30% of the genes were up or down-regulated. Instead of $r(k)$, we also used fixed proportions of deregulated genes from 0.05 to 0.9. We tested that the performance of PenDa on simulated datasets was not affected by the packet size (S11A Fig). The choice of the percentiles (5%, 95%) impacts on the ROC curves while PenDA still remains the best investigated methods in the low FPR range (S11B Fig).

Note that such strategies may be adapted to any data to generate realistic simulated datasets adapted to the user-defined system of interest.

Predictive power on simulated datasets

To test the efficacy of PenDA or of other methods, we generated 10 simulated tumors (see above). For each dataset, in order to realize a fair comparison, we excluded the normal sample from which it was generated to the ensemble of normal samples used to define the reference

properties of each method. For a given method and given parameters, true positive (TPR), false positive (FPR) and false discovery (FDR) rates were computed on these 10 simulations. ROC curves (TPR vs FPR) were obtained by varying one specific parameter for each method (threshold h for PenDA, percentile x for the percentile method, FDR level for Rankcomp and log2 fold change threshold for DESeq2). From each curve, we extracted the maximal informedness defined as the maximal value of the Youden's J statistics defined as the difference between TPR and FPR (TPR-FPR). An ideal predictive method would reach a maximal informedness of 1 while a random-decision method would approach 0 value.

In Fig 2, the effect of the number of control samples in the reference dataset (Fig 2B) and of the number of investigated genes (Fig 2C) were analyzed by randomly choosing a set of control samples or a set of genes from the initial pools and by repeating these operations 10 times. TPR and FPR levels were computed on the ensemble of simulations and of random choices. In Fig 3B, the ROC curves were determined for a set of 10 control samples randomly picked from the original pool. In Fig 3C, effect of normalization was simulated by multiplying RNA-seq counts of control and tumorous samples by random factors uniformly drawn between 1 and 5: the same factor was applied for all the genes of a given sample.

Estimation of the false discovery rate of RankComp from results given in Wang et al

In their original paper [15], Wang et al performed simulations to test the RankComp method. Each simulated sample contains $T = 15000$ genes including $P = 3000$ deregulated genes. In Table 2 of [15], they gave the sensitivity SE and specificity SP of the method for several simulations. From that, we can compute the corresponding false discovery rate $FDR = (T-P)(1-SP)/[(T-P)(1-SP)+P^{*}SE]$. Using this formula, the computed FDRs ranged from 20% to 50%.

PenDA analysis of the lung cancer cohort from the TCGA

The PenDA method was applied on preprocessed expression TCGA data (see Methods section: 'Data and preprocessing'). The PenDA vignette of the penda package version 1.0 was executed on 18143 genes, using 98 control samples and 928 case samples. The data set was pretreated as following: 0 gene and 0 sample were removed during the NA values filtering step, and 1034 gene was removed for low because lowly expressed: under the threshold 'val_min' = 10 in at least 99% of cases. 98 controls were used to generate L and H lists using the following parameters: threshold LH = 0.99 and s_max = 30. The penda method was then applied on 928 cases, with the following set of parameters: quantile = 0.02, factor = 1.2 and threshold = 0.3.

PenDA analysis of the lung cancer cohort from the Grenoble Hospital

The PenDA method was applied on expression data (Affymetrix Human Genome U133 Plus 2.0 Array) of the GSE30219 cohort. The PenDA vignette of the penda package version 1.0 was executed on 19148 genes, using 14 control samples and 293 case samples. The data set was pretreated as following: 0 gene and 0 sample were removed during the NA values filtering step, and 0 gene was removed for low because lowly expressed: under the threshold 'val_min' = 0.5 in at least 99% of cases. 14 controls were used to generate L and H lists using the following parameters: threshold LH = 0.99 and s_max = 100. The penda method was then applied on 293 cases, with the following set of parameters: quantile = 0.05, factor = 1.05 and threshold = 0.8.

Statistical analyses

Statistical analyses were performed on the following PenDA deregulation matrices, for S samples (tumors) and G genes:

- The upregulated matrix U_{mat} with $U_{mat}(g,T) = 1$ if gene g is up-regulated in tumor T (= 0 otherwise), with $T \in (1,\ldots,S)$ and $g \in (1,\ldots,G)$
- The downregulated matrix D_{mat} with $D_{mat}(g,T) = 1$ if gene g is down-regulated in tumor T (= 0 otherwise), with $T \in (1,\ldots,S)$ and $g \in (1,\ldots,G)$
- The matrix of total deregulation $Tot_{mat} \equiv U_{mat} + D_{mat}$.

a. Testing for equality of deregulation proportions (Fig 4) was performed using two-sided two-proportion z-test (prop.test function in R), with a Bonferroni corrected p-value threshold at $2.75.10^{-6}$ (corresponding to 18143 multiple testing).

b. Statistically significant deregulation frequency (Fig 5) was assessed by a t-statistic computed for each gene. The t-statistic was calculated using the R t.test function, with the vector of S values corresponding to the estimated differential expression x_{gT} for the gene g in each tumor T and the true value of the mean defined as $\left\{ mu = \frac{1}{G}\sum_{g=1}^{G}(\frac{1}{S}\sum_{T=1}^{S} x_{gT}),\ x \in \{0,1\} \right\}$. A calibrated p-value associated with the t-statistic and a corresponding q-value were then calculated using the R package fdrtool using the following parameters: cutoff.method = "pct0" and pct0 = 0.90 [53].
The test was applied on the Tot_{mat}. Super-up-regulated genes were defined as follows: (i) $\sum_{T=1}^{S} x_{gT}^{U} > median(\sum_{T=1}^{S} x_{T}^{Tot})$, ii) counts > 10 in at least 80% of the control samples and iii) significant t.test q-value. Super-down-regulated genes were defined as follows: (i) $\sum_{T=1}^{S} x_{gT}^{D} > median(\sum_{T=1}^{S} x_{T}^{Tot})$, ii) counts > 10 in at least 80% of the control samples and iii) significant t.test q-value. Super-conservde genes were defined as follows: (i) $\sum_{T=1}^{S} x_{gT}^{Tot} < median(\sum_{T=1}^{S} x_{T}^{Tot})$, ii) counts > 10 in at least 80% of the control samples and iii) significant t.test q-value.

c. PCA analysis (Fig 6) was performed using the function big_randomSVD of the R package bigstatr[54]. SVM linear regression was performed on the 2 firsts Principle Component of PCA analysis, using the function svm of the R package "e1071", using the following arguments: kernel = linear, cost = 10 and scale = FALSE.

Survival analyses

The R package survival was used to compute Cox-models and create 5-years survival curve (Fig 6 and Fig 7). The *survival*::*coxph* function was used to fit a Cox proportional hazard regression model and the overall likelihood ratio p-value was extracted for further analysis. The *survival*::*survfit* function was used to create survival curves from the Kaplan-Meier estimate.

Gene functional classification

Gene functional classification was performed using the DAVID's Functional Annotation tool of David Bioinformatics Resources 6.8 [55,56]. Enrichment analyses for gene lists of interest were performed against Gene Ontology term–Biological Pathway (direct) repository.

Heatmaps summarizing the results were generated from Functional Annotation Chart, after applying a cutoff of 0.001 on the Modified Fisher Exact P-Value (we used the tutorial kindly provided by Kevin Blighe).

Use of Rankcomp

The original Rankcomp and the RankcompV2 algorithms [15,16] were tested using the Relative Expression Ordering Analysis (REOA) package downloaded from https://github.com/pathint/reoa. We ran the program *reoa* on our simulated datasets using the options*–s 1 –j 2 –a 2* to get individual predictions for both algorithms with default parameters. Results for different FDR levels were obtained using the *–f* option.

Use of DESeq2

The R-package of DESeq2 [7] was imported from Bioconductor3.7. To assess fold changes in expression from simulated datasets, we used DESeq2 default parameters. We performed 10 comparisons between individual simulated tumor sample and 97 independent TCGA control samples (we remove the control sample used for simulating the tumor sample from the reference dataset). As no replicate was available for tumor sample, DESeq2 allowed the variance-mean dependence estimated from control samples to be used for case sample [57]. The log2-foldChange estimation was used for sensitivity and specificity analysis. Performing DESeq2 with or without its internal normalization routine may impact the ROC analysis in particular if data are not standardized (S11C Fig). To assess fold changes in expression from TCGA datasets, we applied DESeq2 methods with default parameters, except for the significance cutoff which was set to 0.01 (alpha value of the *DESeq2*::*results* function).

Supporting information

S1 Fig. Convergence towards a consistent list of deregulated genes is rapidly achieved by the PenDA method. We plotted the evolution of the total number of predicted deregulated genes during the successive iterations of the PenDA method applied to one simulated dataset with $l = 30$ and $h = 0.1$ (red line) or $h = 0.4$ (cyan line).
(PDF)

S2 Fig. Test of the percentile method. (a) ROC curve of the percentile method obtained by varying parameter x for different values of factor f. TPR and FPR were computed on a set of 10 simulations. (b) Maximal informedness of the ROC curve as a function of f.
(PDF)

S3 Fig. PenDA workflow.
(PDF)

S4 Fig. Effect of tumor stages on gene deregulations for ADC and SQCC patients. (a,b) Effect of tumor stages on the total number of gene deregulations in both ADC (a) and SQCC patients (b). (c,d) Effect of tumor stages on down/up ratios in both ADC (c) and SQCC patients (d). Significance was assessed via one-way ANOVA with Tukey's multiple comparison post hoc test, considering the stage as an independent factor, with 5 different levels (stage ia, stage ib, stage ii, stage iii and stage iv). LUAD gene deregulation: Df = 4, F-statistic = 5.18, p-value = 0.0004. LUAD deregulation ratio: Df = 4, F-statistic = 0.99, p-value = 0.4128. LUSC gene deregulation: Df = 4, F-statistic = 3.00, p-value = 0.0182. LUAD deregulation ratio: Df = 4, F-statistic = 1.59, p-value = 0.1767. Dashed red lines represent 1st quartile, median and

3rd quartile of the distributions.
(PDF)

S5 Fig. Gene Ontology (biological pathways) enrichment. GO (biological pathways) enrichment in genes significantly down in ADC compared to SQCC (a), significantly up in ADC compared to SQCC (b), significantly down in SQCC compared to ADC (c) and significantly up in SQCC compared to ADC (d). 1000 top hits of prop.test analysis were used to estimate terms enrichment in each condition. Rows of the heatmap correspond to genes overlapping with at least one enriched term (red). Genes with no overlapping terms were removed from the graphical representation. Columns correspond to enriched terms clustered by Euclidean distance. GO Term significance score corresponds to -log10 of the Modified Fisher Exact P-Value after Benjamini correction (extracted from DAVID's Functional Annotation tool).
(PDF)

S6 Fig. Gene deregulation commitment in ADCs and SQCCs. Genes deregulated in more than 30% of the patient are depicted in the diagram. x-axis corresponds to the % of up-regulation/total-deregulation in ADC, y-axis corresponds to the % of up-regulation/ total-deregulation in SQCC. Each dot (gray cross) corresponds to one gene. Blue points correspond to super-down-regulated genes, red points correspond to super-up-regulated genes (triangles for ADC, circles for SQCC). Diamonds black points represent genes displaying antagonistic commitment behaviour between ADC and SQCC (~5% of the total number of genes depicted).
(PDF)

S7 Fig. Gene Ontology (biological pathways) enrichment in genes super-up-regulated and genes super-down-regulated. GO (biological pathways) analysis for super-up-regulated (a) and super-down-regulated (b) genes in ADC or SQCC. Rows of the heatmap correspond to genes overlapping with at least one enriched term (red). Genes with no overlapping terms were removed from the graphical representation. Columns correspond to enriched terms clustered by Euclidean distance. GO Term significance score corresponds to -log10 of the Modified Fisher Exact P-Value after Benjamini correction (extracted from DAVID's Functional Annotation tool).
(PDF)

S8 Fig. Comparison of ADC subclasses obtained from PenDA analysis with clustering analysis on normalized counts analysis and with ADC iClusters. (a) Graphical representation of the contingency table between ADC iCluster (The Cancer Genome Atlas Research Network) and ADC subclasses (PenDA analysis). (b) Heatmap of normalized counts matrix applied to a specific set of classifier genes (n = 875) in TCGA non-small-cell lung cancers: ADC (orange) and SQCC (purple). Two hierarchical clusterings were performed: using Euclidean distance to sort genes and using Pearson correlation-based distance to classify patients, with a complete linkage function in both cases. ADC subclasses defined by PenDA analysis (colour-coded, class I to III) are defined according to Fig 6 of the main text.
(PDF)

S9 Fig. Effect of putative confounding factors on ADC classification in class II and III. (a) Effect of cancer stage patients (chi-square test p-value = 0.2133). (b) Effect of gender (chi square test p-value = 1). (c) Effect of age patients (chi square test p-value = 0.2133).
(PDF)

S10 Fig. DESeq2 analysis of the ADC and SQCC TCGA cohorts. DESeq2 analysis of the ADC (a) and SQCC (b) TCGA cohorts (green triangles: super-conserved genes, red triangles: super-up-regulated genes, blue triangles: super-down-regulated genes). (c) Genes identified as

deregulated by Tian et al. x-axis corresponds to normalized mean expression in controls. y-axis corresponds to normalized mean expression in tumor. Gene with super patterns identified with PenDA are depicted with triangles (green triangles: super-conserved genes, red triangles: super-up-regulated genes, blue triangles: super-down-regulated genes).
(PDF)

S11 Fig. Effect of the simulation method parameters on PenDA performance. (a) ROC curves (true positive rate TPR vs false positive rate FPR) of the PenDA method on different simulated datasets obtained with different packet sizes. The curves were obtained by varying the proportion threshold h. (b) Comparison with other methods. ROC curves on simulated datasets generated using percentiles 10%,90% (Left) and 20%,80% (Right) in the simulation method (see Methods in the main text) for PenDA, a simple percentile-based method, 2 versions of Rankcomp and DESeq2. (c) ROC curves of the full DESeq2 method (full lines) or with the DESeq2 method skipping the internal routine for normalization (dashed lines) for the same simulated dataset used in Fig 2 and Fig 3A of the main text (green) or with the non-standardized dataset used in Fig 3C of the main text.
(PDF)

S1 Text. PenDA method vignette: Vignette_penda.
(PDF)

S2 Text. PenDA simulation vignette: Vignette_simulation.
(PDF)

S1 Table. Gene deregulation per sample.
(CSV)

S2 Table. Genetic deregulation profiles.
(CSV)

S3 Table. ADC clusters.
(CSV)

S4 Table. The list of 37 biomarkers.
(CSV)

Acknowledgments

We thank the members of the DJ's and SK's groups, in particular Ekaterina Flin, for inspiring discussions during regular joint group meetings. We are grateful to Florian Privé, Michael Blum, Eric Fanchon and members of the BCM team for algorithmic and methodological advices. We acknowledge computational resources from CIMENT infrastructure (supported by the Rhone-Alpes region, Grant CPER07 13 CIRA). EB thanks Centre de ressources (CRB) CHUGA Grenoble, French Ligue contre le cancer for transcriptomic platform and Nicolas Lemaitre for tumor data management.

Author Contributions

Conceptualization: Magali Richard, Daniel Jost.

Data curation: Magali Richard, Clémentine Decamps, Florent Chuffart.

Formal analysis: Magali Richard, Clémentine Decamps, Daniel Jost.

Funding acquisition: Magali Richard, Elisabeth Brambilla, Saadi Khochbin, Daniel Jost.

Investigation: Magali Richard, Clémentine Decamps, Florent Chuffart, Elisabeth Brambilla, Sophie Rousseaux, Saadi Khochbin, Daniel Jost.

Methodology: Magali Richard, Clémentine Decamps, Daniel Jost.

Project administration: Magali Richard, Daniel Jost.

Resources: Magali Richard, Clémentine Decamps, Florent Chuffart.

Software: Magali Richard, Clémentine Decamps, Florent Chuffart.

Supervision: Magali Richard, Daniel Jost.

Validation: Magali Richard, Clémentine Decamps, Florent Chuffart, Elisabeth Brambilla, Sophie Rousseaux, Daniel Jost.

Visualization: Magali Richard, Clémentine Decamps, Florent Chuffart, Daniel Jost.

Writing – original draft: Magali Richard, Clémentine Decamps, Daniel Jost.

Writing – review & editing: Magali Richard, Daniel Jost.

References

1. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2014; 24:14–24. https://doi.org/10.1101/gr.155192.113 PMID: 24092820
2. Lu Y-F, Goldstein DB, Angrist M, Cavalleri G. Personalized Medicine and Human Genetic Diversity. Cold Spring Harb Perspect Med. Cold Spring Harbor Laboratory Press; 2014; 4:a008581–1. https://doi.org/10.1101/cshperspect.a008581 PMID: 25059740
3. Evans WE, science MR, 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. science.sciencemag.org
4. Rousseaux S, Debernardi A, Jacquiau B, Vitte A-L, Vesin A, Nagy-Mignotte H, et al. Ectopic Activation of Germline and Placental Genes Identifies Aggressive Metastasis-Prone Lung Cancers. Sci Transl Med. American Association for the Advancement of Science; 2013; 5:186ra66–6. https://doi.org/10.1126/scitranslmed.3005723 PMID: 23698379
5. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501. https://doi.org/10.1038/nature12912 PMID: 24390350
6. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014; 158:929–44. https://doi.org/10.1016/j.cell.2014.06.049 PMID: 25109877
7. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43:e47–7. https://doi.org/10.1093/nar/gkv007 PMID: 25605792
10. Mutch DM, Berger A, Mansourian R, Rytz A, Roberts M-A. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. BMC Bioinformatics. BioMed Central; 2002; 3:17. https://doi.org/10.1186/1471-2105-3-17 PMID: 12095422
11. Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends in Biotechnology. Elsevier Current Trends; 2017; 35:498–507. https://doi.org/10.1016/j.tibtech.2017.02.012 PMID: 28351613
12. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics. 2018; 19:776–92. https://doi.org/10.1093/bib/bbx008 PMID: 28334202

13. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. BioMed Central; 2015; 16:347.
14. Vitali F, Li Q, Schissler AG, Berghout J, Kenost C, Lussier YA. Developing a "personalome" for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. Briefings in Bioinformatics. 2017; 63:2889.
15. Wang H, Sun Q, Zhao W, Qi L, Gu Y, Li P, et al. Individual-level analysis of differential expression of genes and pathways for personalized medicine. Bioinformatics. 2015; 31:62–8. https://doi.org/10.1093/bioinformatics/btu522 PMID: 25165092
16. Li X, Cai H, Wang X, Ao L, Guo Y, He J, et al. A rank-based algorithm of differential expression analysis for small cell line data with statistical control. Briefings in Bioinformatics. 2019; 20:482–91. https://doi.org/10.1093/bib/bbx135 PMID: 29040359
17. Guan Q, Chen R, Yan H, Cai H, Guo Y, Li M, et al. Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. Oncotarget. 2016; 7:68909–20. https://doi.org/10.18632/oncotarget.11996 PMID: 27634898
18. Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, et al. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. Briefings in Bioinformatics. 2016; 17:233–42. https://doi.org/10.1093/bib/bbv064 PMID: 26254430
19. Zhou X, Li B, Zhang Y, Gu Y, Chen B, Shi T, et al. A relative ordering-based predictor for tamoxifen-treated estrogen receptor-positive breast cancer patients: multi-laboratory cohort validation. Breast Cancer Res. Treat. 2013; 142:505–14. https://doi.org/10.1007/s10549-013-2767-8 PMID: 24253811
20. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010; 26:136–8. https://doi.org/10.1093/bioinformatics/btp612 PMID: 19855105
21. Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research. 2015; 43: e140. https://doi.org/10.1093/nar/gkv711 PMID: 26184878
22. Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics. 2012; 28:2782–8. https://doi.org/10.1093/bioinformatics/bts515 PMID: 22923299
23. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA: A Cancer Journal for Clinicians. 2019; 69:7–34.
24. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543–50. https://doi.org/10.1038/nature13385 PMID: 25079552
25. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. https://doi.org/10.1038/nature11404 PMID: 22960745
26. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016; 48:607–16. https://doi.org/10.1038/ng.3564 PMID: 27158780
27. Chen F, Zhang Y, Parra E, Rodriguez J, Behrens C, Akbani R, et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. Oncogene. 2017; 36:1384–93. https://doi.org/10.1038/onc.2016.303 PMID: 27775076
28. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research. 2012; 22:2109–19. https://doi.org/10.1101/gr.145144.112 PMID: 22975805
29. Testa U, Castelli G, Pelosi E. Lung Cancers: Molecular Characterization, Clonal Heterogeneity and Evolution, and Cancer Stem Cells. Cancers (Basel). Multidisciplinary Digital Publishing Institute; 2018; 10:248.
30. Villalobos P, Wistuba II. Lung Cancer Biomarkers. Hematol. Oncol. Clin. North Am. 2017; 31:13–29. https://doi.org/10.1016/j.hoc.2016.08.006 PMID: 27912828
31. Tang H, Wang S, Xiao G, Schiller J, Papadimitrakopoulou V, Minna J, et al. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. Ann. Oncol. 2017; 28:733–40. https://doi.org/10.1093/annonc/mdw683 PMID: 28200038
32. Shea M, Costa DB, Rangachari D. Management of advanced non-small cell lung cancers with known mutations or rearrangements: latest evidence and treatment approaches. Ther Adv Respir Dis. SAGE PublicationsSage UK: London, England; 2016; 10:113–29. https://doi.org/10.1177/1753465815617871 PMID: 26620497
33. Tsao M-S, Marguet S, Le Teuff G, Lantuejoul S, Shepherd FA, Seymour L, et al. Subtype Classification of Lung Adenocarcinoma Predicts Benefit From Adjuvant Chemotherapy in Patients Undergoing

Complete Resection. J. Clin. Oncol. American Society of Clinical Oncology; 2015; 33:3439–46. https://doi.org/10.1200/JCO.2014.58.8335 PMID: 25918286

34. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol. 2011; 6:244–85. https://doi.org/10.1097/JTO.0b013e318206a221 PMID: 21252716

35. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol. 2015; 10:1243–60. https://doi.org/10.1097/JTO.0000000000000630 PMID: 26291008

36. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin. Cancer Res. 2010; 16:4864–75. https://doi.org/10.1158/1078-0432.CCR-10-0199 PMID: 20643781

37. Hayes DN, Monti S, Parmigiani G, Clinical CGJO, 2006. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. cancer.unc.edu

38. Love M, Anders S, Huber W. Differential analysis of count data–the DESeq2 package. Genome Biology; 2014.

39. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong K-K. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat. Rev. Cancer. 2014; 14:535–46. https://doi.org/10.1038/nrc3775 PMID: 25056707

40. George J, Walter V, Peifer M, Alexandrov LB, Seidel D, Leenders F, et al. Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. Nat Commun. 2018; 9:1048. https://doi.org/10.1038/s41467-018-03099-x PMID: 29535388

41. Zhang S, Li M, Ji H, Fang Z. Landscape of transcriptional deregulation in lung cancer. BMC Genomics. 2018; 19:435. https://doi.org/10.1186/s12864-018-4828-1 PMID: 29866045

42. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009; 41:1238–42. https://doi.org/10.1038/ng.465 PMID: 19801978

43. Massion PP, Taflan PM, Rahman SMJ, Yildiz P, Shyr Y, Edgerton ME, et al. Significance of p63 Amplification and Overexpression in Lung Cancer Development and Prognosis. Cancer Res. American Association for Cancer Research; 2003; 63:7113–21. PMID: 14612504

44. Mar N, Vredenburgh JJ, Wasser JS. Targeting HER2 in the treatment of non-small cell lung cancer. Lung Cancer. 2015; 87:220–5. https://doi.org/10.1016/j.lungcan.2014.12.018 PMID: 25601485

45. Shtivelman E, Hensing T, Simon GR, Dennis PA, Otterson GA, Bueno R, et al. Molecular pathways and therapeutic targets in lung cancer. Oncotarget. Impact Journals, LLC; 2014; 5:1392–433. https://doi.org/10.18632/oncotarget.1891 PMID: 24722523

46. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. BioMed Central; 2015; 16:150. https://doi.org/10.1186/s13059-015-0702-5 PMID: 26201343

47. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Meth. Nature Publishing Group; 2018; 15:255–61.

