Methodological developments for omic data integration : applications to oncology and neurosciences

Galadriel Briere

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

Par **Galadriel BRIÈRE**

DÉVELOPPEMENTS MÉTHODOLOGIQUES POUR L'INTÉGRATION DE DONNÉES OMIQUES : APPLICATIONS À L'ONCOLOGIE ET AUX NEUROSCIENCES

Sous la direction **Patricia THÉBAULT** et **Agnès NADJAR**

Soutenue le 23 novembre 2022

Membres du jury :

M.	Christophe AMBROISE	PR	Uni. Évry Val d'Essonne	Rapporteur
Mme.	Hélène HIRBEC	PR	Uni. Montpellier	Rapportrice
M.	Guillaume BLIN	PR	Uni. Bordeaux	Président du jury
M.	Laurent BRÉHÉLIN	CR	Uni. Montpellier	Examinateur
M.	Jacques VAN HELDEN	PR	Uni. Aix-Marseille	Examinateur
Mme.	Patricia THÉBAULT	PR	Uni. Bordeaux	Directrice de thèse
Mme.	Agnès NADJAR	PR	Uni. Bordeaux	Co-directrice de thèse
Mme.	Raluca URICARU	MCF	Uni. Bordeaux	Co-encadrante, Invitée

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

Par **Galadriel BRIÈRE**

METHODOLOGICAL DEVELOPMENTS FOR OMICS DATA INTEGRATION: APPLICATIONS TO ONCOLOGY AND NEUROSCIENCES

Sous la direction **Patricia THÉBAULT** et **Agnès NADJAR**

Soutenue le 23 novembre 2022

Membres du jury :

M.	Christophe AMBROISE	PR	Uni. Évry Val d'Essonne	Rapporteur
Mme.	Hélène HIRBEC	PR	Uni. Montpellier	Rapportrice
M.	Guillaume BLIN	PR	Uni. Bordeaux	Examinateur
M.	Laurent BRÉHÉLIN	CR	Uni. Montpellier	Examinateur
M.	Jacques VAN HELDEN	PR	Uni. Aix-Marseille	Examinateur
Mme.	Patricia THÉBAULT	PR	Uni. Bordeaux	Directrice de thèse
Mme.	Agnès NADJAR	PR	Uni. Bordeaux	Co-directrice de thèse
Mme.	Raluca URICARU	MCF	Uni. Bordeaux	Co-encadrante, Invitée

Remerciements

Je souhaite avant tout remercier mes directrices et encadrantes de thèses, Agnès Nadjar, Patricia Thébault et Raluca Uricaru, qui m'ont guidée et soutenue avec la plus grande bienveillance tout au long de ces trois années.

Agnès, je te remercie pour la patience dont tu as fait preuve et de m'avoir toujours ramenée à la question biologique lorsque j'avais tendance à m'égarer dans des directions parfois trop éloignées de la réalité pratique. Tu m'as en quelque sorte permis de "garder les pieds sur terre" et de ne pas oublier la vraie raison pour laquelle on effectuait ces recherches.

Patricia, je te remercie pour ton enthousiasme sans faille, qui m'a poussé plus haut et plus loin. Ton optimise et ton énergie, tu me les as toujours transmis et dans les moments difficiles, lorsque je n'y croyais plus, tu y as cru pour moi sans jamais douter.

Raluca, je te remercie pour la confiance que tu m'as accordée, dès nos premières interactions quand j'étais encore complètement étrangère au monde de la bioinformatique. Tu m'as toujours rassurée aux moments où j'étais emprise du syndrome de l'imposteur, tu m'as fait relativiser et prendre confiance en moi. Je te remercie également pour tous tes conseils, tes longues relectures de papiers et de manuscrit qui étaient toujours très minutieuses.

Agnès, Patricia, Raluca, vous avez chacune contribué à cette thèse à votre manière, et vous vous êtes toutes les trois parfaitement complétées. Je n'aurai pas pu rêver meilleur encadrement. Puissiez-vous encore encadrer de nombreuses thèses !

Je remercie également très chaleureusement Élodie Darbo, qui a été très impliquée dans la première partie de cette thèse. Élodie, tout au long de cette thèse, tu as été ma référente technique, celle a qui je pouvais toujours poser une question sur telle ou telle analyse ou tel code en R. Mais plus qu'une collègue, tu es devenue une vraie amie.

Je souhaite également remercier toutes les personnes qui ont contribué à cette

thèse, de près ou de loin: bien sur, mon école doctorale et sa direction, mais également l'ensemble de mon équipe BKB, ainsi que Guillaume Blin et Xavier Fioramonti, qui ont constitué mon comité de suivi, sans oublier tous les collègues du LaBRI avec qui j'ai eu l'occasion d'échanger dans les couloirs ou à la pause café.

Je remercie mon ancienne école, Bordeaux Sciences Agro, qui m'a soutenue dans ma réorientation en bioinformatique, qui m'a laissée partir en année de césure alors que j'étais en retard dans mon dossier, et hors des clous pour certaines démarches administratives. Je sais que sans votre soutiens et votre compréhension, cette réorientation n'aurait pas été possible.

Je remercie mes collègues et amis, rencontrés durant cette thèse. Bien sur, en premier lieu, l'ensemble du bureau 328, et en particulier ces personnes qui sont devenues bien plus que de simples collègues: Myriam (la fro de la brioffe), Elsa (gâteau au chocolat), Claire (sous les tropiques). Nous avons formé un véritable quatuor et je sais que nous le resterons, même avec le temps et la distance. Je n'oublie bien sur pas Cécile, Luc (hinhin), Maxime (Ximema), Chabname (Chab'), Kévin, Tu, Aaron, Loann, Warren, et tous les autres. Merci à tous, c'était un véritable plaisir de faire votre connaissance et de partager ces moments avec vous.

Enfin, je remercie ma famille et mes amis qui m'ont toujours soutenue et encouragée: bien sur ma maman, sur qui je peux toujours compter de manière inconditionnelle. Mon papa, qui m'a transmis son amour pour la science, les énigmes et la logique. Mes soeurs, toujours présentes pour moi, même à l'autre bout du monde. Mon chéri, qui m'a vue dans tous mes états pendant cette thèse mais qui a toujours tout fait pour me remonter le moral, qui ne m'a fait aucun reproche, même quand j'ai du annuler à la dernière minutes des vacances prévues depuis des mois. Et enfin mes amies Manue et Muriel, sur qui je peux compter depuis des années maintenant et mon ami Antoine, qui est même venu deux fois assister à ma soutenance de thèse (hihi).

Dédicace

Je dédie cette thèse mon père, Thierry Brière.

Résumé

Développements méthodologiques pour l'intégration de données omiques: applications à l'oncologie et aux neurosciences

Les données dites "omiques", sont des données massives et hétérogènes, issues de la mesure de différents objets biologiques. Par exemple, la génomique s'intéresse à l'étude du génome (ADN), la transcriptomique à l'étude des transcrits (ARNs), la protéomique à l'étude des protéines, etc. L'interaction de l'ensemble de ces omiques entre elles ainsi qu'avec des facteurs environnementaux produit - à l'échelle d'une cellule, d'un tissu, ou d'un organisme - un ensemble de caractères observables appelé phénotype. Un des objectifs ultimes de la recherche en sciences de la vie est l'élucidation de la diversité du phénome (c'est-à-dire de l'ensemble des phénotypes observables) par l'identification des facteurs internes, environnementaux et de leurs interactions, associés à chaque phénotype.

Ce manuscrit de thèse aborde la question de l'intégration de données - définie comme une solution permettant l'utilisation de multiples sources d'information (données) pour mieux comprendre un système, une situation, une association, etc. - et particulièrement de la question de l'intégration de données omiques, c'est-à-dire tout type d'intégration

de sources de données provenant de différentes omiques, et/ou d'une même omique mesurée dans différents contextes expérimentaux et/ou de données omiques avec un type de données non-omique.

Dans une première contribution, nous proposons une nouvelle stratégie pour le clustering consensus de données multi-omiques pour la détection de sous-types moléculaires de cancers. Cette stratégie permet, à partir de clusterings de cohortes de patients obtenus en considérant diverses données omiques et/ou différents algorithmes de clusterings existants, de produire un clustering consensus en réconciliant l'ensemble des prédictions contenues dans les clusterings soumis en entrée de l'algorithme. Deux scénarios d'intégration ont été testés : une intégration dite "multi-to-multi", produite par intégration de clusterings multi-omiques et une intégration dite "single-to-multi", produite par l'intégration de clusterings générés indépendamment pour différents omiques.

Dans une seconde contribution, nous proposons une stratégie de détection de groupes de liens différentiellement co-exprimés identifiés par la comparaison de plusieurs jeux de données de type cas/contrôle. Elle repose sur la construction et l'analyse de réseaux multi-couches de co-expression différentielle, chaque couche représentant l'ensemble des dérégulations de la co-expression génique observée pour un contexte expérimental donné. La détection de groupes de liens de co-expression différentielle topologiquement similaires (c'est-à-dire impliquant un même ensemble de gènes) et observées dans les mêmes sous-ensembles de couches du réseau permet d'identifier des mécanismes associés à une maladie dans différents contextes expérimentaux (tissus, stade de développement, etc.), ou associés à différentes maladies.

Nous avons appliqué la stratégie développée à la détection de motifs de co-expression différentielle dans l'hippocampe et le cortex de souris modèles de la maladie d'Alzheimer, ce qui nous a permis d'identifier des motifs clés de dérégulation de l'expression génique associés au phénotype pathologique. Certains de ces motifs ont été observés dans le cortex comme dans l'hippocampe, tandis que d'autres apparaissent spécifiques à l'une ou l'autre des deux structures cérébrales. Cette preuve de concept démontre la pertinence de la stratégie pour l'identification de perturbations de la co-régulation génique et la caractérisation transcriptomique de la diversité des phénotypes.

Mots-clés Données omiques ; Intégration de données ; Oncologie ; Neurosciences ; Bioinformatique

Abstract

Methodological developments for omics data integration: applications to oncology and neurosciences

"Omics" data are massive and heterogeneous data types, obtained from the measurement of different biological objects. For example, genomics is the study of the genome (DNA), transcriptomics is the study of transcripts (RNAs), proteomics the study of proteins, etc. The interaction of all these omics with each other and with environmental factors produces - at the scale of a cell, a tissue, or an organism - a set of observable characteristics called phenotype. One of the ultimate goals in life science research is the elucidation of the diversity of the phenome (i.e., the set of observable phenotypes) by identifying the internal and environmental factors and their interactions associated with each phenotype.

This thesis manuscript addresses the issue of data integration - defined as a solution allowing the use of multiple sources of information (or data) to better understand a system, a situation, an association, etc. - and particularly the issue of omics data integration, i.e., any kind of integration of data sources coming from different omics, and/or from the same omics measured in different experimental contexts and/or from omics data with a non-omics data type.

In a first contribution, we propose a novel strategy for the consensus clustering of multi-omics data, designed for the prediction of molecular subtypes of cancers. This strategy aim, from a set of clusterings of a patient cohort obtained by considering various omics data and/or different existing clustering algorithms, to produce a consensus clustering by reconciling all the predictions contained in the clusterings submitted as input to the algorithm. Two integration scenarios were tested: a "multi-to-multi" integration scenario, through the integration multi-omics clusterings obtained from existing integrative clustering strategies, and a "single-to-multi" integration scenario, through the integration of single-omics clustering independently produced for several omics.

In a second contribution, we propose a novel strategy for detecting differentially co-expressed link communities by comparing co-expression patterns in multiple case/control datasets. The strategy is based on the construction and analysis of multi-layer differential co-expression networks, each layer representing a set of dysregulations of gene pairwise co-expression observed in a given experimental context. The detection of topologically similar (i.e., involving the same set of genes) link communities consistently observed across subsets of layers of the network allows the identification of molecular mechanisms associated with a disease in different experimental contexts (tissues, developmental stage, etc.), or associated with multiple diseases. We applied this strategy for the detection of differential co-expression patterns in the hippocampus and the cortex of Alzheimer's disease model mice, allowing the identification of key gene co-expression dysregulation patterns associated

with the pathologic phenotype. Some of these patterns were observed in both the cortex and the hippocampus, while others appeared to be specific to one or the other of the two brain structures. This proof of concept demonstrates the relevance of the strategy for identifying gene co-regulation perturbations and to characterize the transcriptomic diversity of phenotypes associated with disease.

Keywords Omics data ; Data integration ; Oncology ; Neurosciences ; Bioinformatics

Résumé étendu

Développements méthodologiques pour l'intégration de données omiques: applications à l'oncologie et aux neurosciences

Les données dites "omiques", sont des données massives et hétérogènes, issues de la mesure de différents objets biologiques. Par exemple, la génomique s'intéresse à l'étude du génome (ADN), la transcriptomique à l'étude des transcrits (ARNs), la protéomique à l'étude des protéines, etc. L'interaction de l'ensemble de ces omiques entre elles ainsi qu'avec des facteurs environnementaux produit - à l'échelle d'une cellule, d'un tissu, ou d'un organisme - un ensemble de caractères observables appelé phénotype. Un des objectifs ultimes de la recherche en sciences de la vie est l'élucidation de la diversité du phénome (c'est-à-dire de l'ensemble des phénotypes observables) par l'identification des facteurs internes, environnementaux et de leurs interactions, associés à chaque phénotype.

Contexte Les récentes avancées en matière de séquençage et d'acquisition de données ont permis l'augmentation du volume et de la diversité des données collectées en biologie, ainsi que le développement de nouveaux modèles de recherche, comme la médecine de précision. Ainsi, de plus en

plus d'objets biologiques différents, identifiés par type d'*omiques*, sont susceptibles d'être mesurés et analysés. Les données sont également de plus en plus accessibles à la communauté scientifique grâce au développement de nombreuses bases de données spécialisées dans le dépôt de données omiques et non-omiques, comme des données de connaissance (interactions protéiques, voies biologiques, annotations géniques, etc.), des métadonnées telles que des données cliniques, etc.

Les omiques sont aujourd'hui mesurées de façon routinière et de nombreuses méthodes et outils spécialisées pour l'analyse de chaque type d'omique ont été et sont toujours développées. Mais si l'étude individuelle de chaque type de données est très informative, leur analyse simultanée peut permettre de révéler de nouveaux motifs d'interactions entre les omiques, une signature moléculaire, en lien avec un phénotype d'intéret. C'est pourquoi la question de l'intégration de données est une problématique essentielle en bioinformatique et en médecine, afin de mieux comprendre les mécanismes moléculaires associés à différentes maladies.

Nous définissons la notion d'intégration de données en tant que solution permettant l'utilisation de multiples sources d'information (données) pour mieux comprendre un système, une situation, une association, etc., selon la définition proposée par [Gomez-Cabrero et al.]. Nous définissions la notion d'intégration de données omiques comme tout type d'intégration de multiples sources de données provenant de différentes omiques, et/ou d'une même omique mesurée dans différents contextes expérimentaux et/ou de données omiques avec un type de données non-omique.

Dans ce manuscrit de thèse, je considère la question de l'intégration de données omiques par le développement de deux nouvelles stratégies pour l'analyse et l'intégration de données omiques. La première permet l'intégration de différents types données omiques par une analyse de "clustering consensus" et a été appliquée dans un contexte de sous-typage moléculaire de patients atteints de différents types de cancer. La seconde permet la détection de motifs de différence de co-expression génique à travers plusieurs conditions expérimentales et a permis d'identifier des motifs de co-expression différentielle dans l'hippocampe et le cortex de souris modèles de la maladie d'Alzheimer.

Clustering Consensus pour le sous-typage de cancers Un cancer peut se développer sous différents sous-types moléculaires, selon les mutations et les gènes impactés. Ces sous-types moléculaires ne présentent pas les mêmes mécanismes de développement de la maladie, et répondent différemment aux traitements. Ainsi, l'identification du sous-type est cruciale pour la mise en place de traitements adaptés à chaque individu. Cette prédiction du sous-type de cancer a longtemps été réalisée en considérant spécifiquement les profils d'expression génique (transcriptomique) des patients, par une approche de clustering, c'est-à-dire en regroupant les individus de telle sorte que les patients au sein d'un groupe présentent des caractéristiques transcriptomiques similaires entre eux, mais dissimilaires avec les individus des autres groupes. Cependant, d'autres types d'omiques participent à la mise en place du phénotype pathologique, et une intégration d'autres sources de données omiques pourrait permettre une compréhension bien plus exhaustive

des mécanismes moléculaires sous-jacents. Ainsi, récemment, l'effort a été mis sur le développement de méthodes pour le sous-typage multi-omique de cancer, basées sur des stratégies de clustering mutli-omique.

Plusieurs outils ont été développés à cet effet, dont certains ont été testés et comparés par [Rappoport et Shamir] pour le sous-typage de 10 types de cancer par l'analyse de données d'expression génique, de méthylation de l'ADN, et d'expression de micro-ARN. Cette évaluation comparative a mis en évidence qu'aucune des stratégies testées n'était meilleure que toutes les autres sur la base des métriques de qualité évaluées dans cette étude, et qu'il n'était ainsi pas possible de recommander l'utilisation d'une stratégie en particulier. Cela illustre une problématique récurrente en bioinformatique - que ce soit dans un contexte multi-omique ou single-omique (un seul type de données considéré) - qui est la question de l'estimation de la qualité biologique de résultats d'analyse, et celle du choix d'une stratégie d'analyse parmi l'ensemble des méthodes existantes.

Face à l'hétérogénéité des données omiques et la diversité des méthodes de clustering single- et multi-omiques, nous avons développé une stratégie de clustering consensus qui permet de s'attaquer au problème de l'intégration de données omiques tout en tirant partie des stratégies de clustering existantes. La stratégie repose sur la notion de Clustering par Accumulation de Preuve (CAP), introduite par [Fred et Jain]. Dans ce contexte, on considère chaque co-occurrence de patients dans les clusters prédits par un algorithme de clustering existant comme une preuve de l'association de ces patients. C'est l'accumulation de ces preuves par l'inclusion de divers clusterings produits par d'autres méthodes et/ou

sur d'autres jeux de données qui permet de distinguer les associations les plus "supportées" par l'ensemble des jeux de données et des outils de clusterings utilisés sur ces jeux de données.

La stratégie d'intégration par clustering consensus développée au cours de cette thèse est détaillée dans ce manuscrit, et la publication associée est accessible avec la citation suivante :

[Brière et al.] Galadriel Brière, Élodie Darbo, Patricia Thébault, and Raluca Uricaru. Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics*, 22(1):361, July 2021, doi: 10.1186/s12859-021-04279-1.

Dans cette étude, nous avons considéré la question du sous-typage multi-omique par deux approches:

(i) Par une approche "Multi-to-Multi", nous choisissons de tirer partie de l'ensemble des stratégies de clustering multi-omique existantes et de réconcilier les prédictions réalisées par ces différents outils. L'algorithme prend donc en entrée un ensemble de clusterings multi-omiques et produit un consensus mutli-omique en réconciliant les prédictions contenues dans les clusterings soumis.

(ii) Par une approche "Single-to-Multi", nous produisons d'abord un ensemble de clusterings pour chacun des omiques considérés et le consensus mutli-omique est produit en réconciliant les prédictions single-omiques contenues dans chaque clustering single-omique considéré en entrée.

Nous avons testé notre outil sur un ensemble de 10 types de cancer

pour l'intégration de 3 omiques et en comparaison avec des stratégies existentes. Les résultats obtenus démontrent l'intérêt de telles méthodes consensus pour le sous-typage et l'intégration de données omiques.

Intégration de réseaux de co-expression différentielle La question de l'intégration des données ne se limite pas à l'analyse conjointe de divers types de données omiques, puisque la diversité des conditions expérimentales sous lesquelles ces données sont récoltées peut également faire émerger un besoin d'intégration pour élucider les mécanismes moléculaires associés à différents phénotypes.

Nous nous sommes intéressées à la caractérisation des motifs de co-expression génique différentielle associés à l'hippocampe et au cortex d'un modèle murin de la maladie d'Alzheimer. L'objectif était d'identifier des motifs de dérégulation transcriptomique spécifiques à une structure cérébrale, ou observés de manière récurrente dans l'hippocampe et le cortex de ces souris modèles.

L'analyse de co-expression génique permet d'identifier, à travers un ensemble d'échantillons, des motifs d'association des gènes. Cette analyse vise à identifier des groupes (modules) de gènes dont le profil d'expression est similaire dans l'ensemble des échantillons observés. Il a été montré que les gènes co-exprimés présentent généralement une similarité fonctionnelle, ce qui permet d'inférer des informations sur des gènes peu annotés en utilisant un concept de "culpabilité par association" (*guilt by association*). Ces modules de gènes peuvent donc être associés à des fonctions biologiques et des phénotypes d'intérêt. Cependant, pour les jeux de données de type cas/contrôle, l'analyse de co-

expression ne permet pas directement de mener une analyse comparative des phénotypes inclus, et une analyse de co-expression différentielle est plus appropriée. Les stratégies d'analyse de co-expression différentielle permettent d'identifier des associations géniques *conditionnelles*, c'est-à-dire des liens de co-expression qui apparaissent ou disparaissent en fonction selon les conditions expérimentales. Ces approches ont pour but d'identifier les motifs de dérégulation de la co-expression génique et sont particulièrement utilisées dans la comparaison de groupes d'échantillons sains et malades pour identifier les perturbations transcriptomiques associées à une maladie d'intérêt.

Dans notre étude d'un jeu de données issues de mesures transcriptomiques sur l'hippocampe et le cortex de souris témoins et de souris modèle de la maladie d'Alzheimer, nous avons construit un réseaux de co-expression différentielle composé de deux couches, chacune représentant les dérégulations observées respectivement dans l'hippocampe et le cortex de souris modèle en comparaison avec les échantillons témoins correspondants. Puis, nous avons appliqué une stratégie de détection de communauté dans les réseaux multi-couche, inspirée des travaux de [Ahn et al.] et de [Salem et Ozcaglar] portant sur la recherche de communautés de lien dans les réseaux single- et multi-couches. Contrairement aux méthodes traditionnelles de clustering de graphe qui consistent à identifier des groupes de nœuds dans un graphe, notre stratégie identifie des groupes d'arêtes, similaires d'un point de vue topologique et co-occurentes à travers les couches du réseau. Ce type de partitionnement à l'avantage de considérer le contexte des associations observées (c'est-à-dire en tenant compte de la ou des couches dans

lesquelles elles apparaissent). Les sets de genes obtenus en considérant les gènes impliqués dans chaque communauté de lien sont chevauchants, c'est-à-dire qu'un même gène peut être retrouvé dans différentes communautés, ce qui est une force dans ce contexte puisqu'un même gène peut être associé à plusieurs fonctions biologiques, où ne pas intéragir avec les mêmes ensembles de gènes en fonction du contexte expérimental.

A l'issue de cette analyse, nous proposons un ensemble de gènes d'intérêt pour la recherche liée à la maladie d'Alzheimer, potentiels acteurs clés dans le développement du phénotype pathologique.

Références

[Gomez-Cabrero et al.] D. Gomez-Cabrero, et al., *Data integration in the era of omics: current and future challenges*, BMC Systems Biology, **8**, (2014).

[Rappoport et Shamir] N. Rappoport, et R. Shamir, *Multi-omic and multi-view clustering algorithms*, Nucleic Acids Research, **46**, (2018).

[Fred et Jain] A. Fred, et A. Jain, Evidence accumulation clustering based on the k-means algorithm, Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, (2002).

[Brière et al.] G. Brière, et al., *Consensus clustering applied to multi-omics disease subtyping*, BMC Bioinformatics, **22**, (2021).

[Ahn et al.] Y. Ahn et al., *Link communities reveal multiscale complexity in networks.*, Nature, **466**, (2010).

[Salem et Ozcaglar] S. Salem et C. Ozcaglar, *Hybrid coexpression link similarity graph clustering for mining biological modules from multiple gene expression datasets*, BioData Mining, **7**, (2014).

Contents

List of Figures

Chapter 1

Background and definitions

Living organisms can be modeled as complex and open systems, whose functioning depends on many components in constant interaction within themselves and with their environment, and whose understanding and characterization can only be done through an interdisciplinary and experience-based study.

With the recent increase in the diversity and quantity of data measured in biology, the modelling of complex and multi-level processes in the cell has become a new challenge in bioinformatics studies. More specifically, the objective is to provide mathematical and and computational tools to improve our understanding on the relationships between genotype and phenotype.

To answer this question, it is essential to identify the different biological objects in all their diversity and their connections/relations with the aim to address the complexity of a biological system.

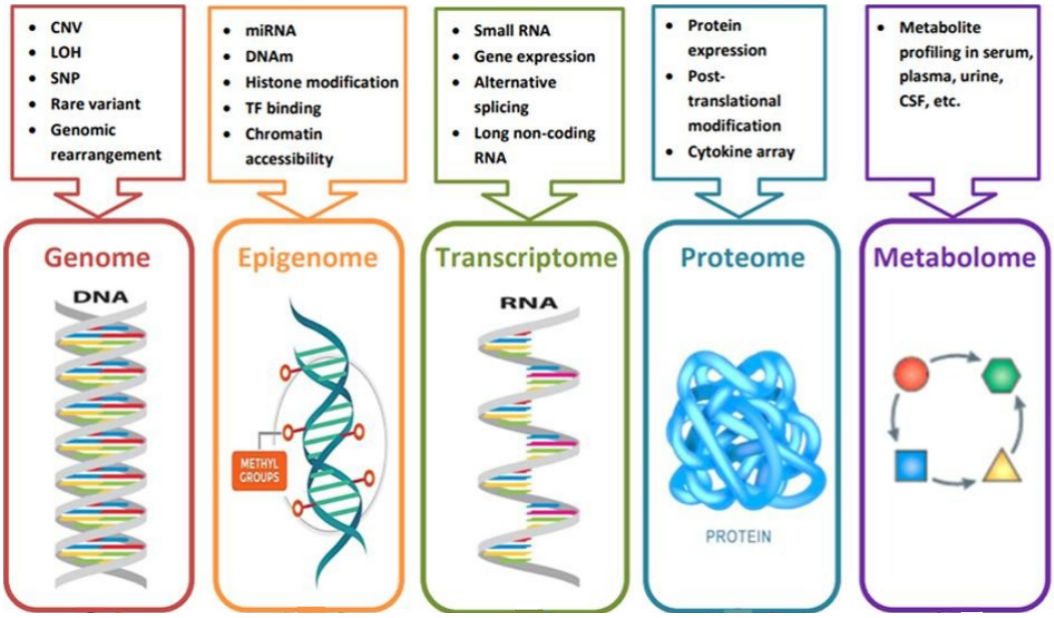


Figure 1.1: Different types of omics data
Adapted from Momeni et al., Journal of Biomedical Informatics, 2020 [1]

1.1 The need for data integration strategies in life sciences

In bioinformatics, the study of the various molecular components of living systems are often called "omics", referring to different families of molecules, from DNA for genomics, to RNA for transcriptomics, proteins for proteomics, and so on... All of these omics participate in the organization of living organisms, and the study of interaction mechanisms within and across omics is essential to characterize a phenotype extensively.

The major omics families are represented in Figure 1.1, adapted from Momeni et al., Journal of Biomedical Informatics, 2020 [1].

To guide the reader, and before addressing the problem of the data integration with its many variations, the following section will provide the essential concepts to understand the richness of biological data that became central in bioinformatics.

1.1.1 "Omics" and their interactions

Each type of omics can generally be defined as measures of a specific biological component. The improvement of data production techniques based on each of these elements has experienced an unprecedented increase since the beginning of this century. In close connection with the ability to measure these biological components, new "omics" terms are regularly suggested to the scientific community. The following paragraphs propose to define and illustrate some of the major omics types and their potential in life science research.

Genomics refers to the study of DNA, the molecule that carries genetic information of all known organisms and DNA-viruses. DNA encodes all the instructions necessary for the essential processes of living beings, from their development to their functioning, growth and reproduction. The structure, variability, evolution and function of DNA are therefore properties that are widely studied. Structural patterns studied in DNA include Copy Number Variations (CNV), i.e., the number of copies of a segment in the genome for a given species, or genomics rearrangements. DNA-variability is also studied via the detection of Single Nucleotide Polymorphism (SNP) and insertion-deletion (indels) events. Functional genomics focuses on identifying and annotating DNA sequences into functional units.

Epigenomics refers to the study of all epigenetic modifications of the DNA, i.e., all the reversible modifications of the DNA which do not affect its sequence but which modify its accessibility, its compaction

and/or its transcription into RNA. The study of the epigenome is a crucial key to understanding the relations between genotype and phenotype. Essential epigenetic events include micro-RNA (miRNA) and Transription Factors (TF) binding to DNA, DNA mehtylation, histone modification and chromatin accessibility.

Transcriptomics refers to the study of RNA, product of DNA transcription. RNA plays multiple roles, acting as a transmitter of genetic information (messenger-RNA or mRNA, translated into protein), as a catalyst for certain metabolic reactions (ribozymes, ribonucleic acid enzymes), but also as a transcriptional regulator (ribosomal-RNA rNRA, transfert-RNA tRNA, ...), post-transcriptional regulator (small-RNA sRNA, in particular micro-RNA, miRNA) and epigenome regulator. Transcriptomic studies aim at characterizing RNA in terms of sequence, structure, function, and expression.

Proteomics refers to the study of proteins, which play many functions, from catalysis of all chemical reactions constituting the metabolism, regulation of gene expression, to transmission of cellular signals, intracellular transport, as well as cellular and tissue structure. Proteomics aims at characterizing proteins by their sequence, localization in the cell, structure, function and expression.

Metabolomics refers to the study of the metabolome (i.e, the set of metabolites intermediate and end products of metabolism) and its characteristics regarding various environments.

1.1.2 Phenome

The five omics presented above already merit a joint study to offer a global characterization of a cell, all other conditions being fixed. Nevertheless, the phenotype of a cell (i.e., the set of its observable traits), is the result of an interaction between its genotype and its environment, and therefore, the characterization of a cell , tissue or organism must also be done according to the set of all its possible phenotypes, also known as its phenome, which depends on both internal and external factors.

Internal factors For multi-cellular organisms, the phenotype can be defined at the cell level or at the organism level. The genome is obviously a main factor of the phenotype. At the cellular level, the phenotype mainly varies according to the tissue considered. At the organism scale, the phenotype also varies according to the development stage (developmental phenotypes and aging phenotypes).

External factors Organisms are in constant interaction with their environment, and are subjected to different types of stress (biotic and abiotic) with diverse effects on the phenotype. They can also evolve in cooperation with other organisms, which will influence their own phenotypic traits (symbiosis, parasitism, microbiota).

All of these factors, genetic and environmental, often in a coordinated manner, can cause perturbations that lead to the development of a pathological phenotype. By describing and highlighting the mech-

anisms responsible for the pathology through all layers of omics, it is possible to identify key factors in the establishment or the regulation of the disease.

The measurement of these different omics profiles has allowed the development of new medical disciplines, such as precision medicine which aim at evaluating the multi-omics characteristics of each patient to propose treatments tailored to each molecular profiles; but also new fields of research in data analysis for the integration of these massive and heterogeneous information.

1.1.3 A plethora of data

As previously described, the diversity of omics data has the great advantage to help researchers to increase the quality of models or predictions. These data can be complemented by a wealth of non-omics data, such as knowledge data already acquired by the scientific community. The following section focuses on assessing the richness of the different types of omics and non-omics data as well as their accessibility.

1.1.3.1 Types of data in life sciences

Omics data There are two main types of data when considering omics:

Sequence data: DNA sequences (genomics - DNA-sequencing), RNA sequences (transcriptomics - RNA-sequencing), transcription sites and other protein binding sites (epigenomics - ChIP-sequencing), DNA methylation sites (epigenomics - bisulfite-sequencing), proteins (proteomics - protein-sequencing), etc.

Abundance data: counts (transcriptomics - RNA-sequencing), arrays (transcriptomics - microarrays; epigenomics - methylation arrays), metabolite quantification (metabolomics - mass spectometry, Nuclear Magnetic Resonance spectroscopy), protein quantification (proteomics - mass spectometry), etc.

Sequence data are used to produce genome assemblies, homology analysis, functional annotation, SNPs and other variants detection, and identifying protein binding sites or methylation sites.

Abundance data are used to perform comparative analysis of RNA, methylation, proteins or metabolites levels with respect to various experimental conditions, or as representations of a system's molecular profile.

Metadata Omics data are often complemented by metadata to describe the overall conditions of the experiment. It includes clinical data, that describe the samples of the omics-analysis. Clinical data can take many forms, from quantitative data from various clinical tests or samples description (age, drug dosage, survival, weight, ..), qualitative data (gender, ethnicity, tumor stage, background information, ...) or even image data (e.g., MRI).

Knowledge data Finally, omics data can be considered with respect to knowledge data, that can also take many forms, from networks (Protein-Protein Interaction PPI network, molecular pathways, ...), ontologies (Gene Ontology GO, Disease ontology, ...), gene-sets (hallmark gene-sets, regulatory gene-sets, ...), genome assemblies and known vari-

ants, etc.

1.1.3.2 Data accessibility

Many efforts have been made by the scientific community to make all of this data, metadata and knowledge increasingly accessible, with the creation of numerous public databases. For example, we can cite knowledge databases on genes and their sequences such as GenBank [2], on protein interactions such as Stringdb [3], on biological processes such as KEGG Pathways [4] and WikiPathways [5], or which propose annotated gene-sets such as the Gene Ontology [6] or the molecular signature database MsigDB [7]. The Gene Expression Omnibus (GEO) database [8] originally hosted RNA-arrays data, but has extended to host all types of gene expression data as well as other omics-sources for a wide range of organisms and studies. Moreover, some databases gather knowledge data, experiment data and metadata for specific research fields such as The Cancer Genome Atlas (TCGA) [9] for cancer research, or Synapse [10] which proposes portals for brain and mental health research or for Alzheimer's disease.

These data, hosted on public databases, are both available to the scientific community and easily accessible, the bioinformatics community being very active in the development of solutions to facilitate the retrieval of these data. Many Application Programming Interfaces (API) have been developed, for example the Bioconductor/R GEOquery packages, to access GEO database data, or msigdb to access the MsigDB molecular signature database, among many others.

1.2 A definition of data integration

Over the past decades, the amount of biological data being gathered has increased enormously thanks to advances in biotechnology, in particular the new sequencing technologies. Today, the amount of data available and its great diversity offer new opportunities for researchers to develop novel, faster, more efficient and sophisticated analysis strategies in order to extract knowledge from these massive and complex data. From this abundance of available data, new areas of research are emerging, such as precision medicine, which aims to provide care tailored to each molecular profile.

With the increase of available data, knowledge is growing, organized in numerous databases and made available to the scientific community: e.g., ontologies, known molecular pathways, protein-protein interactions, specialized atlases, candidate disease genes, ... This knowledge can be used not only to describe analysis results (for example, for annotating gene sets), but also directly during analysis, using supervised learning algorithms for instance.

With this abundance of massive and heterogeneous data comes a need to develop new methods in order to take advantage of all these resources, which taken independently, reveal specific but also complementary patterns of interest. Indeed, a simultaneous analysis of heterogeneous data sources could reveal new interaction patterns in the data, which would not be observed by analyzing each data source independently. This has motivated the development of a research field interested in the implementation of new analysis strategies for data

integration.

Although the issue of data integration is gaining more and more attention in the scientific community, few publications actually define this notion.

Introduced in the context of accession to databases sharing overlapping content, the term "data integration" initially considered only the aspect of data access. In life sciences, this aspect of data integration has been studied: Lapatas and co-authors refer to data integration as *"the computational solution allowing users, from end user (GUI) to power users (API), to fetch data from different sources, combine, manipulate and re-analyze them as well as being able to create new datasets and share these again with the scientific community"* [11].

Many efforts have been made to provide unified access to data gathered from multiple sources through public databases. Moreover, efforts have been made to introduce standard data format for each data type: fastq/fasta files for sequences, SAM files for alignments, etc. Important efforts have also been made to define ontologies, formal systems for modelling concepts and their relationships in diverse domains. These systems are crucial to organize knowledge data and make it widely accessible to the scientific community. There is even an initiative to improve the interoperability of these ontologies (The Open Biomedical Ontologies (OBO) Foundry [12]).

Therefore, currently in life sciences, the critical aspect of data integration is not accessing data from several sources, but rather their joint exploration. Hence, in this thesis, data integration is defined as *"the use of multiple sources of information (or data) to provide a better*

understanding of a system/situation/association/etc", as described by Gomez-Cabrero and co-authors in [13].

Moreover, in this manuscript, the terms "omics data integration" will not only refer to the integration of multiple omics data types but also to any (single- or multi-) omics data integration with other non-omics data types such as clinical data or knowledge data.

1.3 Different types of data integration

Data integration can take many forms, according to the experimental context and biological question it is applied to.

1.3.1 Depending on the structure of the data

Data integration can take three major forms depending on the data to fuse and the relationship between the datasets: vertical data integration, horizontal data integration and a "diagonal" data integration, which is neither vertical nor horizontal. These are represented in Figure 1.2, adapted from Eidem et al., BMC Medical Genomics, 2018 [14].

1.3.1.1 Vertical integration

Vertical data integration considers the same set of samples, for which different data sources have been measured. The various data sources to fuse might consider related objects (e.g., correspondence between genes and proteins when considering genomics and proteomics data) or entirely unrelated objects (e.g., genes from genomics data and clinical metadata), i.e., with or without any bipartite relations between the data sources.

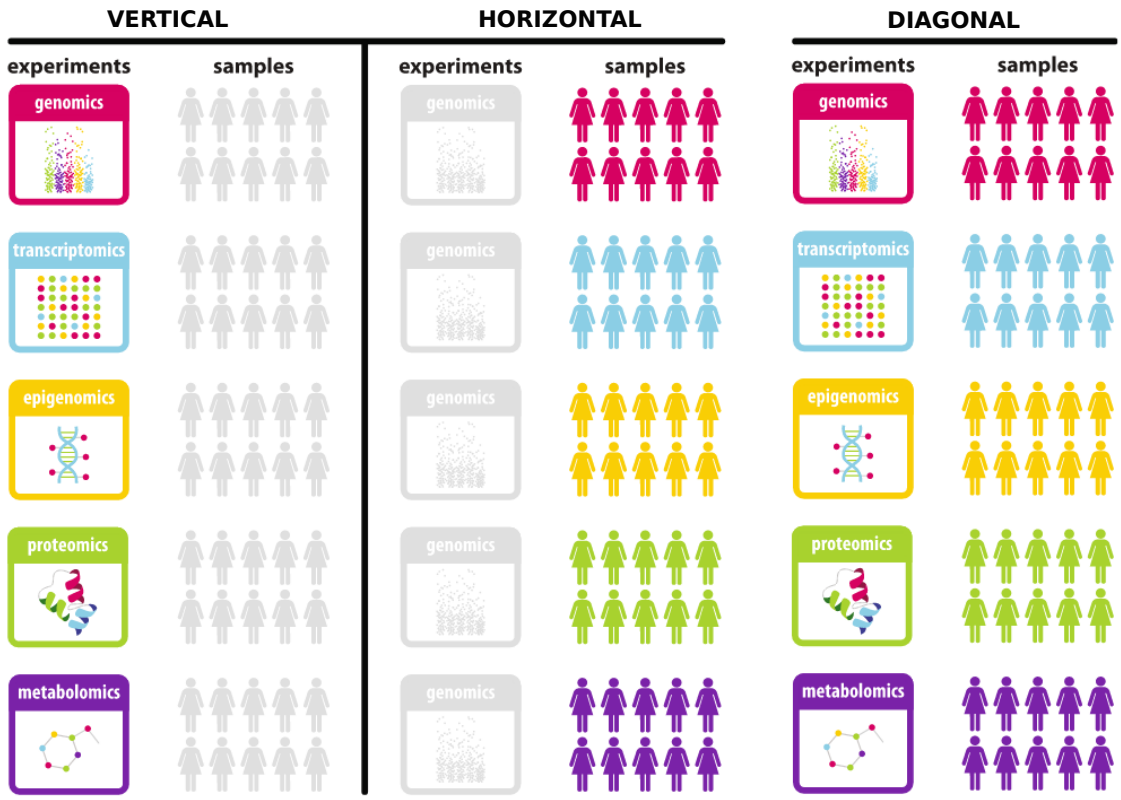


Figure 1.2: Vertical, horizontal and diagonal integration
Adapted from Eidem et al., BMC Medical Genomics, 2018 [14]

Vertical integration is mostly used to predict or characterise a trait on the samples of the experiments, for instance considering the various omics profiles of a set of patients to perform multi-omics-based disease subtype prediction [15], or to predict patients survival [16] based on their multi-omics profiles.

