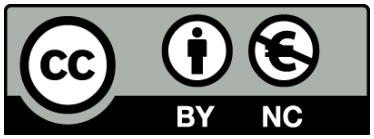
UNIVERSITAT DE BARCELONA

Network-based methods for biological data integration in precision medicine

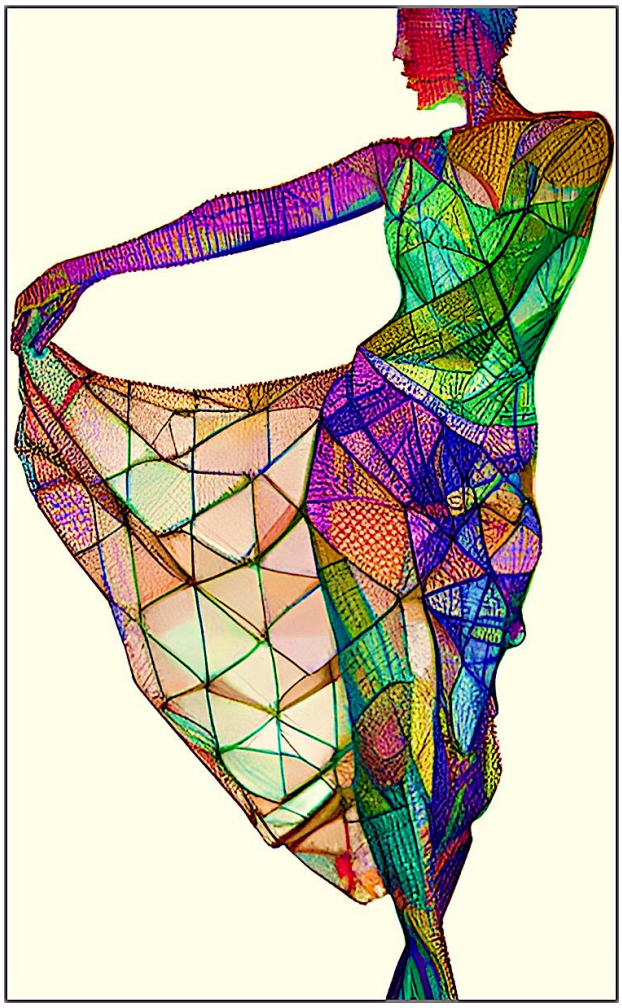
Iker Núñez Carpintero



Doctoral Thesis

Network-based methods for biological data integration in precision medicine

Iker Núñez Carpintero



Barcelona, November 2023

Department of Biochemistry and Molecular Biomedicine

Faculty of Biology

Doctoral program in Biomedicine

Bioinformatics research area

Network-based methods for biological data integration in precision medicine

Report presented by

Iker Núñez Carpintero

to qualify for the PhD degree at the Universitat de Barcelona

Supervised by

Dr. Davide Cirillo and Prof. Alfonso Valencia

Prof. Josep Lluis Gelpí

Tutor

Agradecimientos / Acknowledgements

Tengo que comenzar estos agradecimientos diciendo lo muy afortunado que soy. ¿Es un cliché comenzarlos así? Absolutamente, pero no por ello deja de ser rematadamente cierto. Tengo mucha suerte de estar rodeado de tanta gente que quiere y apoya de forma incondicional, sin la cual sería inconcebible el camino que he recorrido hasta ahora, desde que comencé el grado en Biología en 2013, hasta la finalización de esta tesis doctoral, 10 años más tarde.

Lógicamente los primeros agradecimientos han de ir a mi familia, de la que tengo la enorme suerte y orgullo de formar parte. **A mis padres,** por haberme animado a disfrutar de la vida en todo momento, por todo el cariño y los buenos momentos disfrutados (y los que quedan, ¡ya queda bien poco para Malmö 2024!) y, sobre todo, por el enorme esfuerzo (tanto psicológico como económico) que sé os ha supuesto darme la oportunidad de llegar hasta aquí, siempre anteponiendo mi futuro por delante de cualquier otra cosa. Gracias de todo corazón.

A mi hermano **Jon**, que siempre ha sido, es y será el mejor hermano del mundo y mi mejor amigo. Siempre habrá espacio en esta vida para pasarnos el día viendo otro Mundial, Eurocopa o JJOO o echando un partido más de ese torneo de FIFA por terminar. ¡Siempre nos quedará el chino!

A **Pedro**, que siempre nos ha tratado a mi hermano y a mí como hijos, y que para mí no sólo es como un padre más sino un referente. A **Raquel**, por enseñarme lo interesante que era el mundo desde el primer día (literalmente), la mejor hermana adoptada del mundo, y a su maravillosa familia (**Chiqui, Clara y Nil**). A **Rosa**, **Pablo, Martín** y **Ana**, quiénes me acogieron como uno más de la familia cuando más lo necesitábamos, haciendo posible que hoy esté aquí. Todos ellos saben dónde está, y donde estará siempre, su casa.