48. Tian Z-Q, Li Z-H, Wen S-W, Zhang Y-F, Li Y, Cheng J-G, et al. Identification of Commonly Dysregulated Genes in Non-small-cell Lung Cancer by Integrated Analysis of Microarray Data and qRT-PCR Validation. Lung. 2015; 193:583–92. https://doi.org/10.1007/s00408-015-9726-6 PMID: 25851596

49. MacNeill SA. Structure and function of the GINS complex, a key component of the eukaryotic replisome. Biochem. J. Portland Press Limited; 2010; 425:489–500. https://doi.org/10.1042/BJ20091531 PMID: 20070258

50. Nagahama Y, Ueno M, Miyamoto S, Morii E, Minami T, Mochizuki N, et al. PSF1, a DNA Replication Factor Expressed Widely in Stem and Progenitor Cells, Drives Tumorigenic and Metastatic Properties. Cancer Res. American Association for Cancer Research; 2010; 70:1215–24. https://doi.org/10.1158/0008-5472.CAN-09-3662 PMID: 20103637

51. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia YY, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Research. Oxford University Press; 2015; 43:e44–4. https://doi.org/10.1093/nar/gku1393 PMID: 25572314

52. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173:371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060 PMID: 29625053

53. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics. 2008; 24:1461–2. https://doi.org/10.1093/bioinformatics/btn209 PMID: 18441000

54. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Stegle O, editor. Bioinformatics. 2018; 34:2781–7. https://doi.org/10.1093/bioinformatics/bty185 PMID: 29617937

55. Da Wei Huang, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. Nature Publishing Group; 2009; 4:44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956

56. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009; 37:1–13. https://doi.org/10.1093/nar/gkn923 PMID: 19033363

57. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11: R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

Annexe 2

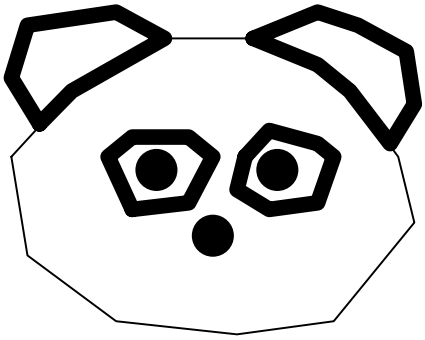
Vignette Penda pour les simulations : Advanced User - Performing simulated personalized data analysis with penda

PENDA: PErsoNalized Differential Analysis

Advanced User - Performing simulated personalized data analysis with `penda`

Magali Richard, Clementine Decamps, Florent Chuffart, Daniel Jost

2019-06-06



Introduction

`penda` (**PE**rso**N**alized **D**ifferential **A**nalysis) is an open-access R package that detects gene deregulation in individual samples compared to a set of reference, control samples. This tutorial aims at providing to non-expert users basic informations and illustrations on how to run the package.

How to cite: Richard et al. (2019) PenDA, a rank-based method for Personalized Differential Analysis: application to lung cancer, in submission.

Dataset and data filtering

Dataset

The dataset used to illustrated the method corresponds to the transcriptomes of 3000 genes (RNAseq counts, normalized with DESeq2) for 40 normal, control samples and 40 tumorous samples taken from the TCGA study of lung adenocarcinoma [PMID:25079552].

`data_ctrl` is a data matrix containing the normalized counts of each control sample. The rownames of the matrix correspond to the gene_symbol, the colnames indicate the sample ID.

```
data_ctrl = penda::penda_data_ctrl
head(data_ctrl[,1:3])
#>       patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> AADAC           347.2489           428.5498           442.0555
#> AAMP            965.2342          1528.3221           968.0266
#> ABCA1             0.0000             0.0000             0.0000
#> ABL1           1508.1784          1227.1325          1747.2431
#> ABL2            582.6719           645.4063           488.5088
#> ACACA             0.0000             0.0000             0.0000
dim(data_ctrl)
#> [1] 3000   40
```

`data_case` is a data matrix containing the normalized counts of each tumor sample. The rownames of the matrix correspond to the gene_symbol, the colnames indicate the sample ID.

```
data_case = penda::penda_data_case
data_case = data_case[rownames(data_ctrl),]
head(data_case[,1:3])
#>       patient_69-7764-01 patient_44-3919-01 patient_86-8278-01
#> AADAC          311.2129           374.9473          445.43169
#> AAMP          1466.5906           979.2256         1059.19225
#> ABCA1            0.0000             0.0000            0.00000
#> ABL1          2676.4306          2065.7474         2503.76905
#> ABL2          1167.0482           678.5603         1263.94317
#> ACACA            0.0000             0.0000           12.79693
dim(data_case)
#> [1] 3000   40
```

Note: this vignette is an example that has been designed for a rapid test of the method. So we limit the number of genes and the number of samples for this purpose. For an optimal utilization of the method, users should however upload all their available data (genes, control and case samples).

Extraction of data for simulations

The optimal choice of parameters (Sec. 4.2) is based on simulations that perturb control samples (Sec. 4.1). Here, we extract three patients (`data_simu`) from `data_ctrl` that will be used later for this purpose. For consistency, we also discard them from the `data_ctrl` matrix that will serve as reference.

```
data_simu = data_ctrl[,1:3]
data_ctrl = data_ctrl[,-(1:3)]
head(data_simu[,1:3])
#>       patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> AADAC          347.2489           428.5498           442.0555
#> AAMP           965.2342          1528.3221           968.0266
#> ABCA1            0.0000             0.0000             0.0000
#> ABL1          1508.1784          1227.1325          1747.2431
#> ABL2           582.6719           645.4063           488.5088
#> ACACA            0.0000             0.0000             0.0000
dim(data_simu)
#> [1] 3000    3
dim(data_ctrl)
#> [1] 3000   37
```

Note: this vignette is an example that has been designed for a rapid test of the method. For a more complete analysis and a better parameter estimation, we recommend users to simulate more cases (10 for example instead of 3).

Data filtering

```
threshold_dataset = 0.99
Penda_dataset = penda::make_dataset(data_ctrl, data_case, detectlowvalue = TRUE,
    detectNA = TRUE, threshold = threshold_dataset)
#> [1] "0 probes are NA in at least 99 % of the samples."
#> [1] "0 patients have NA for at least 99 % of the probes."
#> [1] "Computing of the low threshold"
```

```
#> number of iterations= 97
#> [1] "159 genes have less than 23.11777362263 48 counts in 99 % of the samples."
data_ctrl = Penda_dataset$data_ctrl
data_case = Penda_dataset$data_case
data_simu = data_simu[rownames(data_ctrl), ]
```

The function `make_dataset` contains three steps to prepare the data for the analysis.

- `detect_na_value` removes rows and columns (ie, genes and samples) of the data matrices that contain more than threshold % (default value = 0.99) of NA (Not Available) value.
- `detect_zero_value` removes genes with very low expression in the majority of samples (controls and cases), *ie.* genes whose expression is lower than `val_min` in `threshold`% of all the samples. By default it uses the function `normalmixEM` to estimate the value of `val_min` using all the *log2*-transformed count data but this parameter can also be tuned manually by the user.
- `rank_genes` sorts the genes based on the median value of gene expression in controls. This step is essential for the proper functioning of `penda`.

```
head(data_ctrl[,1:3])
#>         patient_77-8007-11 patient_55-6969-11 patient_22-4609-11
#> CAPZB            0.0000000           1.541098          0.8957864
#> CEACAM4          0.0000000           0.000000          0.0000000
#> DHX8             0.0000000           0.000000          0.0000000
#> DTNB             4.2227541           1.541098          0.0000000
#> EZH1             0.8445508           1.541098          0.0000000
#> GRIA1            0.0000000           0.000000          0.0000000
dim(data_ctrl)
#> [1] 2841   37
head(data_case[,1:3])
#>         patient_69-7764-01 patient_44-3919-01 patient_86-8278-01
#> CAPZB              1.29672          0.5895398          0.4921897
#> CEACAM4            0.00000          0.0000000          0.0000000
#> DHX8               0.00000          0.5895398          0.0000000
#> DTNB               0.00000          6.4849375          0.4921897
#> EZH1               0.00000          0.5895398         16.2422604
#> GRIA1              0.00000          0.5895398          0.0000000
dim(data_case)
#> [1] 2841   40
head(data_simu[,1:3])
#>         patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> CAPZB             1.471394           0.000000           2.996986
#> CEACAM4           0.000000           0.000000           0.000000
#> DHX8              0.000000           0.000000           0.000000
#> DTNB              3.678484          10.326501           1.498493
#> EZH1              1.471394           3.442167           0.000000
#> GRIA1             0.000000           0.000000           0.000000
dim(data_simu)
#> [1] 2841    3
```

Relative gene ordering

```
threshold_LH = 0.99
s_max = 30
```

```
L_H_list = penda::compute_lower_and_higher_lists(data_ctrl, threshold = threshold_LH,
    s_max = s_max)
#> [1] "Computing genes with lower and higher expression"
L = L_H_list$L
H = L_H_list$H
```

The `penda` method uses the relative gene ordering in normal tissue.

The function `compute_lower_and_higher_lists` computes two matrices `L` and `H` based on the filtered control dataset (`data_ctrl`).

Each row of the **L** matrix contains a list of at most `s_max` (default value = 30) genes (characterized by their ids) whose expressions are **lower** than that of the gene associated to the corresponding row, in at least `threshold_LH` (default value = 99 %) of the control samples.

Each row of the **H** matrix contains a list of at most `s_max` (default value = 30) genes (characterized by their ids) whose expressions are **higher** than that of the gene associated to the corresponding row, in at least `threshold_LH` (default value = 99 %) of the control samples.

Below, for some genes (FOXH1, KRTAP2-3, etc.), we show the id of 10 genes of the L and H lists.

```
L[1000:1005,1:10]
#>           [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
#> DDOST      783  705  685  684  677  675  660  652  649   641
#> SMCO1      776  768  760  753  742  740  733  725  721   715
#> KRTAP2-3   868  864  851  849  847  828  816  813  809   804
#> SIGLEC16   813  804  798  797  788  787  770  763  753   746
#> KIAA0895L  827  815  804  798  796  791  788  787  785   782
#> IQSEC1     857  807  799  782  777  775  768  764  763   760
H[1000:1005,1:10]
#>           [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
#> DDOST     1300 1485 1520 1576 1615 1756 1767 1809 1889  1905
#> SMCO1     1304 1336 1343 1346 1349 1351 1355 1363 1368  1376
#> KRTAP2-3  1180 1181 1189 1195 1198 1200 1202 1203 1214  1218
#> SIGLEC16  1271 1282 1319 1323 1328 1332 1333 1334 1339  1343
#> KIAA0895L 1329 1398 1460 1495 1498 1508 1522 1523 1525  1526
#> IQSEC1    1223 1248 1306 1309 1320 1328 1338 1359 1361  1366
dim(L)
#> [1] 2841   30
dim(H)
#> [1] 2841   30
```

Define optimal parameters from simulations

Generation of the simulated dataset

Estimation of optimal parameters adapted to the user data is based on a ROC analysis on simulated datasets. The function `complex_simulation` uses the real distribution of difference between the control and the case samples to simulate the proportion and the value of the dysregulation (see the original paper for details on the method of simulation).

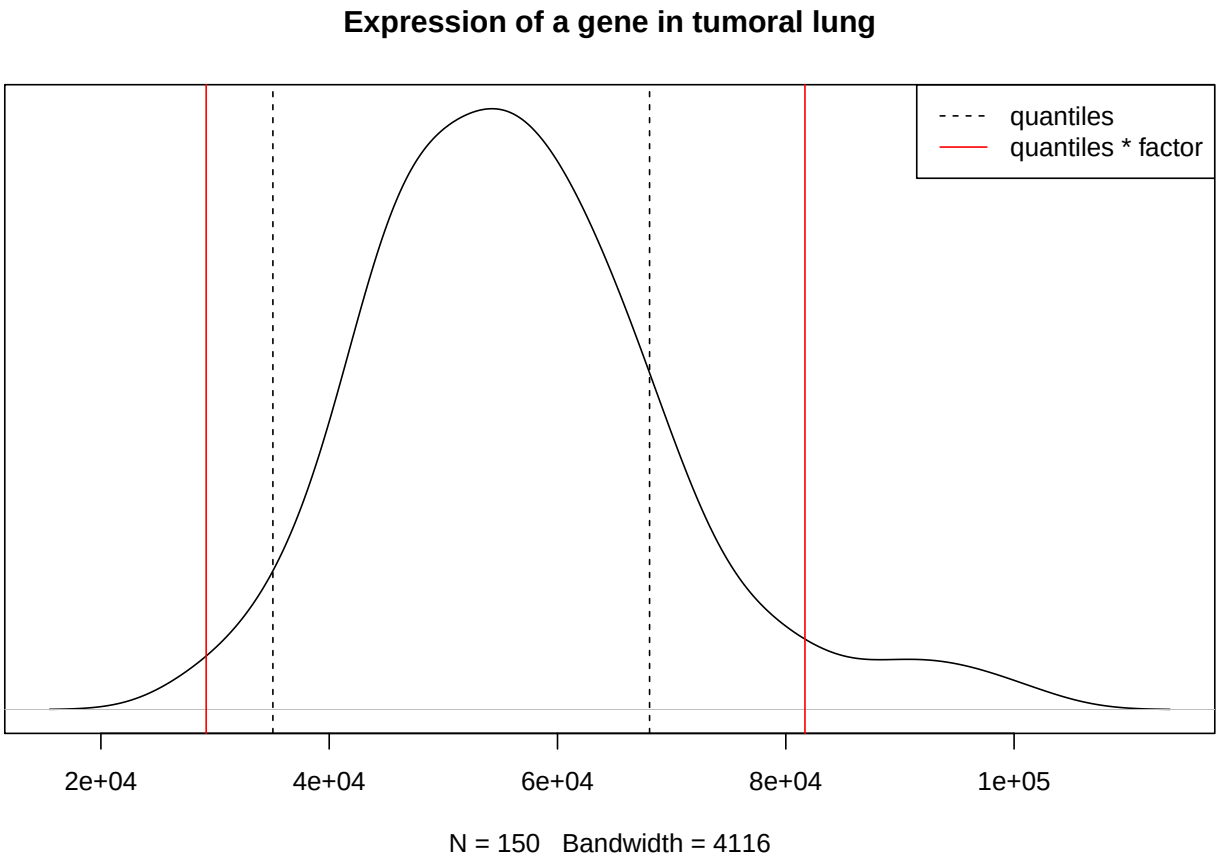
It returns the vector of initial data (`data_simu`, in `simulation$initial_data`), the vector of data with modifications (`simulation$simulated_data`) and the index of modified data (`simulation$changes_idx`).

```r
size_grp = 100
quant_simu = 0.05
simulation = penda::complex_simulation(data_ctrl, data_case,
    data_simu, size_grp, quant = quant_simu)
#> [1] "Computing genes groups"
#> [1] "Simulating dysregulation of 3 patients."
head(simulation$initial_data)
#>         patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> CAPZB             1.471394           0.000000           2.996986
#> CEACAM4           0.000000           0.000000           0.000000
#> DHX8              0.000000           0.000000           0.000000
#> DTNB              3.678484          10.326501           1.498493
#> EZH1              1.471394           3.442167           0.000000
#> GRIA1             0.000000           0.000000           0.000000
head(simulation$simulated_data)
#>         patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> CAPZB             1.471394           0.000000          81.700433
#> CEACAM4         112.867321           0.000000          15.425735
#> DHX8              0.000000           0.000000           0.000000
#> DTNB              3.678484          10.326501           1.498493
#> EZH1              1.471394           3.442167          15.084564
#> GRIA1             0.000000           0.000000           0.000000
```

```r
simulation$changes_idx[1000:1005]
#> [1] 3528 3533 3534 3535 3537 3538
#Before simulation:
simulation$initial_data[simulation$changes_idx[1000:1005]]
#> [1] 332.16910 154.89751 521.48828 103.26501  99.82284 103.26501
#After simulation:
simulation$simulated_data[simulation$changes_idx[1000:1005]]
#> [1]  858.020717   20.398067 1246.969691  973.211691   12.512462    9.348681
```

Optimal parameter choice

For the quantile method



In the rare cases where the lists L or H of a gene are empty, `penda` uses a simpler, less efficient, method based on quantile to determine the deregulation status (see the original paper). This quantile method depends on two parameters:

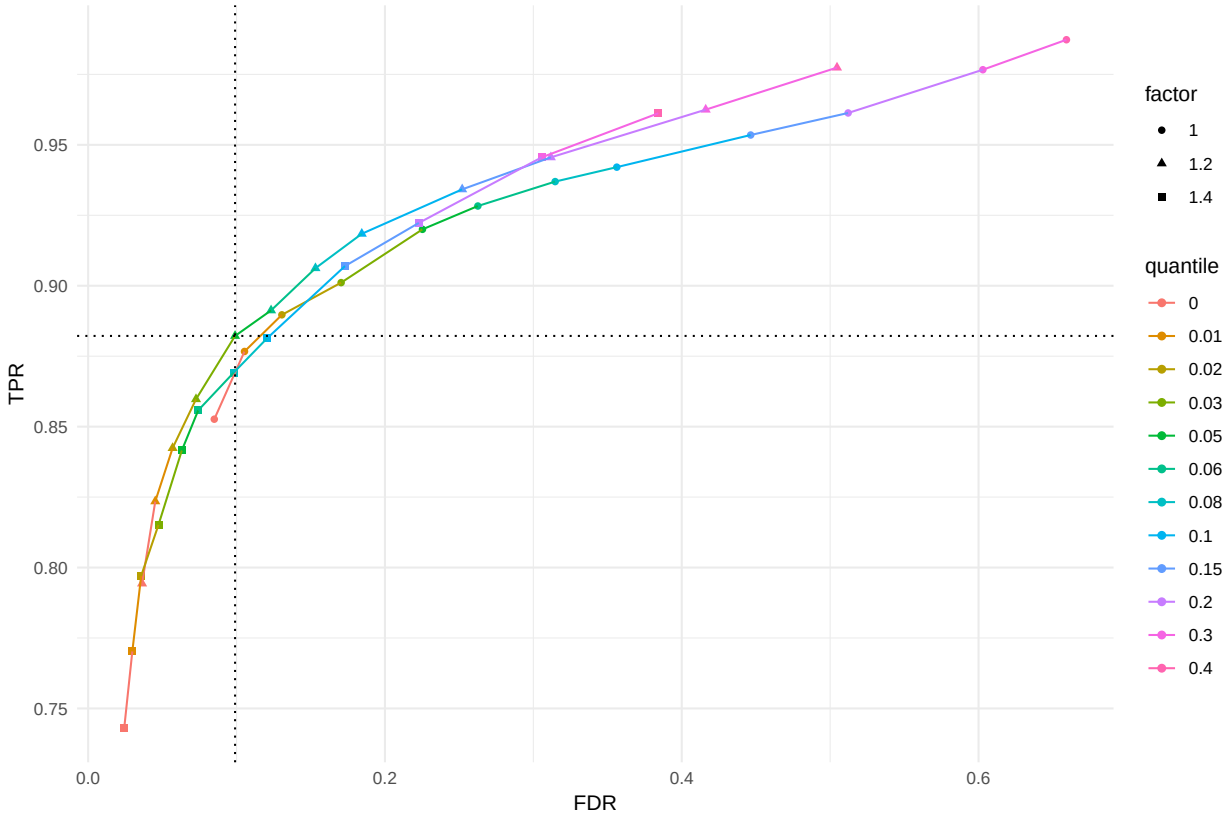
- the `quantile` values of the distribution of expression of a gene in the control samples (dotted line)
- the `factor` parameter which modulates the quantile values and defined the thresholds which determine the deregulation status in case samples (red line).

The function `choose_quantile` applies the quantile method to the simulated dataset for various values of the parameters quantile and factor (grid of values defined by the variables `factor_values` and `quantile_values`, see the R documentation for default parameters). For each set of parameters, it computes the corresponding False Discovery Rate (FDR), True Positive Rate (TPR) and False Positive Rate (FPR). The function `select_quantile_param` then choose the best set that maximize the TPR for a user-specified maximal value of the FDR (defined by the variable `FDR_max`, default value = 0.15).

```
quantile_values = c(0, 0.01, 0.02, 0.03, 0.05, 0.06, 0.08, 0.1,
    0.15, 0.2, 0.3, 0.4)
factor_values = c(1, 1.2, 1.4)
which_quantile = penda::choose_quantile(data_ctrl, simulation,
    factor_values = factor_values, quantile_values = quantile_values)
best_quantile = penda::select_quantile_param(which_quantile,
    FDR_max = 0.1)
```

In this example, optimum quantile method parameters are defined as:

- quantile = 0.05
- factor = 1.2



Note: this vignette is an example that has been designed for a rapid test of the method. For a more complete analysis and a better parameter estimation, we recommend users to simulate more cases (10 for example instead of 3) and test more values for the parameters quantile and factor.

For the PenDA method

The `penda` method infers fo each gene in each case sample its deregulation status (up-regulation, down-regulation or no deregulation) based on the `L_H_list`. It tracks for changes in relative ordering in the sample of interest. If these changes exceed the given threshold, the gene of interest is considered as deregulated.