However, vertical integration can also be used to predict or characterize the behaviour or importance of features or group of features from the experiments [17, 18].

1.3.1.2 Horizontal integration

Horizontal data integration considers the same set of features measured for different samples across independent experiments. Horizontal data integration is used to characterise features according to various exper-

imental conditions of interest.

When the independent experiments to integrate have been generated to answer the same biological question, horizontal integration can be referred to as meta-analysis [19]. Meta-analysis can be used to synthesise the results of multiple experiments, test whether the experiments result in similar conclusions or increase the statistical power and improve the robustness of the results [20, 21].

When the independent experiments to be fused have been generated to answer various biological questions, horizontal integration can be used to construct features cross-experiment "repertoires", groups of features that behave similarly in a subset of experiments [22, 23] or group phenotypes according to the proximity of their molecular profiles [24].

1.3.1.3 Diagonal integration

When looking to integrate different data sources acquired on different samples, a coherent mapping between the entities of the different datasets must be found. This kind of heterogeneous-sample and heterogeneous-feature integration is called diagonal integration, an integration that can exploit the advantages of both vertical and horizontal integration, provided a mapping unifying the set of samples, or the set of omics features.

This is a particularly important problem for single-cell omics data integration, since in this type of dataset, the sample measured is the cell. Existing omics data acquisition techniques being destructive, it is currently impossible to measure different omics for the same cell.

Thus, an integration of single-cells multi-omics data would necessarily be diagonal.

Two types of strategies can be used: feature anchoring or sample anchoring.

Feature anchoring strategies map heterogeneous features to a common factor. For instance, the gene can be used as a key for anchoring features, as it can be associated to a sequence (genomics), a transcript (transcriptomics), a regulatory site (epigenomics) or a protein (proteomics). Most often, the anchoring of features will be done by considering gene-sets or by defining "metagenes", which represent the same biological signal [25, 26, 27]. In this configuration, it becomes possible to apply an horizontal integration strategy.

Sample anchoring consists in identifying groups of similar samples between datasets, i.e., by associating samples with labels. For example, in single-cell, by identifying in each omics datasets different cell types and vertically integrating omics data on these cell types, rather than on each cell [25, 28]. An equivalent in oncology would be to use clinical characteristics and to perform omics integration on this set of common clinical labels, rather than directly on the set of patients.

1.3.2 Depending on the integration strategy

There are numerous ways to perform data integration with different levels of complexity, from the "simple" concatenation of omics to the

inference of complex statistical models.

1.3.2.1 Early integration

Early integration strategies rely on the concatenation of data into a single matrix. The main advantage of early integration strategies is the ease of their deployment, since it is possible to use already existing single-omics analysis strategies on the concatenated dataset.

But this type of method also has many disadvantages:

- It can only consider data that can be concatenated. Generally, it is not possible to directly integrate knowledge or clinical data, which can take very different forms from omics data (i.e., gene-sets, networks, ...).

- Without corrective efforts, it can largely favor the types of data that will present the most features [29]. With unequal dataset sizes, the signal carried by the omics "dominated" in size can be completely lost.

- It further accentuates the problem of the statistical power of omics datasets. There are already dimensionality problems in single-omics data analysis, which is even more amplified by adding new datasets by simple concatenation.

- It does not take into account the interactions, known or measurable, between omics.

- It does not take into account the distributions of the various input omics.

Thus, the methods of early integration will try to answer these problems by various processes. For example, many methods seek to reduce the dimensionality of the data, first by filtering the features of each omics before concatenating them, and then by using dimension reduction methods. For example, the LRAcluster algorithm developed by Wu et al. [30], proposes a low-rank approximation approach before clustering the data projected on the reduced space, rather than directly applying a clustering algorithm on the concatenated data. LRAcluster defines several probabilistic models depending on the type of data and its distribution, and also allows to integrate categorical data, in addition to numerical data. Wang et al. [31] introduce a learning model that learns features weights using joint structured sparsity-inducing norms.

1.3.2.2 Late integration

In contrast to early integration strategies, late integration methods will first make predictions on each of the omics individually and then integrate the results of this analysis. The main advantage of this type of methods is that they can use specialized prediction algorithms for each omics (including for example supervised analysis methods, which allow to take into account known properties and knowledge data). It will also be easier to integrate other types of non-omics data, provided that they can be represented in the same way as the result of a prediction (e.g., clustering or ranking). However, as for early integration methods, interaction patterns between omics are not taken into account. An example of late integration tool is the Cluster Of Clusters Analysis (COCA) [32], which allows to fuse different clusterings produced on

different types of omics data.

1.3.2.3 Transformation-based

Transformation-based strategies rely on transforming the input datasets prior performing integration. This kind of intermediate integration can rely on various transformations of the data, including representing data as networks, or mapping input data into higher dimensional space via feature mapping. The predictions are then made by integrating the transformed data.

Network-based Network-based approaches rely on representing input data as networks and using this representation for integrating datasets and make final predictions. For instance, SNF (which stands for Similarity Network Fusion) [33] integrate multi-omics data by computing a sample-sample similarity network for each omics data type and fuses the resulting networks to obtain a multi-omics patients clustering. To identify co-expressed gene modules in multiple transcriptomics datasets, Salem and Ozcaglar [34] first compute a co-expression network for each transcriptomic dataset and mine the obtained multi-layer network to identify co-expressed communities.

One of the advantages network-based approaches is that the network representation is quite standard and can be applied to various types of data. Knowledge-networks (protein interaction networks, pathways, ...) can easily be added to the omics networks before the integration step.

Network approaches also allow to define (in addition to the relations within a single omics which will constitute a layer of the multi-layer net-

work to integrate) bipartite relations between the different layers, which allows to take into account the relations between the omics. These bipartite relations between layers can also be added between the layers of multi-layer networks considering very different data sources. Very good examples of this versatility are demonstrated in the work of Valdeolidas et al. [35]: making advantage of a random-walk-based algorithm, the authors predicted genes associated with a syndrome using both biological networks (e.g., PPI network, co-expression network) and disease similarity networks, associating bipartite relationships based on protein-transcript matches and on known genes associated with each disease.

Multiple Kernel Learning Multiple Kernel Learning strategies associate a kernel function to each set of omics or non-omics data to infer objects' relative similarities for each data type. Then the kernels are combined by defining an optimization procedure in order to obtain a global similarity of the objects taking into account the different types of input data. The advantage of working with these kernel functions is that they are able to solve non-linear problems by mapping the data into a high-dimensional feature space. Moreover, there are already many kernel functions adapted to different types of data [36, 37, 38], numerical as well as categorical, and applicable to supervised as well as unsupervised analyses, which allows to use multiple kernel learning to integrate very diverse types of data.

There are many Multiple Kernel Learning strategies [39], one of the most popular ones developed for omics data integration being the

approach proposed by Speicher and Pfeifer for cancer subtype prediction [40].

1.3.2.4 Dimension reduction-based

The dimension-reduction approach are among the most classic approaches to data analysis. Principal Component Analysis (PCA) is an unsupervised learning strategy that transforms high-dimensional data into fewer dimensions to simplify data complexity while retaining trends and patterns in the data [41]. PCA has many outcomes and can be used to cluster samples or features, for feature extraction, for visualization purposes, etc. A more recent example of dimension reduction strategies applied in bioinformatics is t-distributed stochastic neighbor embedding (t-SNE), broadly used for clustering and visualizing single-cell transcriptomics data [42]. Traditionally used for single-omics data analysis, dimensionality reduction strategies have also shown their usefulness for heterogeneous data integration, thanks to joint dimensionality reduction strategies, or autoencoders.

Joint Dimensionality Reduction Joint Dimensionality Reduction strategies seek to project the different input data into a common reduced space by decomposing each starting matrix into a product of two matrices: a low-dimensional factor matrix, common to all omics, and a weight or projection matrix, specific to each omics. The advantage of this type of method is that it allows both to infer multi-omics properties on the samples, for example by clustering the factor matrix, but also on the features, by analyzing the weight matrices to identify key

biomarkers or essential pathways.

Many strategies have been developed for Joint Dimensionality Reduction based on various approaches, including Canonical Correlation Analysis, Non-negative Matrix Factorisation or Principal Component Analysis. A number of these methods have been reviewed and benchmarked elsewhere [18, 43].

Autoencoders Autoencoders are artificial neural networks trained to encode data into a low-dimensional latent space. The encoded data can be decoded to reconstruct the original data. Autoencoders have been used to jointly analyze multiple datasets in various studies [44, 45] but their use for omics data integration is relatively recent. A few autoencoders developed specifically for omics data integration have been reviewed in [46]. As for the joint dimension methods, autoencoders have the advantage of being able to infer properties on samples, as well as on features, considering many types of data, numerical as well as categorical or image data.

1.3.2.5 Hierarchical

Hierarchical data integration strategies integrate omics by modeling cross-omics interactions and regulations, combining omics in an unified model. Typically, hierarchical integration based strategies use prior knowledge on omics and their interactions (i.e., they rely on known bipartite relations between omics) to direct a sequential integration. For instance, to better unerstand cis- and trans-regulations in cancer, Aure et al. apply a sequence of sub-analysis on Copy Number Variation data,

expression data and finally a combination of both, to narrow down a list of relevant genes [47]. Similarly, in [48], Chari et al. developed a sequential strategy to identify cancer-associated genes disrupted at multiple omics levels using transcriptomic and epigenetic data, by searching for (1) differentially expressed genes that also show (2) copy number variation, methylation changes or other epigenetic events within the same sample, (3) within multiple samples and (4) at a minimal frequency in the full cohort.

1.3.2.6 Conclusion

Although we have so far identified different classes of integration strategies, it is important to note that, in practice, most omics data integration methods are based on a combination of these different strategies. For example, LRAcluster [30] or concactAE [49] independently project the omics data into a reduced space (using respectively a low-rank approximation or an auto-encoder approach) and then concatenate these projections to produce a prediction, making them early integration strategies while relying on dimension reduction. Another example is the KLIC (Kernel Learning Integrative Ckustering) tool that combines consensus clustering (late integration) and multiple kernel learning [50].

1.3.3 Depending on the task

Strategies developed by the scientific community are generally built to answer a specific type of biological question, although some solutions are more versatile than others.

From a general point of view, the global question is always related

to find some patterns in the data, whether it relates to the samples or to the features of an experiment. Specifying the biological question is the first essential step before developing a new method. In this scope, the following sections propose to illustrate the different tasks that are the most common in bioinformatics.

1.3.3.1 Dimensionality reduction

While multi-omics data projection into a common reduced space is not the final product of dimensionality reduction strategies, such projections allow a variety of downstream analysis, which justifies their classification in a specific task. Indeed, from examining the factor matrix, downstream analysis include sample clustering/classification, biomarker prediction or network inference. From examining omics-specific weight matrices, the omics-specific contribution to detected factors can also be measured.

1.3.3.2 Prioritization

A common task of data integration strategies is the generation of ordered lists of biomarkers (generally at the gene level) associated with a phenotype of interest, most often associated with a disease. The idea is to prioritize features by order of importance in their contribution to the disease phenotype, in order to identify candidate genes that best characterize the multi-omics molecular state associated with the disease and that could be interesting targets for the development of new drugs. Features are associated with a score that determines their rank and the probability of their association with the phenotype of interest.

Most often, prioritization methods consider genes [51, 52, 53], but it is also possible to prioritize variants [54], pathways [55], metabolites [56] or drugs [57].

1.3.3.3 Classification

Classification methods seek to assign labels to objects of a multi-omics dataset by learning on similar datasets with known labels. The classification allows, for example, to predict the molecular subtypes of diseased individuals by considering all available omics layers [58]. Automatic classification of patients via multi-omics strategies is of great interest for the development of personalized medicine.

Almost all classification approaches seek to infer labels on the samples in the experiment. Indeed, annotating samples usually "only" requires clinical expertise, whereas inferring labels on different types of omics features can be very complex, especially since the bipartite relationships between omics layers are not fully resolved. However, very recently, a strategy to classify genes according to their impact on cancer into three distinct classes (neutral, tumor supressor or oncogene), has been introduced and has shown interesting results [59]. This type of method will surely develop as the knowledge of multi-omics molecular mechanisms associated with cancer and other complex diseases will grow.

1.3.3.4 Clustering

Clustering is a widely used analysis, either to identify groups of homogeneous samples, or groups of genes (or modules) associated in some

way (most often, co-expressed). In a multi-omics context, most of the integrative strategies focus on grouping samples showing similar multi-omics molecular profiles. In disease research, such groupings allow the identification of potential molecular subtypes of the disease. Some of these multi-omics clustering strategies have been reviewed in [60].

Because bipartite relationships between features of multi-omics datasets are not always resolved, multi-omics gene cluster prediction is a more complex task, although some strategies have been developed for multi-omics gene module prediction using methylation and expression data [61, 62].

Most of integrative strategies for gene module predictions rather focus on performing meta-analyses using a set of similar experiments to increase the robustness of discovered modules. For instance, MONET [63], which was primarily developed for multi-omics samples clustering was as well used to infer gene modules based on RNA-sequencing and microarray data. In [34], co-expressed genes modules are predicted using transcriptomics data from multiple human tissues.

1.3.3.5 Network inference

The representation of data and knowledge in the form of a network is particularly informative, as it allows to represent the interactions occurring within an omic dataset, but also across omics, and in a condensed way. Network inference, thus, provides a better understanding of the complex regulatory mechanisms within and across omics and their downstream analysis is also quite informative (clustering based on modularity optimisation or random walks; node prioritization based

on their degree or centrality; etc.). A number of strategies for network inference from multi-omics data have been developed, some of which are reviewed in [64], for homogeneous as well as heterogeneous networks.

1.4 Conclusion

This chapter discussed the importance of and recent interest in the integration of omics and non-omics data, which can:

- provide a better understanding of the complex mechanisms of regulation and interaction between different omics layers;
- improve the robustness of predictions through meta-analysis;
- complete our knowledge and characterization of observable phenomes at the cell or organism level;
- take into account the knowledge already available and accessible to the scientific community when analyzing new data.

In light of this potential, many data integration strategies have been developed in the last few years, to address different tasks. These strategies rely on different ways of conceiving integration, whether early, late or intermediate. Each analysis method is often specialized for a specific combination of omics data, and/or a specific task, although some are much more versatile than others. Each also rely on different underlying mathematical concepts, and do not base their predictions on the same type of patterns of interest to look for in the data. Thus, even for the same precise task, and by fixing the nature of the integration, different data integration methods produce different results whose

biological quality is sometimes difficult to estimate or compare. In addition to depicting the interest of data integration, which is no longer to be demonstrated, this brief state of the art also gives the intuition of the importance of identifying ways to reconcile the results produced by a set of integration methods on the same task and the same data, since each one will propose different predictions according to the type of pattern on which they have based their predictions.

Chapter 2

Personal contributions

2.1 Context and objectives

Omics and non-omics data heterogeneity As discussed in the first chapter, the heterogeneity of omics and non-omics data motivates but also, and especially, challenges the issue of data integration. A fundamental question in this field of research is the unification of the information carried by each type of omics and non-omics data.

Various existing analysis strategies Whether it is for single- or multi-omics analysis, numerous analysis strategies exist. These diverse methods consider various types of patterns in the data and make predictions based on these targeted patterns. Similarly to omics data that can be fused to offer a more complete view of a biological phenomenon, taking advantage of all the existing prediction methods should also be highly advantageous. This is all the more true when we do not have a standard measure of quality of a result that is the typical case in life sciences, and when newly published strategies in rapidly evolving fields, like bioinformatics, claim to perform better than existing methods [65].

Objective In this context of abundant heterogeneous data and analysis strategies, the main objective of this thesis is the development of new data analysis strategies allowing to take into account highly heterogeneous data, both omics and non-omics (including omics data derived predictions), in an integrative context (either to infer multi-omics mechanisms or in relation to phenome diversity) and in relation to complex multi-factorial diseases (e.g., cancer and neurodevelopmental and neurodegenerative diseases).

With respect to the types of integration presented in the previous chapter, both vertical and horizontal integration are considered in this work. Though diagonal integration is not specifically discussed in this manuscript, note that it can be tackled from a vertical or a horizontal point of view, provided prior anchoring of samples or features.

Furthermore, it is important to note that the two integration contexts considered here have very different characteristics. In the case of omics data integration for the resolution of multi-omics molecular mechanisms, the objective is the identification of consistent, homogeneous patterns across the various omics datasets. On the other hand, the elucidation of the diversity of molecular mechanisms associated with various phenotypes must be based on a comparative analysis for which it is the heterogeneity of the patterns across the datasets that is investigated. These two biological questions must therefore be tackled using different approaches, based on the search for homogeneous patterns in one case, and heterogeneous ones in the other.

2.2 First contribution: Data integration applied to oncology

2.2.1 Context: multi-omics disease subtyping

Cancer is a genetic and multi-factorial disease that is primarily defined by its organ of origin (brain, breast, kidney, etc.). However, even for cancers affecting the same organ, the molecular mechanisms underlying disease progression and development have been shown to differ and do not have the same impact on the prognosis of affected patients. Therefore, a major objective in oncology research, from a precision medicine perspective, is the identification of "intrinsic" molecular cancer subtypes in order to treat patients accordingly to their molecular alterations.

For instance, for classifying breast cancer, in addition to characterizing clinical parameters (histology grade, tumor size, etc.), molecular characteristics are also evaluated: estrogen and progesterone receptors status, human epidermal growth factor receptor 2 (HER2) status, etc. The combination of these clinical and molecular characteristics are used to classify breast cancer into coherent subtypes, which are increasingly well described in breast cancer research and are associated with very different prognoses, each being treated specifically (for instance, chemotherapy or immunotherapy depending on the cancer subtype). With the advances in microarray technologies and classification methods and in particular the PAM50 strategy which has been developed specifically for breast cancer subtyping and has become a gold-standard for this cancer type, the detection of these breast cancer molecular sub-

types can now be predicted "routinely", based on the analysis of the expression of 50 specific genes [66]. While the classification proposed by PAM50 has become a reference for the classification of breast cancers, it is not without flaws, and other types of classifications exist, based on the analysis of other omics (miRNA arrays, CNVs, etc.) [67]. Moreover, the expression of the genes considered by PAM50 is by nature influenced by genetic and epigenetic factors and therefore a product of multi-omics coordinated regulation mechanisms. Thus, a classification of molecular subtypes based not only on the analysis of a single omic but of several types of heterogeneous omics could reveal homogeneous subtypes within the various omics scales.

This explains why the characterization of cancer (and other diseases) has largely developed in the last few years towards approaches allowing the integration of multi-omics data. The most widely used approaches for the multi-omics prediction of cancer subtypes are based on multi-omics clustering, i.e., the horizontal integration of omics datasets using a clustering approach. These approaches cover all the integration strategies presented in the first chapter of this manuscript: early (e.g., [30]) and late integration (e.g., [32]), network-based (e.g., [33]) and multiple kernel learning approaches (e.g., [40]), joint dimensionality reduction (e.g., [18]), etc.

In a review of nine different multi-omics disease subtyping tools [60], the authors concluded that their benchmark performed on ten cancer types did not identify a strategy that was better than all the others on the basis of the quality metrics evaluated (clinical labels enrichment in clusters and survival analysis), and that therefore the recommendation

of a specific multi-omics clustering method was far from obvious.

2.2.2 Solution: Consensus Clustering applied to multi-omics disease subtyping

Here, the question of multi-omics disease subtyping is tackled with a consensus clustering strategy.

Consensus clustering (later abbreviated to CCl) is the task of combining multiple and potentially conflicting clustering results into a single unified clustering. Because of its versatility, consensus clustering offers many advantages for data integration.

CCl for disease subtyping and beyond Clustering (i.e., the task of grouping objects into clusters in such a way that individuals within a cluster are similar to each other and dissimilar to individuals in other clusters) is a standard task in many fields of research, as it allows to apprehend a wide variety of questions. Although the method developed in this thesis was designed to specifically address the question of multi-omics subtyping and thus the integration of clusterings of patient cohorts, it could in theory be applied to any type of question that can be solved by a clustering analysis.

CCl for integrating heterogeneous predictions from existing integrative clustering strategies Consensus clustering has the potential to reconcile results obtained from running various existing multi-omics clustering tools on the same datasets, and therefore benefits from the various predictions made by looking at different patterns of interest in the

multi-omics data.

CCl for integrating heterogeneous data types Any kind of data that can be clustered (numerical matrices, networks, trees, etc.) can be injected in a consensus clustering analysis, which makes it easy to fuse new types of data. Moreover, categorical data can be used directly as a clustering (e.g., if PAM50 subtypes are known, they can be used directly as an input clustering).

The publication related to this contribution is proposed and discussed in details in Chapter 3 of this manuscript and can be accessed here:

Galadriel Brière, Élodie Darbo, Patricia Thébault, and Raluca Uricaru. Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics*, 22(1):361, July 2021, doi: 10.1186/s12859-021-04279-1.

2.3 Second contribution: Data integration applied to neurosciences

2.3.1 Context: deciphering Alzheimer's disease phenome diversity

Alzheimer's Disease (AD) is a genetic and multi-factorial neurodegenerative disease and the leading cause of dementia in humans. AD is characterized by a progressive atrophy of certain cerebral areas, a pathological accumulation of amyloid-beta plaques in the extracellular

matrix, and a loss of neurons associated with a process of neurofibrillary degeneration. These events in the brain can take place over decades and well before the first symptoms of the disease. They appear, in the early stages of the disease, mainly localized in the hippocampus and entorhinal cortex, then gradually spread to the whole brain [68].

A complete characterization of the molecular mechanisms associated with the development of AD must therefore require a spatio-temporal analysis able to take into account the diversity of phenomes in AD, depending on the type of cells involved (neurons, astrocytes, microglia, etc.) and/or the cerebral structure considered, as well as the pathological stage (which is strongly linked to aging). Thus a study of the transcriptome diversity associated with the different spatio-temporal contexts should reveal context-specific and context-recurrent perturbations at the transcriptomic level.

Classical transcriptomic data analysis include the identification of Differentially Expressed (DE) genes across experimental contexts (often comparing disease and healthy samples) and the identification of co-expressed gene modules, genes showing similar expression profiles across samples. Unlike DE analysis, co-expression analysis was not developed for comparative analysis but rather for identifying co-regulated genes in a specific experimental context. A equivalent of DE analysis applied on co-expression is Differential Co-expression (DC) analysis, which aims at identifying changes in gene co-regulation between two experimental settings (e.g., disease and control samples).

Both co-expression and differential co-expression patterns can be represented using networks, with nodes corresponding to genes and the

weight on the link between two genes corresponding respectively to the intensity of their co-expression or of their differential co-expression. Constructed from various experimental contexts, multiple (i.e., multi-layer) co-expression networks or Differential Co-expression Networks (DCNs) can be analyzed to identify the various co-regulation or, respectively, differential co-regulation patterns observed in the diverse phenotypes considered. Identifying gene communities in such multi-layer networks could help unravel the transcriptomic mechanisms recurrently occurring across phenotypes or specific to a given phenotype.

2.3.2 Solution: Network-based approach for multi-group differential co-expression analysis

Here, the problem of finding co-expression dysregulations in multi-group datasets is addressed by the computation and analysis of multi-layer differential co-expression networks.

Community detection in multi-layer networks Community detection (latter abbreviated to CD), traditionally performed on single-layer networks but here extended to the analysis of multi-layer networks, is the task of clustering the nodes of a network in such a way that nodes from a same community are densely connected while loosely connected to nodes from other communities.

CD in multi-layer networks for phenome characterisation During this thesis, we developed a solution for identifying communities in multi-layer networks that we applied specifically on a multi-layer Differential

Co-expression Network. A layer in this network depicts gene differential co-regulations in a specific experimental context, i.e., in a specific disease phenotype compared to its respective control in the same context. The computed communities may result from dense connectivity observed in multiple layers, or in a specific layer of the network, thus indicating whether the co-expression perturbations are context-dependant or induced by disease in all the considered experimental contexts.

CD in multi-layer networks for integrating heterogeneous data types While applied specifically on multi-layer differential co-expression networks, the strategy could be applied for other biological questions (including multi-layer co-expression network analysis), but also in more diverse contexts, including multi-omics gene community detection, provided that each omics data type can be represented as a network and that there are bijective relations between features of each omics data type (i.e., each omic dataset can be represented as a gene network). Knowledge data such as protein-protein interactions networks or patwhays can also be included a supplementary layers of the multi-layer network.

The article associated to this contribution, though yet unpublished, will be presented and discussed in details in Chapter 4 of this manuscript:

Galadriel Brière, Agnès Nadjar, Raluca Uricaru and Patricia Thébault. Network-based approach for multi-group differential co-expression analysis: application to the 5xFAD mouse model of Alzheimer's Disease *Unpublished work.*

A preliminary version of this work was presented at the 2022 edition of Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM):

Galadriel Brière, Agnès Nadjar, Raluca Uricaru and Patricia Thébault. Condition-specific and recurrent perturbation-communities in multiple differential co-expression networks. In Journées Ouvertes en Biologie, Informatique et Mathématiques, 2022.

Chapter 3

Consensus clustering applied to multi-omics disease subtyping

3.1 Context

Consensus clustering, also referred as ensemble clustering, is of particular interest to perform disease subtyping in a multi-omics context. Indeed, consensus clustering, which consists in producing a consensus from a set of input clusterings, aim at reconciling the predictions made in each of the inputs. An advantage of such strategies is that it allows to consider data integration at two levels:

(i) either by clustering each omic source independently and reconciling the single-omics predictions to produce a multi-omics clustering (a.k.a "single-to-multi" integration strategy),

(ii) or by first clustering the multi-omics data using existing integration strategies (which are numerous, and have been introduced in Chapters 1 and 2 of this manuscript) and then reconciling these multi-omics predictions into a multi-omics consensus (a.k.a "multi-to-multi" integration strategy).

In the first scenario, one may use specialized algorithms for each of the omics in order to produce single-omics clusterings of the highest possible quality before the integration step. In the second scenario, algorithms specialized in multi-omics data integration may be used, each of them giving a different view of the omics interactions patterns observed in the datasets. In other words, consensus clustering has the potential to leverage the strengths of different single or multi-omics analysis strategies.

Moreover, clusterings can be produced from many types of data, omics or non-omics, numerical or categorical, in the form of matrices, networks or trees, facilitating the addition of clusterings from other data sources to be taken into account in the integration.

3.2 State of the art on consensus clustering

Many consensus clustering strategies have been developed, that can be classified in two major families [69, 70]:

Co-occurrence based approaches focus on grouping objects based on their co-occurrences in input clusters. Basically, the idea is that objects that are often clustered together should be classified in the same consensus cluster, and the consensus clustering is obtained using a voting process among the objects.

Median partition based approaches focus on producing a consensus clustering by maximizing its similarity with all input clusterings.

Originally, consensus clustering strategies have been developed for fusing clustering results obtained from the same dataset, and often

using only one clustering algorithm but run using various initialization parameters (e.g., number of clusters). When considering a single dataset it is expected that the different clusterings look similar. This may motivate the choice of using median partition based approaches. However, this is less the case when using consensus clustering to fuse partitions obtained from various datasets. In this case, co-occurrence based approaches seems more likely to resolve contradictory predictions obtained on the different datasets. Therefore, the following literature review focus on co-occurrence based approaches, and on two major publications, namely : (i) the "Consensus Clustering" algorithm proposed by Monti et al. [71] that later inspired the "Cluster of Clusters Analysis" (COCA) strategy introduced by Cabassi and Kirk [50] (specifically designed for integrating multi-omics clusterings), and (ii) the "Evidence Accumulation Clustering" strategy introduced by Fred and Jain [72] which inspired our own contribution, detailed in Section 3.3.

3.2.1 "Consensus Clustering" by Monti et al. and Cluster of Clusters Analysis

3.2.1.1 Strategy

Cluster of Clusters Analysis (COCA) was first introduced in [73] for multi-omics breast tumor subtyping, and has since, been applied in numerous disease subtyping studies [74, 75, 76]. Recently, the approach has been rigorously benchmarked, and an R package introduced in [50].

COCA makes use of the co-occurence based Consensus Clustering strategy introduced by Monti et al. (later referred as CCm) in [71]. The idea behind CCm is that, by perturbing an original dataset and

clustering each of the perturbed datasets, a consensus clustering can be found by examining the co-occurrences of objects in the generated clusterings. The consensus clustering structure obtained is supposed to be robust to perturbations of the data and stochasticity of the clustering algorithm used. Perturbed versions of the original dataset can be obtained by resampling the objects to cluster and/or the features of the dataset. Indeed, a robust clustering structure should not be impacted by such perturbation of the original data.

CCm procedure CCm strategy is illustrated in the red box of Figure 3.1.

After computing clusterings on the perturbed datasets, objects co-occurences in clusters are summarized in co-clustering matrices defined as follow: Let $X = [x_1, ..., x_N]$ be the set of objects and $c = [c_1, ..., c_N]$ their respective cluster labels, the associated $N \times N$ co-clustering matrix C is constructed such that $C_{ij} = 1$ if $c_i = c_j$ and $C_{ij} = 0$ otherwise.

Given a set of H co-clustering matrices $C = [C^1, ..., C^H]$ computed from clustering H perturbed datasets $D = [D^1, ..., D^H]$ with a fixed number of clusters K, a $N \times N$ consensus matrix Δ^K is computed as a properly normalized sum of the co-clustering matrices:

$$\Delta^K_{ij} = \frac{\sum_{h=1}^{H} C^h_{ij}}{max(1, \sum_{h=1}^{H} I^h_{ij})}$$

where I^h is the $N \times N$ indicator matrix such as $I^h_{ij} = 1$ if both objects i and j are present in the perturbed dataset D^h and 0 otherwise.

The consensus matrix is then used to assess the stability of the perturbed clustering results for a given number of clusters K: if all the

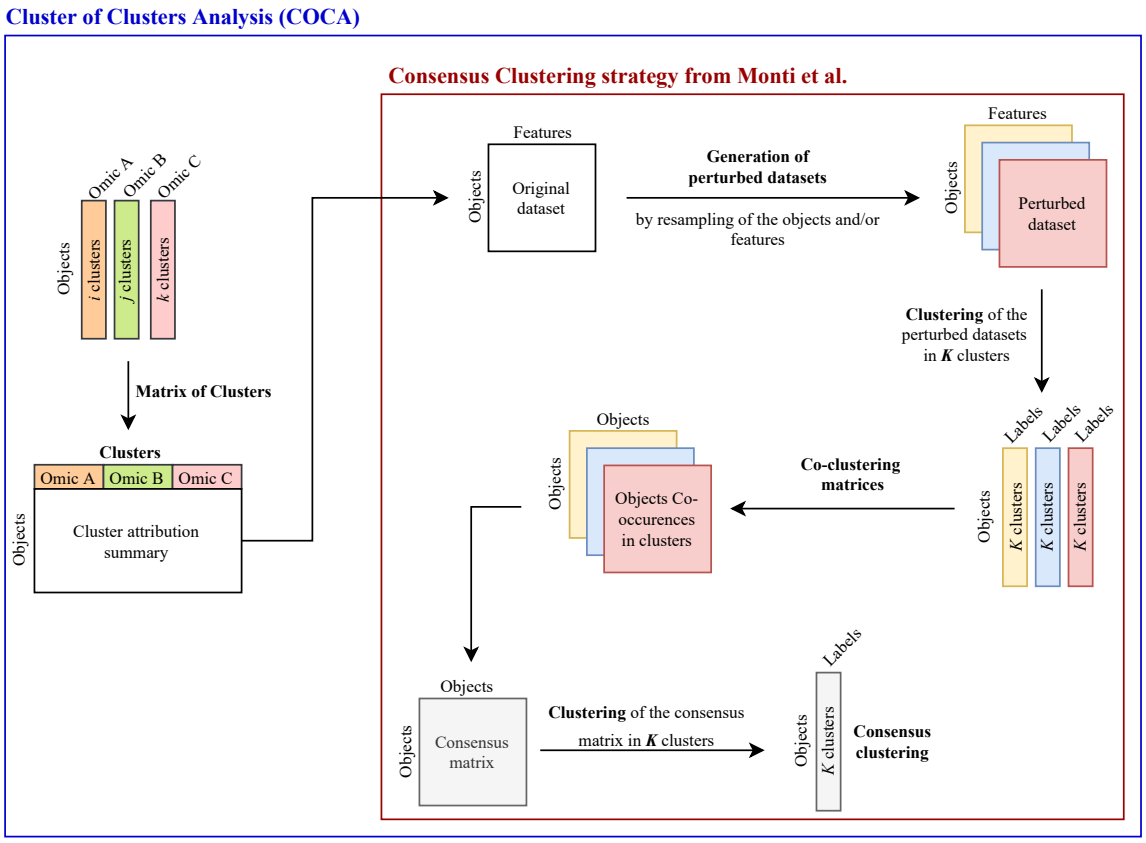


Figure 3.1: Consensus clustering by Monti et al. (red box) and Cluster of Clusters Analysis (COCA) (blue box)

elements in the consensus matrix are close to one or zero, this indicates a consensus in the input predictions because the objects will either have been predominantly classified in the same clusters or predominantly classified in different clusters. Note that the perturbed datasets are all clustered using the same number of clusters K, but that several values can be tested and the optimal number of clusters K_{opt} can be determined by comparing the consensus matrices produced with the different values of K. When the number of cluster is not known, K_{opt} should be estimated before producing the consensus clustering.

Finally, the consensus matrix can be clustered to obtain the final consensus clustering.

COCA procedure COCA strategy is illustrated in the blue box of Figure 3.1. COCA directly uses the CCm procedure but providing a Matrix of Clusters (MOC) as input. The MOC is computed from a set of omics-specific clusterings with no assumption on the number of clusters each omics-specific clustering should contain (i.e., input single-omics clusterings may or may not contain the same number of clusters).

In [50], given a set of clusterings $C = [C_1, ..., C_M]$, the number of clusters obtained in each clustering $K = [K_1, ..., K_M]$ obtained on M omics datasets, $\bar{K} = \sum_{m=1}^{M} K_m$ the total number of obtained clusters and m_k the kth cluster from clustering C_m, the MOC is a $\bar{K} \times N$ matrix defined as:

$$MOC_{n,m_k} = \begin{cases} 1 & \text{if } c_n^m = m_k \\ 0 & \text{otherwise} \end{cases}$$

where c_n^m is the label of object n in clustering C_m.

3.2.1.2 Advantages and limitations of Cluster of Clusters Analysis

Cluster of Clusters Analysis displays several advantages, in addition to those conferred by consensus clustering strategies in general, already discussed in Section 3.1 of this chapter. First, each omics data source can be clustered using a different number of clusters. Moreover, this strategy allows for missing data, since individuals that were not measured for all omics sources can be included in the Matrix of clusters (MOC) by setting the cells corresponding to the missing omics clusters to zero. Finally, even though the number of clusters in the final consensus clustering has to be fixed, an optimal value can be estimated by comparing consensus matrices computed from various numbers of

clusters.

However, this strategy suffers from some limitations, notably a sensitivity to the inclusion of low quality clusterings, since all input clusters have the same influence on the final clustering, as discussed in [50]. Finally, when benchmarking COCA on several cancer datasets, we observed that, in several cases, the clustering returned by COCA was very similar to one of the input clusterings (Adjusted Rand Index close to 1) and weakly similar to all the other clusterings (cf. Figure 4 from Section 3.3). In these cases, one may question the consensus of the final clustering.

3.2.2 Evidence Accumulation Clustering

3.2.2.1 Strategy

Evidence Accumulation Clustering (EAC) was introduced by Fred and Jain in [72] to combine clustering results obtained on the same dataset but using different clustering strategies and/or different parameter configurations of the same clustering algorithm. The motivation behind this work is to take advantage of existing clustering algorithms. Indeed, not all the types of patterns that can be found in the data in practice are detectable by a single clustering strategy. However, we can take advantage of different clustering strategies in order to detect all these patterns.

According to the concept of EAC, each association of objects in the same cluster is considered as an independent evidence of their association, and the set of associations predicted by the input clusterings are considered as an accumulation of evidences that can be used as a new

similarity measure between objects.

The strategy is illustrated in Figure 3.2 and can be formulated as follow: Let $X = [x_1, ..., x_N]$ be the set of objects to cluster and $C = [C^1, ..., C^P]$ a set of P clusterings of those objects, a $N \times N$ co-association matrix can be computed as

$$A_{ij} = \frac{a_{ij}}{P}$$

where a_{ij} is the number of times object i and object j were assigned to the same cluster among the P clusterings. The obtained co-association matrix is used as a novel similarity measure for clustering the objects.

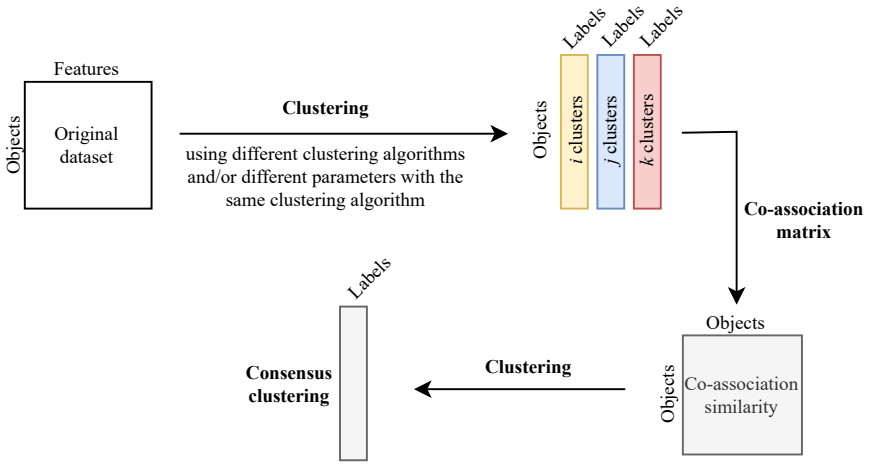


Figure 3.2: Evidence Accumulation Clustering

3.2.2.2 Advantages and limitations of Evidence Accumulation Clustering

Evidence Accumulation Clustering display several advantages, in addition to those conferred by consensus clustering strategies in general, already discussed in Section 3.1 of this chapter.