Quiero agradecer a todos esos amigos que independientemente del día, lugar y tiempo pasado sé que están y estarán siempre ahí para contarnos nuestras

peripecias como si nos hubiésemos visto ayer. A **Eva**, a quien considero como una hermana, por la enorme capacidad que tiene para escuchar y el cariño que da a todo el mundo. A **Pablo**, que es prácticamente parte de mi familia y a **Jose María**, que es tan grande como su corazón. A **Rosa, Muzzy y Vicente**, por la confianza, el cariño y las muchas risas compartidas, a **David**, siempre dispuesto a hablar y ver otro festival de Eurovisión, llenos de EUPHORIA. A **Alejandro** y **Adrián** por las muchas risas y conversaciones existenciales durante los cafés de los sábados, jugando la FUT Champions y pasando las tardes en las valladas. Mi primer año viviendo solo en Barcelona no habría sido igual sin esos momentos. A mis amigos de la universidad, mi primo **Carlos** y a su familia, quienes siempre me han acogido y tratado como uno más de la casa, a **Sebas, Luis, Jose y Marco**, por la sinceridad y honestidad. A **Sergio Talens y** a **Silvia**, con quiénes estoy deseando de ir ya a Malmö el próximo año. A **María**, compañera también en Barcelona. De mi tiempo como estudiante en el Centro Nacional de Biotecnología, también quiero agradecer enormemente a **Mónica**, **Sito** y **Javier**, con quiénes descubrí el maravilloso mundillo de la bioinformática, y de mi tiempo como estudiante de máster a **Eduardo Arranz**, con quien hacer el TFM fue tan sencillo como agradable.

Y con esto (sólo ha costado una página y media) llego a mi tiempo en el BSC. Primero de todo, quiero agradecer a **Alfonso** el haberme dado la oportunidad única de hacer el doctorado en su grupo, y, sobre todo, la enorme capacidad que tiene para darte el consejo que necesitas, en el momento en que lo necesitas. Estos cinco años me han cambiado por completo la vida, sin parar de conocer gente tan fantástica como brillante.

A **Carlos**, probablemente el mejor amigo que me llevo de estos años, a **Hugo**, quién siempre sabe cómo disfrutar de la vida, a **Jon**, por los muchos consejos (y a **Ester** por siempre acogernos con una sonrisa) y por su capacidad para escuchar, a **Victoria**, mi compi de doctorado, por su sinceridad, a **Alba** por el SHUM, a **François** por ser el más bondadoso troll de todos los tiempos, y a todos aquellos que me acogieron al llegar a Barcelona y que me han ayudado y visto crecer no sólo como

científico sino como persona (**Eva**, **Maria**, **Juan**, **Arnau**, **Miguel**, **Eduard**, **Mónica**, **Laure**) a lo largo de estos años. A aquellos que se han ido uniendo a este fantástico grupo más tarde, a **Bea**, por su enorme bondad, a **Jose Carbonell**, **Thaleia** y **Othmane** por las muchas risas y conversaciones random, y a todos los demás miembros del grupo, quiénes son (o han sido en algún momento) parte de esta, gran familia (**Camila**, **María Morales/Kenny**, **Léo**, **Jose Estragués**, **Gonzalo**, **Jorge**, **Paula**, **Xavi**, **Guillermo**, **Fatemeh**, **Marta**, **Miguel Romero**, **Sofía**, **Maxim, Nur, Amhar**, **Alejandro**, **Iria**, **Miriam, Alicia, Núria Saavedra**, **Adam, Asal**).

Dejo lo mejor, claro está, para el final. Primero, a **Núria,** no sólo por el cariño, el apoyo y los buenos momentos durante estos años, sino por ser la mejor compañera de vida que hubiera podido imaginar. Uno nunca puede presagiar lo que nos deparará el futuro, pero espero que los momentos que hemos vivido juntos hasta ahora sólo sean el comienzo de lo que está por venir. También a su familia, por haberme aceptado y acogido desde el primer minuto como uno más.

Segundo y último, a la persona que probablemente más reconocimiento merece en esta historia: **Davide**. Tanto para aquellos que están ya haciendo un doctorado con él (**Bea, Guillermo**, **Fatemeh**) como el futuro lector que se plantee la posibilidad de hacerlo, sólo puedo deciros que para mí siempre será un honor el haber podido ser el primer doctorando de alguien tan brillante, tanto científicamente como en el plano personal, y que habéis acertado por completo al elegir. Como estudiante, uno sólo puede desear haber estado a la altura de alguien que merece tanto lo que le pasando en la actualidad. Davide no sólo es un verdadero referente, sino que, ante todo, es un amigo.

¡UF! Sí que era gente a la que agradecer. Espero no haberme dejado a nadie. A todos vosotros y a los que estén por venir, gracias y Cha Cha Cha, de todo corazón.

Iker

Table of Contents

Abstract

The vast and continuously increasing volume of available biomedical data produced during the last decades opens new opportunities for large-scale modeling of disease biology, facilitating a more comprehensive and integrative understanding of its processes. Nevertheless, this type of modelling requires highly efficient computational systems capable of dealing with such levels of data volumes.