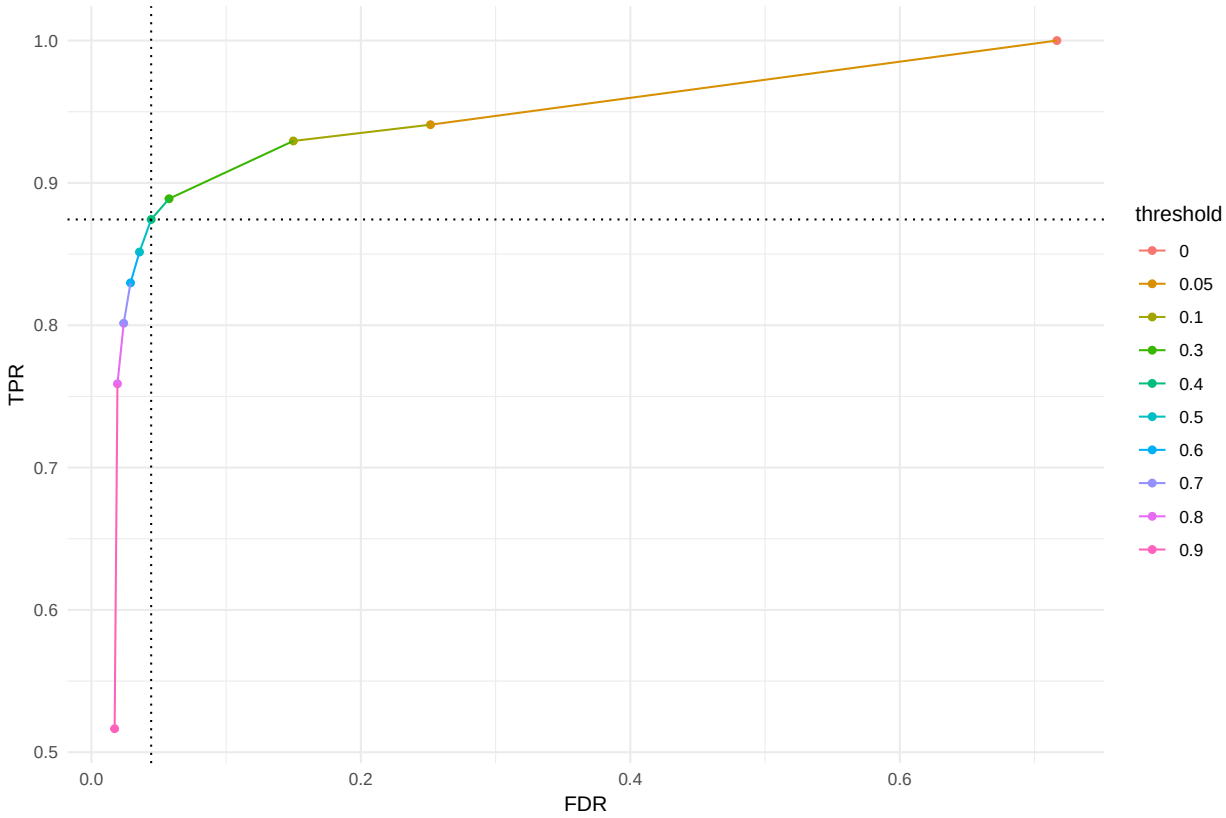
The second step of the parameter choice is therefore to determine the optimal value of `threshold`. The function `choose_threshold` applies PenDA to the simulated dataset for different threshold values (defined by the variable `threshold_values`) and computes the corresponding FDR, TPR and FPR. The function `select_threshold_param` then choose the threshold value that maximize the TPR for a user-specified maximal value of the FDR (defined by the variable `FDR_max`, default value = 0.05).

```
threshold_values = c(0, 0.05, 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8,
    0.9)
best_quant = best_quantile$quantile
best_fact = best_quantile$factor
which_threshold = penda::choose_threshold(data_ctrl, L_H_list,
    30, simulation, threshold_values, quant_test = best_quant,
    factor_test = best_fact)
```

```
best_threshold = penda::select_threshold_param(which_threshold,
    FDR_max = 0.05)
```

In this example, optimum test threshold parameter is defined as:

- threshold = 0.4



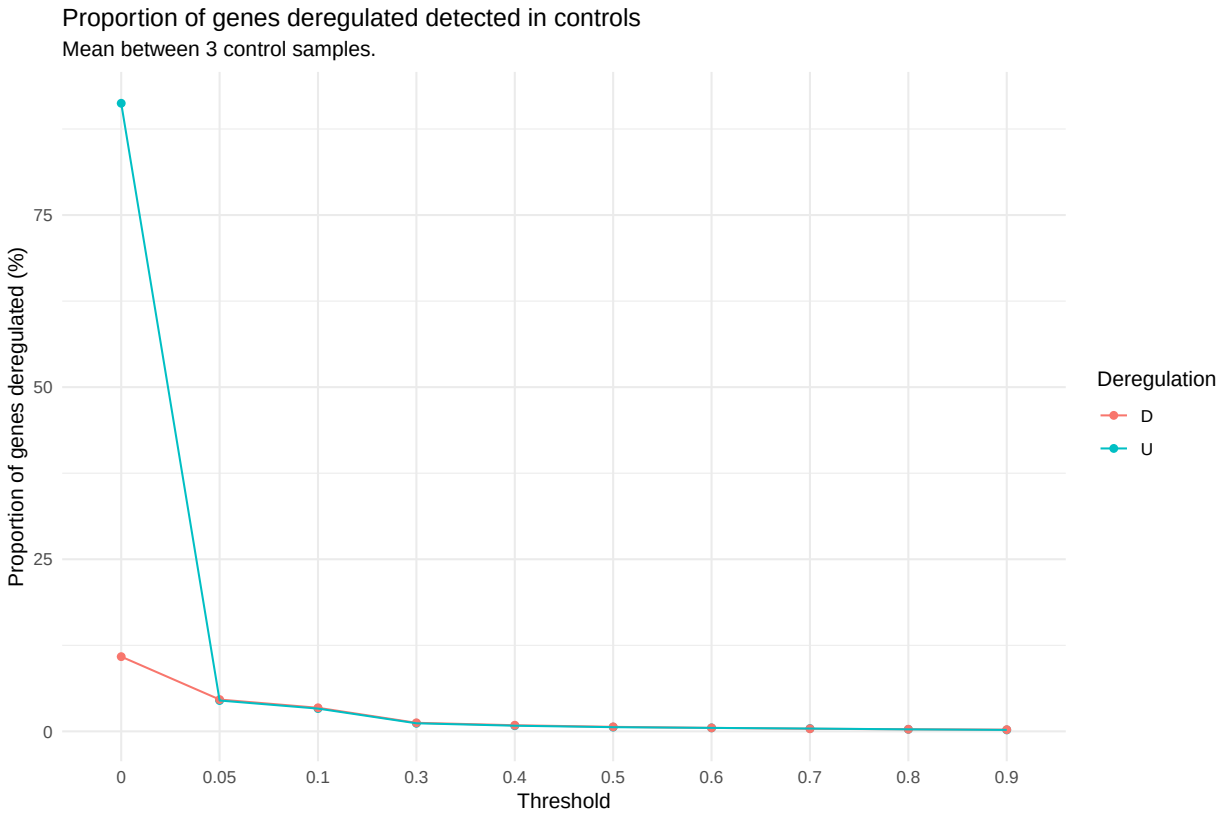
Note: this vignette is an example that has been designed for a rapid test of the method. For a more complete analysis and a better parameter estimation, we recommend users to simulate more cases (10 for example instead of 3) and test more values for the parameter threshold.

Test for false positive in control samples

As a safety check, PenDA is applied to the control samples used for the simulations and estimates the proportion of false positives.

```
threshold_values = c(0, 0.05, 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8,
    0.9)

best_quant = best_quantile$quantile
best_fact = best_quantile$factor
results = c()
for (thres in threshold_values) {
    penda_res = penda::penda_test(samples = data_simu, controls = controls,
        threshold = thres, iterations = 20, L_H_list = L_H_list,
        quant_test = best_quant, factor_test = best_fact)
    results = rbind(results, c("U", thres, colSums(penda_res$up_genes)))
```

```
    results = rbind(results, c("D", thres, colSums(penda_res$down_genes)))
}
```



Summary of simulation results

With these simulations you can now perform analysis on your real data (see `vignette_penda`) using the parameters:

- quantile = 0.05
- factor = 1.2
- threshold = 0.4

Material and methods

This paragraph is automatically generated by the vignette to specify the method and data filtering parameters. It can be directly cut and paste to the "material and methods" section of the user analysis.

The simulation vignette of the `penda` package version 1.0 was executed on 3000 genes, using 37 control samples and 40 case samples.

The data set was pretreated as following: 0 genes and 0 samples were removed during the NA values filtering step, and 159 genes were removed because lowly expressed: under the threshold `val_min` = 23.12 in at least 99 % of cases.

3 cases samples were simulated using the complex simulation function with the following parameters: group size = 100, quantile = 0.05. Theses simulations identified 29.8% of genes as typically deregulated in cases

samples.

37 controls were used to generate L and H lists using the following parameters: threshold LH = 0.99 and s_max = 30.

The quantile method was applied on the 3 simulated cases. We retained a global FDR value of 0.099, with the following set of parameters: quantile = 0.05 and factor = 1.2.

The PenDA method was then applied on these 3 cases. We retained a global FDR value of 0.0444 , with the following set of parameters: quantile = 0.05, factor = 1.2 and threshold = 0.4.

Session Information

```
sessionInfo()
#> R version 3.5.1 (2018-07-02)
#> Platform: x86_64-conda_cos6-linux-gnu (64-bit)
#> Running under: Debian GNU/Linux 8 (jessie)
#>
#> Matrix products: default
#> BLAS/LAPACK: /summer/epistorage/miniconda3/lib/R/lib/libRblas.so
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
#>  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
#>  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
#>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
#> [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] ggplot2_3.1.1
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.1        penda_0.1.0       compiler_3.5.1
#>  [4] pillar_1.4.0      formatR_1.6       plyr_1.8.4
#>  [7] mixtools_1.1.0    tools_3.5.1       digest_0.6.19
#> [10] evaluate_0.13     tibble_2.1.1      gtable_0.3.0
#> [13] lattice_0.20-38   pkgconfig_2.0.2   rlang_0.3.4
#> [16] Matrix_1.2-17     yaml_2.2.0        xfun_0.7
#> [19] withr_2.1.2       stringr_1.4.0     dplyr_0.8.1
#> [22] knitr_1.22        segmented_0.5-4.0 grid_3.5.1
#> [25] tidyselect_0.2.5  glue_1.3.1        R6_2.4.0
#> [28] survival_2.44-1.1 rmarkdown_1.12    purrr_0.3.2
#> [31] magrittr_1.5      scales_1.0.0      htmltools_0.3.6
#> [34] MASS_7.3-51.4     splines_3.5.1     assertthat_0.2.1
#> [37] colorspace_1.4-1  labeling_0.3      stringi_1.4.3
#> [40] lazyeval_0.2.2    munsell_0.5.0     crayon_1.3.4
```

Annexe 3

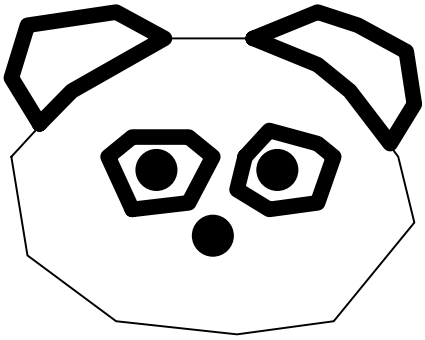
Vignette Penda pour l'analyse : Performing personalized data analysis with Penda

PENDA: PErsoNalized Differential Analysis

Performing personalized data analysis with `penda`

Magali Richard, Clementine Decamps, Florent Chuffart, Daniel Jost

2019-06-06



Introduction

`penda` (**PE**rso**N**alized **D**ifferential **A**nalysis) is an open-access R package that detects gene deregulation in individual samples compared to a set of reference, control samples. This tutorial aims at providing to non-expert users basic informations and illustrations on how to run the package.

How to cite: Richard et al. (2019) PenDA, a rank-based method for Personalized Differential Analysis: application to lung cancer, in submission.

Dataset and data filtering

Dataset

The dataset used to illustrated the method corresponds to the transcriptomes of 3000 genes (RNAseq counts, normalized with DESeq2) for 40 normal, control samples and 40 tumorous samples taken from the TCGA study of lung adenocarcinoma [PMID:25079552].

`data_ctrl` is a data matrix containing the normalized counts of each control sample. The rownames of the matrix correspond to the gene_symbol, the colnames indicate the sample ID.

```
data_ctrl = penda::penda_data_ctrl
head(data_ctrl[,1:3])
#>       patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> AADAC           347.2489           428.5498           442.0555
#> AAMP            965.2342          1528.3221           968.0266
#> ABCA1             0.0000             0.0000             0.0000
#> ABL1           1508.1784          1227.1325          1747.2431
#> ABL2            582.6719           645.4063           488.5088
#> ACACA             0.0000             0.0000             0.0000
dim(data_ctrl)
#> [1] 3000   40
```

`data_case` is a data matrix containing the normalized counts of each tumor sample. The rownames of the matrix correspond to the gene_symbol, the colnames indicate the sample ID.

```
data_case = penda::penda_data_case
data_case = data_case[rownames(data_ctrl),]
head(data_case[,1:3])
#>       patient_69-7764-01 patient_44-3919-01 patient_86-8278-01
#> AADAC           311.2129           374.9473          445.43169
#> AAMP           1466.5906           979.2256         1059.19225
#> ABCA1             0.0000             0.0000            0.00000
#> ABL1           2676.4306          2065.7474         2503.76905
#> ABL2           1167.0482           678.5603         1263.94317
#> ACACA             0.0000             0.0000           12.79693
dim(data_case)
#> [1] 3000   40
```

Note: this vignette is an example that has been designed for a rapid test of the method. So we limit the number of genes and the number of samples for this purpose. For an optimal utilization of the method, users should however upload all their available data (genes, control and case samples).

Method

`penda` performs a 3-steps analysis:

1. Data filtering and creation of the dataset
2. Relative gene ordering
3. Differential expression testing

Data filtering

```
threshold_dataset = 0.99
Penda_dataset = penda::make_dataset(data_ctrl, data_case, detectlowvalue = TRUE,
    detectNA = TRUE, threshold = threshold_dataset)
#> [1] "0 probes are NA in at least 99 % of the samples."
#> [1] "0 patients have NA for at least 99 % of the probes."
#> [1] "Computing of the low threshold"
#> number of iterations= 182
#> [1] "559 genes have less than 483.918718057525 counts in 99 % of the samples."
data_ctrl = Penda_dataset$data_ctrl
data_case = Penda_dataset$data_case
```

The function `make_dataset` contains three steps to prepare the data for the analysis.

- `detect_na_value` removes rows and columns (ie, genes and samples) of the data matrices that contain more than threshold % (default value = 0.99) of NA (Not Available) value.
- `detect_zero_value` removes genes with very low expression in the majority of samples (controls and cases), *ie.* genes whose expression is lower than `val_min` in `threshold`% of all the samples. By default it uses the function `normalmixEM` to estimate the value of `val_min` using all the *log2*-transformed count data but this parameter can also be tuned manually by the user.
- `rank_genes` sorts the genes based on the median value of gene expression in controls. This step is essential for the proper functioning of `penda`.

```
head(data_ctrl[,1:3])
#>        patient_55-6984-11 patient_43-6773-11 patient_55-6978-11
#> GRIA1           0.0000000           0.000000           0.000000
#> POU3F4          0.0000000           1.721083           2.996986
#> KLF10           0.7356968           0.000000           0.000000
#> SPOP            0.7356968          13.768668           4.495480
#> PRMT3           0.0000000           0.000000           0.000000
#> KLF2            0.0000000           0.000000           0.000000
dim(data_ctrl)
#> [1] 2441   40
head(data_case[,1:3])
#>        patient_69-7764-01 patient_44-3919-01 patient_86-8278-01
#> GRIA1             0.00000          0.5895398           0.000000
#> POU3F4            0.00000          0.0000000           0.000000
#> KLF10             0.00000          0.5895398           0.000000
#> SPOP           1989.16884          0.0000000         169.805449
#> PRMT3            85.58354         88.4309664         324.845208
#> KLF2              0.00000         38.3200855           7.382846
dim(data_case)
#> [1] 2441   40
```

Relative gene ordering

```
threshold_LH = 0.99
s_max = 30
L_H_list = penda::compute_lower_and_higher_lists(data_ctrl, threshold = threshold_LH,
    s_max = s_max)
#> [1] "Computing genes with lower and higher expression"
L = L_H_list$L
H = L_H_list$H
```

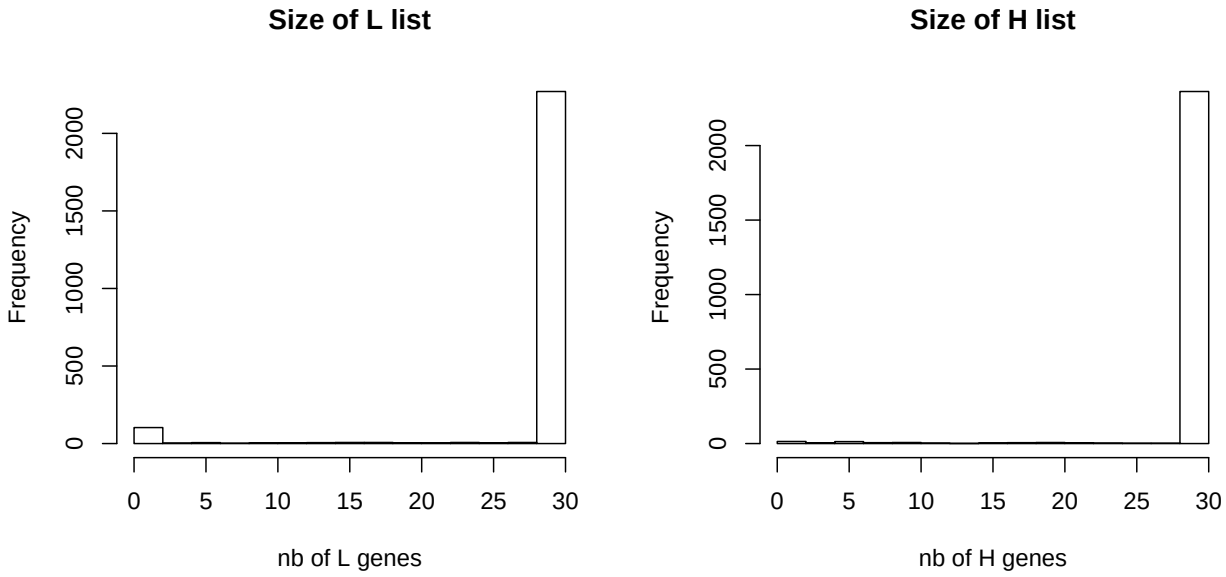
The `penda` method uses the relative gene ordering in normal tissue.

The function `compute_lower_and_higher_lists` computes two matrices `L` and `H` based on the filtered control dataset (`data_ctrl`).

Each row of the **L** matrix contains a list of at most `s_max` (default value = 30) genes (characterized by their ids) whose expressions are **lower** than that of the gene associated to the corresponding row, in at least `threshold_LH` (default value = 99 %) of the control samples.

Each row of the **H** matrix contains a list of at most `s_max` (default value = 30) genes (characterized by their ids) whose expressions are **higher** than that of the gene associated to the corresponding row, in at least `threshold_LH` (default value = 99 %) of the control samples.

Below, for some genes (FOXH1, KRTAP2-3, etc.), we show the id of 10 genes of the L and H lists.



Differential expression testing

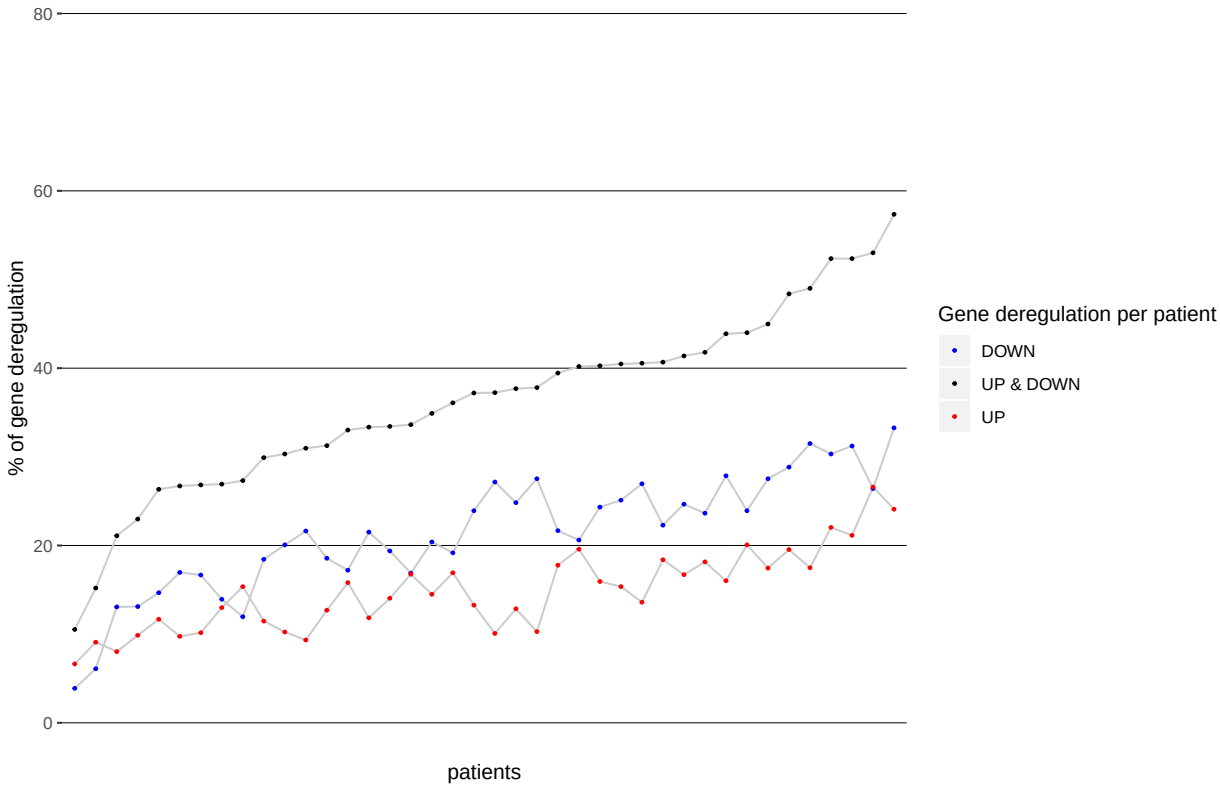
```
threshold = 0.4
iterations = 20
quant_test = 0.05
factor_test = 1.2

penda_res = penda::penda_test(samples = data_case, controls = data_ctrl,
    threshold = threshold, iterations = iterations, L_H_list = L_H_list,
    quant_test = quant_test, factor_test = factor_test)
```

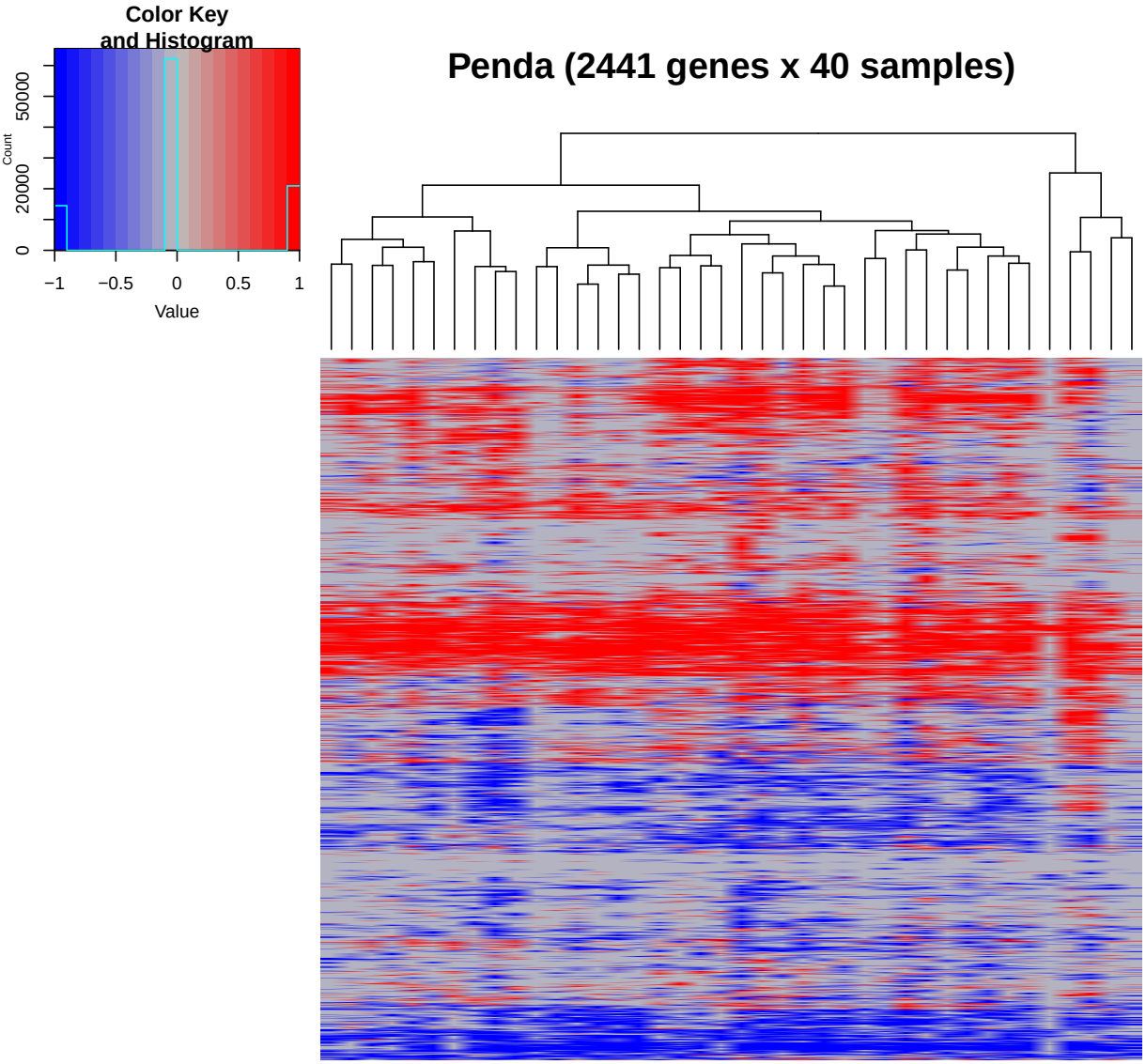
The function `penda_test` infers for each gene and for each sample of the `data_case` matrix its deregulation status (up-regulation, down-regulation or no deregulation). This function analyses case samples one by one. It is based on the `L_H_list` and tracks for changes in relative ordering in the sample of interest. If these changes exceed the given `threshold`, the gene of interest is considered as deregulated.

By default, the `threshold` parameter is set to 0.4 but we strongly advise users to use the vignette `vignette simulation` to adjust this parameter to the user-specific data.

Results are in the form of two matrices `$down_genes` and `$up_genes`. Each row corresponds to a gene and each column to a case sample. A TRUE entry in these matrices means that the corresponding genes are deregulated (down or up-regulated) in the corresponding samples.

```
#> Need help? Try Stackoverflow:
#> https://stackoverflow.com/tags/ggplot2.
```


80
60
40
20
0
% of gene deregulation
patients
Gene deregulation per patient
DOWN
UP & DOWN
UP



Material and methods

This paragraph is automatically generated by the vignette to specify the method and data filtering parameters. It can be directly cut and paste to the "material and methods" section of the user analysis.