First, no assumptions are made regarding the number of clusters, whether it be for generating input clusterings or for generating the consensus clustering. Moreover, although this strategy does not imply

assigning a weight to the input clusterings according to their estimated quality, it is quite robust to the integration of low quality clusterings since these are more likely to propose divergent associations from the other clusterings. Indeed such dissimilar predictions will moderately impact the co-association measure, provided that a large number of clusterings are fused. Finally, the notion of consensus when considering EAC is quite suitable, since the co-association similarity can be seen as the result of a voting mechanism.

However, EAC was developed and applied for integrating clusterings obtained from a single data-source and has not been applied in a multi-source data integration context. For the same reason, there is no procedure to deal with missing data (objects that would not have been measured for all the omics considered).

3.3 Contribution

In the following contribution for multi-omics consensus clustering, we developed a novel EAC-based approach for multi-omics disease subtyping that we applied on 10 cancer datasets composed of gene expression, miRNA expression and methylation data and compare the results with those of COCA generated from the same set of input clusterings. We considered two integration strategies: multi-to-multi and single-to-multi integration scenarios. This novel strategy includes a procedure for handling missing data, allowing the consensus classification of individuals that may not have been measured for all omics considered. Moreover, we show that it is quite robust to the inclusion of low quality

clusterings, as the spurious associations predicted in these clusterings will be counterbalanced by the accumulation of consistent associations across the other clusterings.

Brière *et al. BMC Bioinformatics* (2021) 22:361
https://doi.org/10.1186/s12859-021-04279-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Consensus clustering applied to multi-omics disease subtyping

Galadriel Brière[1,2*], Élodie Darbo[1,3], Patricia Thébault[1†] and Raluca Uricaru[1†]

*Correspondence: marie-galadriel.briere@u-bordeaux.fr
†Patricia Thébault and Raluca Uricaru have contributed equally to this work
[2] INRA, Bordeaux INP, NutriNeuro, UMR 1286, Univ. Bordeaux, 33000 Bordeaux, France
Full list of author information is available at the end of the article

Abstract

Background: Facing the diversity of omics data and the difficulty of selecting one result over all those produced by several methods, consensus strategies have the potential to reconcile multiple inputs and to produce robust results.

Results: Here, we introduce ClustOmics, a generic consensus clustering tool that we use in the context of cancer subtyping. ClustOmics relies on a non-relational graph database, which allows for the simultaneous integration of both multiple omics data and results from various clustering methods. This new tool conciliates input clusterings, regardless of their origin, their number, their size or their shape. ClustOmics implements an intuitive and flexible strategy, based upon the idea of *evidence accumulation clustering*. ClustOmics computes co-occurrences of pairs of samples in input clusters and uses this score as a similarity measure to reorganize data into consensus clusters.

Conclusion: We applied ClustOmics to multi-omics disease subtyping on real TCGA cancer data from ten different cancer types. We showed that ClustOmics is robust to heterogeneous qualities of input partitions, smoothing and reconciling preliminary predictions into high-quality consensus clusters, both from a computational and a biological point of view. The comparison to a state-of-the-art consensus-based integration tool, COCA, further corroborated this statement. However, the main interest of ClustOmics is not to compete with other tools, but rather to make profit from their various predictions when no gold-standard metric is available to assess their significance.

Availability: The ClustOmics source code, released under MIT license, and the results obtained on TCGA cancer data are available on GitHub: https://github.com/galadrielbriere/ClustOmics.

Keywords: Disease subtyping, Multi-omic data, Data integration, Consensus clustering

Background

Recent advances in biological data acquisition have made it possible to measure a wide range of data. Polymorphism data, DNA methylation, RNA expression, and copy number variations as well as other "omics" data are now routinely observed and analyzed. Each omics type has the potential to reveal different molecular mechanisms associated with a phenotype, and making use of all available omics data could decipher complex and multilevel molecular interactions. Though several integrative tools have been developed,

with all of them aiming to answer biological questions by using multiple available data sources, the issue of omics data integration is far from solved. Along with the issue of omics data heterogeneity and integration, scientists are challenged with the diversity of strategies and methods available to answer the same biological question, each approach having its own perks and benefits.

The question of cancer subtyping is particularly representative of this kind of issue. By performing a clustering analysis, disease subtyping aims at detecting subgroups of patients (samples) showing similar characteristics. Even in the single-omics context, such analysis can be challenging, and numerous clustering strategies have been implemented and/or tested to this end: hierarchical clustering strategies, density-, distribution- or centroid-based strategies, supervised and unsupervised strategies, etc. The selection of a clustering method as well as of the optimal parameters to use is generally tricky. Moreover, the various biological mechanisms that are involved may vary from one patient to another: each tumor is different and has its own characteristics, both in the tumor cells themselves and in their interaction with their environment. As these mechanisms are not restricted to a single molecular level, the detection of groups of patients showing similar characteristics across different omics is a key issue to enable personalized medicine, which aims to offer patients a treatment adapted to the characteristics of their tumors.

This detection of groups of patients showing similar characteristics across different omics motivated the development of new computational methods implementing different strategies to analyze several omics datasets simultaneously (for detailed reviews, see [1, 2]). According to the classification proposed in [1], the *early integration* strategy consists of concatenating omics datasets in a large matrix and applying a clustering method conceived for single-omics data [3, 4]. However, *late integration* approaches first cluster each omics dataset independently and fuse single-omics clusterings into one multi-omics clustering [5, 6]. Other approaches perform *intermediate integration*, fusing sample similarities across omics [7–9], using dimension reduction strategies [10, 11], or statistical modeling with Bayesian frameworks [12–14].

To tackle both issues mentioned above, i.e., multi-omics and multi-strategy integration, one may want to apply a particular type of late integration strategy by taking multiple clustering results (using different data, methods and parameters) and fusing all of them into one *consensus clustering*. Such a consensus clustering should benefit from the complementary information carried by various omics data and capitalize upon the strengths of each method while fading their weaknesses.

Note that with respect to classical late integration strategies that start from the raw omics datasets (e.g., PINS [6] uses perturbations of raw omics data to generate the most stable multi-omics clustering), consensus clustering methods rely solely on clustering results. This property is essential, as it allows for any clustering algorithm and any clustering result to be used, regardless of the availability of the raw omics dataset and of its type.

A naive way to compute a consensus clustering would be to perform the *intersection* of the clustering results, i.e., by simply taking the associations on which all methods agree. However, the greater the number of clusterings to fuse, the smaller the intersection is. Moreover, when clusterings show different numbers of clusters, the question of the

intersection is not trivial. Therefore, the issue of consensus clustering requires further methodological developments.

To compute a consensus clustering from a set of input clusterings, two main strategies exist: object co-occurrence-based approaches and median partition-based approaches [15]. In the former strategy, consensus clustering is computed from a matrix counting co-occurrences of objects in the same clusters [5]. The latter strategy focuses on finding a consensus clustering maximizing the similarity with the input partitions. Both strategies raise several nontrivial questions. The choice of a clustering algorithm and its tuning is not straightforward when working with a co-occurrence matrix. However, for the median partition-based approach, the choice of a similarity measure is determinant. Nevertheless, for consensus clustering in a multi-omics multi-method context, comparing co-occurrences of objects is more pertinent than comparing similarities between partitions.

Here, we present ClustOmics, a new graph-based multi-method and multi-source consensus clustering strategy. ClustOmics can be used to fuse multiple input clustering results, obtained with existing clustering methods that were applied on diverse omics datasets, into one *consensus clustering*, regardless of the number of input clusters, the number of individuals clustered, the omics and the methods used to generate the input clusterings.

The co-occurrence strategy implemented in ClustOmics (detailed in the "Methods" section) is based on *evidence accumulation clustering* (EAC), first introduced by Fred and Jain [16]. The idea is to consider each partition as independent evidence of data organization and to combine them using a voting mechanism. Similar to clustering methods that use a distance or a similarity measure to compare objects, EAC considers the co-occurrences of pairs of objects in the same cluster as a vote for their association. The underlying assumption is that objects belonging to a *natural* cluster are more likely to be partitioned in the same groups for different data partitions. Thus, one can use the counting of the co-occurrences of the objects in clusters as a pairwise similarity measure. We further refer to these co-occurrence counts as the *number of supports*. This measure, summarizing the results from the input clusterings, is a good indicator of the agreement between the partitions and allows production of a new partitioning that can be qualified as consensual. Although computationally expensive, this strategy allows exploiting all clustering results, regardless of the number of clusters and their size and shape.

We designed ClustOmics as an exploratory tool to investigate clustering results in order to increase the robustness of predictions, taking advantage of accumulating evidence. To allow the user to tackle a specific question and to explore relationship patterns within input clusterings and generated consensus, we store the data in a non-relational graph-based database implemented with the Neo4j graph platform [17]. The use of a graph native database facilitates the storage, query and visualization of heterogeneous data, hence allowing the development of a solution that is flexible to various integration strategies. Indeed, by fusing clusterings from different clustering methods, different data types, different experimental conditions, or several options at the same time, through the use of what we call *integration scenarios*, ClustOmics can address a wide range of biological questions.

ClustOmics was applied in the context of multi-omics cancer subtyping, with TCGA data from different cancer types and multiple omics datasets. Input clusterings were computed with several single and multi-omics clustering methods and then were fused in a consensus clustering. Further details on the strategy implemented in ClustOmics are given in "Methods" section. To assess the benefit of this novel method, we further explored the robustness of our consensus clusterings with respect to the input clusterings, as well as their biological relevance, based on clinical and survival metadata available for each patient. We compare the ClustOmics results with those of COCA [5], a well-known co-occurrence-based consensus clustering tool that has already been used to combine multiple omics datasets to reveal cancer subtypes [18].

Results

Consensus clustering for disease subtyping in a multi-omics context can be implemented as an a priori solution making a consensus of omics-specific input clusterings or by a posteriori computing a consensus from multi-omics input clusterings. To better understand the perks and benefits of fusing omics data in one way or another, ClustOmics was tested in these two contexts based on two *integration scenarios*.

First, we used ClustOmics to fuse multi-omics clusterings computed with existing integrative methods. In this scenario (multi-to-multi, MtoM), the integration of omics is performed by various existing clustering tools, and ClustOmics computes a consensus result of the different multi-omics clusterings produced. The second scenario (single-to-multi, StoM) involves both methods and omics integration, as only single-omics clusterings computed from various methods are fused into one consensus clustering. See Fig. 1 for a visual representation of these two scenarios.

Below, we analyze and compare ClustOmics and COCA consensus multi-omics clusterings (produced using the same set of input clusterings) for the two integration scenarios, on TCGA data from ten different cancer types, three omics datatypes (gene

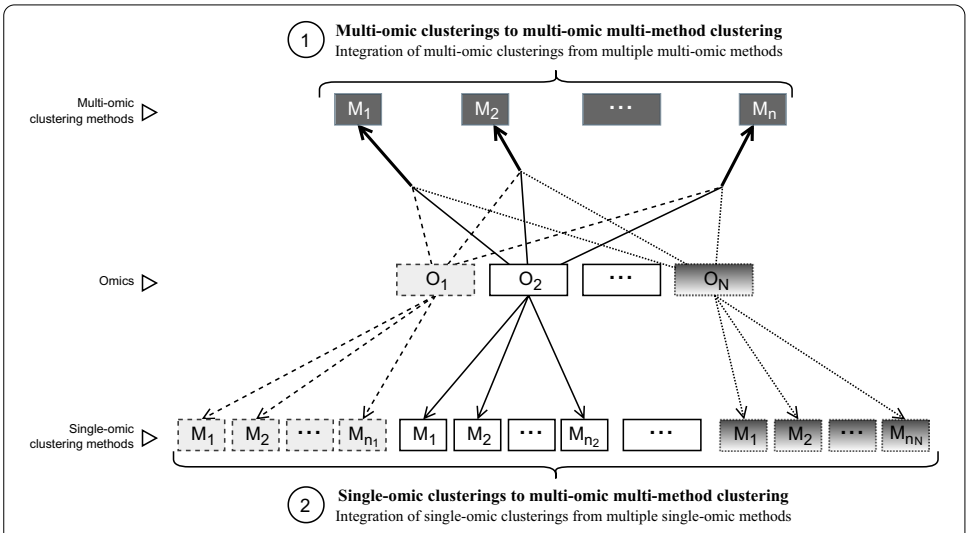


Fig. 1 Two integration scenarios: multi-to-multi consensus clustering and single-to-multi consensus clustering. Arrows are dashed according to the omics considered by each input clustering method

expression, miRNA expression and methylation) and various input clustering strategies (described in "Methods - Datasets and tools used for computing input clusterings" section). We further focus on breast cancer and analyze the ClustOmics results for the single-to-multi scenario on the breast dataset.

Results overview of the ten cancer types for the two integration scenarios

Running ClustOmics and COCA on the ten cancer datasets with respect to the different integration scenarios implies starting by computing single- and multi-omics input clusterings to group patients according to their single- and multi-omics profiles.

For the multi-to-multi (MtoM) scenario, multi-omics input clusterings were obtained with existing multi-omics clustering tools: PINS [6], SNF [7], NEMO [8], rMKL [9] and MultiCCA [10] (see Table 6).

For the single-to-multi (StoM) scenario, the same tools listed above were applied, except for the MultiCCA tool, which can only be used in a multi-omics context and was replaced with the simple yet robust state-of-the-art method, K-means clustering [19]. In this scenario, the tools were applied to each omics dataset independently. Moreover, to evaluate the benefits of including patients with missing data (that were not measured for all of the three omics), two different runs were performed. In the first run, referred to as *StoM OnlyMulti*, only patients measured for the three omics were considered, that is, patients with no missing data. For the second run, named *StoM All*, all available patients for each omics were kept, implying that in this scenario the set of patients clustered in input clusterings was different across omics.

A survival and clinical label enrichment analysis was conducted on ClustOmics and COCA multi-omics consensus clusterings, as well as on the single-omics and multi-omics input clusterings (see "Methods - Biological metrics" section for more details on the biological metrics used). An overview of the results for the ten cancer types is displayed in Fig. 2.

In terms of clinical label enrichment in clusters, the number of clinical labels significantly enriched varies from 19 to 26 (for a total of 79 enrichment p values computed from 32 distinct clinical labels), depending on the clustering tool. The majority of clinical labels found enriched in ClustOmics consensus clusters were also found enriched for at least one input clustering as well as in the corresponding COCA consensus, and clinical labels stably enriched in input clusterings were also found enriched in ClustOmics and COCA consensus clusterings. For details with respect to the distribution of the clinical labels found enriched in input and consensus clusterings for the MtoM and StoM scenarios, see Additional file 1: Figure S1.

The survival analysis results show high heterogeneity, supporting the idea of computing a consensus clustering, especially when no gold-standard metric or ground-truth data are available. In this sense, it is important to stress out that ClustOmics succeeded to compute biologically relevant consensus partitions from input clusterings of variable quality. Indeed, in the multi-to-multi case, ClustOmics managed to find 4 out of 10 significant log-rank p values, counterbalancing the PINS and MultiCCA mitigated results, although they were part of the input clustering results used for this integration scenario. For the same set of input clusterings, COCA MtoM yielded to a 3 survival-wise significant consensus result.

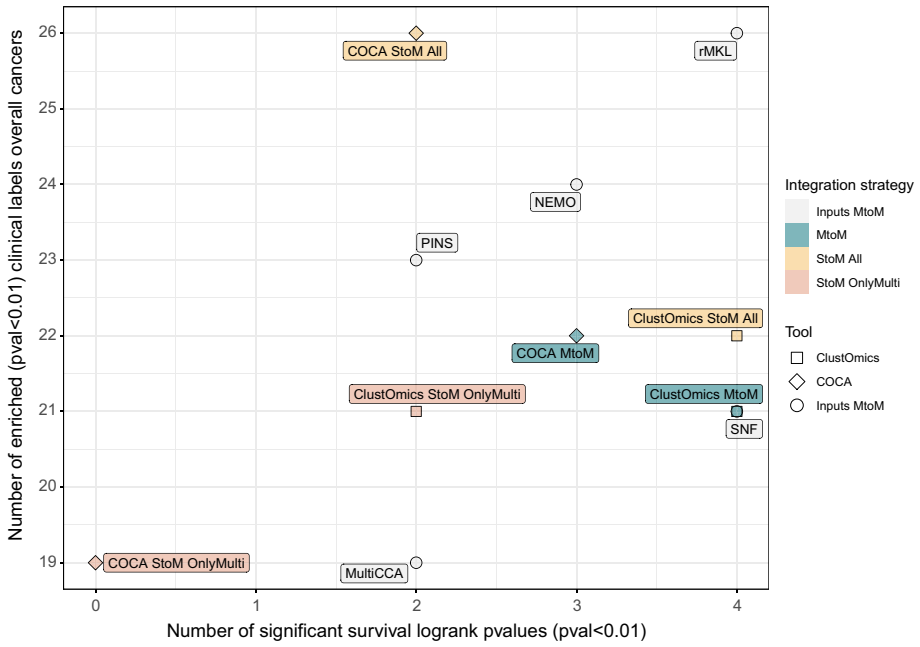


Fig. 2 Overview of survival and clinical label enrichment results for the ten cancer types analyzed. The *x*-axis represents the number of significant survival *p* values (< 0.01) found for each clustering, over all the ten cancer types. The *y*-axis represents the total number of significantly enriched clinical labels (*p* values < 0.01), all cancer types included. In total, 79 enrichment *p* values were computed from 32 distinct clinical labels

Interestingly, for the single-to-multi scenario, Fig. 2 clearly shows that considering individuals with missing data (StoM All) greatly improves the consensus clusterings, both in terms of clinical label enrichment and survival analysis, for ClustOmics as well as for COCA. For the StoM All scenario, COCA found 4 additional enriched clinical labels with respect to ClustOmics but yielded only 2 out of 10 survival-wise quality clusterings, compared to 4 for ClustOmics. The quality results for omics-specific input clusterings used for this scenario do not appear in Fig. 2 but are further detailed in "Results - Integration of single-omics clusterings" section.

Detailed results for the MtoM and StoM integration scenarios are given in the following two sections.

Integration of multi-omics clusterings (multi-to-multi scenario)

Input multi-omics clusterings were computed with the five multi-omics clustering methods presented in "Methods - Datasets and tools used for computing input clusterings" section using default parameters and following recommendations of the authors. The clusterings were produced using the multi-omics patients exclusively (those for which all three omics data are available). To make all input clusterings comparable, we ran NEMO in the same way, though compared to the other five tools, NEMO is able to handle partial data.

ClustOmics was run with the *min_size_cluster* parameter arbitrarily set to 8 nodes for all cancer types, meaning that clusters of size below 8 were removed from the consensus clustering, with the corresponding individuals being reassigned to consensus clusters exceeding the size threshold. We also set the *min_size_consensus* parameter

to 95% of the population to ensure that less than 5% of individuals are being reassigned to consensus clusters, either because of the number of support thresholds on the integration graph or because of the filter on the size of the clusters. The quantitative global measures on the ClustOmics consensus clusterings are detailed in Table 1.

Note that the maximum number of supports promoting the association of two patients in the same consensus cluster is bounded to 5, as five input clusterings (computed from five integrative clustering tools) were used for this integration scenario. After testing all possible thresholds on the number of supports, the optimal filtering threshold was obtained for each cancer type (2, 3 or 4, depending on the cancer type), meaning that only pairs of patients clustered in the same multi-omics cluster by at least 2 to 4 clustering methods were considered to compute the consensus clustering.

When comparing input and ClustOmics consensus clusterings, we observe a certain consistency in terms of the number of clusters. COCA, however, resulted in 2 to 3 clusters independently from the cancer type, which suggests a lower sensitivity to input clustering dissimilarities compared to ClustOmics.

Two cancer datasets, COAD and LUSC, showed the lower consistency between the input predictions and clustered with a number of supports of 2. For LUSC cancer type, the consensus clustering resulted in only 2 clusters, despite the large size of the available cohort (341 individuals). The computation of the adjusted Rand index (ARI) [20] between input clusterings, a measure of similarity between partitions, showed that for these two cancer types, SNF and NEMO clusterings were very similar (with an ARI value of 0.7 for COAD SNF and NEMO clusterings and of 0.9 for LUSC; see Additional file 1: Figure S2) while the 3 other input clusterings showed high pairwise dissimilarity ($ARI \leq 0.4$). The resulting consensus clusterings for both ClustOmics and COCA were very similar to SNF and NEMO and dissimilar to the other input clusterings, failing to compute an actual consensus of all input partitions. For the other cancer types, similarities between input clusterings were more balanced, enabling ClustOmics to reconcile predictions. ARI heatmaps comparing input and

Table 1 Multi-to-Multi Scenario: Number of patients initially clustered by ClustOmics, number of patients reassigned to consensus clusters, number of supports used to filter the integration graph, number of consensus clusters generated by ClustOmics and COCA, and average number of clusters in input clusterings

Cancer	# patients clustered	# patients reassigned	# supports threshold	# clusters ClustOmics	# clusters COCA	Avg # clusters inputs
AML	165	5	4	6	3	4.6
BIC	619	2	4	5	2	4.0
COAD	220	0	2	3	3	5.2
GBM	267	7	4	3	3	3.4
KIRC	176	7	3	3	3	4.2
LIHC	363	4	4	5	2	3.2
LUSC	341	0	2	2	2	3.4
OV	285	2	4	5	2	3.4
SARC	257	0	3	3	3	3.4
SKCM	351	0	3	5	3	4.8

consensus clustering similarities for the ten cancer types in the MtoM scenario are available in Additional file 1: Figure S2.

Unsurprisingly, from Table 1, we remark that filtering the integration graph with a higher number of supports generally results in reassigning individuals (from 2 to 7) to consensus clusters. The predictions regarding these reassigned patients do not necessarily meet the number of supports threshold for which the consensus clusters were computed.

Figure 3 presents survival analysis results for the various multi-omics clusterings given as input to ClustOmics and COCA MtoM and for the resulting consensus clusterings. When looking at the input clustering survival results, we can differentiate two cases:

- For AML, LIHC, SARC and SKCM, the input clusterings show a quite high heterogeneity in terms of survival quality
- For BIC, COAD, GBM, KIRC, LUSC and OV cancer types, the input clusterings show relatively homogeneous survival quality

For the first group of cancer types, the heterogeneity of input clustering survival qualities indicates how the choice of one clustering method can drastically impact the results. For these cancer types, ClustOmics produced consensus clusterings of a survival-wise quality approaching the median quality value, considering the input clusterings. Indeed, from input clusterings of various quality, ClustOmics was able to extract the most stable patterns across the input partitions.

When input partitions show homogeneous survival quality, ClustOmics gives similar results, which is an expected behavior. The largest deviation from the median is found

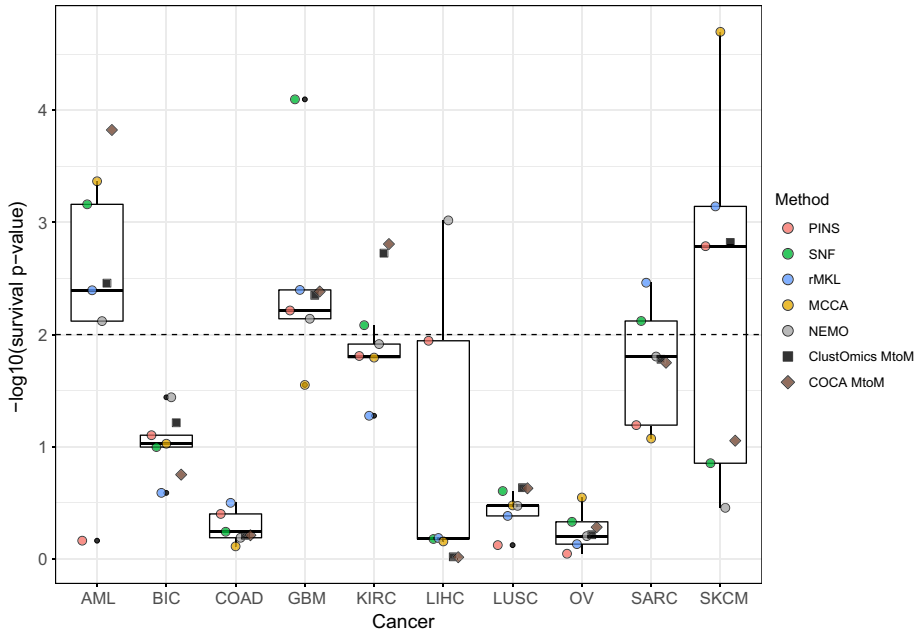


Fig. 3 Survival analysis results for ClustOmics and COCA multi-to-multi consensus clustering and for each input multi-omics clustering. The horizontal dashed line indicates the threshold for significantly different survival rate (p value ≤ 0.01). Boxplots were computed considering input clusterings only

for the KIRC cancer type, for which both COCA and ClustOmics consensus clusterings produced a partition of higher quality than could have been expected.

Consensus clusterings were also investigated for clinical labels enriched in clusters. For the ten cancer types, ClustOmics and COCA found 20 common clinical labels as being enriched, of which 19 were also found enriched in at least one input clustering. However, 16 labels were found as enriched in at least one input clustering but not in the consensus clusterings (see Additional file 1: Figure S1B). Table 2 give complete details on the clinical labels enriched in ClustOmics consensus clusterings for the ten cancer types.

For AML, for example, ClustOmics computed clusters enriched for the CALGB cytogenetics risk category, a risk classification based on the Cancer and Leukemia Group B clinical trial [21], and for the French–American–British (FAB) morphology code, a clinical classification for AML tumors [22]. Reassuringly, BIC consensus clustering was found enriched for the PAM50 classification, a widely used breast-cancer subtype predictor [23].

Integration of single-omics clusterings (single-to-multi scenario)

To assess ClustOmics performance when fusing simultaneously input clusterings computed from different omics data and with different clustering methods, we investigated a second integration scenario, combining single-omics clusterings produced independently on each omics dataset. The overall cancer consensus results for this scenario are displayed in Fig. 2 and discussed in detail in this section.

As stated above, single-omics clusterings were computed using the following five clustering tools: PINS [6], SNF [7], NEMO [8], rMKL [9] and K-means clustering [19] (with an optimal number of clusters computed with the Silhouette index [24]).

Table 2 Clinical labels found enriched in multi-to-multi (MtoM) scenario consensus clusters, in single-to-multi (StoM All) scenario consensus clusters, and for both scenarios

Cancer	Scenarios	Enriched clinical labels
AML	Both	Age at initial pathologic diagnosis
		CALGB cytogenetics risk category
		Leukemia French–American–"British Morphology Code
BIC	Both	Age at initial pathologic diagnosis, PAM50 call
		Pathologic N, Pathologic Stage, Histological type
		Estrogen receptor status, Progesterone receptor status
	StoM All	Pathologic M, Pathologic T
COAD	MtoM	Histological type
	StoM All	Age at initial pathologic diagnosis
GBM	Both	None
KIRC	Both	Pathologic M, Neoplasm histologic grade
	StoM All	Pathologic T
LIHC	Both	Gender, Age at initial pathologic diagnosis, Fetoprotein outcome value
LUSC	Both	None
OV	Both	None
SARC	Both	Gender, Age at initial pathologic diagnosis, Histological type
	MtoM	New neoplasm event type
SKCM	MtoM	Age at initial pathologic diagnosis

To assess the benefit of including individuals with missing data (not measured for all omics), two analysis were performed for this scenario:

- For StoM OnlyMulti, input clusterings were computed using exclusively multi-omics patients. Given that this is the same set of individuals as in the MtoM integration scenario, the same parameters were used for all cancer types, i.e., *min_size_consensus* = 95% and *min_size_cluster* = 8.
- For StoM All, omics clusterings were computed using all available patients. As the proportion of missing data varies between cancer types (up to 66% partial data for KIRC; see Table 5) and between omics, input clusterings do not apply to the same set of patients as in the scenarios previously described. To account for this increase in the number of patients to be clustered, *min_size_cluster* was set to 5% of the *multi-omics* population. The *min_size_consensus* parameter was set to 95% of the *multi-omics* population.

As shown in Fig. 2, fully exploiting the available data (including patients with missing data) greatly improved the consensus clusterings, for both ClustOmics and COCA. Moreover, for 3 cancer types, BIC, GBM and LUSC, capitalizing on all available individuals resulted in increasing the number of supports used to filter the integration graph. The largest increase in the number of supports threshold was observed for BIC, i.e., from 7 supports in the StoM OnlyMulti up to 11 in the StoM All run. For LIHC, OV and SKCM, however, we observe a decrease of -2, -1 and -1, respectively, in the number of supports. For the other cancer types, the threshold on the number of supports is identical between the two runs. In the follow-up of this study, we will focus on the results of the StoM All run.

In this scenario, the maximum possible number of supports is 15 as the five clustering methods were run on three omics datasets for each cancer. Note that for this scenario, the threshold on the number of supports used to filter the integration graph has great influence on the capacity of ClustOmics to produce consensus clusters across omics and on the interpretation of the results. Indeed, the threshold has to be greater than 5 to ensure that all the conserved integration edges rely on an association that is consistent across at least two different omics (one omics type being represented by five input clusterings). To ensure that all integration edges are built upon all three omics, the threshold must be 11 or higher. One should estimate an acceptable threshold depending on the experimental design and the biological question to address.

In our case, as we did not wish to bring any a priori preconceptions on which omics should have a stronger impact on the results (indeed, one omics data type could particularly well explain the disparities in molecular profiles of patients for a cancer type but not for the others), we considered a number of supports of 7 to be sufficient to ensure that selected integration edges are either moderately consistent across the three omics or strongly consistent in one omics type.

Together with the constraint to preserve at least 95% of the multi-omics population (*min_size_consensus* parameter), this gave a number of supports used to filter the integration graph ranging from 7 to 11. The results for this scenario are displayed in Table 3.

Table 3 Single-to-Multi All Scenario: Total population size (which are multi-omics), number of patients clustered or reassigned to consensus clusters (which are multi-omics), number of supports used to filter the graph, number of clusters generated by ClustOmics, number of clusters generated by COCA, and average number of clusters in the input clusterings

Cancer	Total (multi-omics)	Clustered (multi-omics)	Reassigned (multi-omics)	# supports threshold	# clusters ClustOmics	# clusters COCA	Avg # clusters inputs
AML	197 (170)	176 (163)	21 (7)	8	7	4	6.60
BIC	1096 (621)	600 (600)	496 (21)	11	6	2	3.87
COAD	303 (220)	276 (220)	27 (0)	7	6	4	3.47
GBM	578 (274)	434 (262)	144 (12)	9	11	2	4.13
KIRC	534 (183)	316 (174)	218 (9)	9	9	2	4.07
LIHC	377 (367)	363 (354)	14 (13)	8	5	2	4.73
LUSC	501 (341)	337 (321)	164 (20)	8	4	2	4.73
OV	591 (287)	393 (272)	198 (15)	9	9	2	3.47
SARC	261 (257)	261 (257)	0 (0)	8	3	3	4.33
SKCM	368 (351)	345 (329)	23 (22)	8	6	4	4.67

One of the major benefits of this integration scenario (in addition to the fact that single-omics clusterings are easier to compute) is its ability to cluster individuals that did not appear in all input partitions. Interestingly, although multi-omics patients have better chances to show high numbers of supports (as they appear in all input clusterings), some proportion of those multi-omics patients had to be reassigned to consensus clusters, while other individuals who were not measured for the 3 omics were clustered immediately, which suggests a good agreement between the input clusterings for the classification of these individuals.

Input clusterings can show great similarity for a given omics type. If this omics type allows differentiation of groups of individuals in a clear-cut way, it will drive consensus clustering. However, if the omics type is less relevant to partition patients, input clusterings are more likely to show different patient associations. Such clusterings add noise-like integration edges in the integration graph, with low number of supports on edges. Therefore, we expect each omics to have a different impact on the final consensus clustering. To evaluate the impact of omics-specific input clusterings on the consensus result, we used the adjusted Rand index (ARI) [20].

In Fig. 4, ClustOmics and COCA consensus clusterings were compared to each of the input clusterings. The relative proximity of a clustering consensus to the different input clusterings, as measured by the ARI, indicates the ability of the tool to produce a partition that can genuinely be considered as reconciling the input predictions. In that respect, the ARI of a consensus clustering in relation to its inputs should be maximized for a maximum of input clusterings, including for those coming from different omics. The highest similarity between consensus and input clusterings from different omics sources is observed for the SARC cancer dataset (see Fig. 4), with ARI values ranging from 0.4 up to 0.9 for at least one input clustering computed from each of the three omics datasets, suggesting similar associations at different molecular levels. For the other cancer types, the agreement between omics sources is less straightforward. Interestingly, for all cancer types, COCA and ClustOmics consensus clusterings resemble the

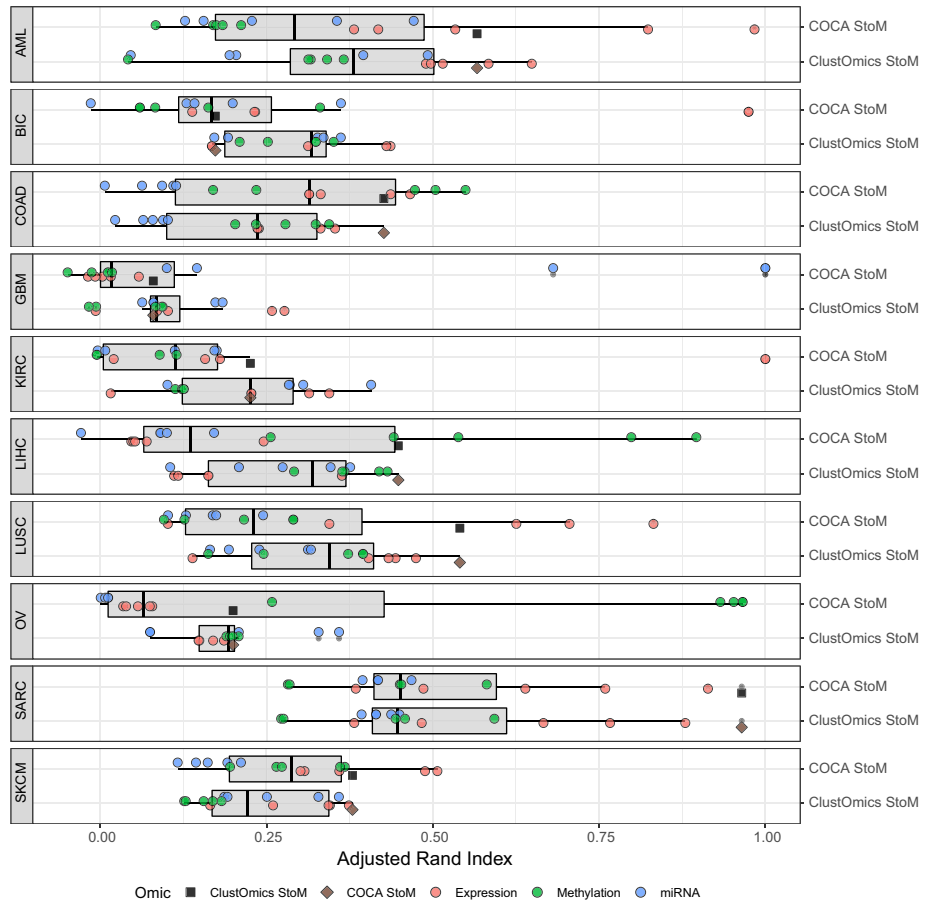


Fig. 4 Adjusted Rand index of input clusterings relative to ClustOmics and COCA StoM consensus multi-omics clusterings. Each point corresponds to one clustering and is colored according to the omics type used. Each omics dataset was clustered using five different clustering tools (PINS, NEMO, SNF, rMKL, K-means), and therefore, it is represented by five input clusterings. ClustOmics and COCA respective consensus clustering similarity is displayed with a black square and a brown diamond

same input clusterings (computed from the same set of omics sources), thus suggesting that some omics are more appropriate to explain molecular differences between individuals. Unsurprisingly, gene expression impacts consensus clustering on most cancer types, but miRNA and methylation data also guided consensus clusterings, especially in COAD, LIHC and OV. Figure 4 also shows that the dispersion of ARI values is much greater for COCA consensus clusterings than for ClustOmics. While COCA consensus clusterings are very similar (if not identical) to a few input clusterings but very dissimilar to the others, ClustOmics produces a consensus that is closer in average to all inputs.

Survival analysis for this integration scenario (see Fig. 5) shows two groups of cancer types, as already noted for the multi-to-multi scenario. For BIC, COAD, LUSC and OV, gene expression, methylation and miRNA input clusterings show homogeneous survival p values. For these cancer types, ClustOmics computed a consensus clustering with similar quality scores. For the cancer datasets showing higher heterogeneity among input clusterings survival quality, ClustOmics found significant survival p values for AML, KIRC, LIHC and SKCM, despite some low-quality clusterings that were given as input.

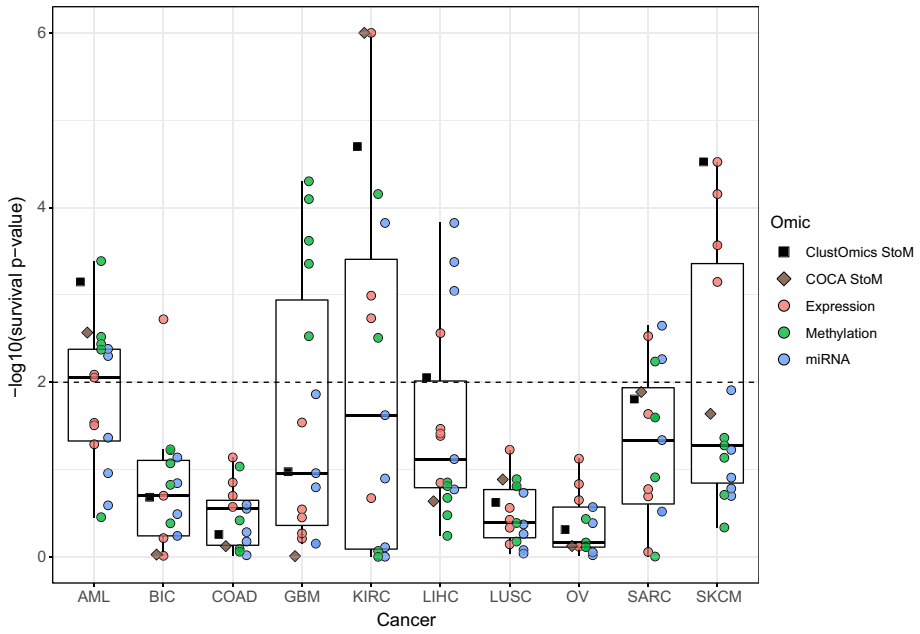


Fig. 5 Survival analysis results for ClustOmics and COCA single-to-multi All consensus clustering and for each input multi-omics clustering. The horizontal dashed line indicates the threshold for significantly different survival rate (p value ≤ 0.01). Boxplots were computed considering only input clusterings

Clinical labels found enriched in consensus clusters are listed in Table 2. AML clusters were found enriched for both CALGB and FAB classifications, KIRC clusters for histologic grade and pathologic M and T (referring to the TNM classification of tumors [25]), LIHC clusters for gender, age at diagnosis and fetoprotein outcome value, while SKCM clusters showed no enriched clinical parameters. While BIC consensus clustering did not show good survival-wise results, pathologic M, N and T labels were found enriched in clusters, as well as pathologic stage, histological type, PAM50 call, and estrogen and progesterone receptor status.