Computational approximations commonly used in machine learning and data analysis, namely dimensionality reduction and network-based approaches, have been developed with the goal of effectively integrating biomedical data. Among these methods, network-based machine learning stands out due to its major advantage in terms of biomedical interpretability. These methodologies provide a highly intuitive framework for the integration and modelling of biological processes.

This PhD thesis aims to explore the potential of integration of complementary available biomedical knowledge with patient-specific data to provide novel computational approaches to solve biomedical scenarios characterized by data scarcity. The primary focus is on studying how high-order graph analysis (i.e., community detection in multiplex and multilayer networks) may help elucidate the interplay of different types of data in contexts where statistical power is heavily impacted by small sample sizes, such as rare diseases and precision oncology.

The central focus of this thesis is to illustrate how network biology, among the several data integration approaches with the potential to achieve this task, can play a pivotal role in addressing this challenge provided its advantages in molecular interpretability. Through its insights and methodologies, it introduces how network biology, and in particular, models based on multilayer networks, facilitates bringing the vision of precision medicine to these complex scenarios, providing a natural approach for the discovery of new biomedical relationships that overcomes the difficulties for the study of cohorts presenting limited sample sizes (**data-scarce scenarios**).

Delving into the potential of current artificial intelligence (AI) and network biology applications to address data granularity issues in the precision medicine field, this PhD thesis presents pivotal research works, based on multilayer networks, for the analysis of two rare disease scenarios with specific data granularities, effectively overcoming the classical constraints hindering rare disease and precision oncology research.

The first research article presents a personalized medicine study of the molecular determinants of severity in congenital myasthenic syndromes (CMS), a group of rare disorders of the neuromuscular junction (NMJ). The analysis of severity in rare diseases, despite its importance, is typically neglected due to data availability. In this study, modelling of biomedical knowledge via multilayer networks allowed understanding the functional implications of individual mutations in the cohort under study, as well as their relationships with the causal mutations of the disease and the different levels of severity observed. Moreover, the study presents experimental evidence of the role of a previously unsuspected gene in NMJ activity, validating the hypothetical role predicted using the newly introduced methodologies.

The second research article focuses on the applicability of multilayer networks for gene priorization. Enhancing concepts for the analysis of different data granularities firstly introduced in the previous article, the presented research provides a methodology based on the persistency of network community structures in a range of modularity resolution, effectively providing a new framework for gene priorization for patient stratification.

In summary, this PhD thesis presents major advances on the use of multilayer network-based approaches for the application of precision medicine to data-scarce scenarios, exploring the potential of integrating extensive available biomedical knowledge with patient-specific data.

List of acronyms

AChR	Acetylcholine receptor
AI	Artificial intelligence
BioGRID	Biological General Repository for Interaction Datasets
CHIP-seq	Chromatin Immunoprecipitation Sequencing
CMS	Congenital Myasthenic Syndromes
COPD	Chronic obstructive pulmonary disease
EGA	European Genome-phenome Archive
EHRs	Electronic Health Records
EMBL-EBI	European Bioinformatics Institute
ENCODE	Encyclopedia of DNA Elements
EWDs	Electronic wearable devices
GEO	Gene Expression Omnibus
GTEx	Genotype-Tissue Expression
GtRNAdb	Genomic tRNA Database
GWAS	Genome wide association studies
HMP	NIH Human Microbiome Project
HPA	Human Protein Atlas
HPC	High-Performance Computing
HPO	Human Phenotype Ontology
IID	Integrated Interactome Database

KEGG	Kyoto Encyclopedia of Genes and Genomes
MB	Medulloblastoma
MCIA	Multiple co-inertia analysis
miRbase	The microRNA database
ML	Machine learning
MOFA	Multi-Omics Factor Analysis
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NLP	Natural Language Processing
NMF	Joint Non-negative Matrix Factorization
NMJ	Neuromuscular Junction
OMIM	Online Inheritance in Man
PCA	Principal Component Analysis
PDB	Protein Data Bank
PPI	Protein-protein interaction
RNA-seq	RNA sequencing
SNF	Similarity Network Fusion
SNVs	Single nucleotide variations
TCGA	The Cancer Genome Atlas
VAEs	Variational Autoencoders
VMH	Recon3D Virtual Metabolic Human
XAI	Explainable artificial intelligence

Chapter 1

Introduction

1. Overcoming data scarcity in precision medicine

Understanding the key factors behind clinical human disease manifestations requires knowledge of the various molecular aspects that may be involved. With the advent of the Big Data revolution, the amount of biomedical information available to study is growing in a constant manner, leading health research and practice to a new and promising era (1).