The PenDA vignette of the `penda` package version 1.0 was executed on 3000 genes, using 40 control samples and 40 case samples.

The data set was pretreated as following: 0 genes and 0 samples were removed during the NA values filtering step, and 559 genes were removed because lowly expressed: under the threshold `val_min` = 483.92 in at least 99 % of cases.

40 controls were used to generate L and H lists using the following parameters: threshold LH = 0.99 and s_max = 30.

The PenDA method was then applied on 40 cases, with the following set of parameters: quantile = 0.05, factor = 1.2 and threshold = 0.4.

Session Information

```
sessionInfo()
#> R version 3.5.1 (2018-07-02)
#> Platform: x86_64-conda_cos6-linux-gnu (64-bit)
#> Running under: Debian GNU/Linux 8 (jessie)
#>
#> Matrix products: default
#> BLAS/LAPACK: /summer/epistorage/miniconda3/lib/R/lib/libRblas.so
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
#>  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
#>  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
#>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
#> [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] ggplot2_3.1.1
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.1         penda_0.1.0        compiler_3.5.1
#>  [4] pillar_1.4.0       formatR_1.6        plyr_1.8.4
#>  [7] bitops_1.0-6       mixtools_1.1.0     tools_3.5.1
#> [10] digest_0.6.19      evaluate_0.13      tibble_2.1.1
#> [13] gtable_0.3.0       lattice_0.20-38    pkgconfig_2.0.2
#> [16] rlang_0.3.4        Matrix_1.2-17      yaml_2.2.0
#> [19] xfun_0.7           withr_2.1.2        stringr_1.4.0
#> [22] dplyr_0.8.1        knitr_1.22         caTools_1.17.1.2
#> [25] gtools_3.8.1       segmented_0.5-4.0  grid_3.5.1
#> [28] tidyselect_0.2.5   glue_1.3.1         R6_2.4.0
#> [31] survival_2.44-1.1  rmarkdown_1.12     gdata_2.18.0
#> [34] purrr_0.3.2        magrittr_1.5       gplots_3.0.1.1
#> [37] scales_1.0.0       htmltools_0.3.6    MASS_7.3-51.4
#> [40] splines_3.5.1      assertthat_0.2.1   colorspace_1.4-1
#> [43] labeling_0.3       KernSmooth_2.23-15 stringi_1.4.3
#> [46] lazyeval_0.2.2     munsell_0.5.0      crayon_1.3.4
```

Annexe 4

Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software

Decamps et al. BMC Bioinformatics (2020) 21:16
https://doi.org/10.1186/s12859-019-3307-2

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software

Clémentine Decamps[1], Florian Privé[1], Raphael Bacher[1], Daniel Jost[1], Arthur Waguet[1], HADACA consortium[2], Eugene Andres Houseman[3], Eugene Lurie[4], Pavlo Lutsik[5], Aleksandar Milosavljevic[4], Michael Scherer[6], Michael G. B. Blum[1] and Magali Richard[1*]

Abstract

Background: Cell-type heterogeneity of tumors is a key factor in tumor progression and response to chemotherapy. Tumor cell-type heterogeneity, defined as the proportion of the various cell-types in a tumor, can be inferred from DNA methylation of surgical specimens. However, confounding factors known to associate with methylation values, such as age and sex, complicate accurate inference of cell-type proportions. While reference-free algorithms have been developed to infer cell-type proportions from DNA methylation, a comparative evaluation of the performance of these methods is still lacking.

Results: Here we use simulations to evaluate several computational pipelines based on the software packages MeDeCom, EDec, and RefFreeEWAS. We identify that accounting for confounders, feature selection, and the choice of the number of estimated cell types are critical steps for inferring cell-type proportions. We find that removal of methylation probes which are correlated with confounder variables reduces the error of inference by 30–35%, and that selection of cell-type informative probes has similar effect. We show that Cattell's rule based on the scree plot is a powerful tool to determine the number of cell-types. Once the pre-processing steps are achieved, the three deconvolution methods provide comparable results. We observe that all the algorithms' performance improves when inter-sample variation of cell-type proportions is large or when the number of available samples is large. We find that under specific circumstances the methods are sensitive to the initialization method, suggesting that averaging different solutions or optimizing initialization is an avenue for future research.

Conclusion: Based on the lessons learned, to facilitate pipeline validation and catalyze further pipeline improvement by the community, we develop a benchmark pipeline for inference of cell-type proportions and implement it in the R package *medepir*.

Keywords: Cell heterogeneity, Deconvolution, DNA methylation, Epigenetics, Matrix factorization, R package/pipeline

* Correspondence: magali.richard@univ-grenoble-alpes.fr
[1]Laboratory TIMC-IMAG, UMR 5525, Univ. Grenoble Alpes, CNRS, F-38700 Grenoble, France
Full list of author information is available at the end of the article

Background

Since the development of high-throughput sequencing technologies, cancer research has focused on characterizing genetic and epigenetic changes that contribute to the disease. However, these studies often neglect the fact that tumors are constituted of cells with different identities and origins (cell heterogeneity) [1]. Quantification of tumor heterogeneity is of utmost interest as multiple components of a tumor are key factors in tumor progression and response to chemotherapy [1].

Advanced microdissection techniques to isolate a population of interest from heterogeneous clinical tissue samples are still not feasible in daily practice (too complicated and costly). An alternative is to rely on computational deconvolution methods that infer cell-type composition. Recently, several "reference-free" algorithms have been proposed to estimate tumor cell-type heterogeneity from global DNA methylation profiling of surgical specimens [2–4]. Indeed, DNA methylation is a stable molecular marker with a cell type-specific profile dynamically acquired during cell differentiation [5] and thus provides valuable information for cell-type heterogeneity characterization and quantification. The "reference-free" algorithms are labelled as such because they do not require a priori information about DNA methylation profiles of cell types found within tumors: they directly infer them from DNA methylation samples using computational methods. While not requiring the a priori reference information, some algorithms are designed to use such information when available (e.g., [2–4]). In the absence of reference information, confounding factors affecting methylation values, such as age and sex, can potentially influence the inference of cell-type proportions. Moreover, the heuristics that are used to estimate the number of underlying cell types may differ between each deconvolution method and the sensitivity of the methods to such variability remains uncharacterized. Therefore, there is an urgent need to comprehensively characterize the current analysis pipelines, identify key features influencing their performance and provide benchmarks and recommendations to guide the application and further development of pipelines that quantify cell-type heterogeneity from reference-free DNA methylation samples [6].

Methods correcting for cell-type heterogeneity have already been compared for their statistical power to detect significant associations between epigenetic variation and biological traits [7, 8]. When associating epigenetic variation to phenotypic traits (Epigenome Wide Association Studies, EWAS), cell-type proportions are considered as confounding factors, their inference is not the main objective, but rather an intermediate step that can contribute to reducing false positive associations [9–12]. In contrast, we here compare reference-free deconvolution methods with the estimation of cell-type proportions as the main objective, as they are directly related to tumorigenesis [1]. This objective excludes several software packages from our comparison that instead return latent or surrogate variables, which are not interpretable in terms of cell-type proportions [7].

We compare three software packages that infer cell type proportions based on methylation data: RefFreeEWAS, MeDeCom and EDec [2–4]. For our comparisons, we rely on simulations where real methylation profiles of different cell types are mixed in differing proportions. While some of the methods include series of steps that may be considered a pipeline, the simulations focus on comparing the core deconvolution step shared by all the three methods (e.g., Stage 1 of EDec) that solves a convolution equation that contains two key variables: (i) the cell-type proportions within the samples, and (ii) the average methylation profiles of constituent cell types. The main outcome of this core deconvolution step are estimates of cell-type proportions and of the methylation profiles of constituent cell types, which are needed to characterize the constituent cell types and quantify tumor heterogeneity. Because accurate references for cell-type specific methylation profiles are sparse, especially for solid tissues and cancer cell types, we further assume that reference data for constituent cell-types is not available, which excludes reference-based methods from our comparative analysis [13, 14].

We here evaluate key factors affecting performance of deconvolution pipelines. We examine to what extent cell-type proportions can be accurately inferred when accounting for measured confounding factors. We determine how feature selection impacts algorithms' performance at inferring cell-type proportions. We study performances variability according to the randomly selected initialization of local optimization involved in solving deconvolution equation. We also test several methods for selecting appropriate number of constituent cell types and ask how sensitive the results are to the variation in cell type number. Based on these, we provide general guidelines for the development of reference-free deconvolution pipelines and define a benchmark pipeline to catalyze further application and improvement of reference-free deconvolution methods.

Results

Evaluation of computational frameworks to estimate cell type composition

We apply MeDeCom, EDec (Stage 1, the core deconvolution step), and RefFreeEWAS to estimate heterogeneity within simulated tumorous tissues (Lutsik et al. 2017; Onuchic et al. 2016; Houseman, Molitor, and Marsit 2014). Simulations are encoded in a matrix D of size

MxN, where M represents the number of CpG probes and N represents the number of samples. All these software packages perform various types of non-negative matrix factorization to infer cell type proportions (matrix A of size KxN, with K as the putative number of cell types) and cell type-specific methylation profiles (matrix T of size MxK) by solving D = TA, or rather by minimizing, under various constraints (that vary between the three tested algorithms), the error term: $\|D - TA\|_2$ (see Material and Methods). We simulate D with 5 cell types (K = 5): 2 cancer-like cells (lung epithelial and mesenchymal), healthy epithelial cells (lung epithelial), immune cells (T lymphocytes), and stromal cells (fibroblasts). These simulations mainly depend on a parameter α_0, which controls the diversity of the generated samples (see Material and Methods): When α_0 is small (~ 1), the simulated proportions of the K cell-types are diverse among samples and as α_0 increases, the variability decreases to the point at which proportions are the same for all samples. Finally, we simulate the effect of confounding factors on these mixtures by using a regression model of methylation data computed from real lung cancer clinical datasets (Additional file 1: Figure S1, see Material and Methods for details).

To evaluate the methods performance, we use Mean Absolute Error (MAE, see Material and Methods) as a metric to compare inferred individual cell type proportions to the ground truth. First, we tested the effect of altering four simulation parameters on the methods performance (Fig. 1 and Additional file 1: Figure S1, 1) the number of simulated samples (N, ranging from 10 to 500, 2) the inter-sample variation in mixture proportions (α_0, from 1 to 10,000, 3) the magnitude of random noise added to the mixture component (ε, from 0.05 to 0.2, 4) the set of K cells profiles used to simulate complex tissues (termed as the cell background, G, which includes all specificities related to the cell profile establishment, such as the donor genetic background or the method used to generate the profile: cell lines or primary cells) (see Material and Methods and Additional file 2: Table S4 for details). As expected, increasing the sample size (Additional file 1: Figure S2) improves the performance of all methods (Fig. 1 columns A to H). Increasing inter-sample proportion variability also substantially improves performance of all methods (Fig. 1 columns I to L). Average error (mean error across the three methods) is 0.074 ($\alpha_0 = 1$, column I) when inter-sample variation is large, increases to 0.147 ($\alpha_0 = 10$, column J) when

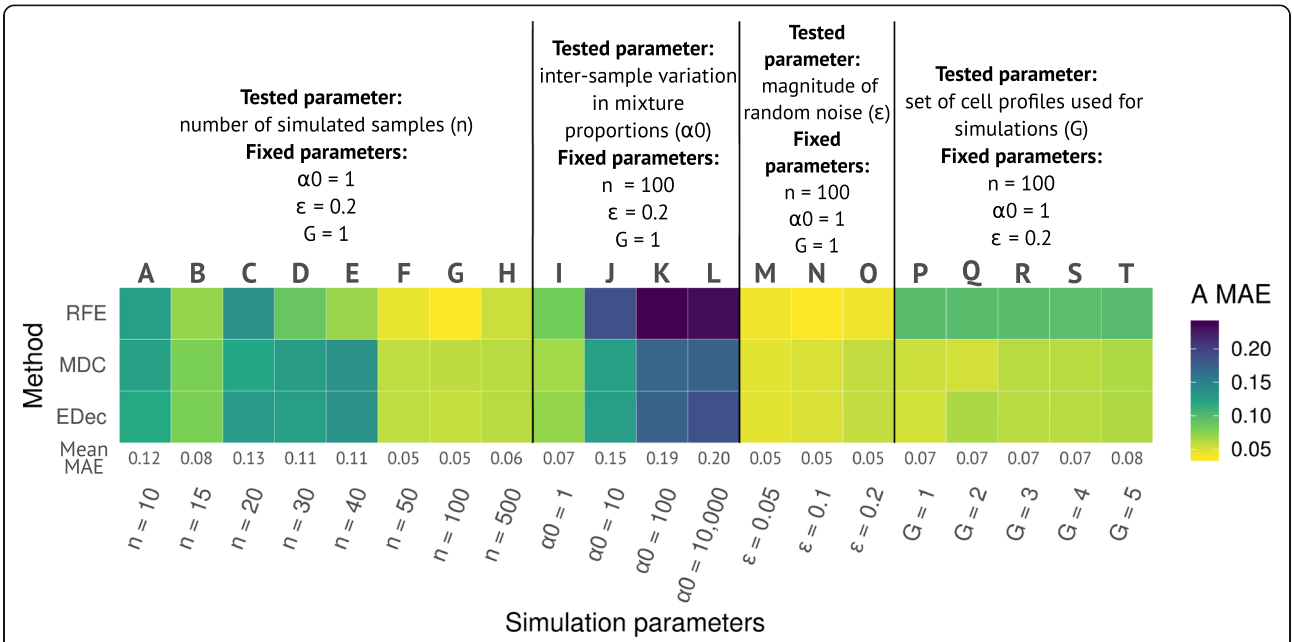


Fig. 1 Performance of the 3 deconvolution methods for different parameter settings. Heatmap of method performance (`A MAE`: Mean Absolute Error on estimated A, the matrix of cell proportions). RFE stands for RefFreeEWAS, MDC for MeDeCom and EDec for EDec stage 1. All algorithms were run on 10 D matrices corresponding to 10 different realizations of the random ε- controlled process on one D matrix computed from one simulated A matrix, each time, with the following parameters n (number of samples), α_0 (inter-sample variation in mixture proportion), ε (magnitude of random noise applied on D) and G (the cell profiles used for simulations). Mean MAE corresponds to the average error of the three methods (computed for each parameter set). A random A matrix was used for testing the effect of G1 and G2, another random A matrix was used for testing the effect of ε magnitude. Testing the effect of n and α_0 required independent simulation of A each time. As a consequence, the four simulations corresponding to the set of parameters $n = 100$, $\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1 have different results, because these simulations are based on different randomly simulated A matrices (see Fig. 7 for a systematic analysis of performance variation according to the random simulations of A)

variation is moderate, and reaches 0.194 ($\alpha_0 = 100$, column K) when variation is almost zero (Fig. 1 and Additional file 1: Figure S3). By contrast, the performances of the three methods are neither sensitive to changes of the cell background (Fig. 1 columns P to T and Additional file 1: Figure S4) nor to variations in the magnitude of the random noise applied during simulations (Fig. 1 columns M to O).

In this first direct comparison, the three deconvolution methods account for all 23,381 probes corresponding to a subset of the Illumina 27 k and 450 k DNA methylation probes, with no specific filtering. To run the algorithms, we used the following functions and parameters: RefFreeEWAS::RefFreeCellMix (5 cell types, 9 iterations), EDec::run_edec_stage_1 (5 cell types, all probes kept as informative loci, maximum iterations = 2000), and MeDeCom::runMeDeCom (5 cell types, lambdas in 0, 0.00001, 0.0001, 0.001, 0.01, 0.1), maximum iterations = 300, 10 random initializations, number of cross-validation folds = 10). Under these not-optimized conditions (i.e. with no pre-processing steps), we observe that all methods provide comparable performance, each algorithm performing best under specific conditions and parameter settings. RefFreeEwas performs best for 9 out of 20 different parameter settings, MeDeCom for 8, and EDec for 3 conditions (lowest MAE on estimated A). Error obtained with EDec is on average 8% larger than the error obtained with RefFreeEwas and 2% larger than MeDeCom. We note that for the purpose of comparison we only performed Stage 1 of EDec and did not perform Stage 0, as recommended in the original EDec publication [4].

These results suggest that the differences between the tested algorithms are minor when default parameters are used and no filters are applied on the provided DNA methylation probes. The main variations in performance are related to simulation parameters, such as sample size (n) or inter-sample proportion variability (α_0).

Different strategies to initialize matrix factorization

Optimization algorithms implemented in MeDeCom, EDec and RefFreeEWAS start with an initial condition for either T, the cell type-specific methylation matrix, or A, the cell type proportion matrix.

RefFreeEWAS initializes the T matrix. To explore the role of T initialization, we run the RefFreeEWAS package on 10 D matrices (generated from 10 random simulations of an A matrix with the following parameters: K = 5, $n = 100$, $\alpha_0 = 1$ and $\varepsilon = 0.2$). We use the following initialization schemes: K averaged methylation profiles are derived from the D matrix either by hierarchical clustering (estimation of the mean methylation of the K first clusters using a complete linkage method) based on (1) Euclidean or (2) Manhattan distances; either by (3) singular value decomposition (SVD) (corresponding to the K highest singular values) with discretized methylation values (0/1, 4) the ground truth corresponding to real T matrix used in the simulations is also tested (4) (Fig. 2, Additional file 2: Table S1, see Material and Methods for details). The RefFreeEWAS outcome varies significantly according to how it is initialized, especially when T is initialized by hierarchical clustering applied on the D matrix (estimation of the mean methylation of the K first clusters). Indeed, these two approaches based on hierarchical clustering display a high variability depending on the random simulations of A. Hierarchical clustering based on Euclidean distance or Manhattan distance performs similarly (error ranging from 0.022 to 0.135 for Euclidean distance, and from 0.022 to 0.138 for Manhattan distance). In some cases, they even outcompete initialization with the real T matrix used for simulations (0.059 mean MAE for real T), whereas the SVD approach perform systematically worse (0.152 mean MAE). Surprisingly, the effect of T initialization is highly dependent on the inter-sample variation in mixture proportion (α_0). When variation of proportion is low (α_0 = 10,000), SVD initialization is more efficient than hierarchical clustering (0.077 mean MAE for SVD-based initialization versus 0.23 mean MAE for clustering-based initialization) (Additional file 1: Figure S5, Additional file 2: Table S2).

MeDeCom performs multiple random initializations of A, the cell type proportion matrix, and EDec performs initialization with random guesses of proportions of cell types in each sample (randomized A generated from a Dirichlet distribution meeting boundary conditions of $0 < A < 1$). To roughly estimate the differences between these two approaches on the outcome of the deconvolution algorithms, we run MeDeCom (without the regularization function: parameter lambda = 0) and EDec on 10 independent D matrices (generated from 10 random simulation of matrices A, $n = 100$, $\alpha_0 = 1$ and $\varepsilon = 0.2$) (Additional file 1: Figure S6). Interestingly, both approaches give similar results. There appears to be a range of errors across 10 different D matrices simulated for both methods tested, with 3 D matrices showing a high standard deviation across 10 random noises (Additional file 1: Figure S6), suggesting these initialization methods may show sensitivity as well in certain situations.

In summary, the strategies of initialization can have important impacts on method performance, some of them being highly dependent on the composition of the original D matrix. Further work is therefore needed to better understand the relationship between the initialization of the algorithm and method performance.

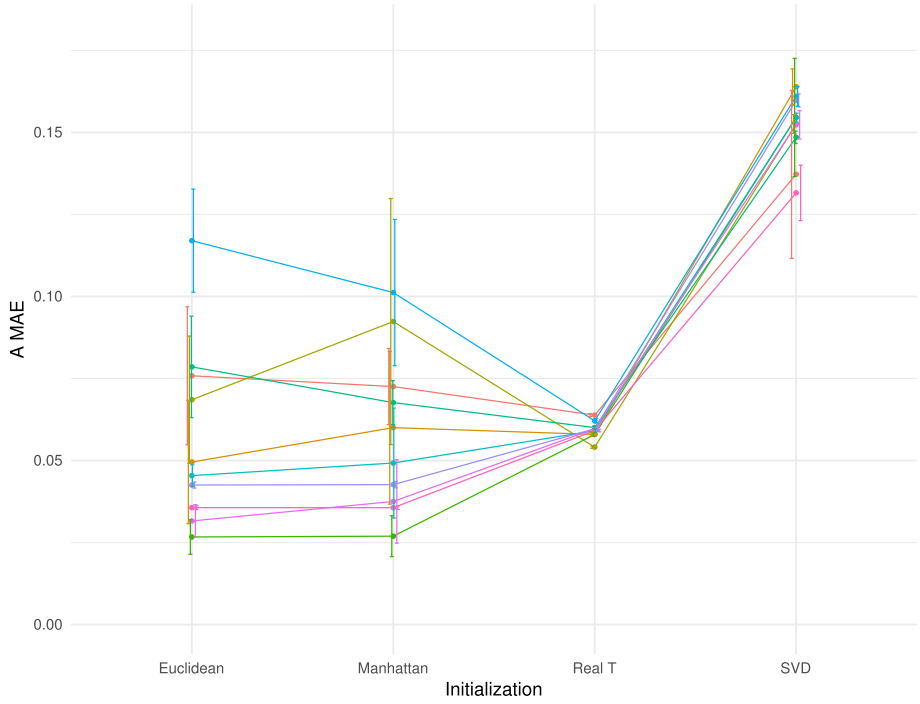


Fig. 2 Impact of algorithm initialization of RefFreeEWAS method performance. `A MAE` is shown for 10 D matrices (mean value of 10 random noises applied on D) computed from 10 random A. Each color represents a different simulated A. Error bars represent standard deviation on 10 random noises. The following parameters were used to simulate D: K = 5, $\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1 and $n = 100$). Euclidean corresponds to RefFreeEWAS::RefFreeCellMixInitialize function applied with the default parameter dist.method = "euclidean". Manhattan corresponds to RefFreeEWAS::RefFreeCellMixInitialize function applied with the parameter dist.method = "manhattan". Real T corresponds to RefFreeEWAS::RefFreeCellMix used with the parameter mu0 = real_T, with real_T the matrix composed of the 5 cell types used to simulate D. SVD corresponds to RefFreeEWAS::RefFreeCellMixInitializeBySVD function with default parameters

Feature selection and accounting for confounders

Variation in DNA methylation is associated with different factors (such as age, sex, batch effects, etc.) that are not always related with cell-type composition. A popular assumption is that removing probes by feature selection will improve performance of deconvolution methods. Yet, such an approach may also discard relevant biological information [13].