In the following section, we further explore the single-to-multi consensus clustering for BIC dataset.

Study case: BIC single-to-multi consensus clustering

In this section, we focus on the consensus clustering of the 15 single-omics clusterings for the BIC dataset (five clustering methods, listed in the previous section, applied on three omics data types) and analyze these results in parallel to the PAM50 classification. As the PAM50 classification is computed from the expression of 50 specific genes, while in this work, we capitalize on three different omics, a certain heterogeneity in the clusters when compared to the PAM50 prediction is expected. Moreover, this heterogeneity is to be further explored, as it could reveal subtypes that are not distinguishable when considering only PAM50 genes but that are heterogeneous when integrating other data sources.

From the 1096 patients available in the BIC dataset (of which only 621 patients are measured for the three omics), ClustOmics succeeded in primarily classify 600 multi-omics patients in consensus clusters, with a number of supports threshold of 11 (see

Table 3). The remaining 21 multi-omics individuals were reassigned to consensus clusters, as well as the 475 individuals with missing data. The consensus clustering resulted in a partition with 6 clusters, with sizes ranging from 115 to 254 individuals (the minimum allowed size for a cluster *min_size_cluster* being set to 5% of the multi-omics population, that is, 31 individuals for BIC).

As the PAM50 clinical labels were missing for 255 patients, we applied the original classifier introduced by Parker et al. [23] to call the missing labels. To estimate the quality of reassessed PAM50 labels, we evaluated the concordance between available PAM50 labels and recomputed PAM50 labels. The F1-scores showed the Basal, Luminal A, Luminal B and Her2 PAM50 labels to be well predicted (F1-score of 0.89, 0.75, 0.74 and 0.64, respectively). Predictions for the Normal-like class are less reliable (F1-score of 0.27) due to the small size of the class (23 individuals).

We further mapped the PAM50 calls to ClustOmics consensus clusters and observed significant concordance, as depicted in Fig. 6.

Indeed, Luminal A samples are overrepresented in the consensus clusters *B* and *E*, Luminal B samples in *A* and *D*, Her2 samples in *A*, and Normal-like samples in consensus cluster *C* (see Fig. 6 and Table 4). The vast majority of basal-like samples were classified in consensus cluster *F*, which gathers 190 of the 197 basal samples, with the remaining 7 being clustered in consensus clusters *B*, *C* and *D*.

This mapping of PAM50 calls on consensus clusters, which seems fuzzy at first glance, is not surprising as it has been shown that separation of Luminal A and B

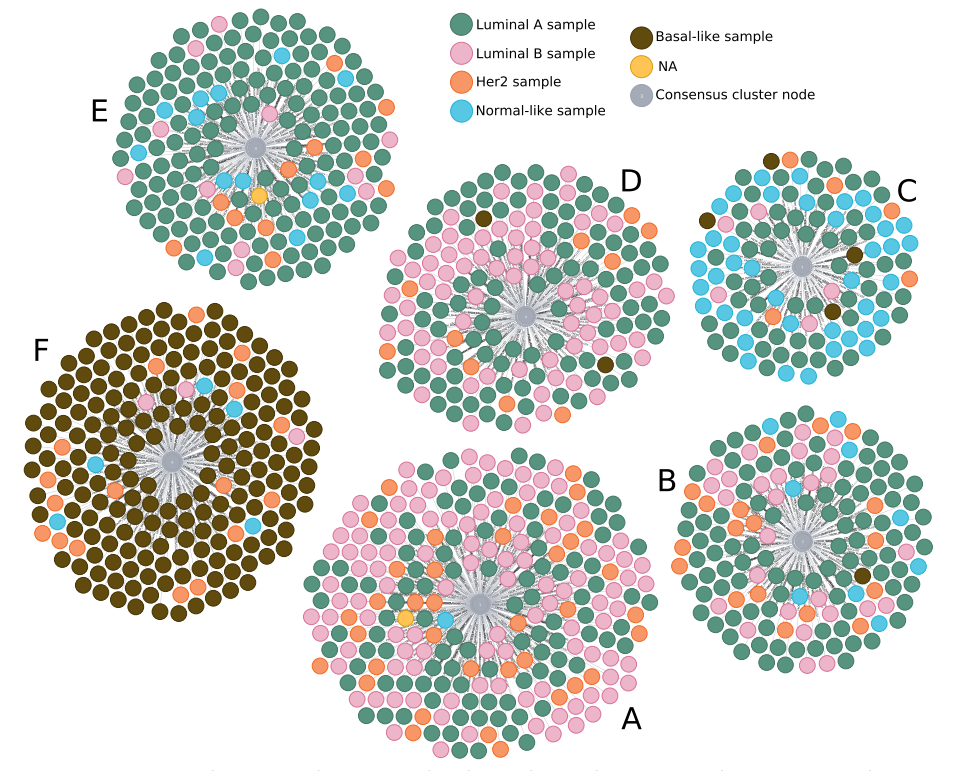


Fig. 6 BIC consensus clustering with patients colored according to the PAM50 prediction. Annotated screenshot from the Neo4j browser for graph visualization

Table 4 Over- and underrepresented clinical labels within BIC consensus clusters. ER+/ER− and PR+/PR−, respectively, correspond to estrogen receptor status and progesterone receptor status, positive and negative. M, N, and T stages refer to the TNM staging system

Cluster	Over-represented labels	Under-represented labels
A	Infiltrating Ductal Carcinoma	Infiltrating Lobular Carcinoma
	Her2, Luminal B	Basal, Normal-like
	ER+	ER−
B	MX, N3, T3	M0
	Stage III	
	Infiltrating Lobular Carcinoma	Infiltrating Ductal Carcinoma
	Luminal A	Basal
	ER+, PR+	ER−, PR−
C	MX, T3	M0, T2
	Infiltrating Lobular Carcinoma	Infiltrating Ductal Carcinoma
	Normal-like	Basal, Luminal B
	ER+, PR+	ER−, PR−
D	Stage X	
	Mucinous Carcinoma	Infiltrating Lobular Carcinoma
	Luminal B	Basal, Normal-like
	ER+, PR+	ER−, PR−
E	T1	
	Stage I	
	Luminal A	Basal, Luminal B
	ER+, PR+	ER−, PR−
F	N0	
	Stage II	Stage III
	Infiltrating Ductal Carcinoma, Medullary Carcinoma, Metaplastic Carcinoma	Infiltrating Lobular Carcinoma
	Basal	Luminal A, Luminal B
	ER−, PR−	ER+, PR+

samples was not reconstructed by RNA-seq unsupervised analysis [26]. Several studies also reported that the separation between Luminal subtypes was not consistent, suggesting that Luminal A and Luminal B samples may represent part of a continuum rather than distinct subgroups [26–28].

Moreover, clinical label enrichment analysis and additional tests applied to describe the clusters show good mapping between ClustOmics clusters and key biological clinical labels such as estrogen receptor and progesterone receptor status (ER/PR) or histological type of tumor (see Table 4).

Finally, we investigated the biological relevance of ClustOmics consensus clustering by comparing gene expression profiles between clusters. We computed the top 1000 genes differentially expressed across groups by applying the Kruskal–Wallis test [29] and selecting FDR adjusted p values below 0.001 (see Additional file 1: Figure S3). We clustered the top 1000 genes in 6 clusters using hierarchical clustering and for each gene list, we looked for overrepresented biological process (BP)-related Gene Ontology terms (GO terms). One of the gene clusters showed no significant results (FDR adjusted p values ≥ 0.05), but the other 5 gene lists were found enriched for cilium organization and assembly, response to transforming growth factor β, tissue

migration, T-cells activation, mitotic nuclear division or other biological processes (see Additional file 1: Figure S4).

More precisely, we found that the gene cluster $X2$, associated with cilium organization and assembly, microtubule bundle formation and regulation of intracellular steroid hormone receptor signaling pathway, was downregulated in consensus cluster F (composed mainly of Basal-like samples), compared to other consensus clusters. Gene cluster $X3$ was associated with response to transforming growth factor β, extracellular organization, transmembrane receptor protein serine/threonine kinase signaling pathway and regulation of muscle cells. Those genes appear downregulated in consensus clusters A (Her2, Luminal B samples), D (Luminal B samples) and F (Basal-like samples). Gene cluster $X4$, associated with epithelium migration and astrocyte differentiation, was found downregulated in consensus clusters A and D (both enriched in Luminal B samples) and upregulated in consensus cluster F (Basal-like samples). Gene cluster $X5$ is associated with T-cell activation, lymphocyte and leukocyte differentiation, membrane raft organization and regulation of peptidase activity and is downregulated in consensus clusters A and D (Luminal B samples). Finally, gene cluster $X6$, related to chromosome segregation and mitotic nuclear division, was found upregulated in consensus cluster F (Basal-like samples) and downregulated in consensus clusters B (Luminal A), C (Normal-like) and E (Luminal A).

Discussion

The novel method that we present in this paper deals with two key issues raised by the present context in biology and medicine and, in parallel, in bioinformatics. Indeed, these domains are witnessing an actual revolution in the acquisition of molecular data and thus facing a flood of various types of omics data. The ultimate goal is to benefit from the diversity and complementarity of these omics data (data on DNA methylation, copy number variations, polymorphism, etc.) by analyzing them simultaneously. However, multi-omics data integration is only one facet, as we also face an outburst of biocomputational approaches meant to deal with this unprecedented variety and quantity of data, and the choice of a method or of the optimal parameters is generally challenging. In this paper, both simultaneous integration of multiple omics and of various methods are tackled in an innovative manner through an original integration strategy based on consensus.

More specifically, in this work, we address the cancer subtyping problem from a personalized medicine-related perspective, which is gaining increasing attention. To treat patients according to their disease profile, one should be able to distinguish between disease subtypes. These disease subtypes can be predicted from omics data (traditionally gene expression but also methylation, miRNA, etc.) by performing patient clustering (hierarchical clustering, density-based clustering, distribution-based clustering, etc.). Our novel graph-based multi-integration method can fuse multiple input clustering results (obtained with existing clustering methods on diverse omics datasets) into one consensus clustering, regardless of the number of input clusters, number of objects clustered, omics and methods used to generate the input clusterings.

To compute a consensus clustering, our method, implemented in a tool called ClustOmics, uses an intuitive strategy based on evidence accumulation. The evidence

accumulation counts (i.e., the number of supports on the integration edges) make the consensus clustering results easier to interpret, as they provide insight into the extent to which the consensus clustering can be considered multi-source (issued from multiple omics) and to the overall agreement of input partitions.

The original EAC strategy as proposed by Fred and Jain [16] uses input partitions obtained by running the K-means algorithm multiple times (≈ 200) with random initialization of cluster centroids. From these partition results, a co-occurrence matrix is computed, and a minimum spanning tree algorithm is applied to find consensus clusters, by cutting weak links between objects at a threshold t defined by the user. The authors recommend that clusterings obtained for several values of t should be analyzed. In ClustOmics, we developed a weighted modularization optimization strategy to automatically select the best filtering threshold. Additionally, rather than generating input clusterings from running the same algorithm multiple times as proposed by Fred and Jain, here, we benefit from using various clustering strategies, each searching for different patterns and giving different insights to the data. This approach also allows the use of algorithms that are specialized for one omics type. Moreover, by taking as input a high number of clusterings obtained with a same tool with varying parameters, the convergence of the consensus clustering, especially in a single-omics context, can be improved, and this improvement can also be achieved by ClustOmics by giving the appropriate input clusterings.

Though our method does not formally weight input datasets (e.g., according to their level of confidence), one can artificially enhance the impact of one or several omics sources by providing supplementary single-omics input clusterings. In the same way, when dealing with missing data, patients measured with all omics are more likely to accumulate supports and therefore more likely to cluster together. In a context of multi-source integration, favoring individuals with the least quantity of missing data makes it possible to highlight the predictions supported by several data sources, which is the desired behavior. In a context of single-source integration, the same set of objects is usually used in all input clusterings (apart from a few specificities of the input clustering tools used).

TCGA real datasets from three different omics and ten cancer types were analyzed with respect to two integration scenarios: (1) fusing multi-omics clusterings obtained with existing integrative clustering tools and (2) fusing omics-specific input clusterings. In both cases, ClustOmics succeeded in computing high-quality multi-omics consensus clusterings, with clusters showing different survival curves and enriched for clinical labels of interest, coherent with what could be found in the cancer literature. Moreover, the results indicate that ClustOmics is robust to heterogeneous input clustering qualities (reconciling and smoothing the disparities of partition) and in comparison with a state-of-the-art consensus-based integration method, COCA.

The overall results show that the rMKL tool outperformed the other tools when considering the survival and clinical label enrichment metrics. However, our method is not meant to compete with existing single or integrative omics clustering methods, as it implements a more generic strategy. While "classical" tools take raw omics data as input, ClustOmics starts from classification results, thus allowing fusing any type of data, as

the classification results may correspond to clustering results (obtained from the analysis of one or several omics), to biological annotations, to clinical data, etc.

In contrast, ClustOmics aims at capitalizing on the preliminary input predictions to increase their robustness by taking advantage of accumulating evidence to reveal sharper patterns in the data. This selection of robust patterns across input partitions renders ClustOmics stable when facing heterogeneous input clusterings and is particularly useful when no gold-standard metric is available to assess the quality of the results. Hence, with a sufficient number of input clusterings, no prior analysis of the input is needed, given that low-quality clusterings, likely to add noise to the integration graph, play a smaller part in the evidence accumulation. Omics for which the separation of samples is clear-cut will drive the consensus clustering, while omics that do not show interesting patterns across samples will be faded via the integration graph filtering step. For the same reason, it is important to highlight that as long as the signals in the available omics are strong, ClustOmics is able to cluster samples that do not appear in all omics datasets, making use of available data and addressing the issue of partial data.

Finally, though presented in a disease subtyping context, one should grasp that our method is not limited to this application case. ClustOmics is generic and adaptable to a wide range of biological questions, as one can use any kind of partitioning of the data, including clinical labels, groups of genes of interest, etc., as an input clustering. A major strength of ClustOmics resides in its exploratory aspect, resulting both from a flexible intrinsic model that gives the user complete power on the integration scenario to investigate and from the use of the graph-oriented database Neo4j. All input data and metadata are stored in this kind of database, which may easily be queried and visualized by a nonspecialist with the Neo4j browser.

Conclusion

Facing the diversity and heterogeneity of omics data and clustering strategies, one might want to make profit from all available data to compute a consensus clustering. ClustOmics is able to fuse any set of input clusterings into one robust consensus, which can easily be interpreted based on the number of supports evidence accumulation scores. ClustOmics can be adapted to answer a wide range of biological questions. The use of integration scenarios allows users to explore various integration strategies, by adding or discarding data sources and/or clustering methods.

Methods

In this section, we detail the strategy implemented in ClustOmics. We then describe the datasets and the metrics that were used to evaluate our new method. For the sake of simplicity as, in this paper, ClustOmics was applied in the context of cancer subtyping, we will further refer to objects of interest as *Patients*. However, ClustOmics can be applied to different biological entities such as genes or cells.

ClustOmics integration strategy

The ClustOmics integration strategy, depicted in Fig. 7, starts from a set of input clusterings generated with various clustering methods and/or from different omics sources.

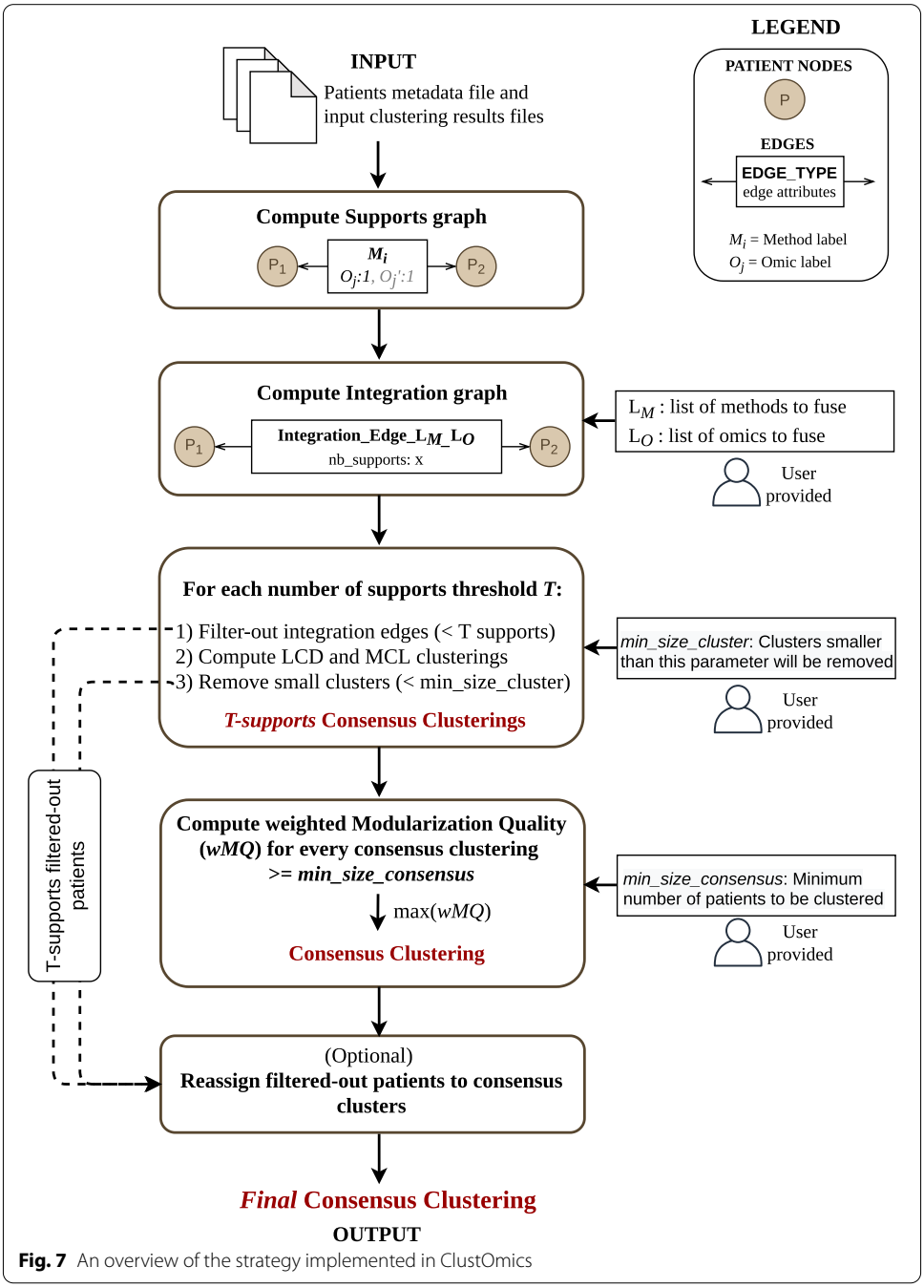


Fig. 7 An overview of the strategy implemented in ClustOmics

First, from a patient metadata file and from available input clusterings, a support graph (SG) is instantiated. In this graph, each *Patient (P)* corresponds to a node and shares a *support edge* with another patient when classified in the same cluster (co-clustered) in at least one input clustering. One support edge relates to one input clustering tool, and each support edge displays one or multiple attributes to indicate the omics sources supporting the co-clustering of the patients.

Next, given an integration scenario, meaning a list of omics and methods to integrate, the corresponding *integration graph (IG)* is computed. Then, the integration graph is

filtered and clustered to produce a consensus clustering according to the given integration scenario.

Below, we detail the integration graph computation, filtering and clustering steps, resulting in a ClustOmics consensus clustering.

Compute the integration graph (IG)

Given an *Integration Scenario* (defined by a set of input clusterings), ClustOmics exploits the information on the support edges to compute the so-called *number of supports*, by counting the considered input clusterings sustaining the association of patients. The numbers of supports are reported on the *Integration Edges* and so, for a given integration scenario, a pair of patients may share at most one integration edge. In this way, heterogeneous data is aggregated into co-occurrence counts that are used as a similarity measure to perform evidence accumulation clustering (EAC) [16].

Filter and cluster the integration graph

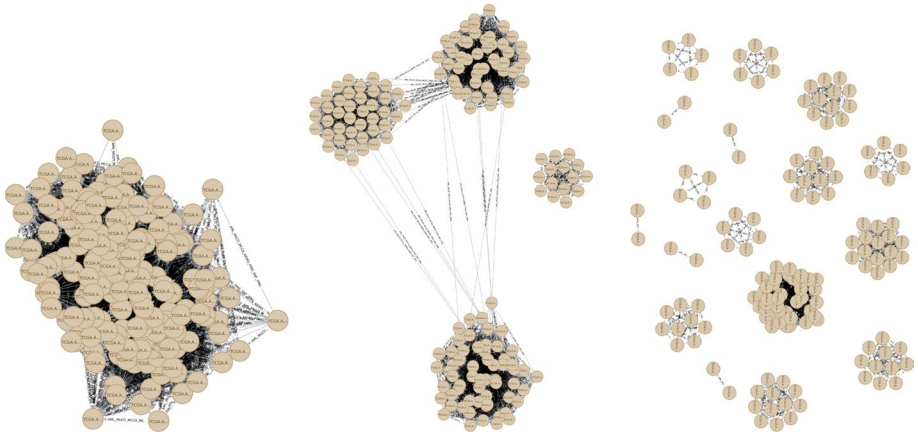
Integration graphs are generally densely connected, as each pair of nodes may have been clustered together at least once over the set of omics and methods. However, as integration edges are weighted with the number of supports agreeing on the corresponding associations, the most robust integration edges can be distinguished from predictions that are not consistent across omics and methods. Hence, ClustOmics filters the graph according to the number of supports by removing non-consistent integration edges. The goal is to obtain a filtered graph foreshadowing *natural* clusters that correspond to a consensus. The choice of a threshold to filter the integration edges is therefore determinant.

Figure 8 depicts the impact of an increasing number of support-filtering thresholds on the internal structure of the integration graph. One can observe that two issues arise from this filtering process:

- First, increasing the threshold generates smaller graphs. Indeed, pairs of nodes that do not share any integration edge with a sufficient number of supports are removed, leading to a partial classification of the input set of patients.
- The second issue is the loss of structure in the filtered integration graph: when filtering at a high threshold, the resulting graph may become too sparse to be considered informative, like the graph in Fig. 8c with numerous small connected components.

Therefore, producing a relevant classification requires finding the best compromise for the support threshold. For this effort, ClustOmics tests all possible configurations by iteratively filtering the integration graph with increasing support thresholds. At each iteration, ClustOmics uses state-of-the-art graph clustering methods (see the subsection below) to compute consensus clusterings for the corresponding filtered integration graph.

The resulting consensus clusterings should be analyzed with respect to the number of supports used to filter the graph prior to clustering. This number of supports indicates the level of agreement between the input clusterings and gives insight on the extent to which the resulting consensus clustering can be considered as being truly multi-omics.



(a) No filtering
197 nodes, 8122 edges

(b) $nb\_supports \geq 3$
170 nodes, 3471 edges

(c) $nb\_supports \geq 5$
148 nodes, 1244 edges

Fig. 8 An integration graph filtered with increasing threshold values: 1, 3, and 5 (the maximum number of supports for an integration edge being 5 in this example). Screenshots from the Neo4j browser for graph visualization

Moreover, to deal with the two issues described above and to keep merely informative results, each integration graph consensus clustering result goes through an additional filtering step with the two following parameters:

- the *min_size_cluster* parameter indicates the minimum accepted size for a cluster that is part of the consensus clustering. Clusters with less than *min_size_cluster* nodes are removed from the analysis.
- the *min_size_consensus* parameter indicates to what extent ClustOmics is allowed to discard nodes, i.e., patients. ClustOmics will further consider only consensus clustering results having at least *min_size_consensus* nodes.

Finally, a quality metric, i.e., the weighted MQ index, is computed for all consensus clusterings that passed the filtering steps.

Optionally, one may want to reconsider the individuals that were discarded during the filtering steps (either when filtering-out integration edges or small clusters) and analyze them with respect to the consensus clusters. With this in mind, ClustOmics is able to reassign filtered-out individuals with respect to the mean number of supports shared with patients from consensus clusters, though such additional predictions do not necessarily meet the threshold with which the consensus clusters were originally computed.

Below, we give insights on the graph clustering algorithms that are used to compute the consensus clusters, as well as on the quality metric employed for the identification of a robust consensus clustering, the *weighted modularization quality*.

Graph clustering algorithms

ClustOmics filters the integration graph for each possible threshold on the number of supports and, for each filtered graph, ClustOmics computes two consensus clusterings with two state-of-the-art, complementary graph clustering algorithms: the Louvain community detection (LCD) algorithm, based on modularity optimization [30], and the Markov clustering (MCL) algorithm, based on the simulation of stochastic

flow in graphs [31]. MCL and LCD are unsupervised clustering algorithms and do not require the number of clusters to be estimated in advance.

Modularity optimization is one of the most popular strategies in graph clustering algorithms [32], while MCL/MCL-based methods have proven highly efficient in various biological network analyses (protein-protein interaction networks [33, 34], protein complex identification [35], detection of protein families [36]). Moreover, modularity optimization algorithms have been shown to present a resolution issue [37, 38]: a tendency to fuse small clusters (even for those that are well defined and have few interconnections), thus favoring the formation of larger clusters than those computed by MCL [39]. Small clusters predicted by MCL can be an issue in the ClustOmics case, as it removes clusters smaller than the user-defined *min_size_cluster* parameter, considering them to be non-informative.

Selection of the best consensus clustering based on the weighted MQ index

The *modularization quality* (*MQ*) was first defined by Mancoridis et al. in the context of software engineering [40]. Compared to the popular *modularity* measure [41] optimized in graph clustering algorithms, which compares the distribution of edges with respect to a random graph with the same number of vertices and edges as the original graph, the *modularization quality* (*MQ*) evaluates the quality of a clustering as the difference between internal and external connectivity ratios; that is, the ratio between the number of connections observed within a given cluster and between two given clusters, and the maximum possible number of such edges. An optimal clustering for this measure should maximize the intraconnectivity ratio (every two nodes belonging to the same cluster share an edge) and minimize the interconnectivity ratio (nodes classified in different clusters do not share edges). Indeed, in the context of consensus clustering based on evidence accumulation, it makes more sense to compare the distribution of the edges in the integration graph to the case where all nodes would have been partitioned in the same optimal way in all input clusterings, i.e., a graph where all intracluster nodes are connected, and all intercluster nodes are disconnected.

Moreover, we adapted the original *MQ* index for weighted undirected graphs with no self-loops (in our case, a *Patient* node cannot share an integration edge with itself). We denote this adaptation of the modularization quality as the *weighted modularization quality* (*wMQ*).

Let $G = (V, E)$ be a graph where V denotes the set of nodes and E the set of edges of G. Let $C = (C_1,..., C_k)$ be a consensus clustering with K clusters and $|C_i|$ the number of nodes classified in cluster C_i. Let us also note $w(e)$ the weight of a given edge, $W(e_{ii})$ the sum of weights of the edges internal to C_i cluster (connecting vertices from C_i), $W(e_{ij})$ the sum of weights between clusters C_i and C_j (connecting a vertex from C_i to a vertex from C_j) and $max(w_{IG})$ the maximum possible weight on the edges of the given integration graph (the maximum possible number of supports, also corresponding to the number of input clusterings being fused). We therefore define the *wMQ* index computed for a consensus clustering C obtained on the integration graph IG as:

$$wMQ(IG, C) = \frac{1}{k \times max(w_{IG})} \sum_{i=1}^{k} \left(\frac{2 \times W(e_{ii})}{|C_i| \times (|C_i| - 1)} - \frac{1}{k-1} \sum_{j \neq i} \frac{W(e_{ij})}{|C_i| \times |C_j|} \right)$$

The first term of the sum corresponds to the weighted internal connectivity ratio for a cluster C_i. Indeed, the sum of the internal edges weights $W(e_{ii})$ is adjusted with the maximum possible value of the sum of the edges linking a set of $|C_i|$ nodes, which would be reached if all C_i nodes were connected with $max(w_{IG})$ weighted edges. Note that for an undirected and no self-loop graph, the maximum number of edges in a subgraph of $|C_i|$ nodes is $\binom{|C_i|}{2}$. Similarly, the second term of the sum represents the weighted external connectivity ratio of a cluster C_i, given by the sum of the weights of the edges linking a node from cluster C_i to a node belonging to a cluster C_j ($\neq C_i$).

The *wMQ* values range from -1 to 1, where a *wMQ* of -1 corresponds to the case where there is no intracluster edge and all intercluster pairs of vertices are connected with edges of weight $max(w_{IG})$. A *wMQ* of 1 corresponds to the case where no intercluster vertices are connected, and all pairs of intracluster vertices are connected with edges of weight $max(w_{IG})$. A high-standard consensus clustering should maximize this index.

ClustOmics computes the *wMQ* for the LCD and MCL consensus clusterings obtained with various numbers of support-thresholds and having passed the filtering steps, and it returns the clustering that maximizes this quality measure.

Datasets and tools used for computing input clusterings

We used ClustOmics to predict cancer subtypes from gene expression, microRNA expression and DNA methylation datasets available in *The Cancer Genome Atlas* (TCGA) [42]. Our case-study is based on the same datasets as in *Rappoport and Shamir*'s review on multi-omics clustering methods [1]. The data cover ten cancer types: leukemia (AML), breast (BIC), colon (COAD), glioblastoma (GBM), kidney (KIRC), liver (LIHC), lung (LUSC), ovarian (OV), sarcoma (SARC) and skin (SKCM). For each cancer type, from 197 and up to 1098 patients were measured for at least one of the three omics (expression, miRNA and methylation), of which 170 to 621 patients were measured for all three. More details on missing data per cancer type are given in Table 5.

To generate input clusterings to be fused by ClustOmics, we used five state-of-the-art integrative clustering tools, summarized in Table 6: PINS [6], SNF [7], NEMO [8], rMKL [9] and MultiCCA [10]. The first four tools can be used in a single-omics context as well as in a multi-omics context, for which they were all designed. Though NEMO is the only tool that can handle partial data, for comparability purposes, this functionality was not used in the analyses we conducted. Each tool has been run with default parameters and based on recommendations of the authors. For the single-to-multi scenario, we also computed single-omics input clusterings with K-means clustering [19], for which the optimal number of clusters K was determined using the Silhouette index [24], with values of K ranging from 2 to 20.

Computation of COCA consensus clusterings

To assess the performance of ClustOmics with respect to a state-of-the-art integration method based on consensus clustering, we applied cluster-of-clusters analysis (COCA)

Table 5 Number of patients measured per omics for each cancer type. Total: Number of patients measured for at least one omics (of which those having been measured for the three omics); proportion of partial data; Exp+Met: Patients measured for expression and methylation only; Exp+miRNA: Patients measured for expression and miRNA only; Met+miRNA: Patients measured for methylation and miRNA only; Exp: Patients measured for expression only; Met: Patients measured for methylation only; miRNA: Patients measured for miRNA only

Cancer	Total (multi-omic)	% partial data	Exp+ Met	Exp+ miRNA	Met+ miRNA	Exp	Met	miRNA
AML	197 (170)	13.71	0	3	15	0	9	0
BIC	1096 (621)	43.34	159	132	2	181	1	0
COAD	303 (220)	27.39	57	0	0	8	18	0
GBM	578 (274)	52.60	4	245	3	5	4	43
KIRC	534 (183)	65.73	135	71	0	144	1	0
LIHC	377 (367)	2.65	4	0	5	0	1	0
LUSC	501 (341)	31.94	29	1	0	130	0	0
OV	591 (287)	51.44	7	0	166	9	122	0
SARC	261 (257)	1.53	2	0	2	0	0	0
SKCM	368 (351)	4.62	16	1	0	0	0	0

Table 6 Methods used to compute input clusterings

Software	Multi-omic context	Single-omic context
PINS [6]	Yes	Yes
SNF [7]	Yes	Yes
NEMO [8]	Yes	Yes
rMKL [9]	Yes	Yes
MultiCCA [10]	Yes	No
k-means [19]	No	Yes

[5] on each integration scenario and from the same set of input clusterings as for ClustOmics. COCA had already been applied to cancer-subtyping in a multi-omics context [43, 44].

COCA is an integrative clustering tool based on the consensus clustering (CC) algorithm introduced by Monti et al. [45]. The CC algorithm implements a resampling- and co-occurrence-based strategy to assess the stability of clusters when analyzing a single dataset. By resampling a single dataset multiple times and applying a clustering algorithm on each perturbed dataset, and from the co-occurrences counts of samples in clusters, a consensus matrix is computed and used as a similarity matrix to compute a final consensus clustering. COCA was run using default parameters, under the same integration scenarios as in ClustOmics, and using the same set of input clusterings.

Clustering pairwise similarity metric

To evaluate the similarity of ClustOmics and COCA consensus clusterings with respect to their inputs or each other, we used the adjusted Rand index (ARI) [20], a measure of similarity between two data clusterings. While the ARI has been used to evaluate the quality of classifications compared to ground-truth data, here, we use it to compare the similarity of various clusterings, without considering any quality aspect.

Biological metrics

To explore the biological relevance of input clusterings and consensus clustering results, we computed the *overall survival rate* of patients. As cancer acuteness is proved to be related to its molecular subtype [46–48], we further investigated whether it was significantly different across clusters using the exact log-rank test for more than two groups, introduced in [49]. For each clustering, the p value of the log-rank test was computed using 100, 000 random permutations of the data.

In addition, we performed an analysis of *clinical labels enrichment* in clusters, using 32 labels available from TCGA metadata (see Table 7). The idea is that patients affected by the same cancer subtype should also share, to a certain extent, the same clinical characteristics. The abundance of clinical labels in clusters and their statistical over-representation provide information on the biological robustness of clusterings. To perform this analysis, we used pancancer (e.g., age at diagnosis or pathologic stage of cancer) and cancer-specific clinical labels for each cancer type (e.g., presence of colon polyps for colon cancer, or smoking history for lung cancer). Clinical labels that were absent for more than half of the patients were removed from the analysis. We used the χ^2 test for independence for discrete parameters and the Kruskal-Wallis test for numeric parameters to assess the enrichment of the clinical labels in a cluster. To increase the robustness of the results, we applied a bootstrapping strategy, computing the test on randomly permuted data to derive an empirical p value (100, 000 permutations).

One must keep in mind that molecular data do not always explain survival or clinical differences between groups of samples. Therefore, in the discussion of the results, we consider survival and clinical analysis as ways to interpret patterns captured by the various clustering results and do not favor one metric over the other.

Finally, to evaluate differentially expressed genes across consensus clusters generated using the StoM scenario on the BIC study case, we applied the Kruskal-Wallis test on each gene available from the BIC expression dataset. The p values were adjusted to control the false discovery rate (FDR) [50], filtered with a 0.001 significance threshold, and top 1000 most significant genes were retained for further

Table 7 Pancancer and cancer-specific labels used for clinical label enrichment analysis. Pathological M, N and T labels refer to the TNM staging system, which describes the anatomical extent of tumor cancers [25]

Pan-cancer	Age at initial pathologic diagnosis, Gender, Pathologic M, Pathologic N, Pathologic T, Pathologic stage, Histological type, New neoplasm event type, Neoplasm histologic grade
AML	CALGB cytogenetics risk category, FAB morphology code
BIC	PAM50Call RNAseq, Estrogen receptor status, Progesterone receptor status, ER level cell percentage category, PR level cell percentage category
COAD	Presence of colon polyps, History of colon polyps
GBM	Prior glioma
KIRC	Hemoglobin result, Platelet qualitative result, Serum calcium result, White cell count result
LIHC	Adjacent hepatic tissue inflammation extent type, Albumin result specified value, Creatinine value, Fetoprotein outcome value, Fibrosis ishak score
LUSC	Tobacco smoking history, Number pack years smoked
OV	No supplementary clinical label
SARC	No supplementary clinical label
SKCM	Melanoma Clark level value, Melanoma ulceration indicator

analysis. Using hierarchical clustering [51], we clustered the top gene list and investigated clusters for enriched Gene Ontology [52] biological process terms with Cluster Profiler [53]. FDR-adjusted p values were filtered with a 0.05 cutoff.

Implementation

ClustOmics is implemented based on the Neo4j graph database management system and uses APOC and Graph Data Science Neo4j libraries. Queries on the graph database are performed in Cypher, Neo4j's graph query language, and are encapsulated in Python scripts. To facilitate its use, ClustOmics can be run through the Snakemake workflow management system.

ClustOmics was tested on a desktop-computer with an Intel Xeon processor (2.70 GHz, 62 GB of RAM) running on Ubuntu 18.04. For the TCGA real datasets that it was applied to, the ClustOmics runtimes range from a few minutes for small datasets (AML, multi-to-multi scenario) up to 2 h for the largest dataset (BIC, single-to-multi scenario), with most of the computation time being consumed for the construction of the integration graph. With the graph stored in a Neo4j database, this step is only to be performed once for each integration scenario, and parameters for graph filtering can be further set and tuned without recomputing the graph.

The ClustOmics source code, released under MIT license, and the results obtained on the ten cancer types with the two integration scenarios described in this paper are available on GitHub: https://github.com/galadrielbriere/ClustOmics.

Abbreviations

AML: Acute myeloid leukemia; ARI: Adjusted Rand index; BIC: Breast invasive carcinoma; CALGB: Cancer and leukemia group B; COAD: Colon adenocarcinoma; EAC: Evidence accumulation clustering; ER: Estrogen receptor; FAB: French–American–British (morphology code); FDR: False discovery rate; GBM: Glioblastoma; KIRC: Kidney renal clear cell carcinoma; LCD: Louvain community detection; LIHC: Liver hepatocellular carcinoma; LUSC: Lung squamous cell carcinoma; MCL: Markov clustering; MtoM: Multi-to-multi (integration scenario); MQ: Modularization quality; OV: Ovarian carcinoma; PR: Progesterone receptor; SARC: Sarcoma; SKCM: Skin cutaneous melanoma; StoM: Single-to-multi (integration scenario); TCGA: The cancer genome Atlas; wMQ: Weighted modularization quality.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04279-1.

Additional file 1. Supplementary Figures 1 to 4.

Acknowledgements

We kindly thank Nora Speicher for sending us the executable of the rMKL tool.

Authors' contributions

GB, PT and RU conceived the project and the method on which ClustOmics is based. GB developed the ClustOmics tool and performed the bioinformatics experiments. GB, ED, PT and RU participated at the analysis of the results. GB, ED, PT and RU wrote, read and approved the final manuscript. All authors read and approved the final manuscript.

Funding

This work was partially done under the NEOMICS project (INS2I PEPS Blanc 0201 2019), supported by the CNRS (France); and under the LaBRI (Laboratoire Bordelais de Recherche en Informatique, France). These funding bodies did not play any role in the design or conclusions of our study.