The potential of such wealth of biomedical data has greatly benefited the field in recent years (2) provided the main purpose of its use: contributing to the modelling of the processes triggering human disease. This increase in biomedical data availability raises a considerable number of challenges to be addressed from the medical point of view (3). Among them, the main challenge consists in the creation of methods that are able to integrate large volumes of heterogeneous information in order to enable the personalized analysis of patient data, the so-called precision medicine (4). Precision medicine though, is a generic term that simplifies the essential objectives of contemporary approaches to understand disease dynamics. Precision medicine aims to shift the medicine paradigm towards a more predictive, preventive, personalized and participatory model ('**P4' medicine**) (5,6). This ongoing revolution in the medical field seeks to unravel the biological processes underlying individual biomedical problems, potentially enhancing its social and clinical impact (e.g., reducing healthcare costs and providing earlier disease detection). In addition to typical biomedical priorities such as disease diagnosis (7,8) and classification (**disease subtyping**) (9,10), precision medicine aims to solve distinct relevant challenges, namely prediction and prevention of clinical outcomes (2,11,12) and the identification of potential patient-specific therapeutical targets (13).

The realization of this new vision for medicine is dependent on the development of integrative and cost-effective methods for the analysis of biomedical knowledge. Indeed, integrative precision medicine studies have proven highly valuable for the analysis of some of the most investigated human conditions in recent years, including

cancer (14), Chronic obstructive pulmonary disease (COPD) (10) and COVID-19 (15). However, the implementation of precision medicine to rare diseases and rare cancers, such as pediatric tumors is classically hindered due to **data scarcity** (16,17).

Rare diseases are defined by their low prevalence within the population. Within the European community, a disease is cataloged as rare if the prevalence is below 5 out 10000 individuals. However, despite their low prevalence, rare diseases actually impact a significant number of people, accounting for up to 7% of the world's population. Unfortunately, most rare diseases (approximately 95%) lack proper described treatments (18). On top of the small fraction of these disorders that has known efficient treatments, some patients may not experience improvement due to the multiple factors driving the heterogeneous manifestations of these diseases. This way, rare disease research means dealing with unique disease biomedical scenarios where precision medicine becomes even more critical (19).

The overall lack of treatments can be attributed to the difficulties in understanding the molecular drivers of rare disease biology, which stem from the limitations in cohort recruitment. Availability of patients is extremely low, and their geographical distribution is highly dispersed (20). In this context, collection of data requires high-level domestic (21) and international collaborative efforts, which become critical for the creation of complete biomedical registries. Initiatives like RD-Connect (22) and individualizedPaediatricCure (https://ipc-project.eu/) foment the integration of biobank platforms for the analysis of this uncommon conditions. Precision medicine aims for the holistic modelling of the relationships among multiple levels of biological data. Data scarcity represents a major obstacle for achieving the effective application of precision medicine to rare diseases. Moreover, overcoming this problem is the key to approach the analysis of biomedical information at the different levels at which it is collected (**data granularity**). Fine-grained data contains detailed and specific information, while coarse-grained data is more generalized and aggregated. To be effective, precision medicine demands a deeper level of granularity to harness its potential for truly personalized healthcare solutions.

This comprehensive view of biomedical knowledge can only be accomplished through computational approaches. As availability of biomedical data increases, so does the importance of computational biology (the study of biological systems by computational means) and bioinformatics (the generation of computational tools for the processing and analysis of biological data).

In terms of providing ways to integrate complex biological data, computational biology relies on the power of machine learning (ML) approaches. The increasing application of ML and artificial intelligence (AI) in biomedicine is fostering the development of new methods that enable explaining opaque models. As a result, ML and AI have become subject of huge interest for the biomedical community during the last decade (23). Nevertheless, such approaches largely benefit from sample sizes greater than feature dimensions, a requirement that is difficult to be met in the biomedical filed, where typically a large number of characteristics are measured and collected for relatively small cohorts (24). For example, approaches based on neural networks easily overfits datasets of small dimensions (25). This **'curse of dimensionality'** greatly affect the a direct applicability of ML and AI in many types of biomedical studies (26).

One promising way to alleviate the limitations coming from data scarcity is to leverage the vast amount of biomedical knowledge available in publicly accessible databases (26,27), which offer a great opportunity to identify different kinds of associations among the limited elements of patient information at hand. In the case of molecular information, this, in turn, yields valuable insights for the interpretation of the underlying biology of the disease under study (**Figure 1**). Confronting data scarcity entails addressing scenarios characterized by small sample sizes, a major factor that significantly impacts statistical power. By simultaneously analyzing the external and complementary biomedical knowledge, it is possible to reveal new relevant relationships that would not be uncovered by using the patient data alone. Moreover, this vision opens the door to a more informed analysis of variations among individuals across different levels of biomedical data.