Because probe selection always involves probe removal, we evaluate to what extent removing probes (i.e. Feature Selection, FS) impacts estimation error. In the previous section, all the 23,381 probes were used by the three non-negative matrix factorization algorithms. Here, we apply different types of feature selection and measure their impacts on deconvolution errors. First, we perform feature selection (FS) without removal of confounding probes (Fig. 3a and b, left panel). When keeping only the most variable probes, or the ones that are the most correlated with principal components (PCs), error remains similar to when using no FS (FS variance and FS PCA, resp. in Fig. 3). When keeping only cell-type variable informative probes based on literature curation and use of publicly available reference cell profiles as surrogates (FS infloci, as suggested in the EDec method, see Material and Methods for probe selection information), we find a large reduction of error (47% error reduction on average with 20 patients and 26% with 100 patients). Then, we remove confounding probes, which corresponds to ~1000–2000 probes significantly correlated with confounder variables (such as age, sex, etc.) with an adjusted false discovery rate (FDR) threshold of 15% [15].We subsequently observe a substantial reduction of error (33% in average with 20 patients and 35% with 100 patients). Other FDR thresholds between 5 and 20% provide similar error measures (Additional file 1: Figure S7). After removal of correlated probes, there is no systematic advantage in filtering additional probes by feature selection (see Fig. 3a and b, right panel). For each deconvolution software, best performances are always obtained, with both 20 and 100 patients, when removing probes correlated with confounders. However, the positive impact of removing confounding probes is only observed when inter-sample variation in mixture proportion is high ($\alpha_0 = 1$). When inter-sample variation is low ($\alpha_0 = 10{,}000$), the deconvolution is much more complicated, which could explain

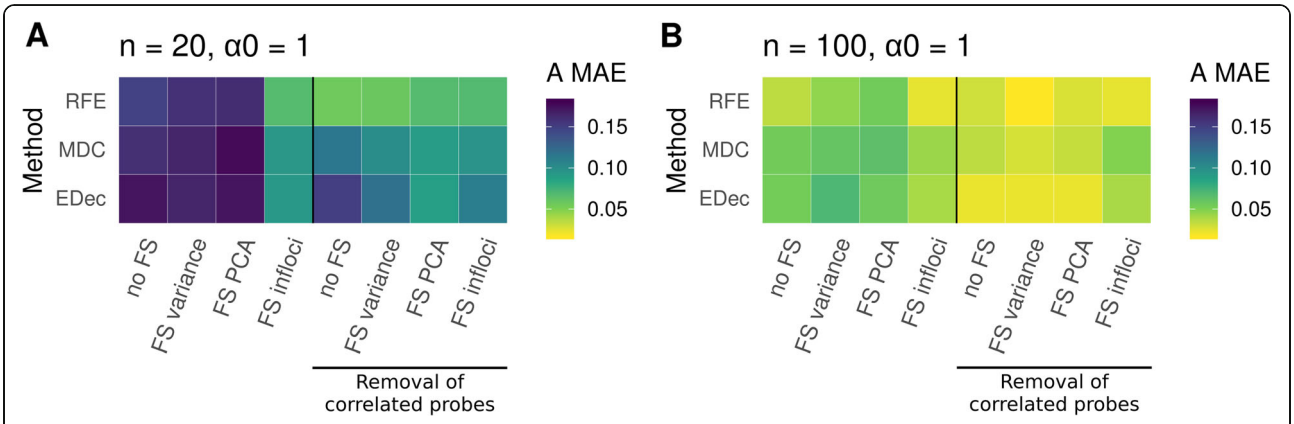


Fig. 3 Impact of pre-processing on method performance. Heatmap of method performances (`A MAE`: Mean Absolute Error on estimated A, the matrix of cell proportions). RFE stands for RefFreeEWAS, MDC for MeDeCom and EDec for EDec stage 1. All algorithms are run on 10 D matrices: 10 different random noises ε were simulated on one matrix D computed from one simulated A matrix. In each heatmap, the left panel corresponds to algorithms run without accounting for confounders (no removal of confounding probes), the right panel corresponds to algorithms run accounting for confounders (removal of confounding probes by linear regression). In each case, different types of feature selection (FS) are tested: no FS = no feature selection, FS variance = selecting probes with high variance (var > 0.02), FS PCA = selecting probes highly correlated with the 4 first PCs (p-value < 0.1), FS infloci = selecting probes expected to biologically vary in methylation levels across constitutive cell types. **a** Simulations were performed with the following parameters K = 5, $n = 20$, $\alpha_0 = 1$, $\varepsilon = 0.2$ and G = 1. **b** Simulations were performed with the following parameters K = 5, $n = 100$, $\alpha_0 = 1$, $\varepsilon = 0.2$ and G = 1. The number of conserved probes is display Additional file 2: Table S1

why we do not observe reproducible improvement of error detection after removing confounding probes in this case (Additional file 1: Figure S8).

In summary, we highly recommend to systematically remove possible confounding probes to account for confounders. Filtering based on biologically informative loci, after removing of confounding probes, can also be an interesting approach, if the investigated biological system is properly defined. In the rest of the paper, removal of confounding probes is always considered before using MeDeCom, and RefFreeEwas whereas EDec infers informative probes from any available references (in EDec Stage 0, as also illustrated in its vignette).

Choice of the number of cell types K

We tested several methods to choose the number of cell types K including the scree plot based on PCA of the D matrix, a cross validation score provided by MeDeCom, a deviance statistic estimated with bootstrap provided by RefFreeEWAS, and contrast those to the "best fit" and stability methods used by EDec.

First, we look at the scree plot by plotting the eigenvalues of the D matrix in descending order (Fig. 4). For choosing K, we use Cattell's rule, which states that components corresponding to eigenvalues to the left of the straight line should be retained [16]. When the actual number of different cell types is equal to K, we expect that there are (K-1) eigenvalues would correspond to the mixture of cell types and that other eigenvalues would correspond to noise (or other unaccounted for confounders). Indeed, one PCA axis is needed to separate two types, two axes for three types, etc. However, when not accounting for confounders, Cattell's rule overestimates the number of PCs and suggests to choose 3 and 5 PCs (i.e. K = 4 and K = 6) whereas the correct answers are K = 3 and K = 5 respectively (Fig. 4a, b). When accounting for confounders by removing probes significantly correlated with confounders, Cattell's rule provides the correct answer of 3 and 5 cell types (2 or 4 PCs) (Fig. 4c, d). Finally, we observe that varying the FDR threshold used to select confounding probes has no impact of the choice of K using Cattell's rule (for all thresholds tested, the estimation of K is correct, Additional file 1: Figure S9).

The cross-validation score provided by MeDeCom gives similar choices of K as the scree plot. When not accounting for confounders, graphical inspection of the decay of cross-validation error, as a function of the number of cell types, suggests to keep K = 4 or K = 6 cell types. When accounting for confounders, it suggests to keep K = 3 or K = 5 cell types, which are the correct answers, yet the distinction is less obvious for K = 5 (Additional file 1: Figure S10 right panel).

The bootstrap estimation of the deviance performed by RefFreeEWAS algorithm provides different answers from the previous two approaches. When the exact value of the number of cell types is K = 3, the minimum value of deviance is reached at the correct value of K = 3, whether confounders were accounted for or not. When the exact value of the number of cell types is K = 5, the deviance statistic indicates to choose a K = 4 or 5 whether confounders were

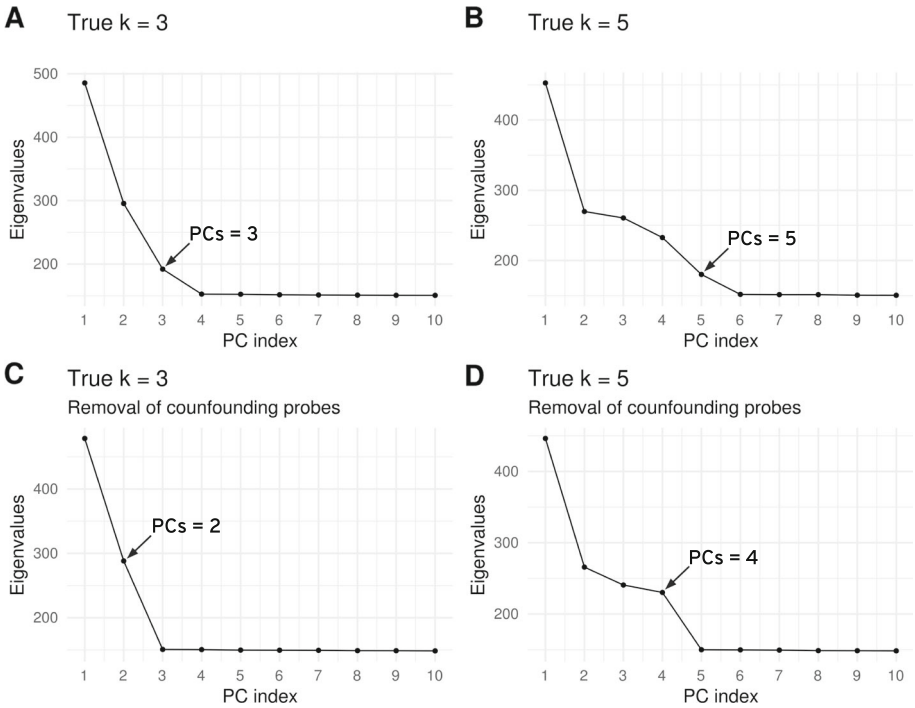


Fig. 4 Determining K with PCA scree plot. To choose K, we recommend to use Cattell's rule, calculating the estimated K as K = PCs + 1. The number of PCs chosen by the Cattel's rule is shown with an arrow. The D matrix was simulated with the following parameters: $n = 100$, $\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1 and K = 3 (**a** and **c**) and n = 100, α0 = 1, $\varepsilon = 0.2$, G = 1 and K = 5 (**b** and **d**). **a, b** Scree plot of PCA applied on D matrix before removal of confounding probes (23,381 probes). **c**, **d** Scree plot of PCA applied on D matrix after removal of confounding probes (22,551 probes in **C**, 22,532 probes in **d**)

accounted for or not (Additional file 1: Figure S10 left panel).

Lastly, we investigate to what extent inference of cell type proportions is robust with respect to the choice of K. We find that the choice of K has a strong influence on inference of cell type proportions. Underestimation of K has a large impact on measured error, and overestimation of K, albeit to a lesser extent, also increases measured error (Fig. 5). For instance, when the correct K is equal to 3, using K = 4 instead of K = 3 provides an error measure that is at least twice as large regardless of the deconvolution method that is used. Overestimation of K leads to large increase of error, even if our MAE-computing algorithm only retains the 3 cell types that minimize MAE error among the 4 inferred cell types.

These results suggest that the choice of K can be reliably guided by a scree plot based on PCA, after accounting for confounders.

Alternatively, the best performance at true K-s in Fig. 5 suggests a different approach where K is selected after running deconvolution for various values of K and choosing the K with best performance. EDec adopts this general idea by selecting the K that explains the most variance (achieves the best fit of D) and the largest K that shows stable estimates of both the A and T matrices. The stability is tested by taking at least 3 subsets (consisting of randomly selected 80% of the sample profiles) and measuring the similarity of estimates for A and for T across the 3 subsets.

Biological interpretation of recovered methylation profiles T

We next compare the estimated matrix of average cell type-specific methylation profiles (matrix T) with the real methylation profiles of cell types used to simulate the datasets.

To assess the similarity of reconstructed methylation profiles, we run the three algorithms on a representative D matrix, with $n = 100$ patients. We then draw a heatmap representing the level of correlation between estimated methylation profiles and reference methylation profiles. When the inter-sample variation is high ($\alpha_0 = 1$), we observe that all the deconvolution methods tested succeed to properly estimate the cell type-specific methylation profiles and to robustly identify corresponding reference cell types (Fig. 6). When the inter-sample variation is low ($\alpha_0 = 10{,}000$), all methods fail to identify reference cell types (Additional file 1: Figure S11), which is consistent with the results observed in Additional file 1: Figure S8. Importantly, in a real setting, true

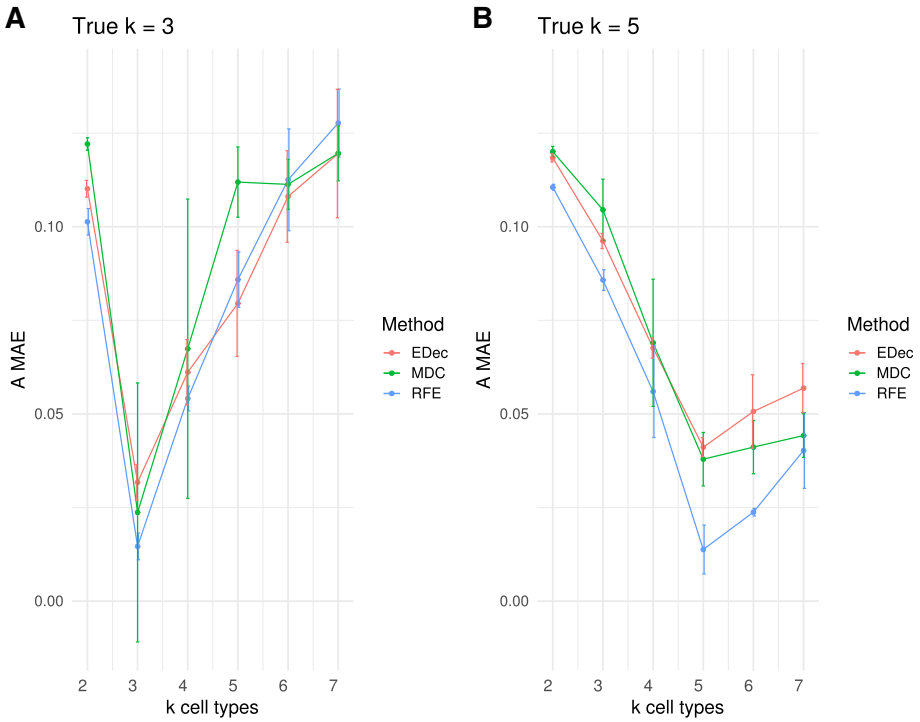


Fig. 5 Impact of K selection on algorithm performance. `A MAE` is shown for D matrices (mean value of 10 random noises applied on D) computed from 1 random A. Each color represents a different method. Error bars represent standard deviation on 10 random noise realizations. RFE stands for RefFreeEWAS, MDC for MeDeCom and EDec for EDec stage 1, each method was applied with various imposed K parameters (from 2 to 7). **a** The following parameters were used to simulate D: K = 3, $\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1 and $n = 100$. RFE and MDC methods were run after removal of confounding probes (between 22,517 and 22,624 remaining probes), EDec was run on informative loci, as recommended by the method's authors (614 remaining probes). **b** The following parameters were used to simulate D: K = 5, $\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1 and n = 100. RFE and MDC methods were run after removal of confounding probes (between 22,551 and 22,602 remaining probes), EDec was run on informative loci, as recommended by the method's authors (614 remaining probes)

methylation reference profiles are not available, and the identity of the cell-types present is unknown, which strongly complexifies the biological interpretation and the annotation of the obtained cell types.

Altogether, this indicates that reducing the inter-sample variability will impact both identification of cell type proportion and the identification of existing cell type-specific methylation profiles.

Recommendation and application of guidelines on simulated and real datasets

We have conducted extensive benchmarking of the core deconvolution step shared by three algorithms dedicated to reference-free cell-type heterogeneity quantification from DNA methylation datasets. We identify the following critical steps influencing method performance: i) accounting for confounders, ii) feature selection and iii) the choice of the number of estimated cell types. To account for confounders, we suggest to remove probes associated with the measured confounders. We suggest to use a large FDR threshold, preferring to remove too many probes rather than keeping probes influenced by confounders, given the large (> 20,000) initial number of probes. When probes associated with confounders are not removed, the number of cell types is overestimated, capturing the additional dimension of confounders (Fig. 4). We recommend that care to be taken to choose the right number K of cell types, as the outcome of deconvolution may strongly depend on K. One such method is the scree plot and Cattell's rule, which states that eigenvalues corresponding to true signal are to the left of the straight line (Fig. 4). The number of cell types to consider is equal to the number of eigenvalues to keep plus one. Choice based on Cattell's rule is robust with respect to the FDR threshold used when discarding probes associated with confounders (Additional file 1: Figure S9). An alternative, implemented by EDec, is to select the largest K that shows stable estimates of both the A and T matrices. Lastly, we suggest to use feature selection and to further reduce the number of probes by focusing on probes that are informative of cell types (as previously implemented in Stage 0 of EDec) or the highest variance markers if no biological information is available. Further reducing the number of probes marginally

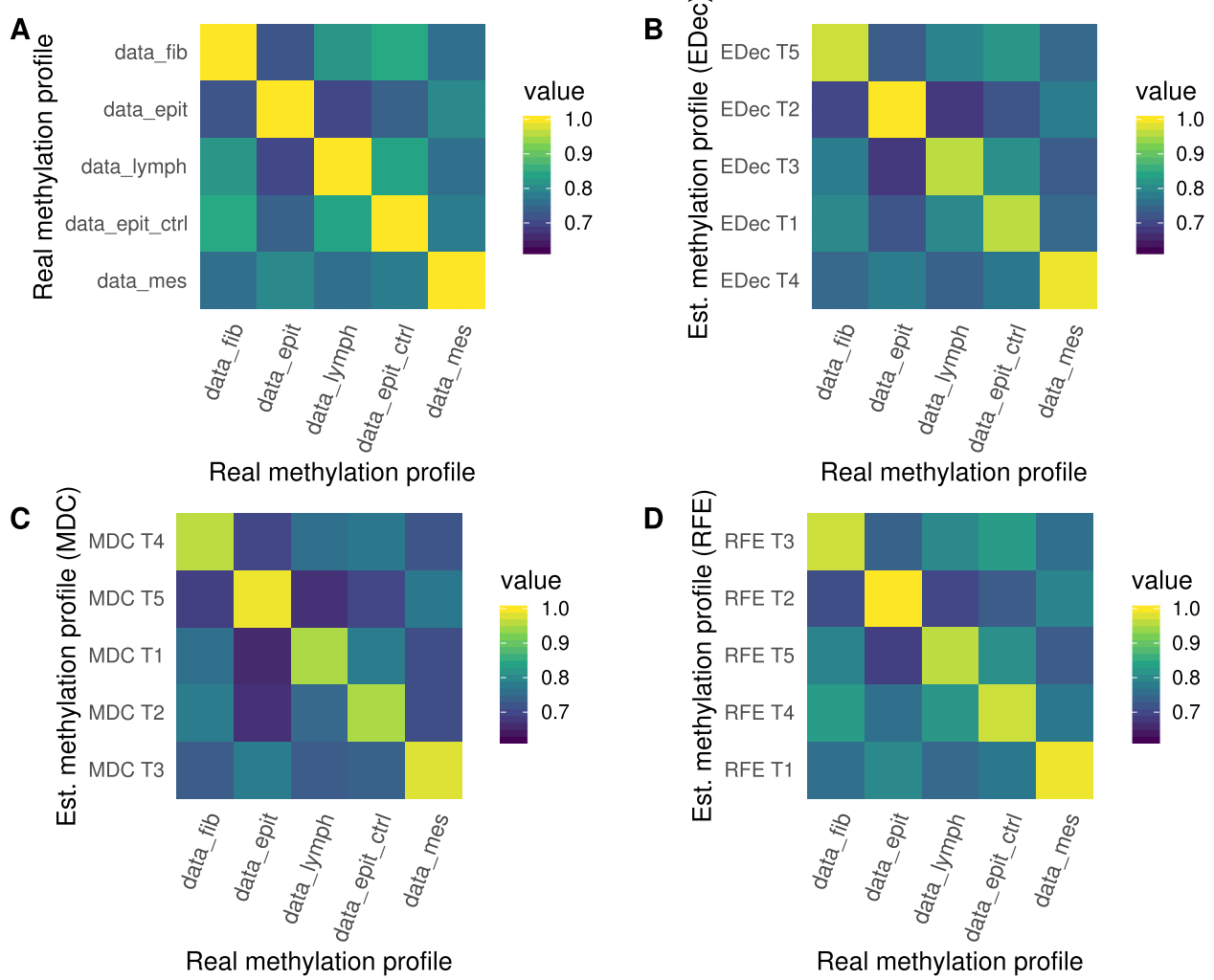


Fig. 6 Correlation between estimated and real cell type-specific methylation profiles. Heatmap of the correlations between cell type-specific methylation profile used for the simulation and cell type-specific methylation profiles estimated (Est.) by different methods. In (**a**), the correlation between different cell types used for the simulation of the T matrix (data_fib = fibroblast, data_epith = cancerous epithelial, data_lymph = T lymphocytes, data_epit_ctrl = healthy epithelial and data_mes = cancerous mesenchymal). We applied EDec (**b**), MeDeCom (**c**) and RefFreeEwas (**d**) on a representative simulation of 100 patients ($\alpha_0 = 1$, $\varepsilon = 0.2$, G = 1, K = 5) after the removal of confounding probes by linear regression (22,483 remaining probes). We used the Pearson method to compute the correlation between the estimated cell type-specific methylation profiles and real cell type-specific methylation profiles used for the simulation

affects estimation error (Fig. 3), but can substantially reduce computational running time (Additional file 2: Table S3).

When applied on simulated datasets, once the markers and the number of cell types have been chosen, the core deconvolution step implemented by one of the three packages would provide comparable estimates sample-specific cell-type proportions (Fig. 7). In order to test the performances of our pipeline on real heterogeneous tumor samples, we applied it to TCGA lung adenocarcinoma cohort (LUAD) and TCGA lung squamous cell carcinoma cohort (LUSC) [17, 18]. We compared the consistency of immune cell (IC) fraction estimated by our pipeline with the IC fraction estimated by the reference-based EpiDISH algorithm [19] and the IC fraction estimated by ESTIMATE algorithm [20] on RNA-seq profiles (Additional file 1: Figure S12 and S13). For both lung cancers, we observed consistency between our reference-free pipeline estimates and independent estimates (average correlation with reference-based estimate of 0.73 for LUAD and 0.78 for LUSC, and averaged correlation with RNA-seq based estimate of 0.66 for LUAD and 0.75 for LUSC). These data demonstrate that our pipeline can achieve good performances on real heterogeneous samples.

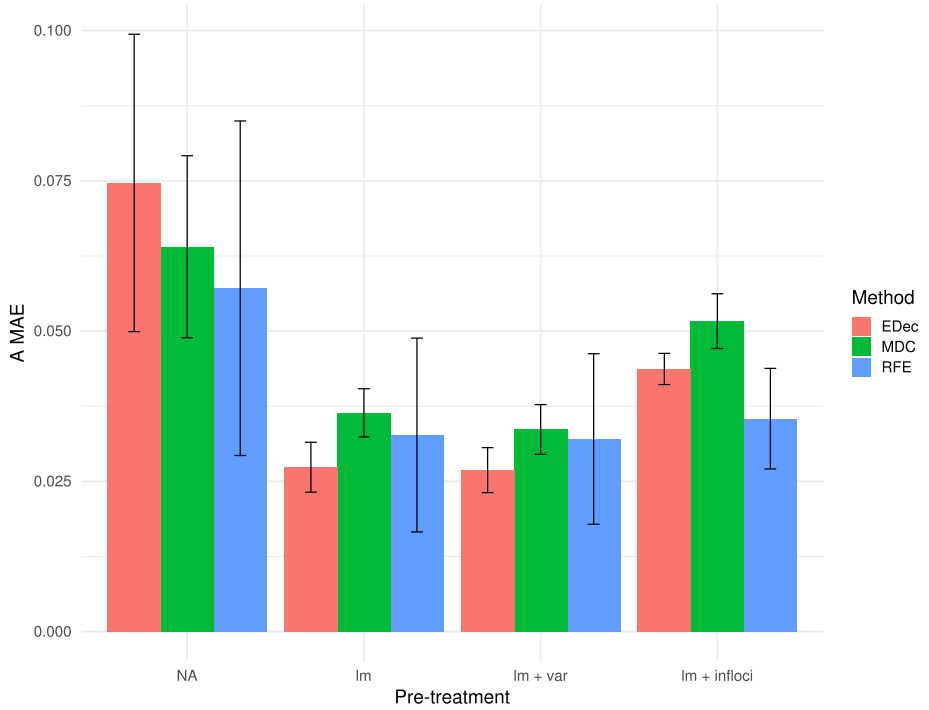


Fig. 7 Comprehensive comparison of the pre-processing pipeline. Histogram of `A MAE` (`A MAE`: Mean Absolute Error on estimated A, the matrix of cell proportions) for 10 D matrices (mean value of 10 random noises applied on D) computed from 10 random A. Each color represents a different method. Error bars represent standard deviation on 10 random noises. RFE stands for RefFreeEWAS, MDC for MeDeCom and EDec for EDec stage 1. The following parameters were used to simulate D: $K = 5$, $\alpha_0 = 1$, $\varepsilon = 0.2$, $G = 1$. The methods were run without pre-processing (NA), after removal of confounding probes by linear regression (lm), after removal of confounding probes and filtering for the most variable probes (lm + var), and after removal of confounding probes and filtering of probes expected to biologically vary in methylation levels across constitutive cell types (lm + infloci)

Conclusion

Based on lessons learned from the simulation experiments, we developed a benchmark pipeline to estimate cell-type proportions that addresses the presence of confounders and other key factors affecting performance of deconvolution algorithms (Fig. 8). We anticipate that this benchmark pipeline will help catalyze wide adoption of deconvolution methods and accelerate improvement of deconvolution pipelines by (1) helping validate other deconvolution pipelines by demonstrating concordant results; (2) serving as a benchmark for demonstrating improved performance of other pipelines; (3) providing a starting point ("toolkit") for development of new pipelines.

We note that the benchmark pipeline is not experimentally validated nor it is systematically compared as a whole against more complex pipelines that include expression data (e.g., all stages of EDec pipeline). In our experience, no deconvolution pipeline can be expected to provide accurate solutions when applied "out of the box" to a new tumor type. Tuning and validation are required in the context of each tumor type, using resources and information that may be tumor-type specific. In that sense, deconvolution may be thought of as a computational modeling approach that goes hand-in-hand with experimentation.