Availability of data and materials

The datasets analysed in the study were obtained from Ron Shamir's lab and are available at http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. ClustOmics' source code, released under MIT licence, as well as the results obtained on TCGA cancer data are available on Github: https://github.com/galadrielbriere/ClustOmics.

Declarations

Ethics approval and consent to participate
No ethics approval was required for the study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]CNRS, Bordeaux INP, LaBRI, UMR 5800, Univ. Bordeaux, 33400 Talence, France. [2]INRA, Bordeaux INP, NutriNeuro, UMR 1286, Univ. Bordeaux, 33000 Bordeaux, France. [3]INSERM U1218, Institut Bergonié, Univ. Bordeaux, 33076 Bordeaux, France.

Received: 20 October 2020 Accepted: 28 June 2021
Published online: 06 July 2021

References

1. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–62. https://doi.org/10.1093/nar/gky889.
2. Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration-a comparison of unsupervised clustering methodologies. Brief Bioinform. 2019;20(4):1269–79. https://doi.org/10.1093/bib/bbx167.
3. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genom. 2015;16:1022. https://doi.org/10.1186/s12864-015-2223-8.
4. Wang H, Nie F, Huang H. Multi-view clustering and feature learning via structured sparsity. In: Proceedings of the 30th international conference on international conference on machine learning—volume 28. ICML'13, pp. 352–360. JMLR.org, Atlanta, GA, USA. 2013.
5. Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of omic datasets. Bioinformatics (Oxford, England). 2020;36(18):4789–96. https://doi.org/10.1093/bioinformatics/btaa593.
6. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. Genome Res. 2017;27(12):2025–39. https://doi.org/10.1101/gr.215129.116.
7. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7. https://doi.org/10.1038/nmeth.2810.
8. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics. 2019;35(18):3348–56. https://doi.org/10.1093/bioinformatics/btz058.
9. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015;31(12):268–75. https://doi.org/10.1093/bioinformatics/btv244.
10. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009;8(1):1–27. https://doi.org/10.2202/1544-6115.1470.
11. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(1):523–42. https://doi.org/10.1093/nar/gky8890.
12. Gabasova E, Reid J, Wernisch L. Clusternomics: integrative context-dependent clustering for heterogeneous datasets. PLoS Comput Biol. 2017;13(10):1005781. https://doi.org/10.1093/nar/gky8891.
13. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics (Oxford, England). 2018;19(1):71–86. https://doi.org/10.1093/nar/gky8892.
14. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29(20):2610–6. https://doi.org/10.1093/nar/gky8893.
15. Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. Int J Pattern Recognit Artif Intell. 2011;25(03):337–72. https://doi.org/10.1093/nar/gky8894.
16. Fred ALN, Jain AK. Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell. 2005;27(6):835–50. https://doi.org/10.1093/nar/gky8895.
17. Neo4j Graph Platform-The Leader in Graph Databases. https://doi.org/10.1093/nar/gky8896. Accessed 28 Sept 2020
18. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, Vee LJv, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Waes CV, Chen Z, Collisson EA, Network TCGAR, Benz C, Perou CM, Stuart JM. Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin. Cell. 2014;158(4):929. https://doi.org/10.1016/j.cell.2014.06.049.
19. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Cam LML, Neyman J, editors. Proceedings of fifth Berkeley symposium on mathematical statistics and probability, vol. 1. Berkeley: University of California Press; 1967. p. 281–97.
20. Steinley D. Properties of the Hubert–Arabie adjusted rand index. Psychol Methods. 2004;9:386–96. https://doi.org/10.1037/1082-989X.9.3.386.
21. Byrd JC, Mrózek K, Dodge RK, Carroll AJ, Edwards CG, Arthur DC, Pettenati MJ, Patil SR, Rao KW, Watson MS, Koduru PRK, Moore JO, Stone RM, Mayer RJ, Feldman EJ, Davey FR, Schiffer CA, Larson RA, Bloomfield CD. Cancer and

Leukemia Group B (CALGB 8461): pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). Blood. 2002;100(13):4325–36. https://doi.org/10.1182/blood-2002-03-0772.

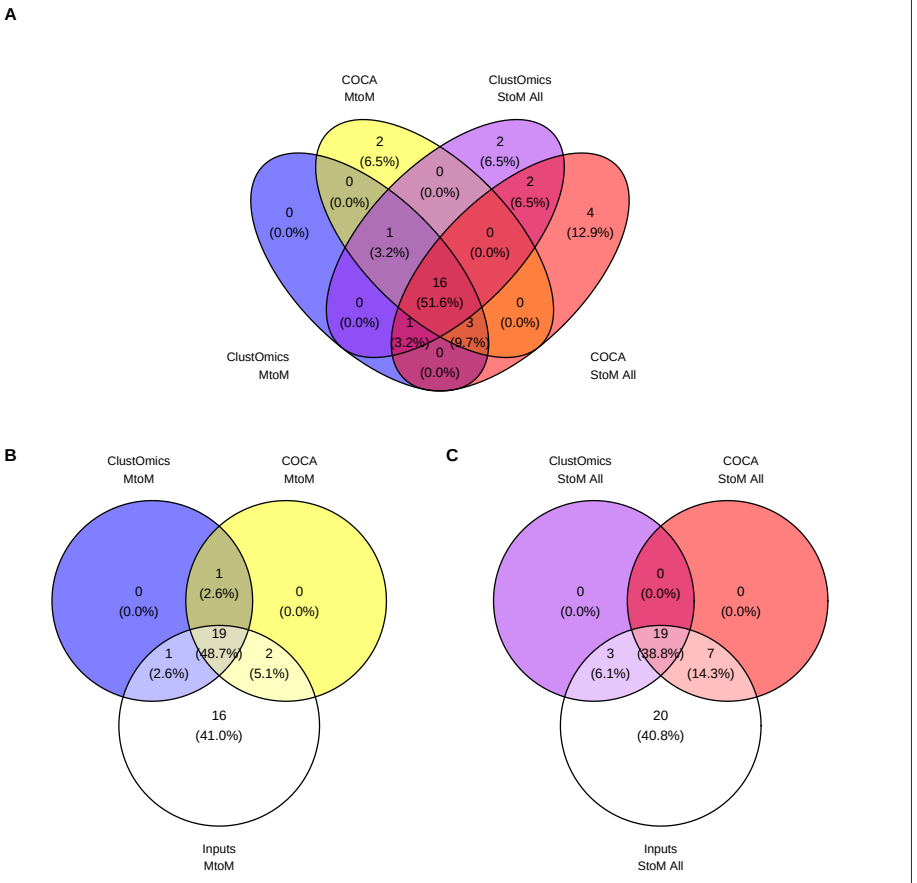
22. Bennett JM, Catovsky D, Daniel M-T, Flandrin G, DaG Galton, Gralnick HR, Sultan C. Proposals for the classification of the acute leukaemias French–American–British (FAB) co-operative group. Br J Haematol. 1976;33(4):451–8. https://doi.org/10.1111/j.1365-2141.1976.tb03563.x.
23. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
24. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken: Wiley; 1990. https://doi.org/10.1002/9780470316801.
25. Sobin LH, Gospodarowicz MK, Wittekind C. International Union against Cancer (eds.): TNM Classification of Malignant Tumours, 7th ed. edn. Wiley-Blackwell, Chichester, West Sussex, UK: Hoboken, NJ; 2009. p. 2010.
26. Netanely D, Avraham A, Ben-Baruch A, Evron E, Shamir R. Expression and methylation patterns partition luminal—a breast tumors into distinct prognostic subgroups. Breast Cancer Res. 2016;18(1):74. https://doi.org/10.1186/s13058-016-0724-2.
27. Alizart M, Saunus J, Cummings M, Lakhani SR. Molecular classification of breast carcinoma. Diagn Histopathol. 2012;18(3):97–103. https://doi.org/10.1016/j.mpdhp.2011.12.003.
28. Weigelt B, Mackay A, A'hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol. 2010;11(4):339–49. https://doi.org/10.1016/S1470-2045(10)70008-5.
29. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441.
30. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
31. Van Dongen SM. Graph clustering by flow simultion. Ph.D. Thesis, University of Utrecht, Netherlands, 2000.
32. Fortunato S. Community detection in graphs. Phys Rep. 2010;486:3. https://doi.org/10.1016/j.physrep.2009.11.002.
33. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinform. 2006;7(1):488. https://doi.org/10.1186/1471-2105-7-488.
34. Vlasblom J, Wodak SJ. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinform. 2009;10(1):99. https://doi.org/10.1186/1471-2105-10-99.
35. Lei X, Wang F, Wu F-X, Zhang A, Pedrycz W. Protein complex identification through Markov clustering with firefly algorithm on dynamic protein-protein interaction networks. Inf Sci. 2016;329:303–16. https://doi.org/10.1016/j.ins.2015.09.028.
36. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84. https://doi.org/10.1093/nar/30.7.1575.
37. Lancichinetti A, Fortunato S. Limits of modularity maximization in community detection. Phys Rev E. 2011;84(6):066122. https://doi.org/10.1103/PhysRevE.84.066122.
38. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc Natl Acad Sci. 2007;104(1):36–41. https://doi.org/10.1016/S1470-2045(10)70008-50.
39. Sardana D, Bhatnagar R. Graph clustering using mutual K-nearest neighbors. In: Active media technology. Lecture notes in computer science. Cham: Springer. 2014. pp. 35–48. https://doi.org/10.1007/978-3-319-09912-5_4
40. Mancoridis S, Mitchell BS, Rorres C, Chen Y, Gansner ER. Using automatic clustering to produce high-level system organizations of source code. In: Proceedings. 6th international workshop on program comprehension. IWPC'98 (Cat. No.98TB100242), pp. 45–52. IEEE Comput. Soc, Ischia, Italy. 1998). https://doi.org/10.1109/WPC.1998.693283
41. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006;103(23):8577–82. https://doi.org/10.1016/S1470-2045(10)70008-51.
42. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8. https://doi.org/10.1038/nature07385.
43. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490(7418):61–70. https://doi.org/10.1038/nature11412.
44. Aure MR, Vitelli V, Jernstrom S, Kumar S, Krohn M, Due EU, Haukaas TH, Leivonen S-K, Vollan HKM, Luders T, Rodland E, Vaske CJ, Zhao W, Moller EK, Nord S, Giskeodegard GF, Bathen TF, Caldas C, Tramm T, Alsner J, Overgaard J, Geisler J, Bukholm IRK, Naume B, Schlichting E, Sauer T, Mills GB, Karesen R, Maelandsmo GM, Lingjaerde OC, Frigessi A, Kristensen VN, Borresen-Dale A-L, Sahlberg KK, Borgen E, Engebraten O, Fodstad O, Fritzman B, Garred O, Geitvik GA, Hofvind S, Russnes HG, Skjerven HK, Sorlie T. OSBREAC: Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Res. 2017;19(1):44. https://doi.org/10.1016/S1470-2045(10)70008-52.
45. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52(1):91–118. https://doi.org/10.1016/S1470-2045(10)70008-53.
46. Noone A-M, Cronin KA, Altekruse SF, Howlader N, Lewis DR, Petkov VI, Penberthy L. Cancer incidence and survival trends by subtype using data from the surveillance epidemiology and end results program, 1992–2013. Cancer Epidemiol Biomark Prev. 2017;26(4):632–41. https://doi.org/10.1016/S1470-2045(10)70008-54.
47. Fallahpour S, Navaneelan T, De P, Borgo A. Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. CMAJ Open. 2017;5(3):734–9. https://doi.org/10.1016/S1470-2045(10)70008-55.
48. Jones JC, Renfro LA, Al-Shamsi HO, Schrock AB, Rankin A, Zhang BY, Kasi PM, Voss JS, Leal AD, Sun J, Ross J, Ali SM, Hubbard JM, Kipp BR, McWilliams RR, Kopetz S, Wolff RA, Grothey A. Non-V600BRAF mutations define a clinically

distinct molecular subtype of metastatic colorectal cancer. J Clin Oncol. 2017;35(23):2624–30. https://doi.org/10.1016/S1470-2045(10)70008-56.
49. Rappoport N, Shamir R. Inaccuracy of the log-rank approximation in cancer data analysis. Mol Syst Biol. 2019;15(8):8754. https://doi.org/10.15252/msb.20188754.
50. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
51. Rokach L, Maimon O. Clustering methods. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. Boston: Springer; 2005. p. 321–52.
52. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Gene Ontol Consort Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
53. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7. https://doi.org/10.1089/omi.2011.0118.

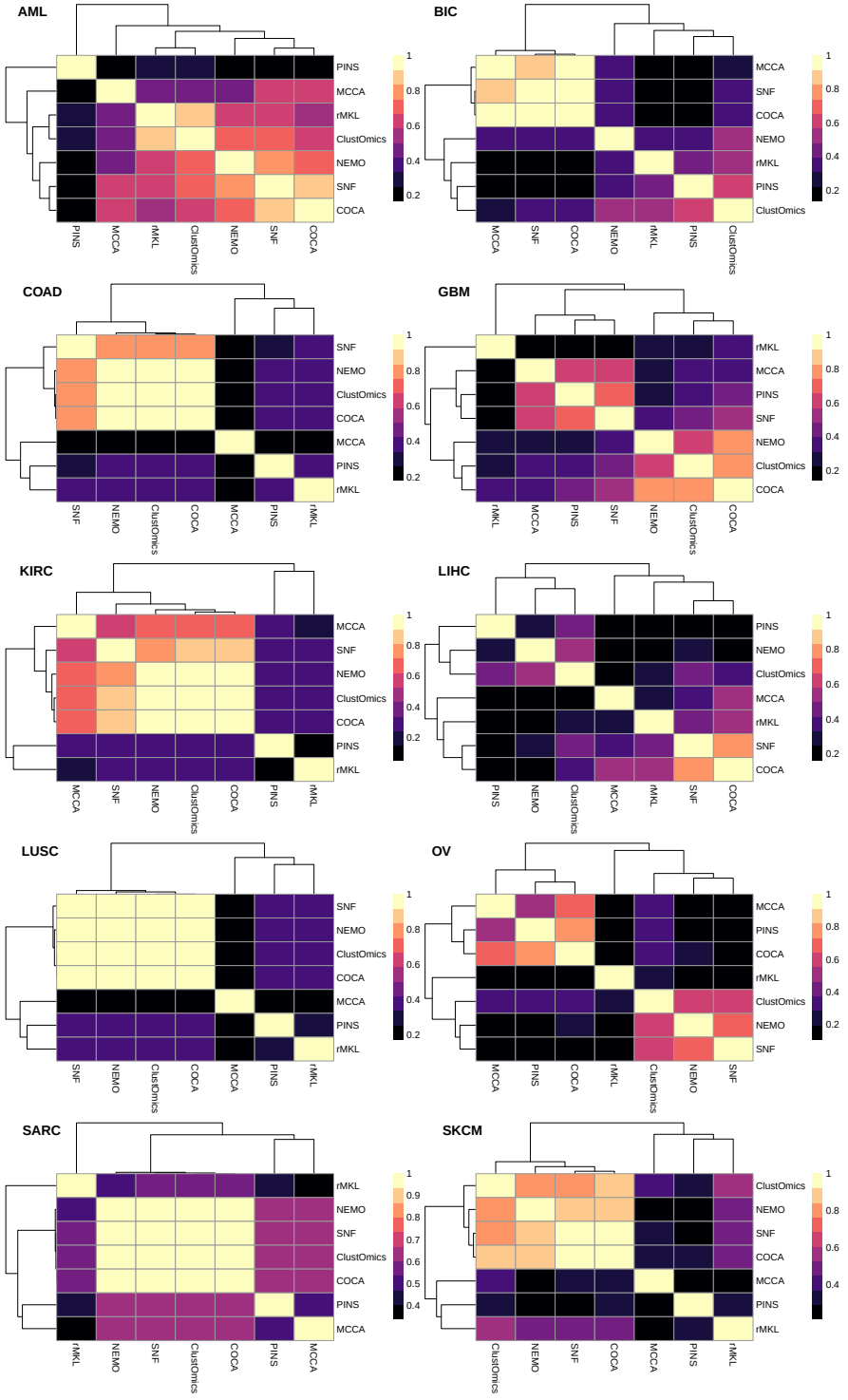
Publisher's Note

Supplementary material

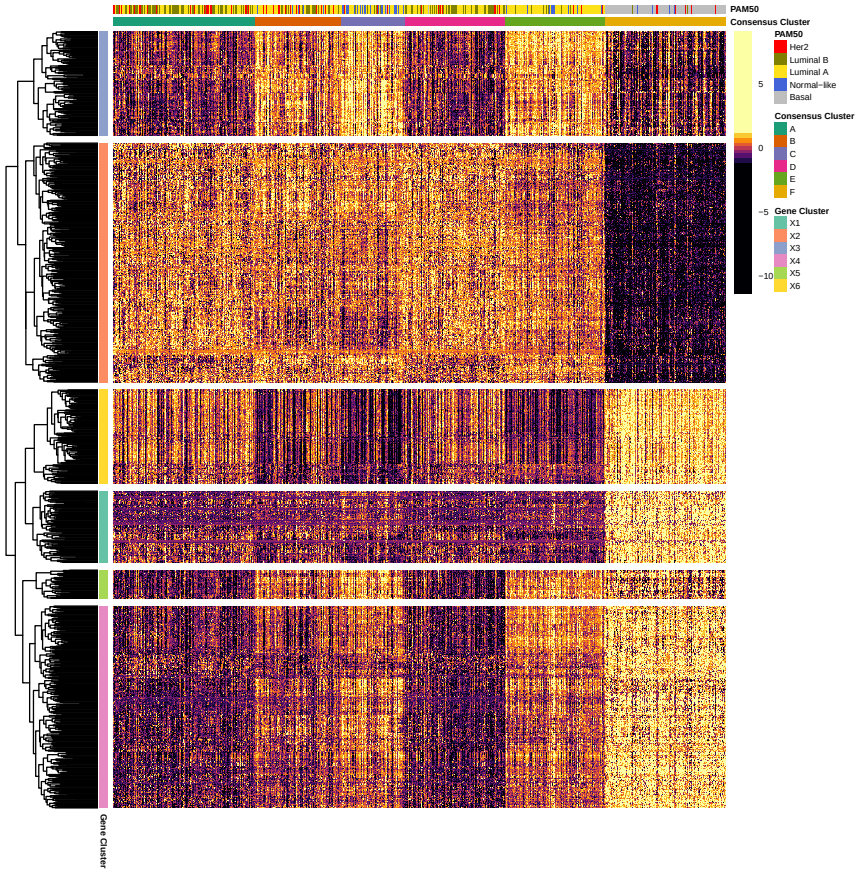
Supplementary Figure 1 Distribution of clinical labels found enriched upon all cancer types : (A) in the consensus clusterings for both MtoM and StoM All scenarios, (B) in MtoM consensus clusterings and the corresponding inputs, (C) in StoM All consensus clusterings and the corresponding inputs.



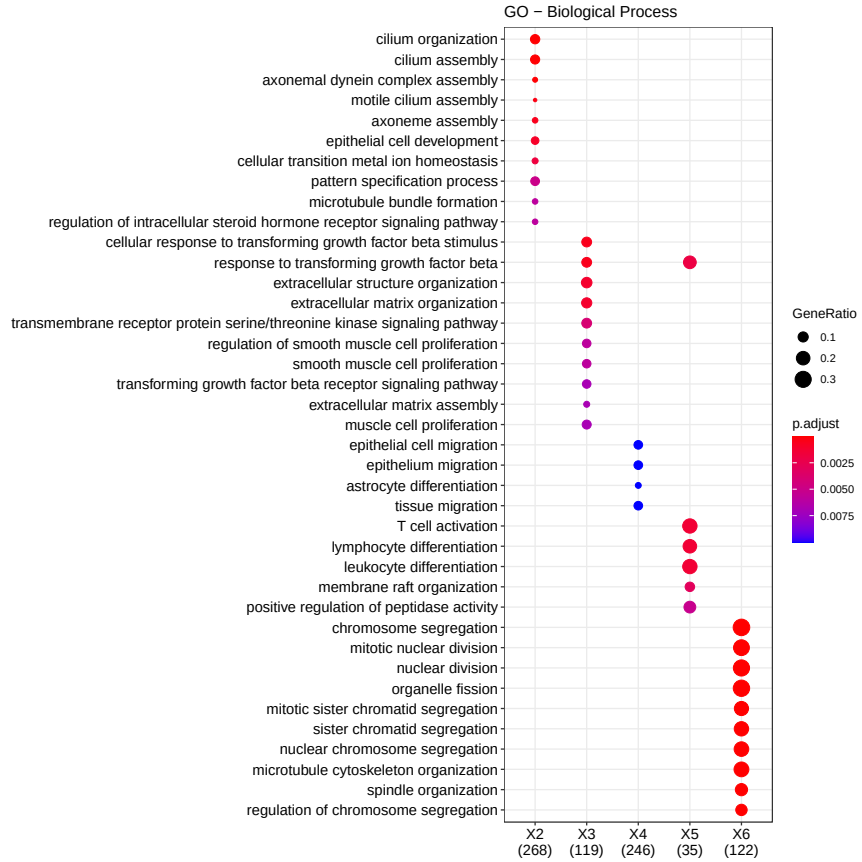
Supplementary Figure 2 ARI heatmaps revealing input and consensus clustering similarities for the MtoM scenario, upon the ten cancer types.



Supplementary Figure 3 Expression heatmap of top 1000 differentially expressed genes across breast-cancer ClustOmics StoM consensus clusters. Genes are displayed in rows and clustered according to their expression profile. Patients are displayed in columns and are ordered according to their consensus cluster attribution. PAM50 labels are given as supplementary annotations for patients.



Supplementary Figure 4 GO (Biological Process) enrichment analysis of 5 gene clusters found differentially expressed across breast-cancer StoM consensus clusters. P-values have been adjusted using the Benjamini-Hochberg correction and filtered with a 0.05 cutoff.



3.4 Discussion

Omics data integration is a challenging task, requiring complex methodological developments to enable the simultaneous analysis of multiple types of interactions. While existing integration strategies can already manage the integration of some types of omics data, the addition of novel omics or non-omics data types may be difficult to implement, notably for algorithms designed to handle specific types of omics data (e.g., using correlation to characterize the interactions between copy number and expression). In this context, consensus clustering presents itself as an interesting alternative. Indeed, the integration of new types of data is straight-forward as clusterings can be easily generated from any type of data. Moreover, this type of strategy allows the use of specialized clustering tools for each considered dataset, as well as taking advantage of the strengths of different clustering algorithms while smoothing their weaknesses. ClustOmics, the tool developed during this thesis for the integration of omics data by consensus clustering, has been applied on multi-omics cancer datasets for the detection of cancer subtypes, thus clustering cohorts of patients. The same methodology could be applied for clustering different objects (for example genes) but one should keep in mind that the search for a consensus should be made in a context where homogeneous patterns (or at least consistent patterns between omics datasets) are expected. This does not mean that only vertical integration (same samples across datasets) can benefit from this type of strategy. Indeed, horizontal integration via consensus clustering would be possible as long as the samples considered through

the analyses present a certain homogeneity: for example, a consensus clustering of genes could be produced by integrating micro-array and RNA-seq data produced independently, but under similar experimental conditions (e.g., for the same disease and in the same tissue). However, when aiming at characterizing the heterogeneity of molecular patterns occurring under various experimental conditions (e.g., various diseases and/or tissues), consensus clustering is no longer appropriate, since it will search for homogeneous patterns in the data. Thus, other types of methods must be introduced for the integration of omics data in a context where the heterogeneity of the data is the pattern of interest on which to base predictions.

Chapter 4

Network-based approach for multi-group differential co-expression analysis: application to the mouse model of Alzheimer's Disease

4.1 Context

With the increasing availability of datasets acquired world-wide, data measured in different experimental contexts have been compared and confronted using meta-analysis strategies, allowing the creation of different atlases [77, 78, 79] and repertoires [80, 81] describing the organization and molecular characteristics associated to different diseases, tissues, organisms, etc. This chapter focuses on the issue of data integration in a meta-analysis context, and aim to describe the molecular features observed across different phenotypes associated with Alzheimer's disease.

4.1.1 Alzheimer's disease and its 5xFAD mouse model

Alzheimer's disease (AD) is a neurodegenerative disease, the leading cause of dementia in humans worldwide, and is characterized by cognitive impairments and neurodegeneration associated with neuron loss and synaptic alterations. AD is a multi-factorial disease and many genetic and non-genetic risk factors have been proposed. In particular, 3 genes (APP, PSEN1 and PSEN2) are regarded as diagnostic biomarkers, as they are known to carry mutations associated with AD and to be involved in the metabolism of amyloid-β (Aβ). It is the aggregation of this molecule in neuritic plaques in brains affected by AD that is considered characteristic of the disease. Furthermore, the $\epsilon 4$ allele of the APOE gene is associated with an increased risk of an early-onset AD, which could be explained by an inhibition of Aβ degradation and a promotion of their aggregation [82]. This allele could also be linked to abnormal and pathological phosphorylation of the tau protein, causing the aggregation of neurofibrillary tangles in neuronal cells and contribute to other pathological dysregulations, such as disruption of microglial homeostasis, alteration of synaptic integrity and plasticity, or metabolic dysfunctions [82]. Indeed, AD seems associated with metabolic disorders, type 2 diabetes and metabolic syndrome being known risk factors of AD [83].

Although several mechanisms have been proposed to explain the onset and development of AD, they are not entirely satisfactory and some are questioned by the scientific community. Recently, the *Aβ hypothesis*, which suggested that AD was caused by Aβ accumulation, has been challenged based on failures of *Aβ hypothesis*-based clinical

trials and the discovery of the presence of amyloid plaques in more than 40% of cognitively healthy elderly persons over 80 years of age [84].

Moreover, recent studies suggested that AD may be an autoimmune disease, as bacterial and viral infections in the brain have been associated with an increased risk of disease development. In this new model of the disease, the formation of beta amyloid plaques could be accelerated by overexposure to microbial infections [84].

The role of microglial cells in AD is also far from being elucidated, and their beneficial and/or detrimental role in AD is still debated [85]. Recently, an evaluation of the gene expression profiles of Aβ plaque-associated and plaque-distant microglia populations revealed several transcriptomics perturbations causing opposite effects on the surrounding cells and further supported the importance of characterizing microglia heterogeneity for the elucidation of their role in the disease [86].

In brief, while some of the effects and risk factors of AD have been well characterized, the overall pathological mechanisms are not yet elucidated and are still a matter of debate in the scientific community. Unraveling the molecular mechanisms of AD is a complex task, all the more so with the interaction of genetic and non-genetic factors causing or promoting the development of the disease. In this context, the development of animal models that encapsulate the pathophysiological processes of the disease facilitates the deciphering of these mechanisms. It allows collecting data for large cohorts, while limiting the factors of variation between samples. Notably, mice models are particularly useful, and several such models of AD have been developed by including mutations or by knockdown of AD-related genes [87].

None of these models are ideal, nevertheless some of them are more sophisticated than others. Notably, the 5xFAD model of AD is interesting because it develops intense and early amyloid pathology, while recapitulating many of the symptoms observed in AD, including cognitive impairment, synaptic loss, neurodegeneration and gliosis [88]. This mouse model of early-onset AD expresses human APP transgene with three mutations and human PSEN1 transgene with two mutations, and was developed with the aim of facilitating the transfer of findings from mouse models to human clinical trials [89].

Several studies have sought to extensively characterize the 5xFAD phenotype, often with a focus on their transcriptomics profiles using differential expression analysis and co-expression analysis [89, 90, 91, 92, 93].

4.1.2 Co-expression and differential co-expression

One key analysis to gain insights from transcripts data is Differential Expression (DE) analysis that seeks to identify genes whose expression changes between two or more sample-groups. DE analysis is often, if not always, performed in transcriptomics studies. However, it only captures abrupt changes in gene expression, and does not account for possible interactions between transcripts. To meet these needs Co-Expression (CE) analysis has been developed, which looks for patterns of association between transcripts and infers gene networks. CE analysis aims at identifying co-expressed genes, i.e., genes that show similar expression profiles for a set of samples. From the computation of expression profiles pairwise similarities, a gene co-expression network is built, in

which nodes represent the genes and weighted edges depict the strength of the gene associations. Thereafter co-expression networks are usually clustered to identify gene modules, i.e., groups of co-expressed genes, that are interpreted as groups of co-regulated genes. Note that if gene pairwise association is most often measured using correlation, other strategies exist [94].

Co-expression analysis is widely used as it may bring to light interesting processes:

(i) genes belonging to the same co-expression module often share a biological function. Based on the principle of "guilt-by-association" [95], this allows identifying new genes involved in a biological process of interest, or alternatively, predicting the function of a gene of interest;

(ii) the expression profile (*eigengene*) of a co-expressed gene module can be correlated to a trait of interest, and thus highlight potential interactions between gene co-expression and phenotype [96];

(iii) the detection of relationships between co-expressed gene modules may reveal a higher-order organization of the transcriptome [97].

However, co-expression analysis does not directly allow for comparative analysis of gene associations in different phenotypes (e.g., disease/control datasets). Therefore, when performing CE analysis on multi-group transcriptomics datasets, co-expressions are often computed either on the full dataset without distinction of experimental sample-groups, or on a subset of the samples. For instance, in [90], co-expression is computed on a transcriptomics dataset gathering 5xFAD

samples and their wild-type (WT) counterparts in various brain tissues and time-points. Expression profiles of each discovered co-expressed module are then correlated to the metadata of the experiment (time-point, strain, tissue and sex) and other clinical traits of interest (Aβ plaques counts and sizes, microglial counts, etc.) to identify co-expressed modules that appear to be associated with the 5xFAD model and the phenotypic traits induced by this model.

As correlation is a relatively noisy measure of gene association, it needs to be calculated on a large set of samples as for it to be robust (some co-expression analysis tools recommend a minimum of 20 samples [98]). Thus, when the experimental groups in a dataset are small, measuring correlation for each sample-group becomes tricky. Nonetheless, whenever possible, correlation should be measured for each sample-group independently, as measuring the correlation on heterogeneous groups of samples can lead to spurious predictions. This phenomenon is called the Simpson's paradox (see description in Figure 4.1) and it refers to a phenomenon in which a trend observed from the union of several sample-groups together is reversed or disappears when the groups of samples are distinguished, or vice versa. Two examples are given in Figure 4.1: (A) a correlation is observed when considering sample-groups 1 and 2 together, yet when considering samples from group 1 and group 2 separately, the correlation trend disappears; (B) the opposite effect is observed, with correlation trends witnessed from each sample-group taken individually, while they disappear when considering the full set of samples.

This phenomenon was already described when comparing transcript

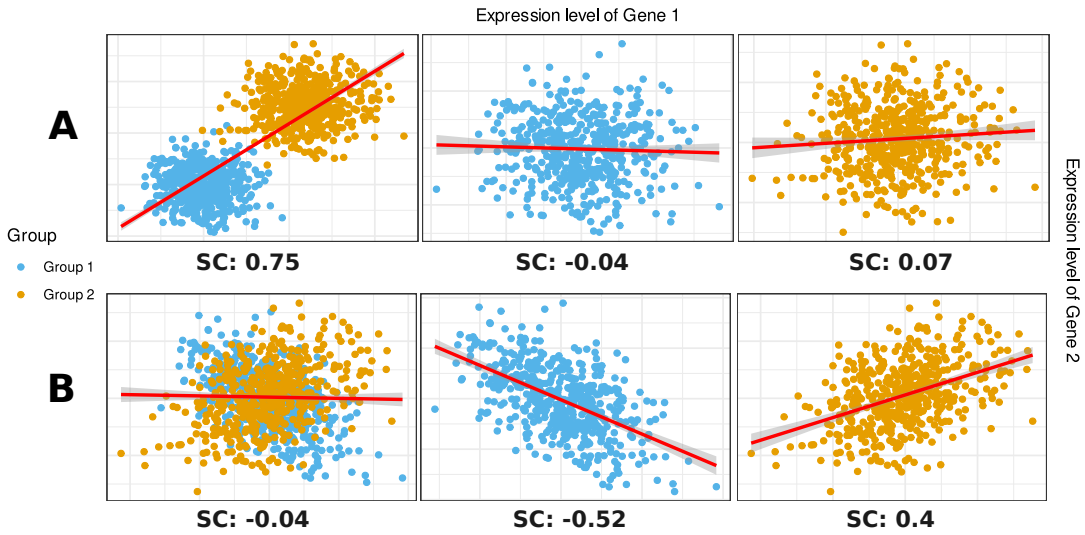


Figure 4.1: Illustration of the Simpson's paradox in two-groups transcriptomics data. Spearman's correlation coefficients (SC) are given under each correlation plot. **A:** Co-expression observed over all samples ($SC = 0.75$), but not within experimental groups ($SC = -0.04$ in Group 1, 0.07 in Group 2) **B:** No co-expression signal over all samples ($SC = -0.04$), while co-expression patterns are observed within each experimental groups ($SC = -0.52$ in Group 1, 0.4 in Group 2)

and protein abundance data [99] or when comparing transcripts expression across various sample-groups [100]. To our knowledge, apart from a few mentions of this phenomenon in co-expression analysis related literature [101, 102], the question of the existence and of the frequency of this paradox in co-expression data has not been addressed. In our opinion, the possibility of Simpson's paradox occurring in multi-group transcriptomics co-expression data should not be ignored, and, when possible, co-expression analysis should be performed on homogeneous sample-groups.

Besides avoiding Simpson's paradox, independent analysis of co-expression patterns in different groups of samples and their comparative evaluation should be highly informative and allow the detection of several differential co-expression patterns. In this regard, several strategies meant to perform Differential Co-expression (DC) analysis

have been developed.

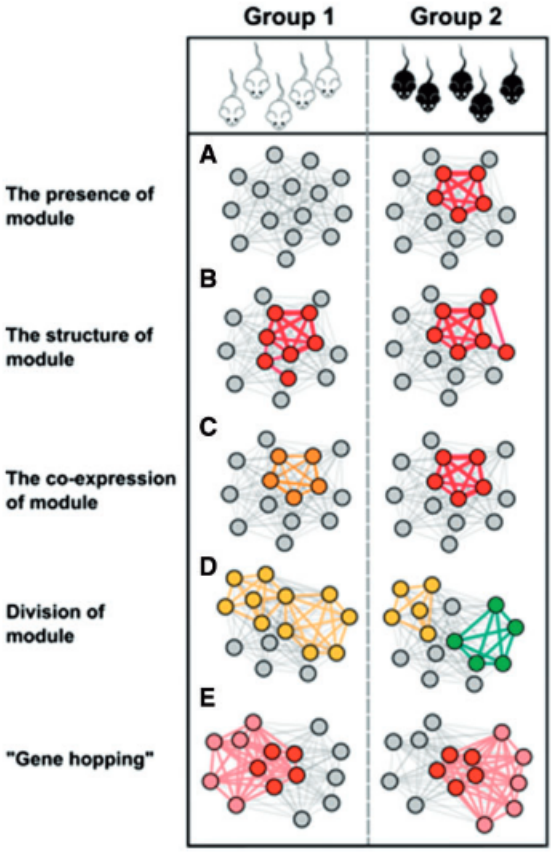


Figure 4.2: Several examples of differential co-expression patterns:
Several DC patterns can occur: **(A)** the presence of a co-expressed module in one sample-group but not in the other; **(B)** some changes in the structure of a co-expressed module; **(C)** a change in the co-expression strength of a module (gain or loss in correlation); **(D)** a co-expressed module that is split into several co-expressed modules in the other sample-group; **(E)** genes "hopping" from one co-expression module to another. (Figure adapted from [103].)

In control/disease experimental designs for example, comparing co-expression patterns in diseased versus control samples may highlight key differential association patterns that are indicators of a deregulation in gene co-expression related to the pathological phenotype. As illustrated in Figure 4.2 (adapted from a review of the literature on CE and DC analysis [103]), several DC patterns can be identified.

DC analysis has been applied in various cancer research studies. In [104], co-expression networks are built and compared between healthy

control and diseased samples from 5 cancer types. The authors observed that a loss of connectivity in the co-expression network was a common topological trait of cancer networks, and proposed novel cancer biomarkers. [105] highlighted a novel differentially co-expressed gene module, specifically co-regulated in ovarian cancer compared to healthy samples and to nine other cancer types. Moreover, in [106], DC analysis was used to uncover a novel critical signalling pathway altered in cancer, while in [107] it allowed to identify new candidate cancer biomarkers. Finally, DC analysis was used for characterizing tissue or timing specificity of co-expression patterns [108, 109]. While most studies focused on the analysis of mRNA expression data, DC analysis has also been applied to unravel mi-RNA [110] and protein[111] differential co-expression.

In neuroscience research, DC analysis was used to identify disease-specific co-expression patterns [112], to uncover disease-specific biomarkers [113] or disease-specific co-expression modules [114].

4.2 State of the art on differential co-expression analysis

Several strategies have been developed for differential co-expression analysis, some measuring DC for known gene sets, while others performing *de novo* DC detection. The strategies also differ according to their object of study: some consider and compare co-expression patterns, while others directly consider the co-expression difference. In this section, the major DC strategies are described.

4.2.1 Targeted approaches

Targeted approaches consider predefined gene sets, for instance gene sets from known pathways, or from co-expression modules inferred from analysing co-expression patterns in one sample-group. Several metrics have been developed to evaluate DC between various sample-groups, for a predefined gene set:

- the Centered Concordance Index (CCI) evaluates the co-expression concordance of a co-expressed gene module in a sample-group [115]. Measured for various sample-groups, CCI values can be compared to identify sample-group specific co-expression;
- Gene Sets Net Correlations Analysis (GSNCA) compares weight factors assigned to genes of a gene set in various sample-groups [116];
- Gene Set Co-expression Analysis (GSCA) approach introduces a novel metric, the Dispersion index, that can be used to characterize DC across two sample-groups, and uses samples permutations between experimental sample-groups to evaluate the strength of the DC [117].

4.2.2 *De novo* approaches

Non-targeted approaches aim at identifying, *de novo*, DC patterns from a set of transcriptomics data. Several approaches focus on the comparison of co-expression networks, while other directly target differential co-expression.

4.2.2.1 Co-expression based

Several co-expression based strategies first build co-expression networks and identify co-expressed gene modules for each sample-group. These group specific modules are then tested for co-expression in other sample-groups, with metrics such as those presented above for targeted approaches (Section 4.2.1). For instance, in CoXpress [118], co-expressed gene modules are inferred independently in control and disease sample-groups, and modules with significant average co-expression in one experimental group with respect to the other are considered differentially co-expressed. In [119], a similar strategy is applied to identify DC modules in AD and healthy samples, using the CCI metric to compare modules obtained for each sample-group. These strategies are mainly able to identify DC patterns A and C in Figure 4.2.

Weighted Gene Co-expression Network Analysis (WGCNA) [96] is the most popular tool for CE analysis. It also proposes ways to evaluate DC by providing a module preservation metric and consensus co-expression networking. In [120], DC patterns between healthy persons and mild and severe COVID-19 cases are being sought. The authors use WGCNA to compute a reference co-expression network from healthy samples, and then test whether the co-expressed modules are preserved in diseased samples.