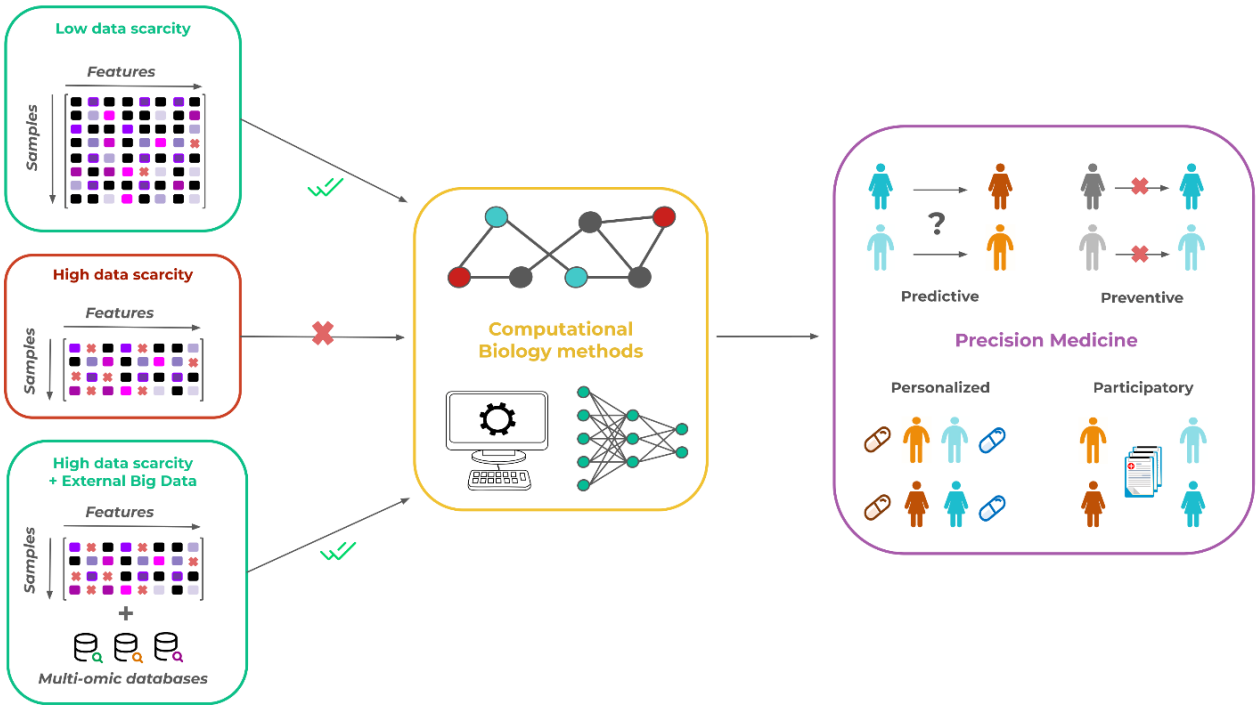


Figure 1. The data scarcity challenge. In situations where patient data availability is limited, generation of computational biology approaches becomes challenging, particularly in the context of precision medicine. To overcome this issue, a solution lies in combining patient data with the vast biomedical knowledge accumulated in available resources. The additional information coming from these resources can help identify associations that enable the application of these methodologies.

Nonetheless, a significant challenge arises due to the diverse nature of such data (e.g., genetics, gene expression, etc.), consisting of effectively managing the inherent differences in this comprehensive information, in order to draw conclusions encompassing the different data layers.

Tables 1 and 2 provide an overview of several data types commonly investigated in precision medicine studies. **Table 1** focuses on omics data, which refers to high-throughput biochemical assays that measure molecules of the same type from a biological sample (28). The joint study of multiple omics data levels (**multi-omics data integration**) has emerged as a crucial practice in understanding the complex interactions between genes, proteins, metabolites, and other quantifiable molecules.

Discipline	Target information	Resources
Genomics	***DNA variation data & chromatin structure*** • WGS and WES • Multiple variation types (e.g. SNVs, CHVs, CNVs)	**ClinVar** **NCBI** **DisGeNET** **OMIM** **TCGA**
Transcriptomics	***Gene expression*** • RNA-seq & scRNA-seq • Multiple RNA functionalities: ○ mRNA, miRNA, siRNA, rRNA, tRNA	**GEO** **Expression atlas** **GTEx** **GtRNAdb** **miRbase**
Proteomics	***Protein dynamics*** • Mass spectrometry & Protein microarrays • Protein-Protein interaction (PPI) • Structure • Cell-specificity & functionality	**UniProt** **HPA** **BioGRID** **IID** **PDB** **AlphaFold**
Metabolomics	***Metabolite dynamics & chemical reactions*** • Mass spectrometry • Metabolite fluxes and levels in cells and tissues	**KEGG** **Reactome** **VMH**
Epigenomics	***Chemical modifications modulating DNA activity*** • ncRNAs • DNA methylation and acetylation • Chromatin accessibility	**ENCODE** **GTEx** **GEO**
Metagenomics	***Organisms inhabiting human ecological niches*** • Involvement in normal human physiology (e.g. immune cell maturation) • Multiple niches (e.g. Oral, gut, skin)	**Human Microbiome Project** **microbioTA**

Table 1. **Omics data disciplines and their targeted biomedical information.** Omics data includes information from measurable high-throughput biochemical assay, including DNA variation and chromatin structure conformation (**genomics**) (30–34), gene expression and RNA biology (**transcriptomics**) (30,35–41), protein information regarding structural data, interactions with other proteins and cell-specific dynamics (**proteomics**) (42–52), chemical reactions and their cellular roles (**metabolomics**) (53–55), other modifications capable of modulating normal DNA activity (**epigenomics**) (56–58) and the key activities played by normal microorganism flora over different human ecological niches (**metagenomics**) (59–69).