Discussion

Initialization

The software MeDeCom, EDec and RefFreeEwas have different methods to initialize the deconvolution algorithm. MeDeCom performs multiple random initialization of the matrix of proportion A. EDec performs initialization with random draws of proportions of cell types in each sample. RefFreeEwas initializes the matrix of cell types using reduction dimension or clustering techniques depending on a user-defined option. For RefFreeEwas, we find that clustering techniques provide the best option in the favorable scenario when cell-type proportions strongly differ between individuals (Fig. 2). However, in the less favorable scenario where cell-type proportions are more similar between individuals, initialization based on singular value decomposition should be preferred (Additional file 1: Figure S5). Differences of performance according to initialization are substantial and depending on initialization, error measures can vary by a factor of two. The fact that deconvolution methods depend on initialization indicates that there is

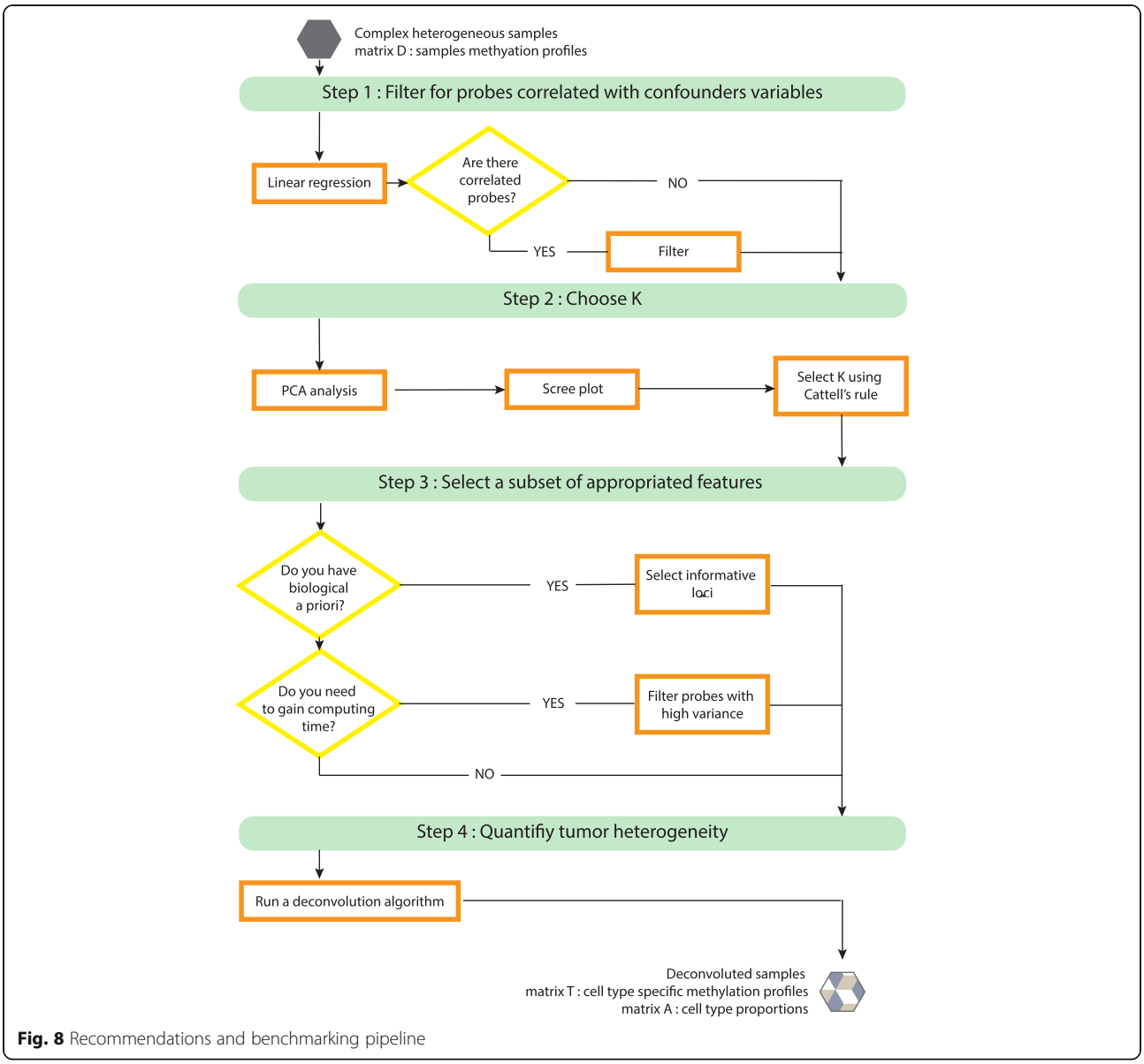


Fig. 8 Recommendations and benchmarking pipeline

room for improvement either by finding an optimal initialization strategy or by using an ensemble method that combines or averages several deconvolution solutions [21, 22].

Evaluation of performance

To estimate performances of methods, we used the MAE metric. A similar alternative metric, the root mean square error (RMSE) gave equivalent performance evaluation (Additional file 1: Figure S14). Internal steps of the MeDeCom algorithm use a leaving-columns-out Cross-Validation Error (CVE) to choose regularization parameter λ and number of cell types K. In our study, we consider a grid of six values for the regularization parameter λ (0, 0.00001, 0.0001, 0.001, 0.01, 0.1). Interestingly, the optimum lambda selected by the CVE approach does not always perform better than λ = 0, when evaluated by the MAE metric (Additional file 1: Figure S15). Although it is difficult to assess the biological relevance of each possible error metric, we would like to emphasize that the choice of evaluation metrics is an important parameter when conducting a benchmarking study. In addition, evaluation based on simulations remain limited compared to evaluation based on real tumor datasets, because the in silico simulation does not model all the biological properties of the system, such as changes in methylation due to cell-cell interaction. However, we are still lacking real tumor dataset with accurate

quantification of tumor heterogeneity. Deconvolution algorithms evaluation will then be significantly improved with the generation of dedicated in vivo benchmarking dataset

Data challenge, collaborative and open science

Our work strongly benefits from a data challenge format where different pipelines were proposed and evaluated. We gathered methylation deconvolution experts for a week of brainstorming on this benchmarking issue. We used the resulting ideas and computational methods to construct several pipelines which we evaluated in the paper. Thus, all challenge participants are referred as consortium authors of the paper. As a key deliverable of the project, to facilitate wide application of reference-free deconvolution and also pipeline development by the community, we develop a benchmark pipeline and release it as an R package (Fig. 8 and R package *medepir*).

Material and methods

Matrix factorization

We assume D is a (M × N) methylation matrix composed of methylation value for N samples, at M CpG methylation sites. Each sample is constituted of K cell types. We assume the following model: D = TA, with T an unknown (M × K) matrix of K cell type-specific methylation reference profiles (composed of M sites), and A an unknown (K × N) proportion matrix composed of K cell type proportions for each sample. In the methods tested here, A and T are found using matrix factorization, which consists of minimizing the error term $||D - TA||_2$, with constraints on methylation values, $0 \le A \le 1$ and $0 \le T \le 1$, and on proportions $\sum_{k=1}^{K} A_{kn} = 1$ (MeDeCom and EDec) or $\sum_{k=1}^{K} A_{kn} \le 1$ (RefFreeEWAS), where A_{kn} is the proportion of the n*th* sample for the k*th* cell type. MeDeCom uses an additional regularization function, which depends on a regularization term, which is weighted by a hyperparameter λ, that favors methylation values close to 0 or 1.

$$\lambda \sum_{k=1}^{K} \sum_{n=1}^{N} \omega(T_{kn}) \text{ with } \omega(x) = x(1-x).$$

TCGA DNA methylation data

We used the The Cancer Genome Atlas (TCGA) lung cancer dataset (LUAD and LUSC) as an example biological dataset. We downloaded and processed level 3 Illumina 450 k and 27 k methylation data (beta values) from TCGA, with associated metadata. To extract confounding factors parameters, we used clinical data associated with normal (non-tumor) samples (LUAD, $n = 56$ and LUSC, $n = 69$). When applying *medepir* pipeline on real heterogeneous dataset, we selected 456 tumor samples of DNAm 450 K for LUAD and 370 tumor samples of DNAm 450 K for LUSC. Difference between type I and type II probes was normalized with the function wateRmelon::BMIQ [23] using [24] as reference for the probe types.

Datasets used for simulations

We simulated synthetic DNA methylation mixtures from cell lines and primary cells (27 K or 450 K DNA methylation, see Additional file 2: Table S4). We used a variety of cell type-specific methylation profiles to simulate lung cancer heterogeneity, including cancerous and normal epithelial cells, cancerous mesenchymal cells, normal fibroblasts for stromal cells, and T lymphocytes for immune cells. We selected M = 23,381 probes using the following criteria: (i) intersect between the Illumina Infinium 27 k and Illumina Infinium 450 k DNA methylation array, and (ii) non-null probes in 100% of LUAD and LUSC methylation datasets. We tested whether Illumina 450 k type I and type II probes have an effect on algorithms' performances using the BMIQ packages that adjusts type II design probes distribution. Algorithms performances were similar on simulations using 450 k data after or before BMIQ correction (Additional file 1: Figure S16), suggesting type I or type II probes design of Illumina 450 k has no significant effect on our specific simulation design.

Simulation models

Simulated D matrix were obtained using D = TA model

Simulation of the A matrix

The cell type proportion A matrix is simulated by a Dirichlet distribution with parameters defined for 10% of fibroblast, 60% of cancerous epithelial, 5% of T lymphocytes, 15% of control epithelial and 10% of cancerous mesenchyme. The mixture proportions are sampled from a Dirichlet distribution with parameters generating sets of proportions more or less variable across the sample population. The parameter α 0, which defines the variability across the sample population, is set to 1 by default. For simulation with only three cell types, we used 20% of fibroblast, 70% of cancerous epithelial and 10% of T lymphocytes.

Simulation of the T matrix

To initiate T, we use the purified cell-types methylome describes in Additional file 2: Table S4. Once the initial T matrix is defined, we apply a series of confounding effects using parameter extracted from the clinical data associated with LUAD and LUSC cohorts' normal samples.

First, we generate an individual-specific T matrix, accounting for two major biological confounders for methylation, which are sex and age.

To identify effects of sex, we performed linear regression of methylation by sex in the TCGA dataset and detected 1397 probes correlated with sex (p-value < 0.01). Given that the majority of cell lines used to construct the initial T matrix were derived from male individuals, we used the corresponding linear regression coefficients to shift accordingly methylation value of female-associated T matrices. We used the same sex coefficient for each cell type represented in the T matrix.

To identify effects of age, we performed linear regression of methylation by age in the TCGA dataset and identified 113 probes correlated with age (p-value < 0.05). We used this linear model to generate an individual-specific methylation profile, according to its age. Then, we arbitrarily decided to assign these 113 methylation values to the normal epithelial cell type. For each probe associated with age, we then applied a normalization coefficient (ratio of each cell type to the epithelial cell type computed in the initial T matrix) to modify methylation values of the remaining cell types. Our simulation scheme implicitly assumes that age has the same effect whatever the cell type.

Simulation of the D matrix

Second, we decided to account for technical confounders. We calculated 22 median plate-effects (TCGA experimental batch effect) using 1000 random probes. For each probe, we modeled plate effects using multiplicative coefficients that measure the ratio of mean methylation values of a plate on mean methylation of the (arbitrarily) 1st plate. Each coefficient is estimated by the median of the 1000 ratios of methylation values. These multiplicative coefficients are then used on all probes to model batch effects on the matrix D of individual convoluted methylation profile.

Finally, we added Gaussian noise on the matrix of convoluted methylation profiles D. By default, we used the Gaussian parameters mean = 0 and sd = 0.2. In case noise generated methylation values larger than 1 or smaller than 0, noise was not added to the methylation value.

The simulation function is accessible using our R package *medepir*.

Software usage

We follow publication guidelines and default parameters for each method (see Material and Methods for details in software usage). RefFreeEwas was used with the function "RefFreeCellMix" with 9 iterations and remaining parameters set to default, unless specified otherwise. MeDeCom was used with the function runMeDeCom with the following parameters: NINIT = 10, NFOLDS = 10, ITERMAX = 300, lambdas = c(0, 0.00001, 0.0001, 0.001, 0.01, 0.1), if unless specified otherwise. EDec was used with the function run_edec_stage_1 with default parameters, unless specified otherwise. When testing the initialization of matrix T in RefFreeEWAS algorithm, we used the following functions: RefFreeEWAS::RefFreeCellMixInitializeBySVD(D, type = 1), RefFreeEWAS::RefFreeCellMixInitialize(method = "ward", dist.method = "euclidean"), RefFreeEWAS:: RefFreeCellMixInitialize(-method = "ward", dist.method = "manhattan"). Each RefFreeEWAS initialization method estimates an averaged methylation profile for K components from the initial D matrix (either by hierarchical clustering on D, or by singular value decomposition of D). The SVD method initializes the reference-free cell mixture deconvolution using an ad-hoc method attempting to obtain T by discretizing to 0/1 the U matrix (left-singular vectors obtained by SVD of D). The threshold used for binarization is the median value of each row of the U matrix.

Confounding factor detection

Accounting for confounding probes was performed using linear regression for each confounding factor (based on clinical metadata extracted from TCGA LUAD and LUSC cohorts). We control for FDR using Benjamini-Hochberg correction. For each confounder variable, we removed probes that are significantly associated with the confounder variable using an FDR threshold of 0.15.

Feature selection

To enhance the precision of the method, we can select the more informative probes. For this, we tested three methods: (i) selecting probes with high variance (var > 0.02), (ii) selecting probes highly correlated with the first four PCs (p-value < 0.1) [25, 26], and (iii) selecting probes expected to biologically vary in methylation levels across constitutive cell types, as described in EDec method (see below for detailed explanation), the corresponding probes are depicted in Additional file 2: Table S5.

Probes on the HM450 array which were previously shown to be cross reactive, sex-specific, contain missing values, or contain SNPs were filtered using http://zwdzwd.github.io/InfiniumAnnotation#current. Additionally, probes with non-zero covariate effects as determined by sparse regression (R package lfmm, function lfmm::lfmm_lasso) with a subset of covariates (t, n, m, age, dead, center, expo and sex) were filtered. Publicly available 450 K and 27 K cell reference profiles were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/) and grouped into 5 representative cell types predicted to constitute the bulk tissue: stroma (5 cancer-associated fibroblast and 5 fibroblast profiles), cancer (21 lung cancer cell profiles), epithelial (7 lung epithelial

references), endothelial (9 references), and immune (19 references consisting of a mixture of T-reg, monocyte, granulocyte, neutrophil, CD4+ T-cell, and CD + 8 T-cell profiles). Using these 5 reference groups, informative loci were chosen as previously described [4]. Briefly, 500 informative loci were chosen using EDec's stage 0 command using the "one.vs.rest" option with a *p*-value of 1e-4. 100 informative loci were then added to this set of 500 using the stage 0 command with the "each.pair" option for each pairwise comparison between the epithelial and stroma groups and the epithelial and immune groups to yield a final set of 614 unique informative probes (unique(500 + 100 epithelial vs. stroma + 100 epithelial vs. immune)).

Performance evaluation

We evaluate algorithm performances by (i) computing the mean absolute error (MAE) on estimated A (the matrix of cell type proportion, of size KxN, with K the putative number of cell types, and N the total number of samples), defined as $MAE = (\sum_{n=1}^{N}\sum_{k=1}^{K}|Aest_{nk} - Areal_{nk}|)/(NK)$, or (ii) computing the root-mean-square error (RMSE) on estimated A, defined as $RMSE = \sqrt{(\sum_{n=1}^{N}\sum_{k=1}^{K}(Aest_{nk} - Areal_{nk})^2)/NK}$, with n the total number of observations.

Application to LUAD and LUSC cell-type heterogeneity deconvolution

medepir pipeline was applied on 456 samples of DNAm 450 K TCGA LUAD and 370 samples of DNAm 450 K TCGA LUSC. First, we removed all probes with NA in at least one sample or a beta-value of 0 in all samples. Then, we filtered probes using confounding factor detection (98,580 probes were removed in LUAD cohort, and 18,826 in LUSC cohort). With used PCA to estimate the number of cell types present in the mixtures (K = 5 in LUAD, k = 4 in LUSC) and we applied feature selection based on probes maximum variance. At the end, we kept 29, 053 probes in LUAD cohort and 97,600 probes in LUSC cohort). EpiDISH algorithm was applied on 456 samples of DNAm 450 K TCGA LUAD and 370 samples of DNAm 450 K TCGA LUSC using the function "epidish" (by default parameters), with centEpiFibIC.m as reference. The proportion of immune cells (IC) was download on Estimate website (https://bioinformatics.mdanderson.org/estimate/disease.html) based on the RNA-seqV2 analysis of TCGA datasets. 449 samples of LUAD and 365 samples of LUSC were common between DNAm 450 K and RNAseq.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3307-2.

Additional file 1: Figure S1. Overview of in silico simulations. **Figure S2.** Variation of the estimated A MAE depending of the number of patients. **Figure S3.** Distribution of cell types proportions according to the parameter α_0. **Figure S4.** Variation of the estimated A MAE depending of the cells used for simulations. **Figure S5.** Impact of the initialization method of RefFreeEwas for a stable Dirichlet simulation. **Figure S6.** Effect of A matrix initialization on algorithms performances. **Figure S7.** Impact of the FDR threshold for the removing of confounding factors. **Figure S8.** Impact of pre-processing for a stable Dirichlet. **Figure S9.** Determining K is robust to variations in accounting for confounders. **Figure S10.** Determining K by RefFreeEWAS and MeDeCom. **Figure S11.** Correlation between estimated and real cell type-specific methylation profiles for a Dirichlet of $\alpha_0 = 10{,}000$. **Figure S12.** Efficiency of the pipeline on heterogeneous clinical LUAD tumor samples**. Figure S13.** Efficiency of the pipeline on heterogeneous clinical LUSC tumor samples. **Figure S14.** Variation of the error metric: Mean Absolute Error and Root-Mean-Square Error. **Figure S15.** Impact of lambda parameter for MeDeCom. **Figure S16:** Effect of probe type on simulations

Additional file 2: Table S1. Number of remaining probes in Fig. 3. **Table S2.** Number of remaining probes in Additional file 1: Figure S8. **Table S3.** Mean execution time in Fig. 7 (minutes). **Table S4.** Table of cells used for simulations. **Table S5.** Informative loci.

Abbreviations

CVE: Cross-validation error; DNAm : DNA methylation; EWAS: Epigenome wide association studies; FDR: False discovery rate; FS: Feature selection; IC: Immune cells; lm : linear model; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; MAE: Mean absolute error; MDC: MeDeCom; PCA: Principal component analysis; PCs: Principal components; RFE: RefFreeEWAS; RMSE: Root mean square error; sd: standard deviation; SNP: Single-nucleotide polymorphism; SVD: Singular value decomposition; TCGA: The Cancer Genome Atlas; var: variance

Acknowledgements

We thank the members of the BCM team for inspiring discussions during regular joint group meetings. The authors gratefully acknowledge the EpiMed core facility (http://epimed.univ-grenoble-alpes.fr) for their support and assistance in this work. We are grateful to the Codalab data challenge open source platform. We thank all members of the HADACA (Health Data Challenge) consortium for helpful discussion and contributions during the methylation deconvolution data challenge (December 2018, Aussois, France). HADACA collaborating authors: Sophie Achard, Elise Amblard, Raphael Bacher, Fabian Bergmann, Michael Blum, Yuna Blum, Guillaume Bottaz-Bosson, Lucile Broseus, Florent Chuffart, Clémentine Decamps, Emilie Devijver, Ghislain Durif, Vassili Feofanov, Eugene Andres Houseman, Melina Gallopin, Paulina Jedynak, Vincent Jonchere, Ellen van de Geer, Basile Jumentier, Tony Kaoma, Eugene Lurie, Pavlo Lutsik, Julia Markowski, Anna Melnykova, Jane Merlevede, Petr Nazarov, Ngoc Ha Nguyen, Olga Permiakova, Florian Privé, Magali Richard, Matthieu Rolland, Michael Scherer, Yannick Spill.

Authors' contributions

MR conceived and designed the project. MR, CD, FP, AW and RB developed the method. CD and FP implemented the R package. MR, CD, FP, and MB analyzed the results. MR and MB wrote the manuscript with the help of CD, FP, DJ, EL, PL, EAH PL, AM and MS. All authors read and approved the final manuscript.

Funding

MR salary has been partially supported by ITMO Cancer (Plan Cancer 2014–2019, Biologie des Systèmes n°BIO2015–08). The data challenge HADACA has been financed by the Univ. Grenoble-Alpes via the Grenoble Alpes Data Institute (which is funded by the French National Research Agency under the "Investissements d'Avenir" program ANR-15-IDEX-02). The publication cost of this article was funded by EIT Health Campus HADACA program, activity 19359. This article did not receive specific sponsorship in the design of the study, analysis, interpretation of data and in writing the manuscript.

Availability of data and materials
The *medepir* R package (DNA MEthylation DEconvolution PIpeline in R) can be downloaded at https://github.com/bcm-uga/medepir. Documentation and usage examples are also available on the same page. All the datasets associated with this publication (lung cancer patient metadata) can be found in the TCGA webpage.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Laboratory TIMC-IMAG, UMR 5525, Univ. Grenoble Alpes, CNRS, F-38700 Grenoble, France. [2]HeAlth DAta ChAllenge (HADACA) collaboration Group, Univ. Grenoble Alpes, CNRS, F-38700 Grenoble, France. [3]Independent Statistical Consultant, La Center, WA, USA. [4]Bioinformatics Research Laboratory, Molecular and Human Genetics Department, Baylor College of Medicine, Houston, TX, USA. [5]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Department of Genetics/Epigenetics, Saarland University, 66123 Saarbruecken, Germany.

Received: 10 July 2019 Accepted: 3 December 2019
Published online: 13 January 2020

References
1. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21:846–53.
2. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics. 2016;17:259.
3. Lutsik P, Slawski M, Gasparoni G, Vedeneev N, Hein M, Walter J. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. BioMed Central. 2017;18:55.
4. Onuchic V, Hartmaier RJ, Boone DN, Samuels ML, Patel RY, White WM, et al. Epigenomic Deconvolution of breast tumors reveals metabolic coupling between constituent cell types. Cell Rep. 2016;17:2075–86.
5. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
6. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum Mol Genet. 2017;26:R216–24.
7. McGregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016;17:84.
8. Kaushal A, Zhang H, Karmaus WJJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. BMC Bioinformatics. 2017;18:216.
9. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
10. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15:R31.
11. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat Meth. 2016;13:443–5.
12. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Meth. 2014;11:309–11.
13. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017;18:105.
14. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Meth. 2018;15:1059–66.
15. Benjamini Y, online YHFTA. Controlling the false discovery rate: A practical and powerful approach to multiple testing. JR Statist Soc B. 1995;57:289–300.
16. Cattell RB. The scree test for the number of factors. Multivariate Behav Res. 2010;1:245–76.
17. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
18. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
19. Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, et al. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. - PubMed - NCBI. Epigenomics. 2018;10:925–40.
20. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
21. Alhamdoosh M, Ng M, Wilson NJ, Sheridan JM, Huynh H, Wilson MJ, et al. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics. 2017;33:414–24.
22. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23:1801–6.
23. Teschendorff AE, Marabita F, Lechner M, 2012. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. academic.oup.com.
24. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2016;45(4):e22. https://doi.org/10.1093/nar/gkw967.
25. Luu K, Bazin E, Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resources. 2017;17:67–77.
26. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Stegle O, editor. Bioinformatics. 2018;34:2781–7.