In [121], the connectivity of co-expression networks from lean and obese mice are directly compared to identify differentially connected genes, considered as body weight-related genes, which corresponds to a search for changes in structure of the co-expression networks, as depicted in Figure 4.2-B.

In [122], a consensus network is computed from white and grey matter brain tissues sample-groups in multiple sclerosis, in order to detect similarities and differences in co-expression patterns of these tissues.

However, one could argue that measuring preservation (or consensus) across co-expression networks is more suitable for identifying shared patterns of co-expression than for detecting differential co-expression patterns. Indeed, preservation and consensus metrics will specifically target similarities in co-expression networks, rather than differences.

4.2.2.2 Differential Co-expression based

In co-expression based strategies, DC is evaluated by comparing group-specific co-expression networks, or by comparing overall module co-expression from one sample-group to another. However, other strategies directly focus on differential co-expression, at the module, or gene-pair level.

Module-based DC strategies aim at discovering differentially co-expressed modules from multi-group transcriptomics data. For instance, in DiffCoEx [123], a matrix of correlation differences is computed from a set of correlation matrices obtained for various sample-groups. This matrix is further processed with a WGCNA-like analysis that directly identifies DC modules, i.e., using a Topological Overlap analysis on the soft-thresholded matrix of correlation differences. The robustness of the DC is then assessed for each module and between modules using the dispersion index presented in Section 4.2.1. This strategy allows DiffCoEx to detect three types of DC patterns in the data:

(i) intra-module DC (i.e., a module co-expressed in one condition but not the other, cf. patterns A and C in Figure 4.2);

(ii) inter-module DC (i.e., two genes modules that are co-expressed in one condition but not the other, cf. pattern D in Figure 4.2);

(iii) genes "hopping" from one module to another (cf. pattern E in Figure 4.2), which are clustered in a separate module.

Similarly, DICER [114] introduces the concept of meta-modules, i.e., pairs of gene sets with no intra-module DC while presenting disrupted inter-module co-expression (cf. pattern D in Figure 4.2). DICER can also detect modules with impaired intra-module DC (cf. pattern A and C in Figure 4.2).

By design, such differential co-expression based methods are able to detect more types of DC patterns in the data than co-expression-based ones. However some of the patterns remain undetectable, such as changes in the structure of a module (pattern B in Figure 4.2), since the DC is assessed at the module level rather than at the gene-pair level.

Network-based strategies aim at building and analysing Differential Co-expression Networks (DCNs). In DCNs, nodes represent genes and edges represent a DC between two genes. The robustness of the DC is directly measured at the gene-pair level, i.e., edges in the network. For example, in [124], the DCN is constructed by estimating the fold-change of the correlation between two co-expression links. Some strategies use the Fisher transformation of the correlation coefficients in order to perform a z-test of the correlation difference to build the DCN [125],

others use Bayesian frameworks to estimate the DC of gene pairs [126].

As the DCN topology can be directly mined in order to identify the set of patterns described in Figure 4.2, DCN-based methods are among the most flexible approaches. On the other hand, since each pair of genes has to be tested to estimate the quality of the DC, network-based DC strategies tend to be highly computationally intensive. Also, if the use of permutation-based strategies over samples from different experimental groups may improve the robustness of the DC computation, applying them in practice is not easy as it requires an intense computational effort. In [127], Bhuva et al. review and benchmark different DCN-based strategies and conclude that z-score based methods perform well, while scaling up to real transcriptomics datasets.

4.2.3 Differential Co-expression analysis for more than two groups

When considering datasets with multiple case/control sample-groups (i.e., diseased and control samples at different time-points and/or for multiple tissues and/or for multiple diseases, etc.), investigating DC allows pruning context-specific co-expression patterns and keeping those patterns directly associated with the disease phenotype. Tissue-specific or time-point-specific co-expression patterns, for example, will not be considered unless they appear deregulated in the disease phenotype.

Moreover, integration of DC patterns observed between control and diseased samples obtained for various experimental contexts has the potential to highlight recurrent dysregulation patterns across the sample-groups, or specific to one (or to a subset) of the experimental conditions.

Only few strategies allow to directly perform multi-group DC analysis. The module-based DiffCoEx method proposes a differential association metric for more than two groups but does not allow to consider the experimental groups as being "paired" (disease/control), meaning that each experimental group is compared to all the others regardless of the experimental context. Similarly, when faced with more than two classes of samples, DICER proposes a DC metric based on "one vs. all" comparisons. However, this type of metric does not allow the integration of multiple paired datasets. Bi-clustering strategies (simultaneous clustering of features and samples) could help to identify DC modules from various case/control conditions, but do not consider sample labels, which are known in such experimental designs.

The comparison of several disease/control groups in various experimental conditions can be achieved using a late integration strategy to fuse DC patterns from each given comparison. For example, by integrating DC modules produced by DC-module-based methods on each paired dataset, or by integrating different DCNs produced by network-based methods. Still, late integration of DC modules from a module-based strategy is not straightforward. Indeed, consensus strategies are not adequate in this context, since they would only detect similarities across modules computed from the various case/control comparisons. Also, a naive integration strategy such as the intersection of the modules of each case/disease analysis, is not effective in practice when integrating more than two lists of modules. Inversely, integrating differential co-expression networks can be conceived by considering each DCN as a layer of a multi-layer DCN network. In such multi-layer networks,

each layer should be composed of the same set of nodes (genes), with potentially different links (differential co-expression). The multi-layer network can then be mined to identify DC gene modules persistent across layers, or specific to a subset of layers of the multi-layer DCN.

4.3 Contribution

In the following contribution for multi-group DC analysis, we propose a novel strategy for building and analyzing multi-layer DCNs. The strategy is based on the identification of link communities, rather than node communities in the network. We applied this strategy to detect recurrent and specific DC-patterns occurring in the hippocampus and the cortex of 5xFAD mice compared to their respective controls.

Network-based approach for multi-group differential co-expression analysis: application to the 5xFAD mouse model of Alzheimer's Disease

Galadriel Brière [1,2,*], Agnès Nadjar [3,4] Raluca Uricaru [1] and Patricia Thébault [1,*]

[1] CNRS, Bordeaux INP, LaBRI, UMR 5800, Univ. Bordeaux, 33400, Talence, France

[2] INRAE, Bordeaux INP, NutriNeuro, UMR 1286, Univ. Bordeaux, 33000, Bordeaux, France

[3] INSERM, Neurocentre Magendie, Physiopathologie de la Plasticité Neuronale, U1215, F-33000, Bordeaux, France

[4] Institut Universitaire de France (IUF)

Abstract

Differential co-expression analysis allows the detection of perturbations in gene co-expression in response to a disease, a stress or any experimental condition of interest, compared to a control state. Comparing co-expression dysregulations across a set of experimental designs could highlight key signatures of genes responses to various perturbations.

In this work, we implemented a strategy to detect groups of differentially co-expressed gene pairs, (i.e., differentially co-expressed links, in multi-layer differential co-expression networks) and applied it to detect and characterize differential co-expression patterns induced in the cortex and the hippocampus of 5xFAD mice, model of Alzheimer's disease.

The strategy is based on the idea of link clustering, an innovative way of defining communities in networks. Differentially co-expressed links are grouped according to their similarity in terms of topological proximity and co-occurrence score in the multi-layer differential co-expression network.

*To whom correspondence should be addressed: marie-galadriel.briere@u-bordeaux.fr ; patricia.thebault@u-bordeaux.fr

GLOSSARY **AD**: Alzheimer's Disease ; **BP**: Biological Process ; **CC**: Co-expressed Cluster ; **CNS**: Central Nervous System ; **DC**: Differential Co-expresion / Differentially Co-expressed ; **DCL**: Differentially Co-expressed Link ; **DCN**: Differential Co-expression Network ; **DE**: Differential Expression / Differentially Expressed ; **FC**: Fold Change ; **FDR**: False Discovery Rate ; **GEO**: Gene Expression Omnibus ; **GO**: Gene Ontology ; **LPS**: Lipopolysaccharide ; **MAMs**: Mitochondria-Associated endoplasmic reticulum Membranes ; **MCL**: Markov Clustering ; **QCH**: Query Composite Hypothesis ; **TF**: Transcription Factor ; **WT**: Wild Type

1 INTRODUCTION

Diseases are complex and dynamic systems whose state varies according to phenotypes, tissues, and time, with eventual pre-symptomatic, symptomatic, and recovery phases. To better characterize a disease, its heterogeneity with respect to time and location must therefore be considered.

Nowadays, tremendous amounts of biological data are measured across experiments and made available to researchers, among which omics data are particularly abundant. It is now possible to capture transcriptomics profiles under many experimental conditions (RNA-sequencing), and even at the scale of the cell (single-cell RNA-sequencing). Simultaneous analysis and comparison of transcriptome profiles from different experimental conditions can therefore contribute to the systematic characterization of complex diseases in a dynamic fashion, considering recurrent or specific disease-induced perturbations observed within and across these different experimental conditions.

One of the key approaches when analyzing transcriptomic data is differential expression (DE) analysis. Although it is essential to capture changes in transcripts' expression levels across two or more sample groups, it does not provide information on whether and how these transcripts interact with each other. Co-expression analysis on the other hand - often performed through transcripts pairwise correlation analysis - is used to detect groups of genes whose expression varies in a similar way across a group of samples. These genes are considered co-regulated or co-activated. The construction and analysis of co-expression networks allow the detection of modules of so-called co-expressed genes, which are more likely to be regulated by the same molecular mechanisms, and whose biological function is generally similar. However, unlike

DE analysis, co-expression analysis as generally performed is not adapted to highlight differences among groups of individuals. Indeed, when measured in a heterogeneous set of samples, e.g., case/control sample groups, correlation only enables the detection of genes whose expressions vary similarly in all samples, and thus, whose co-expression is independent of the disease state. Moreover, observing overall correlation with no distinction of the experimental sample groups might lead to incorrect predictions, as overall co-expression and group-specific co-expression could reveal reverse correlation trends, as described by the Simpson's paradox [1].

In recent years, differential co-expression analysis has therefore gained attention. Differential co-expression (DC) analysis compares gene co-expression in different sample groups to detect genes whose co-expression is dissimilar between experimental groups. For example, when comparing diseased and healthy samples, DC analysis may reveal potential regulators that initiate or participate in the pathogenesis [2, 3]. In addition to providing useful insights on gene interactions, DC strategies can help identifying novel actors of pathogenesis compared to DE analysis, as genes that are not significantly under- or over-expressed may still have significant impact through DC with others [4].

DC analysis can be performed using two main strategies: module- or network-driven approaches. Module-based approaches seek to identify groups of differentially co-expressed genes, exhibiting an overall gain or loss of correlation within and across modules [4, 5]. Network-based approaches, on the other hand, rely on the construction of differential co-expression networks (DCN) by directly searching for co-expression perturbations at the gene pair level (a link in the DCN relates differentially co-expressed gene pairs) in order to capture their conditional association [6, 7].

For the comparative analysis of several diseases, tissues and/or time-points, network-based methods are more adapted than module-based strategies. Indeed, the combinatorics engendered when comparing DC modules produced from multiple case/control comparisons is highly complex, whereas network-based methods facilitate the comparison of DCNs given that they have the same set of nodes (genes/transcripts). DCN computation is generally a two-step process. First, the difference in co-expression between genes is estimated using a conditional association metric. Then, the robustness of the conditional association is estimated to filter out non-significant DC pairs. Several strategies exist for this task and have been reviewed and benchmarked else-

where [3, 8, 9]. Many of these strategies rely on computing a z-score of the correlation difference between gene pairs from two sample groups to estimate the conditional association between genes. Other association models are also used, including F-statistics, generalized linear models, or empirical Bayesian approaches. Statistical tests applied to assess DC score significance include the z-test (for z-score based methods), permutation and modulation tests. However, due to the large number of gene pairs in an expression dataset, DC scoring methods based on repeated measures such as permutations and test modulations are highly computationally intensive.

In this work we introduce a new strategy, based on a multi-layer Differential Co-expression Network (DCN), for the *de novo* detection of recurrent and specific DC patterns in multiple groups, in an effort to systematically characterize co-expression perturbations induced by disease in diverse experimental contexts. A layer in such multi-layer DCN corresponds to the co-expression perturbations observed for a given case/control comparison.

We apply this multi-group DC strategy to detect and describe DC patterns induced in the 5xFAD model of *familial Alzheimer's Disease* (AD), in two cerebral structures: the cortex and the hippocampus. 5xFAD mice express human APP and PSEN1 transgenes and develop several AD-related pathologies, including severe amyloid pathology, astrogliosis and microgliosis, synaptic degeneration and neuron loss. While the phenotypic effects induced in the 5xFAD model are well described, the mechanisms involved in the pathogenesis are unclear. In the present work, we aimed at identifying the critical regulatory disruptions and the key actors in these dysregulations.

2 MATERIALS AND METHODS

2.1 Dataset

In this work we used the RNA-sequencing dataset GSE168137 from the GEO database [10]. This dataset was generated by *Forner et al.* [11] to investigate the transcriptional characteristics of the 5xFAD (C57BL/6 background) Alzheimer model across the lifespan and for two brain regions: hippocampus and cortex. Bulk-RNA-sequencing was performed on 192 5xFAD and corresponding wild-type (WT) C57BL/6 samples at four time-points (4, 8, 12 and 18 months-old) for both tissues. For each brain structure,

about 50 samples from each genotype were sequenced, and for each genotype-cerebral structure combination, 10 samples were mesured for time-points 4, 8 and 12 months-old, and 20 samples for time-point 18 months-old.

Raw data was filtered out to remove lowly expressed genes (normalized counts >= 10 in all samples), keeping about 13000 genes. We further filtered those, keeping only genes exhibiting high variance: top 5676 highest variance genes, including *Psen1* (5676th top variance gene) and *App* (67th top variance gene), the two human transgenes expressed in 5xFAD mice. Count data for these top variance genes was further normalized using the variance stabilizing transformation from the DESeq2 R/Bioconductor package [12].

2.2 Differential co-expression (DC) computation

Correlation is a noisy measure of gene co-association. The fewer the samples in the co-expression computation within an experimental group, the less reliable this measure is. Thus, we preferred to privilege the quality of the measured correlation to the diversity of the experimental settings brought in this experiment by the different time-points measured. Thus, correlations were measured for the cortex and hippocampus of WT and 5xFAD mice considering all available time-points (which represents about 50 samples per structure and genotype) rather than with distinction to the time-points.

Correlations in 5xFAD transgenic mice and C57BL/6 wild-type mice were computed independently in the hippocampus and the cortex samples (see **Figure 1-A: Condition-wise DC-scoring**). This allowed us to test whether a DCL was observed when:

(i) comparing C57BL/6 and 5xFAD correlations in the hippocampus;

(ii) comparing C57BL/6 and 5xFAD correlations in the cortex.

To evaluate genes conditional associations 5xFAD and C57BL/6 sample groups, it is necessary to apply an appropriate transformation on the correlation coefficients, in order to correct the skewness of the distribution observed for values close to ± 1. We apply the Fisher transformation of correlation coefficient on Spearman's correlation coefficients to approximate a normal distribution in each sample group, and in order to compute z-scores of correlation difference for each gene pair across samples groups, following the procedure described in [6]. P-values are then derived from the differential co-expression z-scores to assess the significant of the z-test in each pairwise

5xFAD/WT comparison. Thus, two sets of p-values were derived in this study, one for each cerebral structure.

To identify pairs of differentially co-expressed genes within and across tissues, we applied a p-value integration strategy based on composite-hypothesis testing (see **Figure 1-A: DC-scores integration**), by using the QCH (Query Composite Hypothesis) R package [13]. QCH uses mixture models to classify gene-pairs into different classes and provide a control for Type-I errors. We considered 4 classes to which each gene-pair investigated could belong to:

- C_1 = DC in neither of the cerebral structures,
- C_2 = DC only in the hippocampus,
- C_3 = DC only in the cortex,
- C_4 = DC in both tissues.

The False Discovery Rate (FDR) was controlled using $\alpha = 1e-4$ and significant pairs from classes C_2, C_3 and C_4 were selected.

Using Fisher's transformation of the correlation coefficients to correct the skewness of the correlation coefficients distribution for coefficients close to ± 1, the closer the original coefficients are to ± 1, the more extreme the Fisher-transformed values, and thus the significance of the z-test. In some cases, this can lead to observing correlation differences that are statistically significant but not biologically interpretable. Thus, we apply an additional filter on gene pairs to ensure that their difference in correlation on the untransformed coefficients reaches at least ± 0.4, which we can interpret as an activation or inhibition of co-regulation in one sample group with respect to the other.

2.3 Multi-layer Differential Co-expression Network (DCN) construction

From the set of DC gene pairs identified as described above, we computed a multi-layer DCN with genes corresponding to nodes and pairs of DC genes representing links (DCL) in the network (see **Figure 1-A: DC-scores integration**). Note that the computation of the co-expression difference can be done in an unsigned or in a signed fashion, depending on whether one wants to identify dysregulated biological processes in a broad way, or specifically identify an activated or a deactivated process.

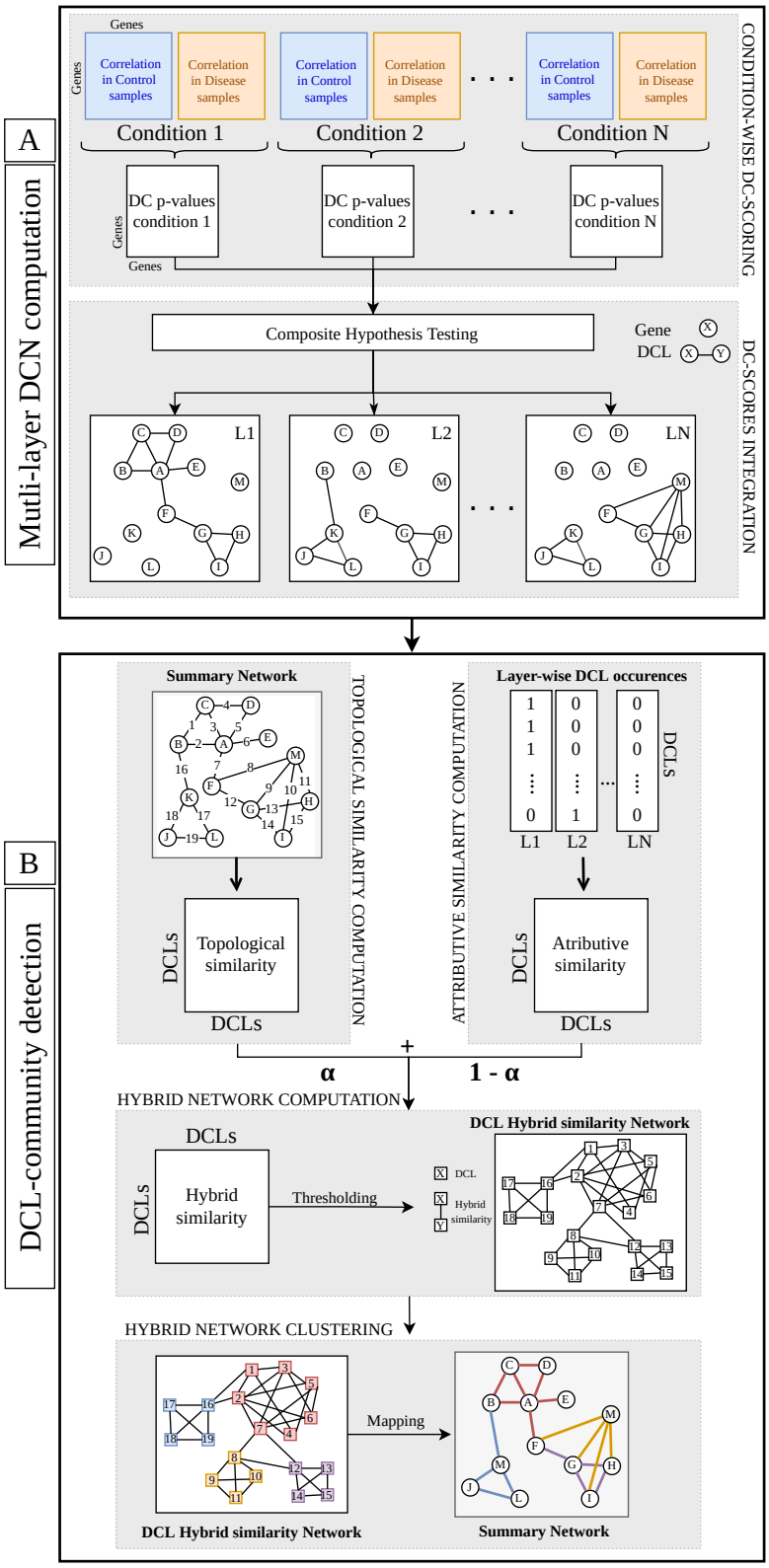


Figure 1: Overview of the approach for detecting DCL-communities from multiple expression datasets. In the set of networks presented in this figure, the nodes represented by circles correspond to genes, and the nodes represented by squares correspond to DCLs. (A) Multi-layer DCN computation: condition-wise DC-scoring for each control/disease conditions comparison and DC-scores integration using Composite Hypothesis Testing. (B) DCL-community detection in the multi-layer DCN: computation of topological and attributive DCLs similarities, hybrid DCL network computation and clustering.

The choice between a strategy or the other will impact the organisation of DCLs in the different layers of the DCN. Indeed, the signed approach will contain twice as many layers as the unsigned one, as the set of DCs observed in one sample group will be split in two based on the sign of the correlation (positive or negative).

Both strategies were tested below on the dataset described in Section 2.1. For the unsigned strategy, DCLs were organised in two layers: one gathering all DCLs observed in the hippocampus and the other in the cortex, regardless of the direction of the correlation.

When taking the sign of the DC into account, four layers were considered: two layers composed of those DCLs with a positive DC score (i.e., gain of correlation in 5xFAD samples) in the hippocampus and the cortex respectively, and two considering DCLs with a negative DC score (i.e., loss of correlation in 5xFAD samples) in the hippocampus and the cortex respectively.

2.4 Link clustering of the multi-layer DCN

Given a multi-layer DCN built as described above, we look for groups of genes whose co-expression seems to be disrupted in a consistent way from the point of view of the affected genes (i.e., high density of DCLs within the group of genes) and/or of the affected layers (i.e., groups occurring across one or more layers of the multi-layer DCN).

To this end, we implement an approach based on the identification of so-called *link communities*.

The concept of link communities was first introduced by *Ahn et al.* [14] in single-layer networks. Rather than identifying groups of strongly interconnected nodes (as in classical node clustering strategies), link clustering aims at identifying groups of links that have high topological similarity, i.e., that are connected to the same sets of nodes. In cases where communities are highly overlapping in a network (with more external connections than internal), classical node community detection strategies fail at detecting structure, while link-based communities have the advantage of capturing relationships between overlapping groups. In biological networks, as a gene may be involved in several biological processes and therefore belong to several overlapping communities, link clustering is indeed of high relevance. Moreover, in the context of multi-layer differential co-expression network analysis, identifying link communities

makes even more sense, as a gene may display highly different interactions among the layers of the network. Being able to capture these different interactions as distinct link communities rather than forcing a gene to be part of a unique community, may help identifying those genes that can switch interactions depending on the layer/experimental condition considered.

Link community clustering has been extended to multi-layer networks by *Salem et al.* who applied it to analyse multi-tissue co-expression networks [15]. Here we implement a similar strategy to explore multi-layer DCNs (see **Figure 1-B**). The link clustering is performed on a DCL-network (that we call hybrid similarity DCL-network below), in which an edge between two DCLs exhibits a hybrid measure of similarity computed as a weighted average between a topological similarity and an attributive similarity measures.

Topological similarity A single-layer summary network is computed from the multi-layer DCN, resuming the DCLs observed in at least one of the layers (see **Figure 1-B: Topological similarity computation**) . Note that DCLs observed in multiple layers are only present once in the summary network, and all links are unweighted. The topological similarity is computed for all pairs of DCLs in the summary network as follows.

Let us consider two DCLs in the summary network, e_{ik} as a DCL between gene i and gene k, and e_{jl}, as a DCL between gene j and gene l. The topological similarity between e_{ik} and e_{jl} is given by:

$$S_t(e_{ik}, e_{jl}) = \begin{cases} \dfrac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} & \text{if } k = l \\ \dfrac{|((N(i) \cup N(k)) \cap ((N(j) \cup N(l))|}{|((N(i) \cup N(k)) \cup ((N(j) \cup N(l))|} & \text{if } k \neq l \end{cases}$$

where $N(i)$ represents the set of neighbours of i in the summary network, including i itself.

The topological similarity is computed as the ratio of shared neighbors between the two DCLs and the total of neighboring genes. When two DCLs share a gene k, the neighborhood of the common gene is not taken into account, as neighboring nodes from k would automatically be shared between the two DCLs, thus artificially increasing their topological similarity. The topological similarity ranges from 0 (the two DCLs do not share any neighbors) to 1 (all neighbors of one of the two DCLs are

also neighbors of the other).

Attributive similarity In the summary network, two DCLs can be similar in terms of topological structure, but never co-occur in the same layer of the multi-layer DCN. To avoid such *chimeric* associations, an attributive similarity is introduced. Unlike the topological similarity, the attributive one is computed on the initial multi-layer DCN (see **Figure 1-B: Attributive similarity computation**), as the ratio between the number of layers in which the two DCLs coexist and the total number of layers L:

$$S_a(e_{ik}, e_{jl}) = \frac{|I(e_{ik}) \cap I(e_{jl})|}{L}$$

where $I(e_{ik})$ is the set of layers of the multi-layer DCN in which the DCL e_{ik} occurs.

The attributive similarity ranges from 0 (the two DCLs never co-occur in the set of layers) to 1 (the two DCLs always co-occur in the set of layers).

Hybrid similarity Finally, in order to ensure that two DCLs are similar both in terms of (i) topology in the summary network and (ii) co-occurrence in the input layers of the multi-layer DCN, we combine the two similarity measures to compute a hybrid similarity measure. For each pair of DCLs e_{ik} and e_{jl}:

$$S_h(e_{ik}, e_{jl}) = \alpha \times S_t(e_{ik}, e_{jl}) + (1 - \alpha) \times S_a(e_{ik}, e_{jl})$$

where α is a user-defined parameter (real number between 0 and 1) that corresponds to the weight given to the topological similarity, and $1 - \alpha$ the weight given to the attributive similarity.

When tuning the α parameter, one should consider the heterogeneity of the input layers and the density of the summary graph as they can respectively influence the distribution of attributive and topological similarities. Indeed, when the layers are highly heterogeneous and most DCLs occur in a single layer of the multi-layer DCN, the attributive similarity is skewed towards low similarities. Inversely, when the majority of DCLs are shared across layers, the attributive similarity is skewed towards high similarities. On the other hand, the distribution of topological similarities is skewed toward low values when the summary graph is very sparse (most DC-genes

will not share neighbors in the summary graph as they display few DCLs) or skewed towards high values when the summary graph is very dense (most DC-genes will share many neighbors in the summary graph since they display many DCLs). Therefore, one should tune α such that both weighted distributions approximately have the same mean, so that both metrics can participate in the signal expressed by the hybrid similarity.

Clustering the hybrid similarity DCL network A hybrid similarity DCL-network is built (with or without prior thresholding of hybdrid similarities), with DCLs as nodes (squares in **Figure 1-B: Hybrid network computation**), and edges being weighted with the hybrid similarity between their respective DCLs.

We cluster this DCL network using the Markov Clustering (MCL) algorithm [16] in order to produce groups of DCLs, *i.e.*, link communities, rather than groups of genes. It is straightforward to see that given that the clusters are computed at the edge level, a gene can be part of several DCL communities (see **Figure 1-B: Hybrid network clustering**).

The link (DCL) communities detected by clustering the hybrid similarity DCL-network are mapped on each layer of the initial multi-layer DCN. This allows computing the occurrence score of a DCL-community C_i in a layer L_j as the proportion of DCLs in the community C_i that occur in the layer L_j:

$$O(C_i, L_j) = \frac{|E(C_i) \cap E(L_j)|}{|E(C_i)|}$$

where $E(C_i)$ is the set of DCLs composing DCL-community C_i, and $E(L_j)$ the set of DCLs on layer L_j.

2.5 Gene Ontology Enrichment Analysis

In order to assess the biological relevance and the biological function of gene sets revealed by the DC procedure, we performed a Gene Ontology (GO) enrichment analysis. With the R/Bioconductor package clusterProfiler [17], we searched for enrichment in biological processes (GO-BP) with a 0.05 p-adjust cutoff and only level 10 terms in the ontology were retained.

3 RESULTS

To characterize and compare potential perturbation patterns induced in the 5xFAD mouse model of Alzheimer's disease, we investigated differential co-expression in the hippocampus and cortex of these mice, in comparison to their wild-type (WT) C57BL/6 counterparts. DCLs computed in each sample group were organised into multi-layer DCNs in both an unsigned and signed fashion in order to identify (i) broadly dysregulated processes induced by 5xFAD mutations, (ii) specifically activated and deactivated processes in 5xFAD mice compared to the WT samples.

3.1 DCL detection and multi-layer DCN construction

After computing z-scores of correlation difference between the hippocampus and the cortex of 5xFAD and WT samples, two sets of p-values were obtained, characterizing the significance of the correlation change in each tissue. DCLs were selected using a composite-hypothesis-based strategy with the QCH R package to directly identify the class of each gene pair (not DC, DC in hippocampus, DC in cortex or DC in hippocampus and cortex), followed by a filtering step to control both the FDR ($\alpha = 1e-4$) and the intensity of the correlation difference (keeping DCLs showing $abs(DC) >= 0.4$). See Section 2.2 above for complete details on the DC computation procedure.

In total, 3527 DCLs were identified, involving 1258 genes. Most of the DCLs (2466) occurred in both the cortex and hippocampus, while some of them were found only in one tissue (917 in the hippocampus and 144 in the cortex). All DCLs identified in both brain structures showed the same sign of correlation difference, which indicates that the dysregulations induced in the 5xFAD genotype operate in the same direction, independently from the cerebral structure. However, regardless of the sign of the DC, the intensity was higher in the hippocampus, suggesting that the impairments induced in the 5xFAD model are stronger in the hippocampus than in the cortex (see Figure 2).

Notably, App and Psen1 human-transgenes expressed in 5xFAD mice were not found differentially co-expressed with any other gene, and additional DE analysis also showed that neither was significantly differentially expressed. This suggests that the set of dysregulations observed in the data is not the result of an abrupt difference in expression or co-expression of Psen1 transgenes and/or App detectable over multiple

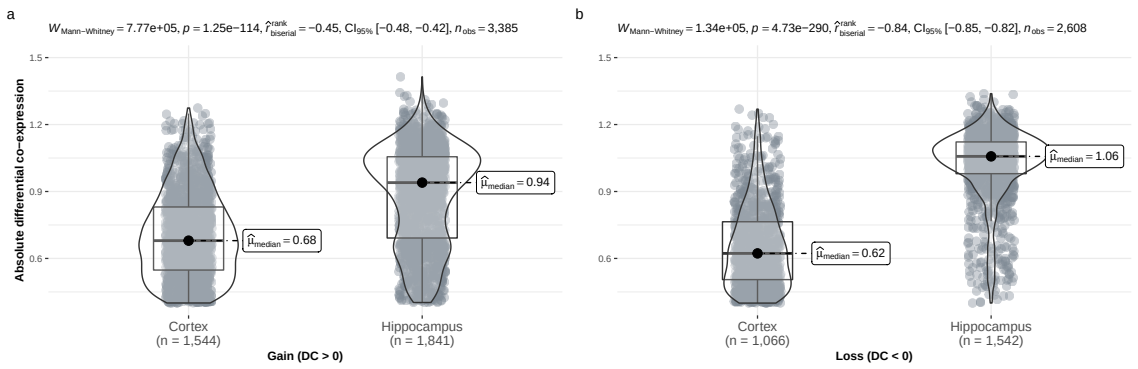


Figure 2: DCLs absolute differential correlation for (a) correlation gain in 5xFAD samples; (b) correlation loss in 5xFAD samples. Reported p-values were computed with the Wilcoxon-Mann-Whitney test, using the R packages *ggstatsplot* [18].

development stages in hippocampus and cortex, but of a more subtle and/or trans-omic mechanism and/or operating at a given time-point but not extending over time.

Further on, the set of DCLs was used to build two multi-level networks: one unsigned, composed of two layers (hippocampus and cortex, regardless of the DC direction) and one signed. For the signed version, the set of DCLs was split with respect to the sign of the correlation, thus generating four layers (each tissue-specific layer being split in two, respectively for DCLs gaining in correlation in 5xFAD samples or losing in correlation in 5xFAD samples).

3.2 Unsigned communities reveal dysregulated biological processes

After building the unsigned two-layers DCN, the summary network was used to compute the topological similarity for each pair of DCLs, while the pairwise attributive similarity was obtained from the per-layer DCL occurrences matrix. DCLs hybrid similarities were computed by combining both similarities using $\alpha = 0.7$, i.e., giving more weight to the topological similarity (see Section 2.4). Indeed, because the multi-layer DCN is composed of two layers only, DCLs attributives similarities can only take three values: 0 (DCLs never co-occurring in any of the 2 layers), 0.5 (DCLs co-occurring in one of the two layers) or 1 (DCLs co-occurring in both layers). As most of the DCLs occur in both layers, the attributive similarity distribution is skewed, with very few DCL-pairs never co-occurring in an input layer ($\approx 2\%$), most DCL-pairs co-occurring in one out of two layers ($\approx 49\%$) or co-occurring in both cortex

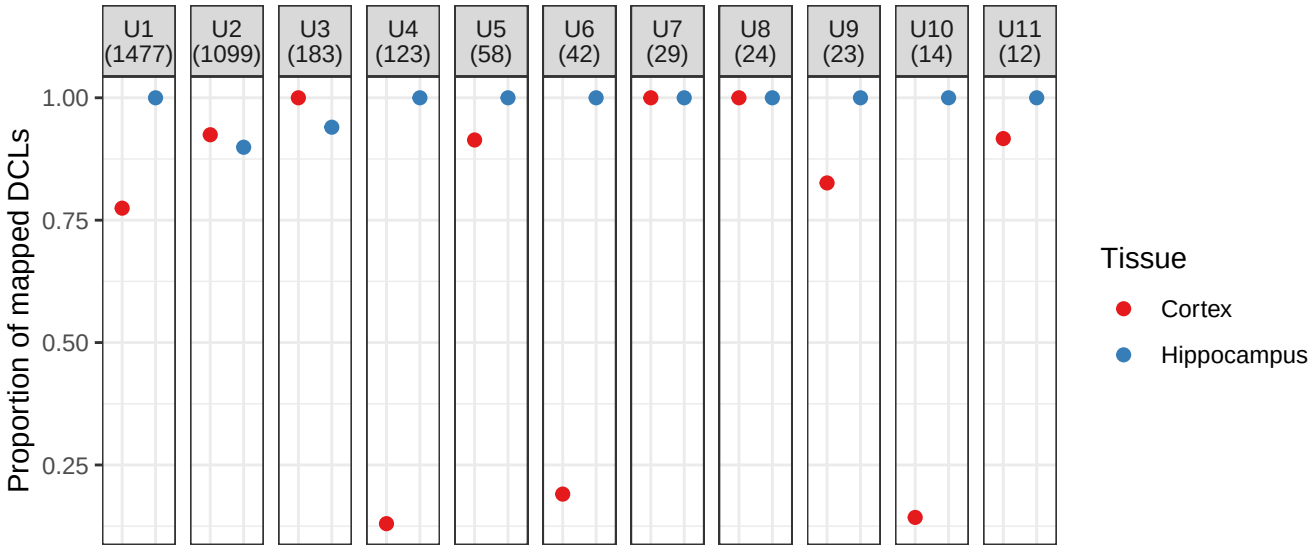


Figure 3: Layer-wise unsigned DCL-communities occurrence scores (and sizes).

and hippocampus ($\approx$ 49%). On the other hand, topological similarities may range from 0 to 1, but 75% of the observed topological similarities were below 0.25 .

Only DCL pairs with hybrid similarities above 0.5 were further kept and clustered using MCL with default parameters. When considering only DCL-communities with at least 10 DCLs, we obtained 11 such DCL-sets with sizes ranging from 12 to 1477 DCLs (corresponding to a number of genes ranging from 13 to respectively 844 genes).

The DCL-communities occurrence scores computed for the input layers showed that except for 3 of them (communities 4, 6 and 10) that occurred exclusively in the hippocampus, the majority of the communities were occurring in both the cortex and the hippocampus (a community is considered to be common to both structures if at least 75% of the DCLs from this community are present in both layers) (see Figure 3).

Unsigned DCL-communities 1 to 6 are represented in Figure 4, with genes (nodes) colored according to expression Fold Change (FC) and DCLs (edges) colored according to their Differential Co-expression (DC) in hippocampus and cortex.

A first general observation is that the computed DCL communities display different topological properties, below we thoroughly analyze the six largest among them:

Unsigned DCL-community 1 (U1) contains genes showing positive and negative DC in both cortical and hippocampal tissues with three hub genes, namely *S1pr3, Sbno2* and *Dhrs1*, showing relatively modest differentially expression (from 0.4 to 1.3 log2-FC). While Dehydrogenase/Reductase Member 1 (*Dhrs1*) was not reported directly to be involved in AD, it has been associated to Specific Language

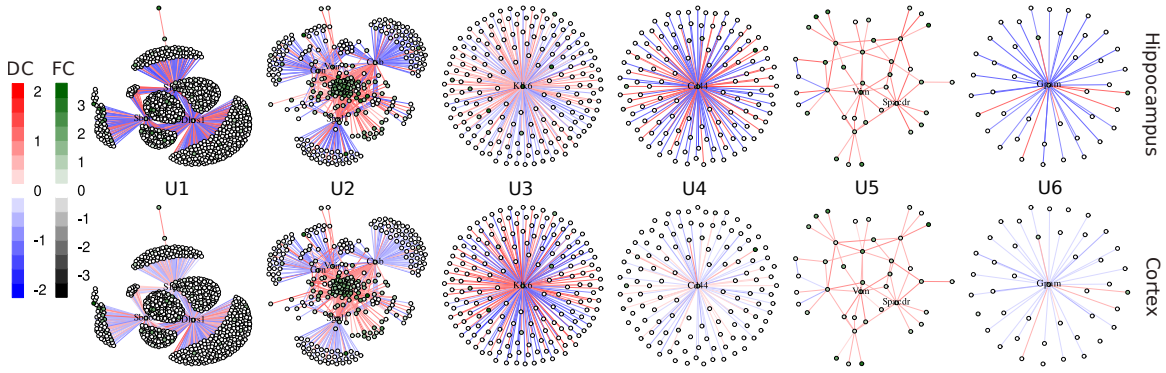


Figure 4: Representation of 6 largest unsigned DCL-communities. Nodes (corresponding to genes) are colored by tissue-specific expression Fold Change (FC) in 5xFAD samples compared to WT samples (greener for genes up-regulated in 5xFAD). Edges (DCLs) are colored by tissue-specific differential co-expression (DC) in 5xFAD samples compared to WT samples (redder for gain in co-expression in 5xFAD). The first row considers DC and FC in the hippocampus, while the second row considers DC and FC in the cortex.