Discipline	Target information	Resources
Imaging data	***Biomedical images*** • Radiology & Magnetic resonance imaging • Histopathology • ECG	**Imaging Data Commons** **MedMNIST**
Treatments	***Drugs and other chemical compounds*** • Potential new drug targets (**Drug discovery**) • Potential alternative usage (**Drug repurposing**) • Known compound with potential usage (**Drug screening**)	**DrugBank** **PubChem** **ChemBL** **DrugMAP**
Epidemiology & EHRs	***Demographic annotation for the study of disease dynamics*** • Disease triggering life habits • Phenotypic traits • Clinical signs	**OMIM** **Human Phenotype Ontology**
Pathways	***High-level arrangement of biochemical events*** • Metabolic routes & their regulative processes • Cell signaling • Membrane transport	**KEGG** **Reactome** **WIKIPathways**
Electronic wearable devices	***Non-invasive physiological recordings*** • Holters • Smart watches & Smart clothing	**PTB-XL** **Physionet**

Table 2. Non-omics data disciplines and their targeted biomedical information. Non-omics data sources covered by state-of-the-art research include images from radiological and magnetic resonance studies (**imaging data**) (70–74), drug administration and their molecular targets (**treatments**) (75–80), clinically pertinent demographic annotations (**epidemiology and electronic health records -EHRs-**) (81–83), the arrangement of related sets of biochemical events (**pathways**) (84–87) and non-invasive physiological measurements (**electronic wearable devices**) (88–93).

This integration allows researchers to explore the underlying biology of diseases by contextualizing these interactions within unified frameworks. Recognizing the potential of providing a comprehensive view of omics information, scientists are now devoted to the development of complete omics resources (29).

In contrast, **non-omics data encompasses** a spectrum of biomedical information that is not acquired through high-throughput biochemical assays (**Table 2**).

Integrating these diverse non-omics data types presents an even more significant challenge compared to omics data integration (94), given their diverse characteristics and frequent utilization within specialized communities (95).

In this sense, **the integrative view provided by systems biology** is crucial for the development of novel methodologies to deal with such large and diverse amounts of biomedical information and it **is a central topic of this PhD thesis**. One major goal of systems biology consists in understanding how the different components of each biological system can be represented, integrated, and analyzed in a unifying manner. The application of systems biology methods to medicine (**Systems medicine**) tries to model the biological aspects driving disease manifestations, giving robust interpretations of their dynamics both at the phenotypic and the molecular level (96).

However, the simultaneous integration of multiple biomedical levels comes with a significant number of challenges (97). Summarizing, a number of these challenges are related to the scarcity of required information (e.g., handling of missing values and class imbalance), while others have to do with the inherent nature of the different biomedical variables (e.g., joint analysis of heterogenous data types and information resources). In this regard, **network biology** solutions (98) are becoming increasingly valued as they considerably ease the representation and integration of these multiple data types and the transparency of systems biology models (99).

Indeed, the challenges related to data scarcity and data granularity can be tackled by making use of the capability of these methods to intuitively integrate patient information with complementary biomedical knowledge from external resources. This PhD thesis aims to provide novel integrative network biology methods for applying precision medicine to data-scarce biomedical contexts, in an effort to contribute to the personalized analysis of the complexities displayed by these cases. Furthermore, it explores the potential of such methodologies for the efficient analysis of the multiple granularity levels existing within biomedical data, specifically providing solutions for scenarios characterized by data scarcity.

2. Integrative approaches in precision medicine

So far, we have discussed how data scarcity hinders the application of precision medicine, particularly in the context of rare disease research. Furthermore, we have introduced how the integration of patient data and complementary biomedical knowledge from external resources holds great potential for overcoming these limitations, while gaining a holistic understanding of the disease. This fact raises several important questions: How can we address simultaneously all these meaningful biomedical aspects? How should we account for all interactions and relationships among the different layers? This is where **data integration** comes into play.

By combining multiple layers of omics and non-omics data it is possible to overcome potential biases and discover biomedical relationships that are not apparent when analyzing datasets with limited content and scope. Over the last decade, diverse integrative data approaches have been developed (100), ranging from consecutive analysis of relevant omics data to more complex approaches accounting for the synergisms existing between the different biological layers (24).

Integrating biomedical data presents various challenges due to its high dimensionality but typically limited sample size. Therefore, a primary objective of data integration for precision medicine is to provide methodologies to reduce such complexity (i.e., **data-driven feature selection**) (24).