Publisher's Note

Annexe 5

DECONbench : a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification

Decamps *et al. BMC Bioinformatics* (2021) 22:473
https://doi.org/10.1186/s12859-021-04381-4

BMC Bioinformatics

SOFTWARE Open Access

DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification

Clémentine Decamps[1†], Alexis Arnaud[2†], Florent Petitprez[3], Mira Ayadi[3], Aurélia Baurès[3], Lucile Armenoult[3], HADACA consortium, Sergio Escalera[4], Isabelle Guyon[5], Rémy Nicolle[3], Richard Tomasini[6], Aurélien de Reyniès[3], Jérôme Cros[7], Yuna Blum[3,8*†] and Magali Richard[1*†]

*Correspondence: yuna.blum@univ-rennes1.fr; magali.richard@univ-grenoble-alpes.fr
†Clémentine Decamps and Alexis Arnaud should be regarded as joint First Authors
†Yuna Blum and Magali Richard should be regarded as joint Last Authors
[1] Laboratory TIMC-IMAG, UMR 5525, CNRS, Univ. Grenoble Alpes, Grenoble, France[8] IGDR UMR 6290, CNRS, Université de Rennes 1, Rennes, France
Full list of author information is available at the end of the article

A full list of Consortium members and their affiliations is available at the end of the text.

Abstract

Background: Quantification of tumor heterogeneity is essential to better understand cancer progression and to adapt therapeutic treatments to patient specificities. Bioinformatic tools to assess the different cell populations from single-omic datasets as bulk transcriptome or methylome samples have been recently developed, including reference-based and reference-free methods. Improved methods using multi-omic datasets are yet to be developed in the future and the community would need systematic tools to perform a comparative evaluation of these algorithms on controlled data.

Results: We present DECONbench, a standardized unbiased benchmarking resource, applied to the evaluation of computational methods quantifying cell-type heterogeneity in cancer. DECONbench includes gold standard simulated benchmark datasets, consisting of transcriptome and methylome profiles mimicking pancreatic adenocarcinoma molecular heterogeneity, and a set of baseline deconvolution methods (reference-free algorithms inferring cell-type proportions). DECONbench performs a systematic performance evaluation of each new methodological contribution and provides the possibility to publicly share source code and scoring.

Conclusion: DECONbench allows continuous submission of new methods in a user-friendly fashion, each novel contribution being automatically compared to the reference baseline methods, which enables crowdsourced benchmarking. DECONbench is designed to serve as a reference platform for the benchmarking of deconvolution methods in the evaluation of cancer heterogeneity. We believe it will contribute to leverage the benchmarking practices in the biomedical and life science communities. DECONbench is hosted on the open source Codalab competition platform. It is freely available at: https://competitions.codalab.org/competitions/27453.

Keywords: Benchmarking platform, Deconvolution, Transcriptome, DNA methylation, Omics integration, Cellular heterogeneity, Cancer

Background

The recent development of high-throughput sequencing technologies has enabled the characterization of the genetic regulations underlying diseases such as cancer. Important advances have been made but studies often overlook the fact that tumors are made up of cells from different identities and origins. The quantification of tumor heterogeneity is of great interest to the biomedical research community because the various components of a tumor are key factors in tumor progression, clinical outcome and response to therapy. To isolate a cell population of interest, microdissection techniques can be performed on clinically heterogeneous tissue samples, but these advanced techniques are not feasible in clinical routine. In addition, single-cell technologies, while promising, have intensive protocols and require expensive and specialized resources, currently hindering their establishment in a clinical setting [1]. Instead, deconvolution methods can be used to infer cell-type composition in silico from bulk measurements, which enable the analysis of a large number of publicly available omic datasets. Bioinformatics tools that assess the different cell populations from bulk transcriptome [2–5] and methylome [6–9] samples have been recently developed, including reference-based and reference-free methods.

Recent efforts have been made to objectively compare existing tools in order to guide the users. In particular, two recent benchmark studies proposed a comprehensive comparison of transcriptome-based deconvolution methods using various parameters and simulation settings [10, 11]. In the same vein, the DREAM challenge proposed in 2019 [12] a data challenge dedicated to the prediction of immune cell types, showing the emerging spirit towards reproducibility and benchmarking. Although interesting, all these efforts are time-bound and cannot take into account upcoming novel methods. Moreover, the possibility to integrate different types of omic data to infer cell-type proportions is currently under-studied.

Standardized unbiased benchmarking resources are essential to evaluate the performances of computational methods. Indeed, these resources should avoid falling into the 'self-assessment trap', in which researchers are unrealistically expected to fairly compare their own computational method with other similar algorithms [13, 14]. In addition, unbiased attempts to benchmark computational methods are often static in space and time, preventing further contributions of other scientists or the assessment of new methods developed after the publication of the benchmark [15]. Recent collective initiatives provided formal guidelines and unified frameworks to improve unbiased performance evaluation [16]. For instance, the Global Alliance for Genomic and Health (GA4GH) published an open access benchmarking tool to assess germline small variant calls in human genomes [17]. More recently, BEELINE, a uniform interface to evaluate Gene Regulatory Network inference from single-cell data, was published and made freely accessible in the form of a docker image [18].

In this project, we built on a previous HADACA (Health Data Challenge consortium) benchmarking study [7] to develop a standardized benchmark framework for accurately evaluating quantification of tumor intra-heterogeneity from a multi-omic dataset. First, we built in silico 10 paired methylome and transcriptome benchmark datasets, using pancreatic cancer (PDAC, pancreatic adenocarcinoma) as a case study. These benchmark datasets were made realistic by the integration of the latest knowledge on PDAC biology [19–21] in the simulation models and can be used as 'truth' to evaluate computational

methods quantifying tumor heterogeneity. Second, we defined Mean Absolute Error (MAE) on estimated cell-type proportions and computational time as standard performance metrics. Third, we embedded the benchmark dataset and the scoring algorithm into a web platform called DECONbench. This web platform enables continuous and crowdsourced benchmarking, by asking participants to submit source code of their algorithm. Each submission is therefore run by the platform on the benchmark dataset and results generated in a reproducible way. Fourth, we implemented on the platform baseline methods based on some previously published deconvolution algorithms and tools. Therefore, DECONbench is an open resource to evaluate novel computational methods in an unbiased way. It provides a private general report on the overall performances of the method submitted by any participant and offers the possibility to share all source code of the contributing methods, as well as performance evaluation on a public leaderboard.

Here we present DECONbench, an innovative public benchmarking platform, open source and freely available, aiming at comparing integrative deconvolution methods for tumor heterogeneity quantification. This framework supports both crowdsourcing benchmarking (collaborative and competitive assessment of the methods) and continuous benchmarking (possibility to continuously integrate novel methods), two features that should contribute to the widespread community adoption of benchmarking good practices [15, 22]. To conclude, DECONbench is an open online benchmark framework including gold standard benchmarking datasets from different types of omic data, state-of-the-art baseline computational methods and it enables the submission of new methods for evaluation.

Implementation

The benchmarking platform infrastructure

DECONbench takes advantage of the Codalab web-based platform (https://competitions.codalab.org/) to provide a software environment for evaluating deconvolution methods. Users submit a full program that is applied to the provided benchmark datasets and compared to the ground truth. DECONbench outputs a performance score displayed on the leaderboard (Fig. 1).

Usage

DECONbench is optimized to execute methods developed in R statistical programming language, using a docker image provided on our website. The benchmark is structured around an ingestion program used as a wrapper object to execute an R program. Should anyone wish to benchmark a method coded in another language, R could then be used as a script language to execute the given program by invoking a System Command. A list of R packages installed on the docker image is as well provided. Users need: (i) to register to DECONbench on the participate tab and to download the starting kit and the public datasets; (ii) to develop an algorithm according to DECONbench guidelines; (iii) to submit their code (as a zip file) in the participate tab. Submitted algorithms are evaluated on DECONbench datasets and benchmarked with the other baseline methods. Users should note that methods relying on stochastic algorithms will give slightly variable performance on each run,

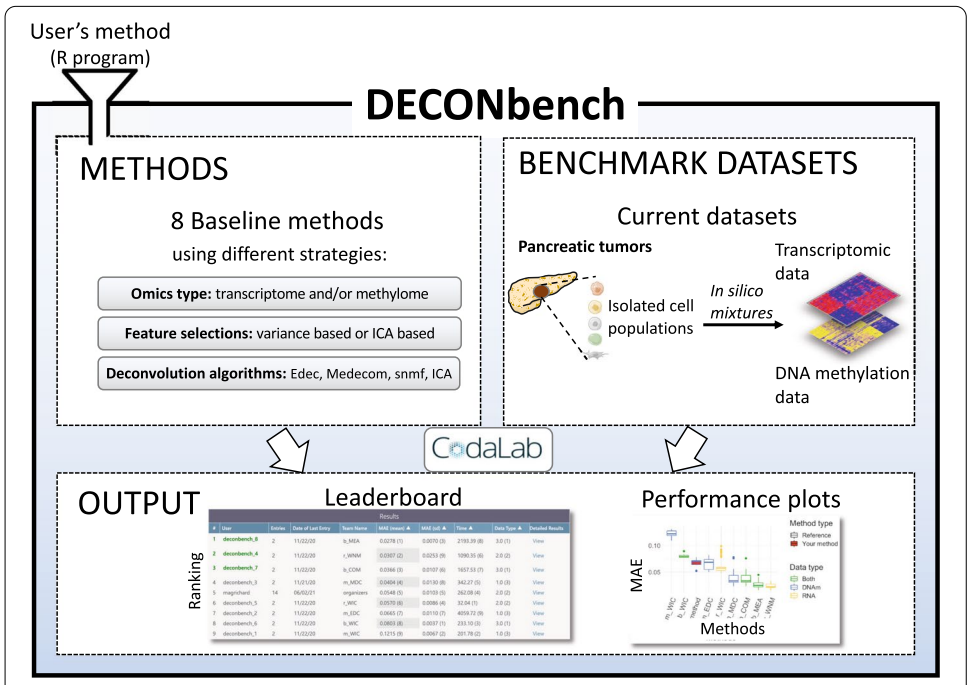


Fig. 1 Overview of the DECONbench platform. The platform proposes a set of 8 baseline deconvolution methods and benchmark datasets consisting of paired methylome and transcriptome of in silico mixtures from pancreatic tumors. The platform releases the performance of each method on a leaderboard and provides plots for deeper evaluation. New methods are automatically compared to the existing ones

unless an initialization is specified in the source code. Resulting scores appear on the leaderboard and a fact sheet is edited summarizing the performances. Importantly, users can choose whether they want their algorithm to be public or private.

Results

Provided benchmark datasets

We have generated paired transcriptome and methylome benchmarking datasets from primary cells from pancreatic tumors and sorted cells from public datasets (Fig. 2). Gold standard heterogeneous samples were simulated using mixtures of individual cell populations (fibroblast, immune cells, normal epithelial cells and cancer cells, see "Methods" section). Exact sample compositions are not accessible to the users. Participants are facing a deconvolution problem to solve the following model: $D = TA$, with D the complex matrix of molecular profiles measured on heterogeneous samples; T, a reference matrix of cell-type specific molecular profiles; and A, a proportion matrix of cell-type abundance in each sample. The aim of the competition is to find the best estimate of the proportion matrix A. Methods are evaluated on their accuracy to estimate the cell-type proportions per sample from transcriptome and/or methylome heterogeneous profiles. The discriminating metric is the mean absolute error (MAE, see "Methods" section) between the estimate and the ground truth.

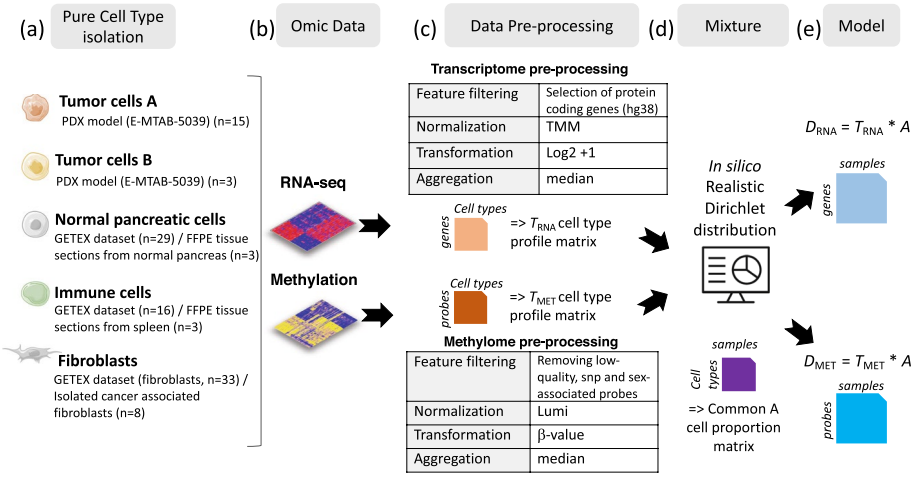


Fig. 2 Benchmark dataset construction: **a** 5 different cell populations present in pancreatic tumors were considered. **b** Raw transcriptome and methylome profiles of these different cell populations were extracted from various sources (PDX model, tissues or isolated cells). **c** Raw cell type profile matrices were preprocessed together (Feature filtering, normalization, signal transformation, sample aggregation) to avoid any batch effect. After pre-processing, transcriptomic data are constituted of log2-transformed expression counts on 21,566 genes and methylome data of beta-values on 772,316 EPIC array CpG sites. **d** In silico Dirichlet distributions have been used based on realistic proportions defined by the anatomopathologist expertise (J. Cros). **e** Paired methylome and transcriptome of in silico mixtures from pancreatic tumors were obtained by considering $D = T \times A$, with T the cell-type profiles (matrix of size $M \times K$, with M the number of features and K = 5 the number of cell types) and A the cell-type proportion per patient (matrix of size $K \times N$, with $N = 30$ the number of samples) common between both omics. One training set (D_{MET} *and* D_{RNA}) is accessible to the users (obtained by one realization of A). The algorithms are compared on 10 test sets (obtained from 10 other realizations of A) that are hidden on the platform

Selection of baseline methods from a data challenge

We used these unreleased benchmark datasets in a data challenge aiming at inferring cell-type proportions from a cancer dataset including both transcriptome and methylome profiles (https://tinyurl.com/hadaca2019). Baseline methods provided on DECONbench were collectively designed, tested and implemented during the challenge. They are composed of two steps: first, we operate a feature selection process to reduce the dimensions of the dataset, second, we apply a deconvolution algorithm. These algorithms consist of various statistical tools already published, based on unsupervised source separation approaches: ICA-based (Independent Component Analysis) [23–25] or NMF-based (Non-negative Matrix Factorization) [8, 9, 26]. Each baseline method was designed to be applied either on single-omic (see Table 1, Data type "RNA" or "DNAm") or in an integrated fashion on both the transcriptome and the methylome dataset (see Table 1, Data type "both" and Multi-omic integration strategy). As baseline on DECONbench, we implemented the eight methods that predict the real cell proportions with the highest accuracy (i.e. lowest MAE between the estimate and the ground truth) (Table 1). All baseline methods source code are publicly accessible on the platform.

Table 1 Description of each baseline method included in the benchmark

Name	RNA_wICA	RNA_wNMF	DNAm_EDec	DNAm_MeDeCom	DNAm_wICA	both_wICA	both_wNMFMeDeCom	both_meanwNMFMeDeCom
Acronym	**r_WIC**	**r_WNM**	**m_EDC**	**m_MDC**	**m_WIC**	**b_WIC**	**b_COM**	**b_MEA**
Data type	RNA	RNA	DNAm	DNAm	DNAm	both	both	both
Feature Selection DNAm	/	/	5.000 most variable probes	5.000 most variable probes	5.000 most variable probes	5.000 most variable probes	5.000 most variable probes	5.000 most variable probes
Feature Selection RNA	ICA, selection of top-contributing genes and filtering of duplicated genes	ICA, selection of top-contributing genes	/	/	/	/	ICA, selection of top-contributing genes	ICA, selection of top-contributing genes
Deconvolution algorithm DNAm	/	/	Edec	MeDeCom	ICA weighted by top-contributing probes	ICA weighted by top-contributing probes	MeDeCom with the A matrix computed on RNA as startA parameter	MeDeCom
Deconvolution algorithm RNA	ICA weighted by top-contributing genes	NMF with snmf/r method	/	/	/	ICA weighted by top-contributing genes	NMF with snmf/r method	NMF with snmf/r method
Multi-omic integration strategy	/	/	/	/	/	Averaged DNAm and RNAm proportion matrix	DNAm deconvolution uses RNA deconvolution as input	Averaged DNAm and RNAm proportion matrix
Time 10 A	~ 10 min	~ 20 min	~ 3 h	~ 17 h	~ 10 min	~ 10 min	~ 17 h	~ 17 h 30 min
Time 1 A	~ 1 min	~ 2 min	~ 20 min	~ 1 h 40	~ 1 min	~ 1 min	~ 1 h 40 min	~ 1 h 45 min
Reference of the tools/algorithms used	Hyvarinen [25]	Frichot et al. [26]	Onuchic et al. [9]	Lutsik et al. [8]	Hyvarinen [25]	Hyvarinen [25]	Lutsik et al. [8] and Frichot et al. [26]	Lutsik et al. [8] and Frichot et al. [26]

A baseline method is composed of two steps: [1] feature selection and [2] deconvolution algorithm. All deconvolution algorithms used as baseline are already published and documented in the literature (see Reference of the tools/algorithms used). A detailed description of the coding instruction and a mathematical description of the algorithms can be found in the "Methods" section. Source code is publicly available on the DECONbench platform. Time 10 A corresponds to the approximated computation time to estimate 10 proportion matrices A (corresponding to the test sets hidden on the platform). Time 1 A corresponds to the approximated computation time to estimate 1 proportion matrix A (closer to real applications on one dataset)

Bold acrononyms are used to identify methods in Figs. 3 and 4

Performance of the baseline single-omic methods

We run all the baseline methods on 10 different simulated datasets and computed the corresponding MAEs (Fig. 3). The best algorithms based on single-omic datasets were the r_WNM method for RNA-based data (mean MAE of 0.024) and the m_MDC method for DNAm-based data (mean MAE of 0.038). Both are NMF-based algorithms, details on the methods can be found in the "Methods" section. DECONbench provides also the computing time for each method, as an indicator of algorithms optimization. It is worth underlying that the computation time of m_MDC algorithm is significantly higher than the other DNAm-based methods we explored, suggesting that even high performance single-omic algorithms might be further optimized.

Performance of the integrative multi-omics methods

Next, we tested basic multi-omic approaches averaging the results of single-omic methods: (i) the b_WIC method averages the proportion matrices given by the independent applications of independent component analysis (ICA) based deconvolution approach to transcriptome and methylome data, (ii) the b_MEA method computes an average proportion matrix from the output of the two best single-omic methods r_WNM and m_MDC. Averaging the ICA based approaches (b_WIC) gave intermediate performances (multi-omic accuracy equivalent to the mean of single-omic

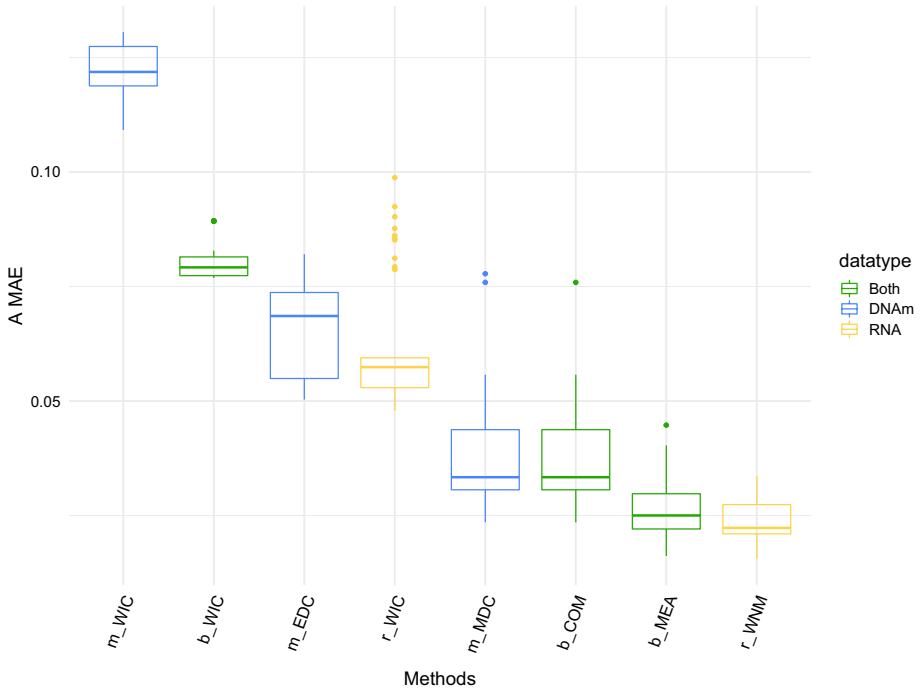


Fig. 3 Baseline algorithm performances: Boxplots of 'MAE A' obtained for reference algorithms ('MAE A': Mean Absolute Error on estimated A, the matrix of cell proportion per patient). Each reference algorithm was run 5 times on each simulated data, for 10 independent simulations (50 measures were used to construct the boxplots). Each color represents the type of data used to infer A ('DNAm' for D_{MET} training set, 'RNA' for D_{MET} training set and 'Both' for methods integrating D_{MET} and D_{RNA}). Details of the algorithms can be found in the "Methods" section

accuracies). Similarly, we did not observe increased performances when averaging the predicted proportion matrices of the two best methods (b_MEA).

We also proposed an integrative method (b_COM) based on transcriptome and methylome data. The best performing methylome-based method (m_MDC) relies on the MeDeCom tool which is a NMF-based deconvolution algorithm that performs multiple random initializations of the cell-type proportion matrix. Instead of using random initialization, we initialized the MeDeCom algorithm with the proportion matrix obtained from a NMF-based deconvolution of transcriptome data. Surprisingly, we did not observe a substantial performance improvement when integrating RNA deconvolution output into DNAm deconvolution algorithm (b_COM method, resulted in an average error decrease of 2.12% compared to m_MDC). These results highlight the need to further develop new methods to improve integration of multi-omic deconvolution algorithms.

Toward crowdsourced and continuous benchmarking

As an example of continuous benchmarking, we used DECONbench to assess the performances of two recently evaluated single-omic algorithms in a comprehensive benchmark of reference-based deconvolution pipelines. We selected the Ordinary Least Square (OLS) and Robust Linear Regression (RLR) approaches, which have been shown to be effective in estimating cellular composition of simulated bulk healthy pancreatic transcriptomes [10]. We implemented the methods as recommended by Avila Cobos et al., including the generation of cell-type reference profiles from a pancreatic single-cell dataset [27] (see supplemental information for source code: Additional file 1: Source code). Interestingly, the performance of these methods is not better than the baseline methods, possibly due to the use of healthy pancreatic cells as a reference to estimate the composition of a simulated pancreatic adenocarcinoma (Additional file 1: Figure S1). These results suggest that further optimization should be considered to properly assess the performance of the OLS and RLR methods. This crowdsourced and continuous integration is now made possible thanks to our DECONbench platform.

Conclusion

The DECONBench platform is a unique opportunity to compare the performance of deconvolution methods on different omics data. It can be used to assess the performance of newly developed methods by applying them on high quality benchmark datasets in a user-friendly fashion. Currently, the accuracy of new methods can be compared with the eight baseline methods that have been included in the benchmarking platform. As compared with previous time-bound comprehensive benchmarks of deconvolution methods (see Avila Cobos et al. [10]), our platform provides the possibility to continuously test and integrate newly developed methods, rather than focusing on an exhaustive comparison of existing tools. The baseline methods and user's methods performances are reported on the leaderboard and on the graphical output of DECONBench (Fig. 4). The source code of the baseline methods can be downloaded directly on the DECONbench platform. The structure of DECONbench is open to evolution. Work is ongoing to generate new benchmark datasets including other omic types that will be added to the platform. In the near future, we plan to expand the usability of DECONbench by

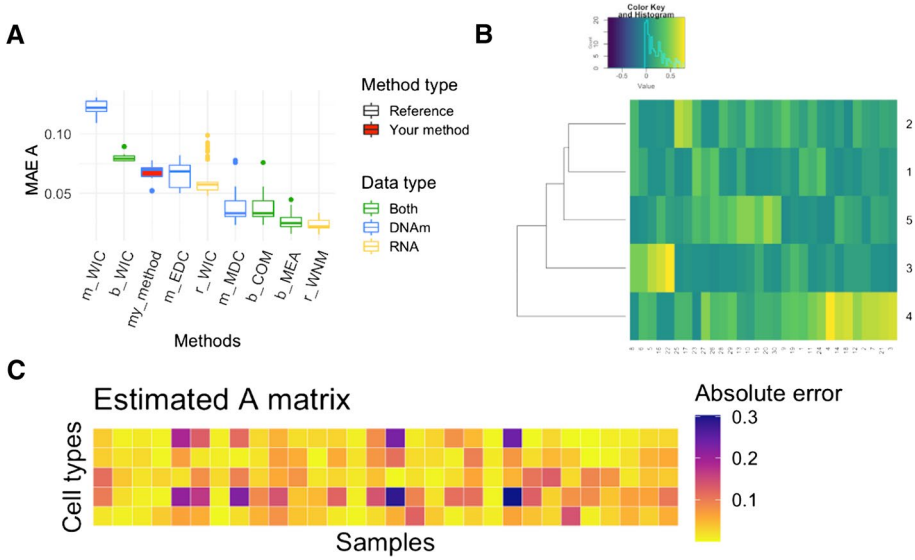


Fig. 4 DECONbench graphical outputs. **a** Boxplots of the Mean Absolute Errors (MAE) of the estimations of the A matrices (i.e. proportion matrices) obtained for each method that uses the transcriptome only (yellow), the methylome only (blue) or both omics (green). Boxplots of the baseline methods and other existing methods are shown in white, whereas the user's method is shown in red. **b** Heatmap of each A-matrix estimate are generated. The cell populations are in rows and the samples in columns. **c** Heatmaps of Absolute Error of each proportion estimate are generated. The cell populations are in rows, and the samples in columns

offering the possibility for owners of benchmark datasets to directly upload them on the platform.