Impairment, another brain disorder [19]. Moreover, Strawberry Notch Homolog 2 (*Sbno2*) is reported to be involved in macrophages activation and the inflammatory response in the Central Nervous System (CNS) [20]. Even more interestingly, *S1pr3* is a receptor for the lysosphingolipid sphingosine 1-phosphate (*S1P*). S1P has neuroprotecive properties and is regulated by the Apolipoprotein E (*Apoe*), whose $\epsilon 4$ allele is known to be a major risk factor for AD [21, 22]. It has been shown to inhibit the classical complement cascade by binding to activated complement component 1q (*C1q*) in $A\beta$ plaques [23]. While *Apoe* was not DC with *S1pr3, Sbno2* or *Dhrs1*, all 3 *C1q* chains (*C1qa, C1qb* and *C1qc*) as well as other *C1q*-related genes (*C1qtnf4* and *C1qbp*) showed positive DC with *S1pr3, Sbno2* and *Dhrs1*, and are part of DCL-community U1. GO analysis of genes involved in U1 showed significant enrichment for axonogenesis, synaptic transport and neuron and dendrite development, amongst other biological processes (see Figure 5).

Unsigned DCL-community 2 (U2) is composed of up-regulated genes in 5xFAD samples, while displaying dense and inter-connected DCLs. Four genes, namely *Grn, Sgpl1, Ctsb* and *Vsir* were found switching co-expression patterns and gaining in correlation with those densely DC and up-regulated genes, while themselves not being as significantly up-regulated (between 0.4 to 1.3 log2-FC). *Grn, Sgpl1, Ctsb* and *Vsir* are annotated in the Mammalian Phenotype ontology [24] as involved in abnormal and increased inflammatory response (MP:0001845 and MP:0001846). The overall

DCL-community was found enriched for GO terms related to leukocyte and myeloid cell differentiation as well as Lipopolysaccharide (LPS)-mediated signalling pathway (see Figure 5). The subset of densely DC and up-regulated genes included genes involved in autoimmune diseases and amyloid plaque formation, notably the well described AD-related genes *Trem2* and *Apoe*. All 3 *C1q* chains also occur in this DCL-community, in addition to being found in the first DCL-community.

Unsigned DCL-community 3 (U3), 4 (U4) and 6 (U6) each are composed of DCLs originating from a single gene, respectively *Klk6, Cd44* and *Gpam*. DCLs for U4 were observed in both the cortex and the hippocampus, while DCLs from U5 and U6 mainly occurred in the hippocampus (see Figure 3 and Figure 4). Kallikrein-related peptidase 6 *Klk6* has previously been described as contributing to vascular abnormalities in AD [25], *Cd44* antigen variants have been linked to $A\beta$-induced neuronal toxicity and AD [26] and Glycerol-3-phosphate acyltransferase 1 *Gpam*, while not directly described as involved in AD, is related to mitochondria-associated endoplasmic reticulum membranes (MAMs), known to regulate autophagy and to be associated to AD [27, 28].

Unsigned DCL-community 5 (U5) gathers about 40 genes, mostly displaying positive pairwise DC, both in the cortex and the hippocampus, including gliogenesis related genes such as Vimentin *Vim*, Glial fibrillary acidic protein *Gfap* and C-X3-C motif chemokine receptor 1 *Cx3cr1*, among others.

3.3 Signed communities highlight activated and deactivated biological processes

Unsigned DCL communities can be hard to interpret because they consider in a equal manner DCLs exhibiting gain and loss of co-expression in 5xFAD compared to WT samples. To, first, specifically identify DCL-communities co-expressed in 5xFAD but not in WT and, second, DCL-communities co-expressed in WT but not in 5xFAD on the other, we further conducted a signed analysis of the DCN (thus splitting the previously described two-layer DCN into four layers showing gain and loss of correlation for each tissue). Hybrid similarities were computed giving the same weight to topological and attributive similarities (i.e., $\alpha = 0.5$), and only similarities

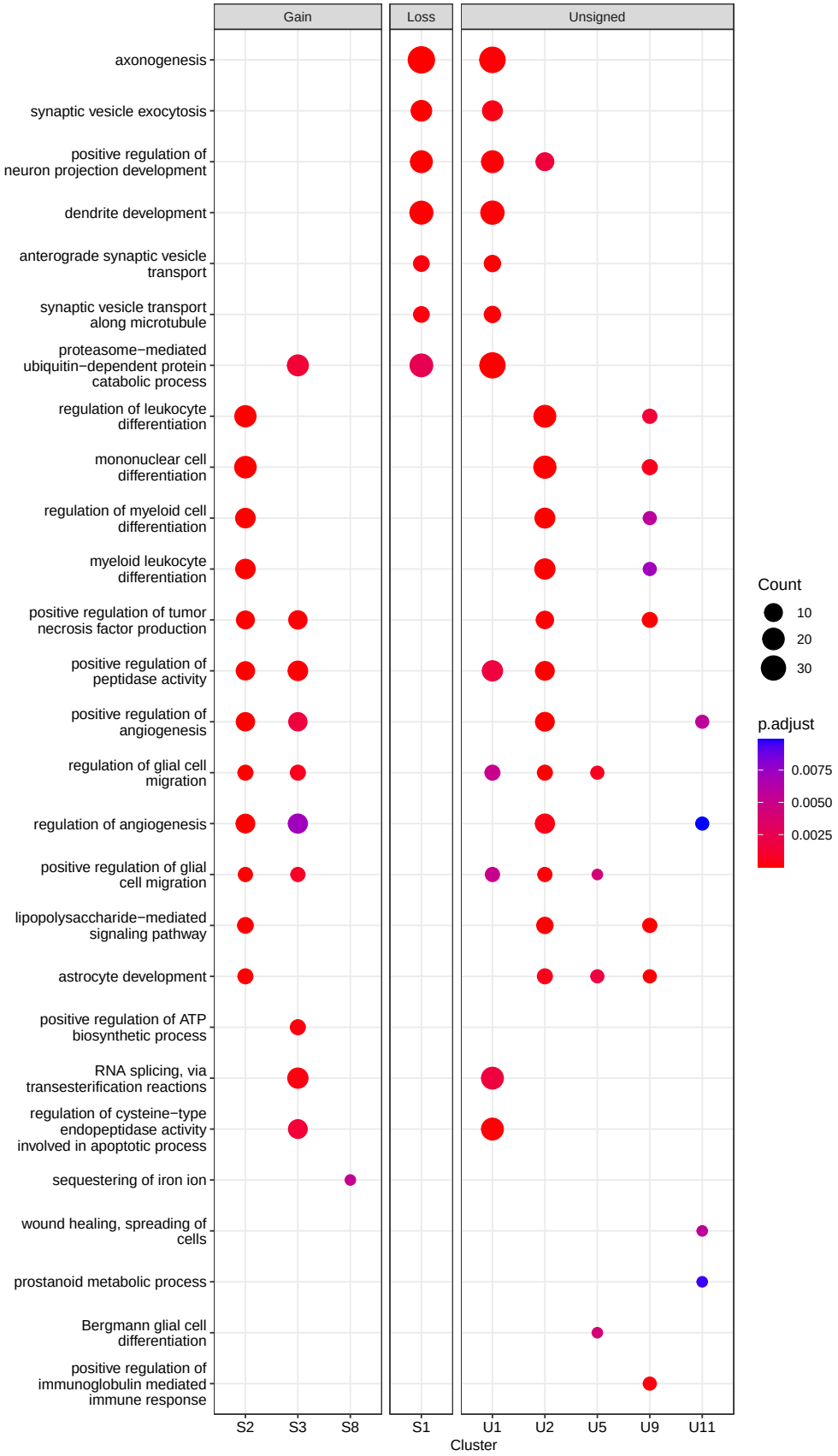


Figure 5: Level 10 GO:BP enrichment analysis in unsigned and signed DCL-communities.

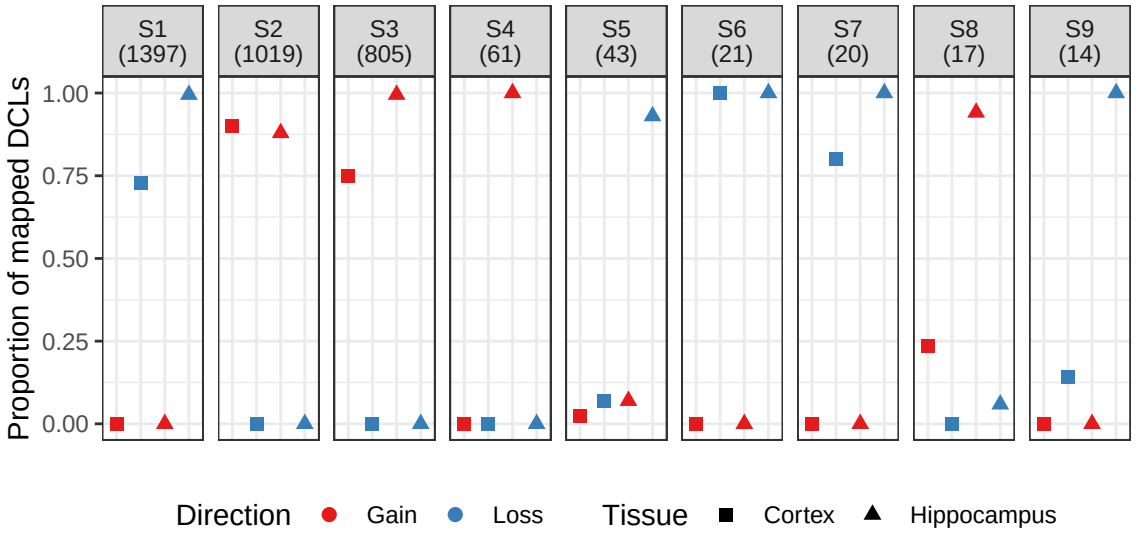


Figure 6: Layer-wise signed DCL-communities occurrence scores (and sizes).

above 0.3 were kept in the hybrid similarity DCL network. Markov clustering was run with default parameters and 9 DCL-communities with at least 10 DCLs were kept (sizes ranging from 15 to 1397 DCLs, and respectively 15 to 695 genes). Because the DCLs identified in both cerebral structures had the same direction (either gain in 5xFAD or loss in 5xFAD for both cerebral structures), DCL-communities detected as occurring in two layers were found on layers displaying the same sign (see Figure 6). Note that for other experimental designs, this is not necessarily expected (for instance when including a layer considering response to a drug or a treatment meant to restore normal co-expression).

Among the 9 DCL-communities, 5 were found in both the cortex and the hippocampus, and 4 occurred specifically in the hippocampus. Signed DCL-communities 1 to 6 are represented in Figure 7, with DCLs (edges) colored according to their Differential Co-expression (DC) in the hippocampus and the cortex. We also colored genes (nodes) according to their expression Fold Change (FC) to provide additional information on gene expression patterns observed when comparing 5xFAD and WT mice.

Signed DCL-community 1 (S1) shows impaired co-expression, with a loss of correlation in the 5xFAD genotype compared to WT mice. Genes involved in this community were found to be associated with axonogenesis and neuronal development (see Figure 5). This DCL-community is composed of two connected components:

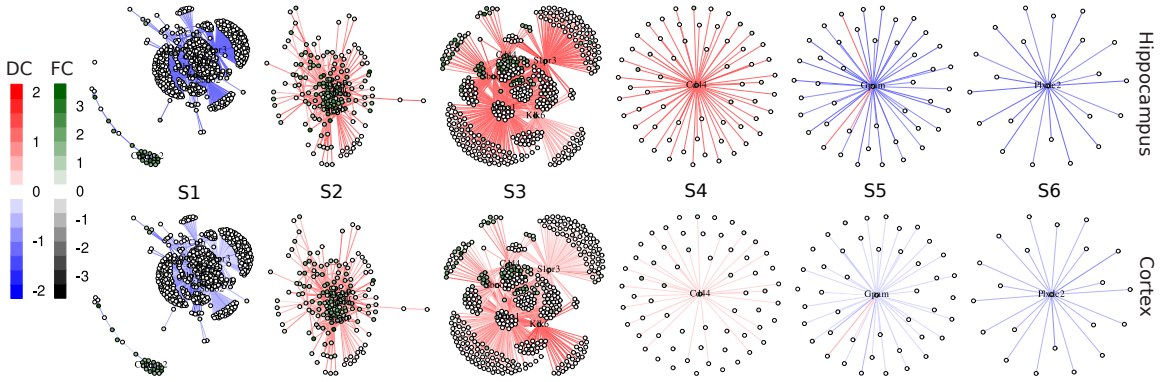


Figure 7: Representation of 6 largest signed DCL-communities. Nodes (corresponding to genes) are colored by tissue-specific expression Fold Change (FC) in 5xFAD samples compared to WT samples (greener for genes up-regulated in 5xFAD). Edges (DCLs) are colored by tissue-specific differential co-expression (DC) in 5xFAD samples compared to WT samples (redder for gain in co-expression in 5xFAD). The first row considers DC and FC in the hippocampus, while the second row considers DC and FC in the cortex.

the largest one includes *S1pr3, Sbno2* and *Dhrs1* as the main actors in dysregulation, while the second connected component is composed of DE genes and highlights Contactin-associated protein-like 2 (*Cntnap2*) as being dysregulated with most of those up-regulated genes, though not being itself particularly differentially expressed (-0.2 log2-FC in cortex and hippocampus). *Cntnap2* has already been reported as being down-regulated in AD [29]. Here we also find that it appears to lose in correlation with genes involved in leukocyte-mediated immunity, including *C1q*-chains.

Signed DCL-community 2 (S2) reveals a co-activation of genes involved in leukocyte and myeloid cell differentiation associated to the 5xFAD genotype. These genes were also up-regulated in 5xFAD samples compared to WT. Interestingly, the LPS-mediated signaling pathway was also found enriched in this community, indicating possible inflammation in the cortex and hippocampus of 5xFAD mice. Moreover, the genes composing this community are densely DC with each other and correspond to those found in the unsigned DCL-community U2 4.

In signed DCL-community 3 (S3) we observed an activation in the co-expression of genes involved in the positive regulation of angiogenesis, a biological process known to be impaired in AD [30]. Once again, *Dhrs1, S1pr3, Sbno2, Klk6* and *Cd44* genes were the most affected.

Signed DCL-communities 4 (S4), 5 (S5) and 6 (S6) present a similar structure, with S4 and S5 being specific to the hippocampus, while S6 appears in both cerebral structures. They are composed of DCLs originating from a single gene for each one of them, *Cd44, Gpam* (both of them having already been identified as hubs in unsigned DCL-communities) and *Plxcd2*, a mitogen for neural progenitors involved in proliferation and differentiation in the developing nervous system [31].

3.4 The role of Transcription Factors in DC communities

Transcription Factors (TF) play a key role in the regulation of gene expression. Using the AnimalTFDB3.0 database [32] we identified several TFs that might be implicated in differential co-expression, in signed and unsigned DCL-communities. In particular 4 of these TFs, namely *Irf8, Spacdr, Plek* and *Mafb*, retained our attention as they shared a high number of DCLs with other genes and were consistently clustered in the same DCL-communities (enriched with genes participating in inflammation) in both signed (S2 community) and unsigned strategies (U2 community).

Furthermore *Irf8, Spacdr, Plek* and *Mafb* show DC with, respectively, 21, 22, 28 and 81 targets, with which they gain correlation in 5xFAD samples compared to WT, in both cortex and hippocampus (Figure 8). While these TF were not found DC with every gene from the DCL-community U2, we show that their overall correlation significantly changes, as displayed in Figure 9, and especially in the hippocampus.This seems to indicate that *Irf8, Spacdr, Plek* and *Mafb* could play a major role in the regulation of inflammation and myeloid cell differentiation.

More specifically, *Mafb* DC-targets include the 3 *C1q*-chains, an interaction that has already been described in the literature [34, 35] in diverse macrophage phenotypes, including in the context of autoimmune diseases. On the other hand, *Irf8* is known to regulate myeloid lineage diversification [36]. Also, the product of *Plek*, *Pleckstrin*, has been reported to be involved in various inflammatory diseases, including diabetes, cardiovascular diseases and rheumatoid arthritis [37, 38]. *Spacdr* is orthologous to several human genes including *TSC22D4*, a transcriptional regulator that could be involved in AD by enhancing the assembly of *NRBP1* and *CRL*, that target and degrade *BRI2* and *BRI3*, two inhibitors of *APP* processing and amyloid β oligomerization [39].

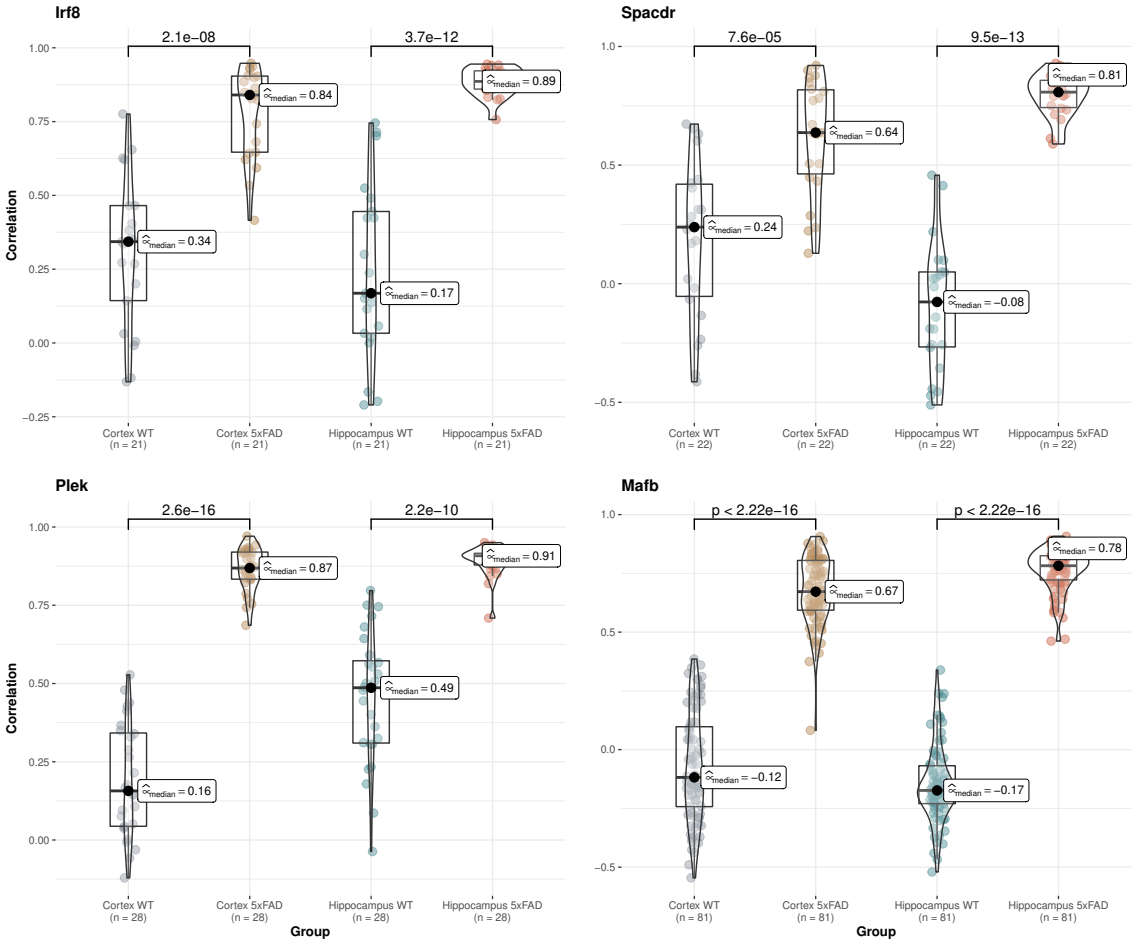


Figure 8: TF correlation with DC targets. Reported p-values were computed with the Wilcoxon-Mann-Whitney test, using the R packages *ggstatsplot* [18] and *ggsignif* [33].

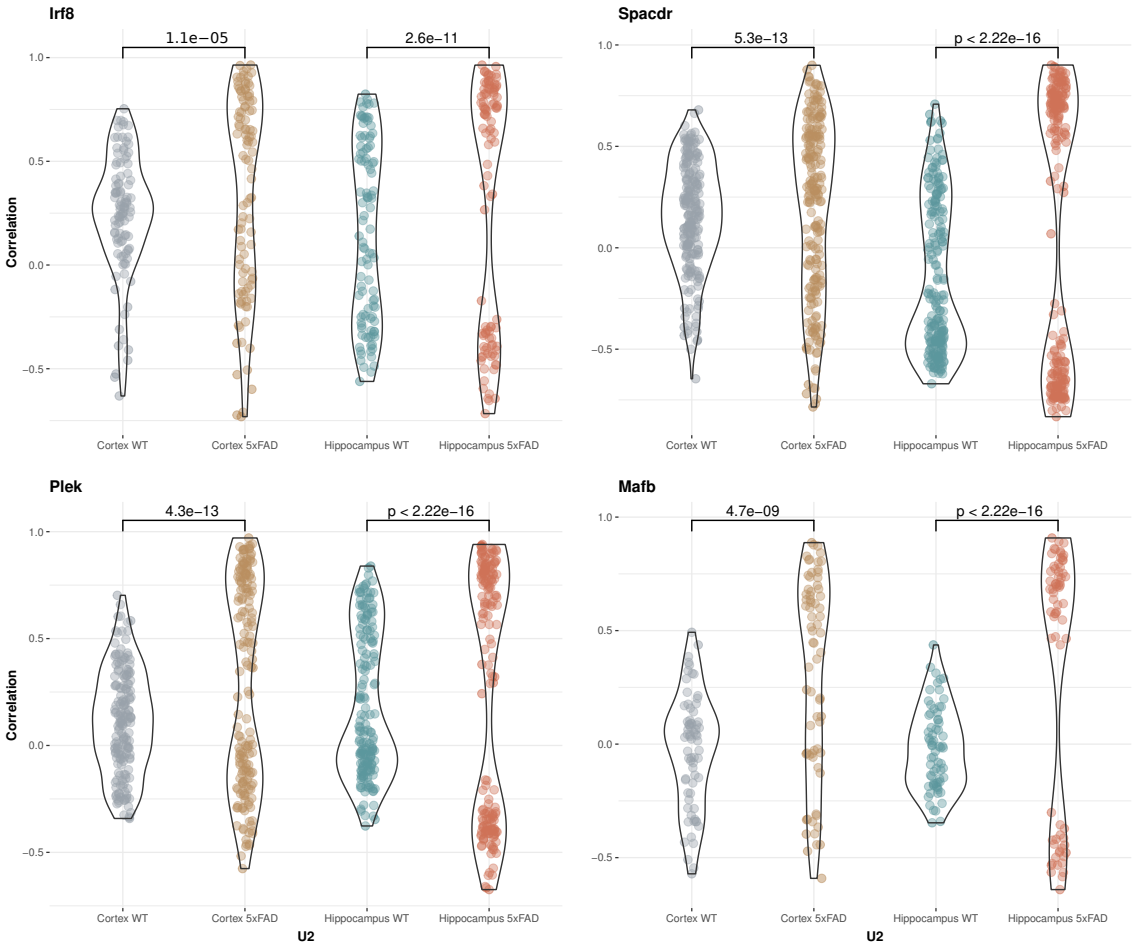


Figure 9: TF correlation with genes in U2 community. Reported p-values have been computed with the Wilcoxon-Mann-Whitney test on absolute correlation values, using the R packages *ggstatsplot* [18] and *ggsignif* [33].

3.5 Co-expression in DC communities

Although there is a clear link between co-expression and differential co-expression, the detection of DCL-communities as performed here does not allow to conclude regarding gene co-expression within these communities. Indeed, a set of DC targets with the same hub gene could be so in an independent fashion, i.e., without exhibiting any particular association between themselves. However, co-expression analysis within identified signed and unsigned DCL-communities showed clear correlation patterns within these gene sets, thus supporting the relevance of identifying DCL-communities in multi-layer DCNs.

We illustrate this by revealing patterns of correlations observed in the U1 and U2 communities of the unsigned strategy (see Figures 10 and 11). Here we present co-expression patterns in WT and 5xFAD hippocampus samples, as DC was more intense in the hippocampus (cf. Figure 2), but we have made similar observations in the cortex.

As previously described, **DCL-community U1**, is composed of DCLs incident to 3 hub genes: *S1pr3, Sbno2* and *Dhrs1*. These genes show losses as well as gains of correlations with their targets in 5xFAD compared to WT samples. No assumption can be made on potential interactions within these targets, since we did not identified DCLs among them, other than those with *S1pr3, Sbno2* and *Dhrs1*. However, when investigating co-expression patterns within the gene set (see Figure 10), we found that

(i) genes displaying negative DC with *S1pr3, Sbno2* and *Dhrs1* were co-expressed in WT and 5xFAD samples (Co-expressed Cluster CC1 in Figure 10),

(ii) genes displaying positive DC with *S1pr3, Sbno2* and *Dhrs1* were co-expressed in WT and 5xFAD samples (Co-expressed Cluster CC2 in Figure 10), and

(iii) genes displaying negative DC with *S1pr3, Sbno2* and *Dhrs1* were found inversely co-expressed with genes displaying positive DC with those 3 hub genes (inverse correlation trend between CC1 and CC2 in Figure 10).

This co-expression analysis reveals that the topology observed in this community is the result of a switch in the co-expression of *S1pr3, Sbno2* and *Dhrs1* with one group of co-expressed genes (negative-DC targets, CC1) to another (positive-DC

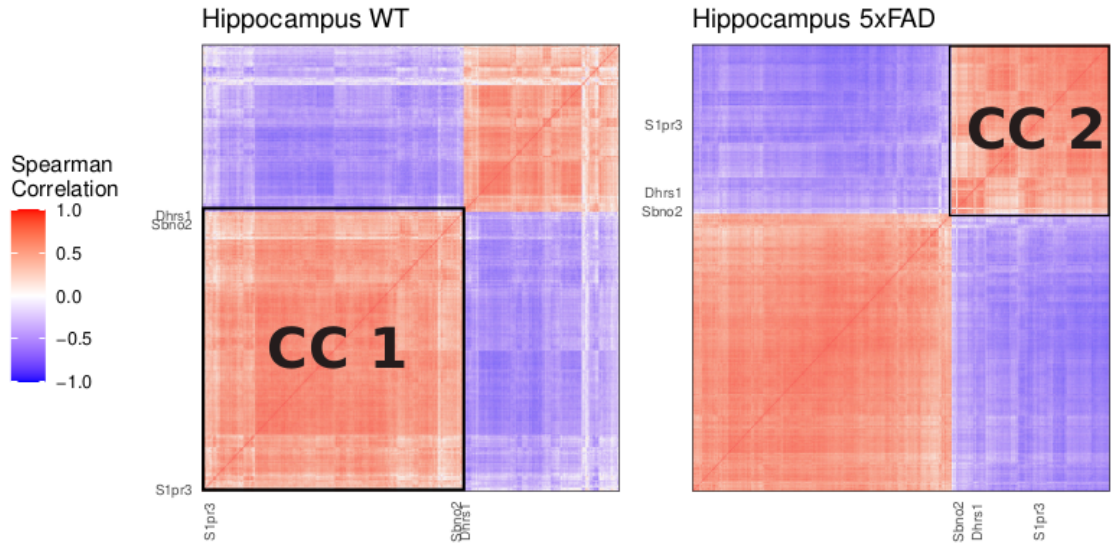


Figure 10: Correlation heatmap of genes from DCL-community U1 in WT and 5xFAD hippocampus samples, showing a switch of genes *S1pr3, Sbno2* and *Dhrs1* from co-expression cluster CC1 in WT samples to co-expression cluster CC2 in 5xFAD samples.

targets, CC2), depending on the genotype. When performing GO:BP enrichment analysis on both co-expressed clusters (CC), we found that CC1 is related to neuron development (GO:0048666, $p.adjust = 2.313E-8$) and axonogenesis (GO:0007409, $p.adjust = 1.552E-5$), while CC2 is involved in macrophage activation (GO:0042116, $p.adjust = 2.814E-3$), inflammatory response (GO:0006954, $p.adjust = 5.634E-3$) and leukocyte mediated immunity (GO:0002443, $p.adjust = 5.491E-3$).

We did not observe an overall change in correlation of the two co-expressed clusters between the WT and 5xFAD conditions, as each of the CCs remained co-expressed in both conditions (excluding *S1pr3, Sbno2* and *Dhrs1*), which was expected from the topology of the community (no DCLs other than those incident to *S1pr3, Sbno2* and *Dhrs1*). This suggests that the co-activation of the CC1 and CC2 gene-sets is not dependent from the co-regulation of *S1pr3, Sbno2* and *Dhrs1*.

For the **DCL-community U2**, which, as previously described gathered much more interconnected genes, the gain of co-expression is clearer, which was expected from the topology of the DCL-community. Inter-DC genes gained in correlation in the 5xFAD genotype, forming a co-expressed cluster (CC2 in Figure 11) involved in the regulation of cytokine production (GO:0001817 , $p.adjust = 1.527E-11$) and leukocyte activation (GO:0045321, $p.adjust = 1.446E-14$). The function of CC1 is not clear, as only high-level GO:BP terms were found enriched, but seems related

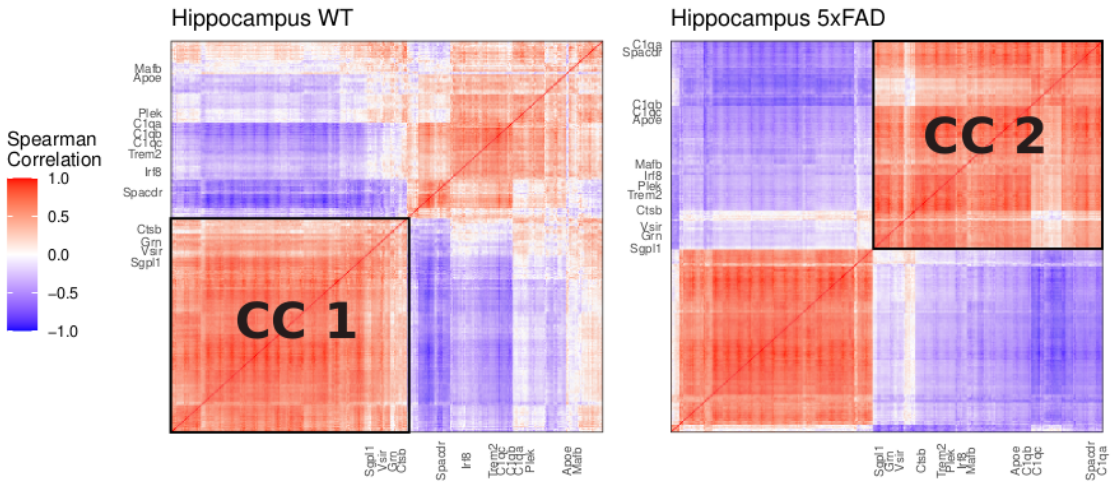


Figure 11: Correlation heatmap of genes from DCL-community U2 in WT and 5xFAD hippocampus samples showing a switch of genes *Grn, Sgpl1, Ctsb* and *Vsir* from co-expressed cluster CC1 in WT samples to co-expressed cluster CC2 in 5xFAD samples, and an overall gain of co-expression in CC2 in 5xFAD samples.

to protein transport (GO:0015031, $p.adjust = 4.633E-6$). *Grn, Sgpl1, Ctsb* and *Vsir*, which were at the periphery of the DCL-community, losing co-expression with a subset of their targets, were found co-expressed with CC1 in the WT condition, but co-expressed with CC2 in the 5xFAD condition. The clear gain in co-expression of genes from CC2 along with a switch of *Grn, Sgpl1, Ctsb* and *Vsir* co-expression indicate those 4 genes may play a role in the establishment of the co-regulation of genes involved in the the inflammatory response in the 5xFAD model.

When considering signed DCL-communities, similar observations were made, with the difference that the various co-expression groups to which a gene can belong, conditionally to the genotype of the mice, are no longer classified in the same DCL-community, but in different communities (loss on one side, gain on the other). The application of the signed analysis is advantageous for the biological interpretation of the communities, allowing the identification of both deactivated and activated processes. However, the unsigned analysis has the advantage of producing more synthetic results on the different biological processes in which a same gene can participate conditionally to the different experimental conditions of interest.

4 DISCUSSION

In this work we introduce a novel differential co-expression (DC) analysis strategy based on the detection of differential co-expression link (DCL) communities in a multi-layer differential co-expression network (DCN). By triggering groups of links (DCLs) in the graph rather than groups of nodes (genes) and also thanks to our hybrid similarity metric, we can ensure that the identified gene sets share DCLs that are similar both in terms of topology on a layer of the DCN (*i.e.*, act similarly in a specific experimental condition) but also in terms of co-occurrence (*i.e.*, are present in several layers of the network, thus in several experimental conditions). This allows us to reveal recurrent as well as specific perturbations induced in various experimental conditions. Note that the detection of recurrent DCL-communities is flexible, as DCLs that were not found in all layers of the network can still be classified as belonging to a recurrent community if presenting strong enough topological similarity with the other DCLs of the community. This reduces the impact of potential false-negatives during the construction of the multi-layer DCN.

Moreover, our link clustering strategy has the advantage of producing overlapping gene communities, which are more relevant from a biological point of view than forcing the assignment of a gene to a unique cluster. For instance, the 3 *C1q* chains were found in community U1, associated to axonogenesis and neuron development, as well as in community U2, associated to leukocyte and myeloid cells differentiation and inflammation. This is particularly interesting in the context of multi-group differential co-expression analysis, since a gene is expected to interact with different sets of genes depending on the experimental condition considered.

We applied the approach on a transcriptomic dataset to explore the variability of co-expression perturbations induced by the 5xFAD genotype on the cortex and the hippocampus. We observed that the two brain structures responded similarly, presenting a majority of DCLs in common. However, the hippocampus seems to exhibit more intense dysregulations than the cortex. The major DCL-commmunities detected were identified on both tissues of the multi-layer DCN, and DCLs that are specific to one or the other structure were not sufficiently numerous to detect hippocampus- or cortex-specific DCL-commmunities that are easily biologically interpretable, the few layer-specific DCL-communities found having been poorly annotated by GO enrichment analysis. However, although the enrichment analysis did not find a clear

functional link for these tissue-specific DCL-communities, the involved hub genes do appear to play a role in AD (notably, Cd44, Gpam and Plxcd2).

GO-term enrichment analysis of the other major DCL-communities showed their biological relevance, since they were found enriched with AD-relevant biological processes, including inflammation-related pathways or neuron-development-related processes. From a topological point of view, the computed communities revealed hub genes and interactions that are consistent with the published literature on Alzheimer's disease, including interaction patterns between *Mafb* and *C1q*-chains or the central role of *S1P* receptors.

Moreover, we also found clear co-expression patterns between the members of DCL-communities, which was not particularly expected but suggests that different targets of a same DC-gene are likely to be functionally linked or co-regulated, even if they do not share DCLs with one another.

We thus propose several genes that seem to be particularly involved in the co-expression perturbations induced by the 5xFAD model (of which, *S1pr3, Sbno2, Dhrs1, Grn, Sgpl1, Ctsb, Vsir, Klk6, Cd44, Gpam* and *C1q*-related genes), including some transcription factors (particularly, *Mafb, Irf8, Plek* and *Spacdr*). While some of these genes showed differences in expression, we observed that many genes sharing DCLs were not particularly, if at all, differentially expressed. Thus, compared to a classical differential expression analysis, differential co-expression not only proposes a topology of the interactions between genes, but also allows the detection of key actors of the disease that would have been missed if considering the expression level solely.

Two strategies for DCL-community detection were tested: an unsigned and a signed one. The former approach is particularly interesting for identifying the different biological processes to which a gene can participate in different experimental contexts. The latter allows to finely annotate gene-communities by precisely identifying the biological processes that are deactivated or activated in 5xFAD samples compared to their WT counterparts.

In this study, we did not explore DC accross time-points. Indeed, because the correlation is a noisy estimation of co-expression which needs to be measured on many samples to be reliable, we preferred to take advantage of all the samples available for a given brain structure. However, since most DCLs were identified in both the hippocampus and the cortex, it would be interesting to perform a complementary

DC analysis, considering in the same experimental group the samples from the two brain structures, but distinguishing the time-points in order to identify potential DCL-communities occurring at a specific development stage and their evolution over time.

Finally, the approach would benefit from being applied to a dataset gathering more heterogeneous sample groups, in order to demonstrate its potential on multi-layer DCNs that are more dissimilar. For example, an application to different CNS diseases could reveal specific or shared responses to a disease or a family of diseases.

5 CONCLUSION

We propose a new strategy for the construction and analysis of multi-layer differential co-expression networks, based on the detection of communities of differentially co-expressed links. The resulting communities are composed of genes with topologically close differential co-expression patterns, occurring in one or several layers of the network. The detection of such link communities allows to better understand the dysregulation patterns induced by one or more experimental conditions of interest compared to their respective control states.

References

[1] Colin R. Blyth. On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association*, 67(338):364–366, June 1972. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1972.10482387.

[2] Nicholas J. Hudson, Antonio Reverter, and Brian P. Dalrymple. A Differential Wiring Analysis of Expression Data Correctly Identifies the Gene Containing the Causal Mutation. *PLOS Computational Biology*, 5(5):e1000382, May 2009. Publisher: Public Library of Science.

[3] Aurora Savino, Paolo Provero, and Valeria Poli. Differential Co-Expression Analyses Allow the Identification of Critical Signalling Pathways Altered during Tumour Transformation and Progression. *International Journal of Molecular Sciences*, 21(24):9461, December 2020.

[4] Bruno M. Tesson, Rainer Breitling, and Ritsert C. Jansen. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(1):497, October 2010.

[5] David Amar, Hershel Safer, and Ron Shamir. Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. *PLOS Computational Biology*, 9(3):e1002955, March 2013. Publisher: Public Library of Science.

[6] Jiexin Zhang, Yuan Ji, and Li Zhang. Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23(21):2903–2909, November 2007.

[7] Andrew T. McKenzie, Igor Katsyv, Won-Min Song, Minghui Wang, and Bin Zhang. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Systems Biology*, 10(1):106, November 2016.

[8] Dharmesh D. Bhuva, Joseph Cursons, Gordon K. Smyth, and Melissa J. Davis. Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome Biology*, 20(1):236, December 2019.

[9] Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. (Differential) Co-Expression Analysis of Gene Expression: A Survey of

Best Practices. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(4):1154–1173, August 2020.

[10] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, January 2013.

[11] Stefania Forner, Shimako Kawauchi, Gabriela Balderrama-Gutierrez, Enikö A. Kramár, Dina P. Matheos, Jimmy Phan, Dominic I. Javonillo, Kristine M. Tran, Edna Hingco, Celia da Cunha, Narges Rezaie, Joshua A. Alcantara, David Baglietto-Vargas, Camden Jansen, Jonathan Neumann, Marcelo A. Wood, Grant R. MacGregor, Ali Mortazavi, Andrea J. Tenner, Frank M. LaFerla, and Kim N. Green. Systematic phenotyping and characterization of the 5xFAD mouse model of Alzheimer's disease. *Scientific Data*, 8(1):270, October 2021.

[12] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.