This section introduces several state-of-the-art approaches commonly applied in data integration and their successful application into biomedical research. Provided the biological focus of the thesis dissertation, a selection of in-depth review articles that explore the formal description of these methods is provided for the interested reader: Cantini et al. 2021 (100), Pierre-Jean et al., 2020 (101), Huang et al., 2017 (102). Instead, we will emphasize the successful applications of each approach, keeping an eye on their interpretability capabilities. Additionally, we comment on the current state of application of these methodologies to data-scarce scenarios.

2.1 Dimensionality reduction methods

Integrative dimensionality reduction methods aim to project the various datasets of interest into lower-dimensional spaces. By merging these datasets into common representations, these methodologies allow for the detection of coherent patterns among the different data levels, effectively providing a reduced number of representative features (**Figure 4**).

The majority of dimensionality reduction algorithms for the analysis of biomedical data are **unsupervised learning** techniques, meaning that they do not rely on labelled target variables for training. Among these techniques, latent variable approaches have gained popularity due to their effectiveness and versatility. However, while commonly applied machine learning dimensionality reduction techniques are able to generate useful and accurate representations of the data (103), they often lack intuitive ways to extract interpretable knowledge from the generated latent variables. This is currently one of the main challenges to be addressed in the field (104).

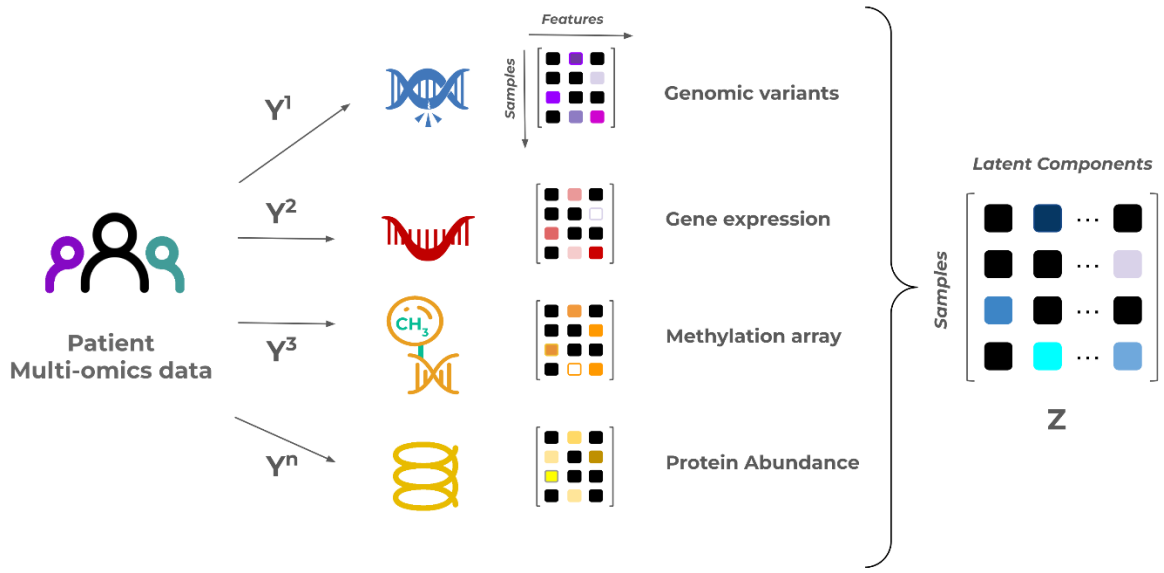


Figure 2. Basic workflow of an integrative dimensionality reduction approach. Measurements from multiple biomedical data types can be represented in the format of a matrix ($\mathbf{Y^n}$). Dimensionality reduction techniques are then applied to obtain common reduced mathematical spaces, with a reduced number of **latent components** (**Z**) that capture consistent patterns across the different data types.

2.1.1 Joint Non-negative Matrix Factorization (NMF)

Matrix factorization models map cases and features to a joint latent factor space of lower dimensionality (105). As a framework for medical data integration, NMF assumes the existence of a common basis matrix between the two decomposed factors of each data type measurement (**latent variables**). The objective function of NMF optimizes the difference between the original data matrices, their specific matrix and the common latent factor, assuming non-negative constraints (106). Despite its computational requirements, NMF has become a very popular dimensionality reduction approach, particularly in the field of recommender systems. Major e-commerce leaders, such as Netflix or Amazon, use NMF-based methods to provide product recommendations, benefiting from their high performance (107).

In the biomedical context, one of the most intuitive applications of matrix factorization is the prediction of drug-disease associations (108). Other recent biomedical implementations include the integration of multi-omic single-cell datasets (109) and the analysis of oncological cohorts (110). NMF has additionally shown high performances in sample clustering problems (111,112), demonstrating its effectiveness in these domains.

Regarding its application to rare disease research, an interesting application of matrix factorization is MultiPLIER (113). Making use of gene expression datasets, MultiPLIER provided valuable latent factors for the analysis of rare disease dynamics, namely antineutrophil cytoplasmic autoantibody-associated vasculitis (a rare autoimmune disease) and medulloblastoma, a rare childhood cancer.