DECONbench evaluation framework presents standard benchmark limitations [15, 16], such as the use of artificial in silico simulated data that do not capture the real experimental complexity, or the ranking of the methods based on a single performance metric. We would like to emphasize that MAE as scoring metric is only an imperfect proxy to evaluate quantification of tumor heterogeneity, as it does neither reflect the accuracy of cell-type specific molecular profile prediction (i.e. biological significance of inferred components), nor the correlation of estimated heterogeneity with real clinical outputs (such as prognosis or survival).

Overall, our platform will guide computational biologists to use the best proposed deconvolution algorithms and allow health professionals and biologists to obtain more accurate information regarding the composition of their samples, an important step towards personalized healthcare.

Methods

Data collection and preprocessing

For both transcriptome and methylome in silico mixtures, the same five cell types present in pancreatic tumors were considered (Fig. 2a, b): tumor cells A, tumor cells B, normal pancreatic cells, immune cells and fibroblasts. Pure cell type transcriptome profiles were retrieved from the GTEX RNA-seq dataset for the immune and normal pancreatic cell types (https://gtexportal.org/) and a previously published pancreatic tumor patient derived xenograft (PDX) RNA-seq dataset (E-MTAB-5039) for the two tumor cell types

(total of 96 pure transcriptome profiles with 3 to 33 replicates per cell type) (Fig. 2c). Pure cell type methylation profiles were retrieved from the same samples of the PDX dataset for the two tumor cell types and tissue or isolated cell profiles were used for the microenvironment cell types (total of 32 methylome pure profiles with 3 to 15 replicates per cell type). Transcriptome dataset was restricted to protein coding genes and subjected to TMM normalization using the edgeR R package and log2 transformation. For the methylation data, we used the beta-value DNA methylation scores and removed probes with low-quality, that contained SNPs or located on sex chromosomes. Data were then adjusted for color balance bias and normalized between samples using the SSN (shift and scaling normalization) method using the lumi package functions (Fig. 2c). For both omics, the median of the replicate profiles for each cell type was calculated to compute the T_{RNA} and T_{MET} matrices, representing the cell type specific profiles for each omic. The median calculation may prevent underlying germline differences. These matrices were used for the in-silico mixtures, as detailed in the next sections (Fig. 2d, e).

Formulation of the deconvolution problem

When a sample is constituted of K cell types, we assume that the level of methylation or gene expression observed in a bulk measurement of this biological sample (containing different cell types) results from a linear mixture of the K cell-type specific molecular profile weighted by the true cell-types proportions present in the sample. This assumption leads to the following models:

$$D_{MET} = T_{MET}A \tag{1}$$

$$D_{RNA} = T_{RNA}A \tag{2}$$

where D_{MET} is a (M $\times$ N) methylation matrix from N bulk heterogeneous samples with $D_{MET_{\{m,n\}}}$ the measured measured methylation (beta-value) of the m*th* CpG site for the n*th* sample representing the measured methylation (beta-values) for N samples; D_{RNA} is a (G $\times$ N) gene expression matrix from the same N bulk heterogeneous samples with $D_{RNA_{\{g,n\}}}$ the measured gene expression (normalized pseudo-log counts) of the g*th* gene for the n*th* sample; T_{MET} is an unknown (M $\times$ K) reference-profile matrix with $T_{MET_{\{m,k\}}}$ representing the average methylation beta-value of CpG site m for the cell-type k; T_{RNA} is an unknown (G $\times$ K) reference-profile matrix with $T_{RNA_{\{g,k\}}}$ representing the average expression value (normalized pseudo-log counts) of gene g for the cell-type k; and A a (K $\times$ N) matrix representing the cell-type composition of the N heterogeneous samples for K cell types (i.e. the cell-type proportions), with $A_{\{k,n\}}$ the proportion of the n*th* sample for the k*th* cell type. Specifically, the A proportion matrix is shared between the two models, as D_{MET} and D_{RNA} bulk molecular profiles are measured on the same biological samples. In the methods tested, A is estimated with the following constrain: $\sum_{k=1}^{K} A_{kn} = 1$.

Data modeling

The benchmark simulated bulk molecular profiles are constituted of 10 paired D_{MET} and D_{RNA} matrices. Simulations are processed as follows:

Step 1: Simulation of the shared proportion matrices

The mixture proportions of the matrices A were sampled from a Dirichlet distribution based on realistic biological composition of a pancreatic tumor, with the variation of Dirichlet parameters set to $\alpha_0 = 10$ for global cell composition (fibroblasts, immune, normal epithelial and cancer epithelial), and a variation of Dirichlet parameters set to $\alpha_0 = 1$ for cancer cells subpopulations (cancer basal-like and cancer classic). Exact proportion parameters are kept private to ensure unbiased evaluation of the methods.

Step 2: Simulation of the bulk D bulk matrices

We use the mathematical models (1) and (2) to simulate the bulk matrices, as previously described in Decamps et al. (2020) [7]. D_{MET} is a methylation matrix composed of 772,316 methylation values (EPIC array CpG sites) for N = 30 samples, D_{MET} was constructed as follows: $D_{MET} = T_{MET}\, A$, with T_{MET} a matrix of K = 5 cell type-specific methylation reference profiles (methylation beta-values for each cell type considered: tumor cells A, tumor cells B, normal pancreatic cells, immune cells and fibroblasts), and A a (K × N) proportion matrix composed of K = 5 cell type proportions for each N = 30 sample. D_{RNA} is a transcriptome matrix composed of 21,566 gene expression values (normalized log-2 transformed RNA-seq counts values for each cell type) for N = 30 samples. D_{RNA} was constructed according to the following model: $D_{RNA} = T_{RNA}\, A$, with T_{RNA} a matrix of the K = 5 cell type-specific transcriptome reference profiles (21,566 gene expression values for each cell type: tumor cells A, tumor cells B, normal pancreatic cells, immune cells and fibroblasts), and A the same (K × N) proportion matrix used to simulate D_{MET}.

Step 3: Simulation of a technical noise

We added a generic Gaussian noise on each bulk simulated matrix using the following parameters: mu = 0 and sd = 0.05.

Step 4: Replication of the simulations

To ensure robustness of the method's evaluation, we generated 10 replications of paired D_{MET} and D_{RNA} matrices, using independent simulation of A proportions matrices. For each pair of D_{MET} and D_{RNA} matrices, the same T_{MET} and T_{RNA} reference matrices were used.

Performance evaluation

The aim of deconvolution algorithms is to correctly estimate the proportion matrix A. We evaluated algorithm performances by computing the mean absolute error (MAE), as previously described in Decamps et al. (2020) [7]:

$$\mathrm{MAE} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}|\mathrm{Aest}_{nk} - \mathrm{Areal}_{nk}|}{NK} \tag{3}$$

One training set (D_{MET} *and* D_{RNA}) is publicly available (the A, T_{RNA} and T_{MET} matrices used for compute D_{MET} *and* D_{RNA} matrices remain private, as they are directly involved in performance evaluation). The algorithms are evaluated on 10 test sets (D_{MET} *and* D_{RNA}), obtained from 10 independent realizations of A, given the simulation models $D_{MET} = T_{MET}\, A$ *and* $D_{RNA} = T_{RNA}\, A$. These test sets are hidden on the platform to avoid

overfitting. During evaluation of baseline algorithms, each algorithm was run 5 times on each simulated set of data, to account for randomness in algorithm outputs.

Description of the baseline methods

The baseline methods we propose here are wrappers of already published unsupervised deconvolution algorithms (ICA-based or NMF-based). We assume here that A, T_{MET} and T_{RNA} are unknown and need to be estimated, either independently (single-omic pipelines) or integratively (double-omic pipelines). Before deconvolution, we systematically apply a pre-treatment step of dimensionality reduction based on feature selection. All baseline methods source code is downloadable on the DECONbench platform.

All baselines relies on unsupervised deconvolution algorithms, which consists in solving $D = TA$, either by ICA-based (i) or NMF-based (ii) approaches. (i) ICA-based approaches (r_WIC, m_WIC and b_WIC) consist of minimizing mutual information of sources by defining independent components. It is based on the fixed-point FastICA algorithm developed by Aapo Hyvärinen [24, 25]. (ii) NMF-based approaches (r_WNM, m_EDC, m_MDC, b_COM, b_MEA) aims to minimizing $||D - TA||_2$.

RNA_wICA (r_WIC, ICA-based deconvolution on RNA)

The method RNA_wICA (r_WIC) uses transcriptomic data as input and is based on the ICA algorithm for both feature selection and deconvolution. It relies on the use of the functions "runICA" and "getGenesICA" developed by P. Nazarov (sablab.net/scripts/LibICA.r) and the deconica R package [23].

STEP1: feature selection For the ICA-based feature selection, the function "runICA" is run at first with the parameters ncomp = 10 and ntry = 50. Then, the function "getGenesICA" selects top-contributing genes with a FDR of 0.2, the feature selection is done on these contributing genes belonging to a component having an average stability greater than 0.8. Finally, duplicated genes are removed.

STEP2: deconvolution First, we perform FastICA unsupervised deconvolution *(*deconica::run_fastica is run with the parameters overdecompose = FALSE and n.comp = 5; remaining parameters are set to default). Second, we compute the abundance of the identified components, using the weighted-mean of the 30-top genes of each Independent Component (IC), in each sample as, a surrogate of the component signal. The 30 most important genes of each ICA component are extracted by the function deconica::generate_markers with the parameter return = "gene.ranked". These genes are used to weight the component scores in each patient (the weighted-score of a given IC in patient p corresponds to the weighted mean expression of the 30-top genes on that component. We used, in the function deconica::get_scores, the log counts of the ICA as "df" parameter, the list of 30 genes as "markers.list" parameter, and the parameter summary = "weighted.mean". Finally, the estimated proportions are calculated from the inferred weighted-score with the function deconica::stacked_proportions_plot on the transpose of the deconica::get_scores output.

DNAm_wICA (m_WIC, ICA-based deconvolution on DNAm)

The method DNAm_wICA (m_WIC) uses DNA methylation data as input.

STEP1: feature selection It has no feature selection step.
STEP2: deconvolution The deconvolution step is based on ICA, similarly to what was described for the second step of RNA_wICA, but applied on the DNA methylation matrix.

both_wICA (b_WIC. ICA-based deconvolution on RNAD and DNAm)

The method both_wICA (b_WIC) combines transcriptomics and DNA methylation information.

STEP1: feature selection It has no feature selection step.
STEP2: deconvolution The deconvolution is in two steps, one on each data type. The transcriptomics and DNA methylation data are separately deconvoluted with the same deconvolution step as in r_WIC and m_WIC respectively to estimate A_{MET} and A_{RNA}.
STEP3: integration Finally, the mean of both A_{MET} and A_{RNA} estimated proportion matrices is computed as the final method output. To compute the average, the cell types of the both deconvolution matrices are matched by iteration. The cell types of the methylation result matrix are reordered 1000 times, and the one that best correlates with the transcriptomic result matrix is kept.

RNA_wNMF (r_WNM, NMF-based deconvolution on RNA)

The method RNA_wNMF (r_WNM), is a two step-approach that uses transcriptomic data as input.

STEP1: feature selection The first step uses ICA to perform a feature selection as described for RNA_wICA, although duplicated genes are kept. This step therefore allows genes that contribute to several components to be present several times in the data.
STEP2: deconvolution The deconvolution is based on sparse NMF and least-squares optimization to minimize $||D - TA||_2$ [26]. It is called by the NMF::nmf function, with the parameter method = "snmf/r".

DNAm_EDec (m_EDC, NMF-based deconvolution on DNAm)

STEP1: feature selection The method DNAm_EDec (m_EDC), uses DNA methylation data as input and follows the pipeline implemented in the R package medepir [7]. The feature selection is performed by medepir::feature_selection for keeping highly variable probes (5000 most variable probes).

STEP2: deconvolution The NMF-based algorithm of the method EDec [9] is used for the deconvolution part, with the function medepir::Edec and all the selected probes as "infloci" parameter. The algorithm consist in minimizing the error term $||D - TA||_2$ with constraints on methylation values: $0 \leq A \leq 1$ and $0 \leq T \leq 1$ and constraints on proportions $\sum_{k-1}^{K} A_{kn} = 1$ where A_{kn} is the proportion of the n_{th} sample for the k_{th} cell type.

DNAm_MeDeCom (m_MDC, NMF-based deconvolution on DNAm)

STEP1: feature selection The method DNAm_MeDeCom (m_MDC), uses DNA methylation data as input and is based on the pipeline of the R package medepir. The feature selection is performed as for DNAm_EDec above to select the 5000 most variable probes.
STEP2: deconvolution The deconvolution step, however, uses the MeDeCom R package [8]. It is run with the function MeDeCom::runMeDeCom, with the lambda parameter set to 0.01. As EDec implementation of NMF algorithm, MeDeCom algorithm consists in minimizing the error term $||D - TA||_2$ with constraints on methylation values: $0 \leq A \leq 1$ and $0 \leq T \leq 1$; and constraints on proportions $\sum_{k-1}^{K} A_{kn} = 1$ where A_{kn} is the proportion of the n_{th} sample for the k_{th} cell type. It also uses a regularization function that favors methylation values close to 0 or 1.

both_wNMFMeDeCom (b_COM, NMF-based deconvolution on RNA and DNAm)

The method both_wNMFMeDeCom (b_COM) combines transcriptomics and DNA methylation information. It is the combination of the two methods RNA_wNMF and DNAm_MeDeCom. The method r_WNM is first applied to the RNAseq matrix.

STEP1: feature selection The DNA methylation matrix is pre-treated as described in the m_MDC method, with the selection of 5000 most variable probes.
STEP2-3: deconvolution-integration Finally, the MeDeCom algorithm is run on the DNAm data, with the result of r_WNM as the initialization parameter startA.

both_meanwNMFMeDeCom (b_MEA, NMF-based deconvolution on RNA and DNAm)

The method both_meanwNMFMeDeCom (b_MEA), which integrates transcriptomics and DNA methylation, applies r_WNM to the transcriptomics matrix, m_MDC to the DNA methylation matrix.

STEP1: feature selection Feature selection is performed on D_{MET} and D_{RNA} matrices as described in r_WNM and m_MDC sections.
STEP2: deconvolution Deconvolution is performed on D_{MET} and D_{RNA} matrices as described in r_WNM and m_MDC sections to estimate A_{MET} and A_{RNA} matrices.
STEP3: integration We computed the mean of the two estimated A_{MET} and A_{RNA} matrices, similarly to b_WIC.

Availability and requirements

Project name: DECONbench
Project home page: https://competitions.codalab.org/competitions/27453
Operating system(s): Linux (CodaLab platform)/Debian (DECONbench)
Programming language: Python (CodaLab platform)/R (DECONbench)
Other requirements: none
License: Apache 2.0 (CodaLab platform)/CeCILL (DECONbench)
Any restrictions to use by non-academics: none

Abbreviations
DNAm: DNA methylation; FDR: False discovery rate; ICA: Independent component analysis; MAE: Mean absolute error; NMF: Non-negative matrix factorization; sd: Standard deviation; var: Variance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04381-4.

Additional file 1: Figure S1: DECONbench benchmark of OLS and RLR methods: an example of graphical outputs of new contributions to the benchmark. **Source code**.

Acknowledgements
We thank all members of the HADACA consortium for helpful discussion and contributions during the HADACA data challenge 2nd edition (November 2019, Aussois, France). We also thank Daniel Jost and the members of the BCM team for inspiring discussions during regular joint group meetings. We are grateful to the Codalab data challenge open source platform. The authors gratefully acknowledge the EpiMed core facility for their support and assistance in this work. This work is part of the national program Cartes d'Identité des Tumeurs supported by the Ligue Nationale Contre le Cancer. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

HADACA (Health Data Challenge) Consortium
Nicolas Alcala[6], Alexis Arnaud[2], Francisco Avila Cobos[7], Luciana Batista[8], Anne-Françoise Batto[9], Yuna Blum[3], Florent Chuffart[10], Jérôme Cros[5], Clémentine Decamps[1], Lara Dirian[11], Daria Doncevic[12], Ghislain Durif[13], Silvia Yahel Bahena Hernandez[14], Milan Jakobi[10], Rémy Jardillier[15], Marine Jeanmougin[16], Paulina Jedynak[10], Basile Jumentier[1], Aliaksandra Kakoichankava[17], Maria Kondili[18], Jing Liu[19], Tiago Maie[20], Jules Marécaille[11], Jane Merlevede[21], Maxime Meylan[3,22], Petr Nazarov[23], Kapil Newar[1], Karl Nyrén[14], Florent Petitprez[3], Claudio Novella Rausell[14], Magali Richard[1], Michael Scherer[24], Nicolas Sompairac[21], Katharina Waury[14], Ting Xie[25] and Markella-Achilleia Zacharouli[14]

1. Laboratory TIMC-IMAG, UMR 5525, Univ. Grenoble Alpes, CNRS, Grenoble, France. 2. Data Institute, Univ. Grenoble Alpes, Grenoble, France. 3. Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France. 4. INSERM U1068 CRCM, Marseille, France. 6. Section of Genetics, International Agency for Research on Cancer (IARC-WHO), Lyon, France. 7. Center for Medical Genetics Ghent, Department of Biomolecular Medicine, Ghent University, Ghent, Belgium. 8. Innate Pharma, Marseille, France. 9. Equipe Cancer et Immunité- INSERM Centre de Recherche des Cordeliers, Paris, France. 10. Institute for Advanced Biosciences, CNRS UMR 5309, Inserm, U1209, Univ. Grenoble Alpes, F-38700 Grenoble, France. 11. Verteego, Paris, France. 12. Health Data Science Unit, BioQuant Center and Medical Faculty Heidelberg, Germany. 13. Université de Montpellier, CNRS, IMAG UMR 5149, Montpellier, France. 14. Uppsala University, SE-751 05, Uppsala, Sweden. 15. University Grenoble Alpes, CEA, INSERM, IRIG, Biology of Cancer Infection UMR_S 1036, 38000 Grenoble, France & University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Institute of Engineering University Grenoble Alpes, 38000 Grenoble, France. 16. Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital - Oslo, Norway. 17. Vitebsk State Medical University & NatiVita, Vitebsk, Belarus. 18. Centre de Recherche de St. Antoine, Paris, AP-HP. 19. Institut Curie, PSL Research University, Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR144, Equipe Labellisée Ligue contre le Cancer, 75005 Paris, France. 20. Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, Aachen, Germany. 21. Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, F—75005, Paris, France. 22. INSERM U1138 Centre de Recherche des Cordeliers, France. 23. Quantitative Biology Unit, Luxembourg Institute of Health, L-1445 Strassen, Luxembourg. 14. Uppsala University, SE-751 05, Uppsala, Sweden. 24. Department of Genetics/Epigenetics, Saarland University, Saarbrücken, Germany. 25. Centre de Recherche en Cancérologie de Toulouse, Inserm UMR 1037, F-31037, Toulouse, France.

Authors' contributions
MR and YB conceived and designed the project. MR, YB, CD, AA, FP implemented the DECONbench platform. JC, AB, LA prepared the tissue samples and extracted the biological material for the benchmark dataset generation. IG and SE contributed to the development of the platform. MR, YB, CD, AA, FP, MA, RN, RT, AdR contributed to the benchmark dataset generation. HADACA consortium proposed and implemented the reference methods. MR, YB, CD, AA, FP wrote the manuscript. All authors read and approved the final manuscript.

Funding
The research leading to these results was supported by Univ. Grenoble-Alpes via the Grenoble Alpes Data Institute [MR, AA] (ANR-15-IDEX-02), EIT Health Campus HADACA and COMETH programs [MR, YB], activities 19359 and 20377 and the Ligue Nationale Contre le Cancer. Other fundings: South-Eastern Norway Regional Health Authority (project number 2019030 [MJ]), European IMI IMMUCAN project [NS], European Union's Horizon 2020 program (Grant 826121, iPC project, [JM, FAC]). This article did not receive specific sponsorship in the design of the study, analysis, interpretation of data and in writing the manuscript.

Availability of data and materials
DECONbench is hosted on the open source Codalab competition platform. It is freely available at: https://competitions.codalab.org/competitions/27453. Further documentation (online demo) is available at: https://deconbench.github.io/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Laboratory TIMC-IMAG, UMR 5525, CNRS, Univ. Grenoble Alpes, Grenoble, France. [2]Data Institute, Univ. Grenoble Alpes, Grenoble, France. [3]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France. [4]Universitat de Barcelona and Computer Vision Center, Barcelona, Spain. [5]LISN (INRIA/CNRS), Université Paris-Saclay, Gif-sur-Yvette, France. [6]INSERM U1068 CRCM, Marseille, France. [7]Dpt of Pathology, Beaujon Hospital, Univ. Paris-INSERM U1149, Clichy, France. [8]IGDR UMR 6290, CNRS, Université de Rennes 1, Rennes, France.

Received: 30 October 2020 Accepted: 20 September 2021
Published online: 02 October 2021

References

1. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34:1969–79.
2. Becht E, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17:1–20.
3. Nazarov PV, et al. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. BMC Med Genomics. 2019;12:1–17.
4. Blum Y, et al. Dissecting heterogeneity in malignant pleural mesothelioma through histo-molecular gradients for clinical applications. Nat Commun. 2019;10:1333.
5. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. Methods Mol Biol Clifton NJ. 2020;2117:135–57.
6. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30:1431–9.
7. HADACA Consortium, et al. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. BMC Bioinform. 2020;21:16.
8. Lutsik P, et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. 2017;18:1–20.
9. Onuchic V, et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. Cell Rep. 2016;17:2075–86.
10. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11:5650.
11. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol. 2021;22:102.
12. White BS, et al. Abstract 1690: A tumor deconvolution DREAM challenge: inferring immune infiltration from bulk gene expression data. Cancer Res. 2019;79:1690–1690.
13. Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average? Mol Syst Biol. 2011;7:537.
14. Buchka S, Hapfelmeier A, Gardner PP, Wilson R, Boulesteix A-L. On the optimistic performance evaluation of newly introduced bioinformatic methods. Genome Biol. 2021;22:1–8.
15. Mangul S, et al. Systematic benchmarking of omics computational tools. Nat Commun. 2019;10:1393.

16. Marx V. Bench pressing with genomics benchmarkers. Nat Methods. 2020;17:255–8.
17. Krusche P, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60.
18. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17:147–54.
19. Puleo F, et al. Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. Gastroenterology. 2018;155:1999-2013.e3.
20. Maurer C, et al. Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. Gut. 2019;68:1034–43.
21. Collisson EA, Bailey P, Chang DK, Biankin AV. Molecular subtypes of pancreatic cancer. Nat Rev Gastroenterol Hepatol. 2019;16:207–20.
22. Ellrott K, et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. Genome Biol. 2019;20:195.
23. Czerwinska U. UrszulaCzerwinska/DeconICA: DeconICA first release. Zenodo. 2018. https://doi.org/10.5281/zenodo.1250070.
24. fastICA: FastICA algorithms to perform ICA and projection pursuit. https://CRAN.R-project.org/package=fastICA.
25. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw. 1999;10:626–34.
26. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014;196:973–83.
27. Baron M, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 2016;3:346–360.e4.

Publisher's Note