[13] Tristan Mary-Huard, Sarmistha Das, Indranil Mukhopadhyay, and Stephane Robin. Querying multiple sets of p-values through composed hypothesis testing. *Bioinformatics (Oxford, England)*, page btab592, September 2021.

[14] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, August 2010.

[15] Saeed Salem and Cagri Ozcaglar. Hybrid coexpression link similarity graph clustering for mining biological modules from multiple gene expression datasets. *BioData Mining*, 7(1):16, December 2014.

[16] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, April 2002.

[17] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, 16(5):284–287, May 2012.

[18] Indrajeet Patil. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61):3167, 2021.

[19] R Nudel, NH Simpson, G Baird, A O'Hare, G Conti-Ramsden, PF Bolton, and ER Hennessy. Consortium sli, ring sm, davey smith g, francks c, paracchini s, monaco ap, fisher se, newbury df. genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes Brain Behav*, 13:418–429, 2014.

[20] Magdalena Grill, Taylor E. Syme, Aline L. Noçon, Andy Z. X. Lu, Dale Hancock, Stefan Rose-John, and Iain L. Campbell. Strawberry notch homolog 2 is a novel inflammatory response factor predominantly but not exclusively expressed by astrocytes in the central nervous system. *Glia*, 63(10):1738–1752, October 2015.

[21] Timothy A. Couttas, Nupur Kain, Benjamin Daniels, Xin Ying Lim, Claire Shepherd, Jillian Kril, Russell Pickford, Hongyun Li, Brett Garner, and Anthony S. Don. Loss of the neuroprotective factor Sphingosine 1-phosphate early in Alzheimer's disease pathogenesis. *Acta Neuropathologica Communications*, 2(1):9, January 2014.

[22] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106–118, February 2013.

[23] Changjun Yin, Susanne Ackermann, Zhe Ma, Sarajo K. Mohanta, Chuankai Zhang, Yuanfang Li, Sandor Nietzsche, Martin Westermann, Li Peng, Desheng Hu, Sai Vineela Bontha, Prasad Srikakulapu, Michael Beer, Remco T. A. Megens, Sabine Steffens, Markus Hildner, Luke D. Halder, Hans-Henning Eckstein, Jaroslav Pelisek, Jochen Herms, Sigrun Roeber, Thomas Arzberger, Anna Borodovsky, Livia Habenicht, Christoph J. Binder, Christian Weber, Peter F. Zipfel, Christine Skerka, and Andreas J. R. Habenicht. ApoE attenuates unresolvable inflammation by complex formation with activated C1q. *Nature Medicine*, 25(3):496–506, March 2019.

[24] Cynthia L. Smith and Janan T. Eppig. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome*, 23(9):653–668, October 2012.

[25] Emma L. Ashby, Patrick G. Kehoe, and Seth Love. Kallikrein-related peptidase 6 in Alzheimer's disease and vascular dementia. *Brain Research*, 1363:1–10, December 2010.

[26] Elhanan Pinner, Yaron Gruper, Micha Ben Zimra, Don Kristt, Moshe Laudon, David Naor, and Nava Zisapel. CD44 Splice Variants as Potential Players in Alzheimer's Disease Pathology. *Journal of Alzheimer's disease: JAD*, 58(4):1137–1149, 2017.

[27] Eric A. Schon and Estela Area-Gomez. Mitochondria-associated ER membranes in Alzheimer disease. *Molecular and Cellular Neuroscience*, 55:26–36, July 2013.

[28] Ming Yang, Chenrui Li, Shikun Yang, Ying Xiao, Xiaofen Xiong, Wei Chen, Hao Zhao, Qin Zhang, Yachun Han, and Lin Sun. Mitochondria-Associated ER Membranes – The Origin Site of Autophagy. *Frontiers in Cell and Developmental Biology*, 8, 2020.

[29] Daan van Abel, Omar Michel, Rob Veerhuis, Marlies Jacobs, Marie van Dijk, and Cees B. M. Oudejans. Direct Downregulation of CNTNAP2 by STOX1A is Associated with Alzheimer's Disease. *Journal of Alzheimer's Disease*, 31(4):793–800, January 2012. Publisher: IOS Press.

[30] Wilfred A Jefferies, Katherine A Price, Kaan E Biron, Franz Fenninger, Cheryl G Pfeifer, and Dara L Dickstein. Adjusting the compass: new insights into the role of angiogenesis in Alzheimer's disease. *Alzheimer's Research & Therapy*, 5(6):64, 2013.

[31] Suzanne F. C. Miller-Delaney, Ivo Lieberam, Paula Murphy, and Kevin J. Mitchell. Plxdc2 Is a Mitogen for Neural Progenitors. *PLoS ONE*, 6(1):e14565, January 2011.

[32] Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, 47(D1):D33–D38, January 2019.

[33] Ahlmann-Eltze Constantin and Indrajeet Patil. ggsignif: R package for displaying significance brackets for 'ggplot2'. *PsyArxiv*, 2021.

[34] Michito Hamada, Yuki Tsunakawa, Hyojung Jeon, Manoj Kumar Yadav, and Satoru Takahashi. Role of MafB in macrophages. *Experimental Animals*, 69(1):1–10, 2020.

[35] Mai Thi Nhu Tran, Michito Hamada, Hyojung Jeon, Risako Shiraishi, Keigo Asano, Motochika Hattori, Megumi Nakamura, Yuki Imamura, Yuki Tsunakawa, Risa Fujii, Toshiaki Usui, Kaushalya Kulathunga, Christina-Sylvia Andrea, Ryusuke Koshida, Risa Kamei, Yurina Matsunaga, Makoto Kobayashi, Hisashi Oishi, Takashi Kudo, and Satoru Takahashi. MafB is a critical regulator of complement component C1q. *Nature Communications*, 8(1):1700, November 2017. Number: 1 Publisher: Nature Publishing Group.

[36] Hongsheng Wang and Herbert C. Morse. IRF8 regulates myeloid and B lymphoid lineage diversification. *Immunologic research*, 43(1-3):109–117, 2009.

[37] Yong Ding, Alpdogan Kantarci, John A. Badwey, Hatice Hasturk, Alan Malabanan, and Thomas E. Van Dyke. Phosphorylation of Pleckstrin Increases Proinflammatory Cytokine Secretion by Mononuclear Phagocytes in Diabetes Mellitus. *The Journal of Immunology*, 179(1):647–654, July 2007.

[38] M. Abdul Alim, Duncan Njenda, Anna Lundmark, Marta Kaminska, Leif Jansson, Kaja Eriksson, Anna Kats, Gunnar Johannsen, Catalin Koro Arvidsson, Piotr M. Mydel, and Tülay Yucel-Lindberg. Pleckstrin Levels Are Increased in Patients with Chronic Periodontitis and Regulated via the MAP Kinase-p38α Signaling Pathway in Gingival Fibroblasts. *Frontiers in Immunology*, 12, 2022.

[39] Takashi Yasukawa, Aya Tsutsui, Chieri Tomomori-Sato, Shigeo Sato, Anita Saraf, Michael P Washburn, Laurence Florens, Tohru Terada, Kentaro Shimizu, Ronald C Conaway, et al. Nrbp1-containing crl2/crl4a regulates amyloid β production by targeting bri2 and bri3 for degradation. *Cell Reports*, 30(10):3478–3491, 2020.

4.4 Discussion

Multi-layer DCN link-clustering has the potential to highlight key dysregulations in genes co-regulation occurring in all or a subset of investigated experimental contexts. Moreover, it has the advantage to produce overlapping gene communities, and can thus resolve the different implications that the same gene can present under different experimental conditions. Applying this strategy to a two-layer DCN build from DC analysis in the cortex and the hippocampus of 5xFAD mice compared to their respective control, we found key deregulation communities that occurred in both cerebral structures and several structure-specific DC patterns. This network-based strategy can detect various types of DC-patterns across layers. DC-modules with loss or gains in correlation in hippocampus and cortex 5xFAD samples compared to controls samples (patterns A and C from Figure 4.2) were identified, as well as "gene hopping" patterns occurring in both cerebral structures (pattern E from Figure 4.2). We also propose several genes, and specifically transcription factors, that appear as key actors of DC.

Since we observed a large overlap of hippocampus- and cortex-DC patterns, a complementary analysis on the dataset studied in this work could consider time-points rather than brain structure to compute layers of the multi-DCN to investigate the time-course behaviour of DC in the 5xFAD model.

More broadly, the approach would benefit in considering more heterogeneous DC-layers, for instance by integrating several neurodevelopmental or degenerative diseases, or datasets from various cancer types.

The limiting factor in the choice of data sets considered remains the number of replicates per sample-group, in order to obtain robust measurement of gene association.

Moreover, knowledge data such as protein-protein interactions or pathways could be included as supplementary layers in the analysis. Layers could also be computed from other omics, for instance using proteins or miRNA abundance, to offer a multi-omics view of differential regulation induced in a disease.

Chapter 5

Conclusion

The integration of omics and non-omics data is currently a major challenge in biological data analysis. The advantages of integrative data analysis for identifying and describing complex omics interactions are clear. After the development of new data acquisition technologies, it is the development of omics integrative strategies that is now in full swing, and there is no doubt that these strategies will continue to develop and become increasingly sophisticated.

In this thesis manuscript, we focused on two major biological questions associated with data integration:

(i) The identification of interaction patterns between multiple types of omics data in order to resolve multi-omics mechanisms associated with a phenotype of interest (particularly in a multi-omics subtyping context).

(ii) The elucidation of phenome heterogeneity by characterizing molecular patterns associated with different phenotypes of interest (especially in the context of analysis of differential co-expression patterns associated with Alzheimer's disease).

Our first contribution on the subject of omics data integration addresses the first type of question, i.e., the resolution of multi-omics mechanisms associated to a phenotype of interest, and in our case the prediction of novel cancer subtypes through the analysis of multi-omics datasets. The solution computes a consensus clustering by reconciling the predictions contained in various input clusterings. In addition to tackling the issue of integrating heterogeneous omics and non-omics data, consensus clustering is able to reconcile conflicting predictions obtained through to different analysis strategies, and thus take advantage of existing integration strategies. Thus, we have shown the efficiency of consensus clustering for the integration of omics data at two scales: in a multi-to-multi integration scenario, for the integration of multi-omics predictions, and in a single-to-multi integration scenario, for the integration of single-omics predictions into a multi-omics consensus.

Our second contribution addresses the second type of question, i.e., the evaluation and characterization of the diversity of molecular patterns associated with different phenotypes. Aiming at characterizing differential co-expression patterns induced in the hippocampus and the cortex of a mouse model of Alzheimer's disease, we developed a novel strategy for the computation and analysis of multi-layer Differential Co-expression Networks. Based on the detection of groups of differentially co-expressed gene pairs, the strategy allowed the detection of several patterns occurring in both brain structures and tissue-specific differential co-expression patterns. Applied on different datasets, the strategy could be used to detect gene conditional associations across a wider range of experimental conditions, including tissues, time-points

or diseases. Moreover, we particularly focused on the detection of differential co-expression patterns but the same strategy could be applied to the co-expression and differential co-expression of genes, micro-RNAs and proteins. Since the strategy is network-based, knowledge data such as pathways or protein-protein interactions could also be included as supplementary layers.

In the course of these methodological developments, we were confronted with the problem of estimating the quality of the predictions obtained from integrating heterogeneous data sources (and which motivated the development of a consensus approach for our first contribution). Indeed, we noticed that the traditional quality metrics were not always adapted to the estimation of the quality of results produced by the integration of several data sets. Indeed, the seemingly simple choice of an appropriate metric to assess the quality of results, remains a real challenge, usually overlooked.

When "ground-truth" data is available an objective quality evaluation can be made, however, in bioinformatics studies it is rarely the case. Models of omics data organisation are more easily available, but often predicted in other contexts. For example, for the multi-omics subtyping of breast cancer, one may estimate the quality of the results by comparing them to the subtypes predicted by the PAM50, which became a reference in breast cancer subtyping. However, this classification remains a prediction made on the basis of the expression of 50 genes, and it is not expected to be equally consistent when considering other types of omics, which has been demonstrated in [128]. Although other quality metrics can be used (e.g., survival or clinical similarities

of patients), they represent only one interpretation of the predictions, and predicted subtypes could be of high quality without showing differences in survival rates, for example. The question of the choice of quality metrics is therefore crucial, and the assessment of the quality of multi-omics predictions should be based on quality metrics adapted to this context.

Another example of bias in quality assessment is the use of gene ontology enrichment analysis for estimating gene sets robustness. Gene ontology enrichment analysis is traditionally used to annotate gene sets but also to evaluate their quality, gene sets with a high functional enrichment being considered of high quality. However, as pointed in [129], 58% of the gene ontology annotations concern only 16% of the human genes, which is the source of a strong bias in the interpretation of results. Gene sets that might otherwise be coherent but which are composed of poorly annotated genes, will thus be perceived as being of poor quality, which is a brake on new discoveries. Similarly, the annotation of these gene sets is also impacted by the updates of databases. A gene set considered to be qualitative at a given time will not necessarily remain so, or inversely, a gene set considered low-quality could be better annotated in the future by means of novel annotations added to the Gene Ontology. A similar quality metric for biological networks is the computation of their enrichment for known protein-protein interactions, which suffers the same bias as Gene Ontology enrichment strategies. Such knowledge-based metrics, while incredibly relevant for annotating gene sets and biological networks, should be employed with consideration to their potential biais. The question of evaluating pre-

dictions obtained from the integration of omics and knowledge data may also arise, since it cannot be resolved with these same knowledge-based evaluation metrics.

Thus, although integrative methods have become widely available in recent years, there is a lack of quality metrics adapted to this type of context. This observation has motivated some research work on the harmonization of "Figures of Merit", novel quality descriptors applicable to different omics data types in 2020 [130]. Such developments for the standardization of quality metrics in bioinformatics could, in the long run, offer a coherent framework adapted for each type of omics data, as well as knowledge data integration.

Bibliography

[1] Zahra Momeni, Esmail Hassanzadeh, Mohammad Saniee Abadeh, and Riccardo Bellazzi. A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, 107:103466, July 2020.

[2] Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 44(Database issue):D67–D72, January 2016.

[3] Lars J. Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database issue):D412–D416, January 2009.

[4] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[5] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, January 2021.

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[7] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011.

[8] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002.

[9] The Cancer Genome Atlas Program - NCI, June 2018. URL: https://www.cancer.gov/tcga.

[10] Synapse | Sage Bionetworks. URL: https://www.synapse.org/.

[11] Vasileios Lapatas, Michalis Stefanidakis, Rafael C. Jimenez, Allegra Via, and Maria Victoria Schneider. Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, 22(1):9, September 2015.

[12] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen

Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251, November 2007.

[13] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):I1, March 2014.

[14] Haley R. Eidem, Jacob L. Steenwyk, Jennifer H. Wisecaver, John A. Capra, Patrick Abbot, and Antonis Rokas. integRATE: a desirability-based data integration framework for the prioritization of candidate genes across heterogeneous omics and its application to preterm birth. *BMC Medical Genomics*, 11(1):107, November 2018.

[15] Suchi Saria and Anna Goldenberg. Subtyping: What It is and Its Role in Precision Medicine. *IEEE Intelligent Systems*, 30(4):70–75, July 2015. Conference Name: IEEE Intelligent Systems.

[16] Iliyan Mihaylov, Maciej Kańduła, Milko Krachunov, and Dimitar Vassilev. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology Direct*, 14(1):22, November 2019.

[17] Mengyun Wu, Huangdi Yi, and Shuangge Ma. Vertical integration methods for gene expression data analysis. *Briefings in Bioinformatics*, 22(3):bbaa169, May 2021.

[18] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1):124, January 2021. Number: 1 Publisher: Nature Publishing Group.

[19] Meta-analysis in basic biology. *Nature Methods*, 13(12):959–959, December 2016. Number: 12 Publisher: Nature Publishing Group.

[20] Andre Franke, Dermot P. B. McGovern, Jeffrey C. Barrett, Kai Wang, Graham L. Radford-Smith, Tariq Ahmad, Charlie W. Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A. Anderson, Joshua C. Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M. Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G. Mathew, Grant W. Montgomery, Natalie J. Prescott, Soumya Raychaudhuri, Jerome I. Rotter, Philip Schumm, Yashoda Sharma, Lisa A. Simms, Kent D. Taylor, David Whiteman, Cisca Wijmenga, Robert N. Baldassano, Murray Barclay, Theodore M. Bayless, Stephan Brand, Carsten Büning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jürgen

Glas, Andre Van Gossum, Stephen L. Guthery, Jonas Halfvarson, Hein W. Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julián Panés, Anne Phillips, Deborah D. Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A. Hillary Steinhart, Pieter C. F. Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R. Targan, Steven R. Brant, John D. Rioux, Mauro D'Amato, Rinse K. Weersma, Subra Kugathasan, Anne M. Griffiths, John C. Mansfield, Severine Vermeire, Richard H. Duerr, Mark S. Silverberg, Jack Satsangi, Stefan Schreiber, Judy H. Cho, Vito Annese, Hakon Hakonarson, Mark J. Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, 42(12):1118–1125, December 2010.

[21] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research*, 14(6):1085–1094, June 2004.

[22] Matthew C. Altman, Darawan Rinchai, Nicole Baldwin, Mohammed Toufiq, Elizabeth Whalen, Mathieu Garand, Basirudeen Syed Ahamed Kabeer, Mohamed Alfaki, Scott R. Presnell, Prasong Khaenam, Aaron Ayllón-Benítez, Fleur Mougin, Patricia Thébault, Laurent Chiche, Noemie Jourde-Chiche, J. Theodore

Phillips, Goran Klintmalm, Anne O'Garra, Matthew Berry, Chloe Bloom, Robert J. Wilkinson, Christine M. Graham, Marc Lipman, Ganjana Lertmemongkolchai, Davide Bedognetti, Rodolphe Thiebaut, Farrah Kheradmand, Asuncion Mejias, Octavio Ramilo, Karolina Palucka, Virginia Pascual, Jacques Banchereau, and Damien Chaussabel. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications*, 12(1):4385, July 2021. Number: 1 Publisher: Nature Publishing Group.

[23] Daniel Toro-Domínguez, Juan Antonio Villatoro-García, Jordi Martorell-Marugán, Yolanda Román-Montoya, Marta E Alarcón-Riquelme, and Pedro Carmona-Sáez. A survey of gene expression meta-analysis: methods and applications. *Briefings in Bioinformatics*, 22(2):1694–1705, March 2021.

[24] Peter H. Sudmant, Maria S. Alexis, and Christopher B. Burge. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology*, 16(1):287, December 2015.

[25] Yang Xu and Rachel Patton McCord. Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nature Communications*, 13(1):3505, June 2022. Number: 1 Publisher: Nature Publishing Group.

[26] Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-Cell Multiomic Integration Compares and Contrasts Features of Brain Cell

Identity. *Cell*, 177(7):1873–1887.e17, June 2019. Publisher: Elsevier.

[27] Chao Gao, Jialin Liu, April R. Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R. Nery, Margarita M. Behrens, Joseph R. Ecker, Bing Ren, and Joshua D. Welch. Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*, 39(8):1000–1007, August 2021.

[28] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, 38(1):211–219, January 2022.

[29] Jeppe S. Spicker, Søren Brunak, Klaus S. Frederiksen, and Henrik Toft. Integration of Clinical Chemistry, Expression, and Metabolite Data Leads to Better Toxicological Class Separation. *Toxicological Sciences*, 102(2):444–454, April 2008.

[30] Dingming Wu, Dongfang Wang, Michael Q. Zhang, and Jin Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*, 16(1):1022, December 2015.

[31] Hua Wang, Feiping Nie, and Heng Huang. Multi-View Clustering and Feature Learning via Structured Sparsity. In *Proceedings of the 30th International Conference on Machine Learning*, pages 352–360. PMLR, May 2013. ISSN: 1938-7228.

[32] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D. M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A. Margolin, Laura J. van't Veer, Nuria Lopez-Bigas, Peter W. Laird, Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, Christopher C. Benz, Charles M. Perou, and Joshua M. Stuart. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944, August 2014.

[33] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, March 2014. Number: 3 Publisher: Nature Publishing Group.

[34] Saeed Salem and Cagri Ozcaglar. Hybrid coexpression link similarity graph clustering for mining biological modules from multiple gene expression datasets. *BioData Mining*, 7(1):16, August 2014.

[35] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaëlle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. Random walk with restart on multiplex and het-

erogeneous biological networks. *Bioinformatics*, 35(3):497–505, February 2019.

[36] X. Wang, E. P. Xing, and D. J. Schaid. Kernel methods for large-scale genomic data analysis. *Briefings in Bioinformatics*, 16(2):183–192, March 2015.

[37] Lluís A Belanche and Marco A Villegas. Kernel functions for categorical variables with application to problems in the life sciences. In *Artificial Intelligence Research and Development*, pages 171–180. IOS Press, 2013.

[38] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, June 2008. Publisher: Institute of Mathematical Statistics.

[39] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[40] Nora K. Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, June 2015.

[41] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[42] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14, 2019.

[43] Chen Meng, Oana A. Zeleznik, Gerhard G. Thallinger, Bernhard Kuster, Amin M. Gholami, and Aedín C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, July 2016.

[44] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On Deep Multi-View Representation Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1083–1092. PMLR, June 2015. ISSN: 1938-7228.

[45] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[46] Marta Lovino, Vincenzo Randazzo, Gabriele Ciravegna, Pietro Barbiero, Elisa Ficarra, and Giansalvo Cirrincione. A survey on data integration for multi-omics sample clustering. *Neurocomputing*, 488:494–508, June 2022.

[47] Miriam Ragle Aure, Israel Steinfeld, Lars Oliver Baumbusch, Knut Liestøl, Doron Lipson, Sandra Nyberg, Bjørn Naume, Kristine Kleivi Sahlberg, Vessela N. Kristensen, Anne-Lise Børresen-Dale, Ole Christian Lingjærde, and Zohar Yakhini. Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data. *PLoS ONE*, 8(1):e53014, January 2013.

[48] Raj Chari, Bradley P. Coe, Emily A. Vucic, William W. Lockwood, and Wan L. Lam. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology*, 4(1):67, May 2010.

[49] Li Tong, Jonathan Mitchel, Kevin Chatlin, and May D. Wang. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Medical Informatics and Decision Making*, 20(1):225, September 2020.

[50] Alessandra Cabassi and Paul DW Kirk. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics*, 36(18):4789–4796, 2020.

[51] Rosario M. Piro and Ferdinando Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *The FEBS journal*, 279(5):678–696, March 2012.

[52] Yong Chen, Xuebing Wu, and Rui Jiang. Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, 6(1):57, December 2013.

[53] Michele Gentili, Leonardo Martini, Marialuisa Sponziello, and Luca Becchetti. Biological Random Walks: multi-omics integration for disease gene prioritization. *Bioinformatics*, 38, July 2022.

[54] Nam D Nguyen, Ting Jin, and Daifeng Wang. Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics*, 37(12):1772–1775, June 2021.

[55] Yan Li, Teng Huang, Yun Xiao, Shangwei Ning, Peng Wang, Qianghu Wang, Xin Chen, Xu Chaohan, Donglin Sun, Xia Li, and Yixue Li. Prioritising risk pathways of complex human diseases based on functional profiling. *European Journal of Human Genetics*, 21(6):666–672, June 2013. Number: 6 Publisher: Nature Publishing Group.

[56] Chi Zhang, Yan Wang, Chun-Lei Zhang, and Hua-Rong Wu. Prioritization of candidate metabolites for postmenopausal osteoporosis using multi-omics composite network. *Experimental and Therapeutic Medicine*, 17(4):3155–3161, April 2019.

[57] Krishna R. Kalari, Jason P. Sinnwell, Kevin J. Thompson, Xiaojia Tang, Erin E. Carlson, Jia Yu, Peter T. Vedell, James N. Ingle, Richard M. Weinshilboum, Judy C. Boughey, Liewei Wang, Matthew P. Goetz, and Vera Suman. PANOPLY: Omics-Guided Drug Prioritization Method Tailored to an Individual Patient. *JCO clinical cancer informatics*, 2:1–11, December 2018.

[58] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, June 2021. Number: 1 Publisher: Nature Publishing Group.

[59] Malvika Sudhakar, Raghunathan Rengaswamy, and Karthik Raman. Multi-Omic Data Improve Prediction of Personalized Tumor

Suppressors and Oncogenes. *Frontiers in Genetics*, 13:854190, May 2022.

[60] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546–10562, November 2018.

[61] Tejaswi V. S. Badam, Hendrik A. de Weerd, David Martínez-Enguita, Tomas Olsson, Lars Alfredsson, Ingrid Kockum, Maja Jagodic, Zelmina Lubovac-Pilav, and Mika Gustafsson. A validated generally applicable approach using the systematic assessment of disease modules by GWAS reveals a multi-omic module strongly associated with risk factors in multiple sclerosis. *BMC Genomics*, 22(1):631, August 2021.

[62] Min Chen, Jianying Yan, Qing Han, Jinying Luo, and Qinjian Zhang. Identification of hub-methylated differentially expressed genes in patients with gestational diabetes mellitus by multi-omic WGCNA basing epigenome-wide and transcriptome-wide profiling. *Journal of Cellular Biochemistry*, 121(5-6):3173–3184, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcb.29584.

[63] Nimrod Rappoport, Roy Safra, and Ron Shamir. MONET: Multi-omic module discovery by omic selection. *PLOS Computational Biology*, 16(9):e1008182, September 2020.

[64] Eric Bonnet, Laurence Calzone, and Tom Michoel. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLOS*

Computational Biology, 11(2):e1003983, February 2015. Publisher: Public Library of Science.

[65] Stefan Buchka, Alexander Hapfelmeier, Paul P. Gardner, Rory Wilson, and Anne-Laure Boulesteix. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, 22(1):152, May 2021.

[66] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[67] Soledad Ochoa, Guillermo de Anda-Jáuregui, and Enrique Hernández-Lemus. Multi-omic regulation of the pam50 gene signature in breast cancer molecular subtypes. *Frontiers in oncology*, 10:845, 2020.

[68] Vincent Planche, José V Manjon, Boris Mansencal, Enrique Lanuza, Thomas Tourdias, Gwenaëlle Catheline, and Pierrick Coupé. Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain Communications*, 4(3):fcac109, 2022.

[69] Sandro Vega-Pons and José Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011.

[70] Tahani Alqurashi and Wenjia Wang. Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, 10(6):1227–1246, 2019.

[71] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[72] Ana LN Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850, 2005.

[73] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[74] Yang Yuan, Pan Qi, Wang Xiang, Liu Yanhui, Li Yu, and Mao Qing. Multi-omics analysis reveals novel subtypes and driver genes in glioblastoma. *Frontiers in genetics*, 11:565341, 2020.

[75] Miriam Ragle Aure, Valeria Vitelli, Sandra Jernström, Surendra Kumar, Marit Krohn, Eldri U Due, Tonje Husby Haukaas, Suvi-Katri Leivonen, Hans Kristian Moen Vollan, Torben Lüders, et al.

Integrative clustering reveals a novel split in the luminal a subtype of breast cancer with impact on outcome. *Breast Cancer Research*, 19(1):1–18, 2017.

[76] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[77] Ying-Wooi Wan, Rami Al-Ouran, Carl G Mangleburg, Thanneer M Perumal, Tom V Lee, Katherine Allison, Vivek Swarup, Cory C Funk, Chris Gaiteri, Mariet Allen, et al. Meta-analysis of the alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell reports*, 32(2):107908, 2020.

[78] Said Assou, Tanguy Le Carrour, Sylvie Tondeur, Susanne Ström, Audrey Gabelle, Sophie Marty, Laure Nadal, Véronique Pantesco, Thierry Réme, Jean-Philippe Hugnot, et al. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem cells*, 25(4):961–973, 2007.

[79] Irene Pérez-Díez, Marta R Hidalgo, Pablo Malmierca-Merlo, Zoraida Andreu, Sergio Romera-Giner, Rosa Farràs, María de la Iglesia-Vayá, Mariano Provencio, Atocha Romero, and Francisco García-García. Functional signatures in non-small-cell lung can-

cer: A systematic review and meta-analysis of sex-based differences in transcriptomic studies. *Cancers*, 13(1):143, 2021.

[80] Damien Chaussabel and Nicole Baldwin. Democratizing systems immunology with modular transcriptional repertoire analyses. *Nature Reviews Immunology*, 14(4):271–280, 2014.

[81] Matthew C Altman, Darawan Rinchai, Nicole Baldwin, Mohammed Toufiq, Elizabeth Whalen, Mathieu Garand, Basirudeen Syed Ahamed Kabeer, Mohamed Alfaki, Scott R Presnell, Prasong Khaenam, et al. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications*, 12(1):1–19, 2021.

[82] Yu Yamazaki, Na Zhao, Thomas R Caulfield, Chia-Chen Liu, and Guojun Bu. Apolipoprotein e and alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15(9):501–518, 2019.

[83] Christiane Reitz and Richard Mayeux. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. *Biochemical pharmacology*, 88(4):640–651, 2014.

[84] Bryant Lim, Ioannis Prassas, and Eleftherios P Diamandis. Alzheimer disease pathogenesis: The role of autoimmunity. *The Journal of Applied Laboratory Medicine*, 6(3):756–764, 2021.

[85] Anne-Laure Hemonnot, Jennifer Hua, Lauriane Ulmann, and Hélène Hirbec. Microglia in alzheimer disease: well-known targets

and new opportunities. *Frontiers in aging neuroscience*, 11:233, 2019.

[86] Anne-Laure Hemonnot-Girard, Cédric Meersseman, Manuela Pastore, Valentin Garcia, Nathalie Linck, Catherine Rey, Amine Chebbi, Freddy Jeanneteau, Stephen D Ginsberg, Joël Lachuer, et al. Comparative analysis of transcriptome remodeling in plaque-associated and plaque-distant microglia during amyloid-β pathology progression in mice. *Journal of Neuroinflammation*, 19(1):1–26, 2022.

[87] Miyabishara Yokoyama, Honoka Kobayashi, Lisa Tatsumi, and Taisuke Tomita. Mouse models of alzheimer's disease. *Frontiers in Molecular Neuroscience*, 15, 2022.

[88] Caroline Ismeurt, Patrizia Giannoni, and Sylvie Claeysen. The 5xfad mouse model of alzheimer's disease. In *Diagnosis and Management in Dementia*, pages 207–221. Elsevier, 2020.

[89] Adrian L Oblak, Peter B Lin, Kevin P Kotredes, Ravi S Pandey, Dylan Garceau, Harriet M Williams, Asli Uyar, Rita O'Rourke, Sarah O'Rourke, Cynthia Ingraham, et al. Comprehensive evaluation of the 5xfad mouse model for preclinical testing applications: a model-ad study. *Frontiers in aging neuroscience*, 13:713726, 2021.

[90] Stefania Forner, Shimako Kawauchi, Gabriela Balderrama-Gutierrez, Enikö A Kramár, Dina P Matheos, Jimmy Phan, Dominic I Javonillo, Kristine M Tran, Edna Hingco, Celia da Cunha,

et al. Systematic phenotyping and characterization of the 5xfad mouse model of alzheimer's disease. *Scientific data*, 8(1):1–16, 2021.

[91] Asli Uyar, Ravi Pandey, Christoph Preuss, Kevin Kotredes, Gareth Howell, Michael Sasner, and Gregory Carter. Aging related transcriptomic changes in the mouse models of alzheimer's disease. *Innovation in Aging*, 4(Suppl 1):117, 2020.

[92] Christoph Preuss, Ravi Pandey, Erin Piazza, Alexander Fine, Asli Uyar, Thanneer Perumal, Dylan Garceau, Kevin P Kotredes, Harriet Williams, Lara M Mangravite, et al. A novel systems biology approach to evaluate mouse models of late-onset alzheimer's disease. *Molecular neurodegeneration*, 15(1):1–16, 2020.

[93] Lukas da Cruz Carvalho Iohan, Jean-Charles Lambert, and Marcos R Costa. Analysis of modular gene co-expression networks reveals molecular pathways underlying alzheimer's disease and progressive supranuclear palsy. *Plos one*, 17(4):e0266405, 2022.

[94] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):1–21, 2012.

[95] Sara Movahedi, Michiel Van Bel, Ken S Heyndrickx, and Klaas Vandepoele. Comparative co-expression analysis in plant biology. *Plant, cell & environment*, 35(10):1787–1798, 2012.

[96] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

[97] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1(1):1–17, 2007.

[98] Peter Langfelder and Steve Horvath. WGCNA package: Frequently Asked Questions. `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html`.

[99] Alexander Franks, Edoardo Airoldi, and Nikolai Slavov. Post-transcriptional regulation across human tissues. *PLoS computational biology*, 13(5):e1005535, 2017.

[100] Michael Brimacombe. Genomic aggregation effects and simpson's paradox. 2014.

[101] Benjamin D Harris, Megan Crow, Stephan Fischer, and Jesse Gillis. Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Systems*, 12(7):748–756, 2021.

[102] Li Yieng Lau, Antonio Reverter, Nicholas J Hudson, Marina Naval-Sanchez, Marina RS Fortes, and Pâmela A Alexandre. Dynamics of gene co-expression networks in time-series data: A case study in drosophila melanogaster embryogenesis. *Frontiers in genetics*, 11:517, 2020.

[103] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

[104] Roberto Anglani, Teresa M Creanza, Vania C Liuzzi, Ada Piepoli, Anna Panza, Angelo Andriulli, and Nicola Ancona. Loss of connectivity in cancer co-expression networks. *PloS one*, 9(1):e87075, 2014.

[105] Esra Gov and Kazim Yalcin Arga. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Scientific reports*, 7(1):1–10, 2017.

[106] Aurora Savino, Paolo Provero, and Valeria Poli. Differential co-expression analyses allow the identification of critical signalling pathways altered during tumour transformation and progression. *International Journal of Molecular Sciences*, 21(24):9461, 2020.

[107] Medi Kori, Esra Gov, and Kazım Yalçın Arga. Novel genomic biomarker candidates for cervical cancer as identified by differential co-expression network analysis. *Omics: a journal of integrative biology*, 23(5):261–273, 2019.

[108] Emma Pierson, GTEx Consortium, Daphne Koller, Alexis Battle, and Sara Mostafavi. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS computational biology*, 11(5):e1004220, 2015.

[109] Yongchun Zuo, Guanghua Su, Shanshan Wang, Lei Yang, Mingzhi Liao, Zhuying Wei, Chunling Bai, and Guangpeng Li. Exploring timing activation of functional pathway based on differential co-expression analysis in preimplantation embryogenesis. *Oncotarget*, 7(45):74120, 2016.

[110] Malay Bhattacharyya and Sanghamitra Bandyopadhyay. Studying the differential co-expression of micrornas reveals significant role of white matter in early alzheimer's progression. *Molecular BioSystems*, 9(3):457–466, 2013.

[111] Nicholas T Seyfried, Eric B Dammer, Vivek Swarup, Divya Nandakumar, Duc M Duong, Luming Yin, Qiudong Deng, Tram Nguyen, Chadwick M Hales, Thomas Wingo, et al. A multi-network approach identifies protein-specific co-expression in asymptomatic and symptomatic alzheimer's disease. *Cell systems*, 4(1):60–72, 2017.

[112] Fan Xu, Jing Yang, Jin Chen, Qingyuan Wu, Wei Gong, Jianguo Zhang, Weihua Shao, Jun Mu, Deyu Yang, Yongtao Yang, et al. Differential co-expression and regulation analyses reveal different mechanisms underlying major depressive disorder and subsyndromal symptomatic depression. *BMC bioinformatics*, 16(1):1–10, 2015.

[113] Fereshteh Izadi and Mohammad Hasan Soheilifar. Exploring potential biomarkers underlying pathogenesis of alzheimer's disease by differential co-expression analysis. *Avicenna Journal of Medical Biotechnology*, 10(4):233, 2018.

[114] David Amar, Hershel Safer, and Ron Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS computational biology*, 9(3):e1002955, 2013.

[115] Zhi Han, Jie Zhang, Guoyuan Sun, Gang Liu, and Kun Huang. A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules. *BMC genomics*, 17(7):303–315, 2016.

[116] Yasir Rahmatallah, Frank Emmert-Streib, and Galina Glazko. Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368, 2014.

[117] YounJeong Choi and Christina Kendziorski. Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21):2780–2786, 2009.

[118] Michael Watson. Coxpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7(1):1–12, 2006.

[119] Shunian Xiang, Zhi Huang, Tianfu Wang, Zhi Han, Christina Y Yu, Dong Ni, Kun Huang, and Jie Zhang. Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of alzheimer's disease patients. *BMC medical genomics*, 11(6):39–51, 2018.

[120] Aliakbar Hasankhani, Abolfazl Bahrami, Negin Sheybani, Behzad Aria, Behzad Hemati, Farhang Fatehi, Hamid Ghaem Maghami Farahani, Ghazaleh Javanmard, Mahsa Rezaee, John P Kastelic,

et al. Differential co-expression network analysis reveals key hub-high traffic genes as potential therapeutic targets for covid-19 pandemic. *Frontiers in Immunology*, 12, 2021.

[121] Tova F Fuller, Anatole Ghazalpour, Jason E Aten, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Weighted gene co-expression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6):463–472, 2007.

[122] Keping Chai, Xiaolin Zhang, Huitao Tang, Huaqian Gu, Weiping Ye, Gangqiang Wang, Shufang Chen, Feng Wan, Jiawei Liang, and Daojiang Shen. The application of consensus weighted gene co-expression network analysis to comparative transcriptome meta-datasets of multiple sclerosis in gray and white matter. *Frontiers in neurology*, 13:807349, 2022.

[123] Bruno M Tesson, Rainer Breitling, and Ritsert C Jansen. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics*, 11(1):1–9, 2010.

[124] Hui Yu, Bao-Hong Liu, Zhi-Qiang Ye, Chun Li, Yi-Xue Li, and Yuan-Yuan Li. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. *BMC bioinformatics*, 12(1):1–11, 2011.

[125] Jiexin Zhang, Yuan Ji, and Li Zhang. Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23(21):2903–2909, 2007.

[126] Duolin Wang, Juexin Wang, Yuexu Jiang, Yanchun Liang, and Dong Xu. Bfdca: A comprehensive tool of using bayes factor for differential co-expression analysis. *Journal of molecular biology*, 429(3):446–453, 2017.

[127] Dharmesh D Bhuva, Joseph Cursons, Gordon K Smyth, and Melissa J Davis. Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome biology*, 20(1):1–21, 2019.

[128] Dvir Netanely, Ayelet Avraham, Adit Ben-Baruch, Ella Evron, and Ron Shamir. Expression and methylation patterns partition luminal-a breast tumors into distinct prognostic subgroups. *Breast Cancer Research*, 18(1):1–16, 2016.

[129] Aurelie Tomczak, Jonathan M Mortensen, Rainer Winnenburg, Charles Liu, Dominique T Alessi, Varsha Swamy, Francesco Vallania, Shane Lofgren, Winston Haynes, Nigam H Shah, et al. Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific reports*, 8(1):1–10, 2018.

[130] Sonia Tarazona, Leandro Balzano-Nogueira, David Gómez-Cabrero, Andreas Schmidt, Axel Imhof, Thomas Hankemeier, Jesper Tegnér, Johan A Westerhuis, and Ana Conesa. Harmonization of quality metrics and power calculation in multi-omic studies. *Nature communications*, 11(1):1–13, 2020.