2.1.2 Multiple co-inertia analysis

Multiple co-inertia analysis (MCIA) is a two-step data integration process that identifies co-relationships between multiple datasets (114). The first step of MCIA consists of applying a table ordination method, such as Principal Component Analysis (PCA), to transform the data layers into matrices of similar dimensions.

By assessing the relative contribution of each individual to each dataset, MCIA maximizes the covariance between the different matrices, resulting in a space of a single latent variable. Interpretation of MCIA-based analysis can be challenging (115). However, applications for data visualization have proven promising (116).

Other recent effective MCIA applications to the multi-omics field include taxonomic analysis of human gut microbiome populations (117) and disease modelling in Chron's disease (118). While the literature on MCIA's application to rare disease scenarios is considerably limited (119), its promising results on the analysis of cancer data (100) offer a positive outlook for its potential.

2.1.3. iCluster and iCluster+

iCluster is an integrative framework specifically designed for disease subtyping (120). Like NMF, iCluster assumes the existence of a common latent variable that connects the different datasets, but without requiring non-negative inputs. In addition, iCluster adds an independent error factor for each data type, capturing the variances after accounting for inter-dataset correlations. However, the computational complexity of this methodology makes selecting the final model a complex task (121).

The iCluster framework uses a likelihood-based formulation to estimate the latent features. Disease subtypes are determined using k-means clustering based on the joint latent variable matrix. The upgraded version of iCluster, iCluster+, allows for the combination of generalized linear modelling of heterogeneous data types assuming varied distributions (121). This extension enables the simultaneous integration of binary, continuous and categorical data.

The original iCluster algorithm demonstrated its efficacy in discovering disease subtypes in breast and lung cancer (120). iCluster+ was originally applied on a colorectal cancer cohort from TCGA, helping detect 2 new subtypes of the disease. Concerning rare diseases, disease subtyping using iCluster has been successfully applied to hepatocellular carcinoma (HCC) (122) and glioblastoma (123).

2.1.4. Multi-Omics Factor Analysis (MOFA)

Multi-Omics Factor Analysis (MOFA) was introduced in 2018 by Argelaguet et al., as a generalization of PCA for multi-omics data analysis (124). MOFA follows a similar matrix decomposition approach as iCluster+, assuming the existence of a latent common variable shared across all data matrices, along with weight and error matrices specific to each data type. However, instead of using a likelihood-based formulation to obtain these variables, MOFA assumes randomness for these matrices, formulating the model as a probabilistic Bayesian Framework, placing prior hierarchical distributions to all unobserved variables of the model to initiate the learning process.

One notable feature of MOFA is its flexibility in supporting a variety of omics-specific distributions for the error variables, rather than assuming normal distributions. MOFA also provides a model regularization technique that assesses the degree to which factors are specific to each single data layer, enabling a more detailed analysis of the individual contributions to the whole system. MOFA has been applied for the identification of clinical markers for chronic lymphocytic leukemia (CLL) (125) and pathways related to the rare disease methylmalonic aciduria (126), showcasing its potential in uncovering meaningful insights also in data-scarce scenarios.

2.1.5. Neural Network Autoencoders

AI research has placed strong emphasis on neural network-based models, which serve as the foundation for various applications. A neural network operates as a graph-like system, consisting of three types of node layers: an input layer that receives the values of the dataset variables, hidden node layers, that vary the dimensionality of the original input layer, and an output layer that encodes the different possible outcomes. Nodes within each layer are interconnected with nodes in the previous and the following layer, with a given weight value.

Neural network systems are trained using the backpropagation algorithm, a supervised learning technique that iteratively updates the interlayer weights during

training. During the learning steps, neural networks aim to minimize a designated loss function (127), which serves as an evaluation metric, adjusting the weights while assessing the performance of the model.

Variational Autoencoders (VAEs) have raised as a promising methodology for the task of multi-omics dimensionality reduction (128). VAEs present a unique architecture, consisting of two neural networks. The first network, known as the encoder, contains hidden layers that perform the dimensionality reduction by recursively reducing the set of nodes between each layer, ultimately producing a limited set of nodes as the output of the network. The second network, the decoder, is trained to accurately recover the original input from the encoder. The minimized output of the encoder system represents a new low-dimensional latent representation of the input data that encapsulates the information to reconstruct the input data (129).

VAEs, like other dimensionality reduction techniques, offer useful modelling solutions for multi-omics tasks due to their accuracy and efficiency. They have been successfully applied in various areas, including classification of tumor subtypes (130), clinical disease endotyping (131) and the identification of interactions between long non-coding RNAs and protein-coding genes (132). As for rare diseases, a recent application of VAEs has been the prediction of patient severity state in glioblastoma from medical images (123).

2.2. Network-based methods

Up until this point, we have discussed various dimensionality reduction techniques, putting a focus on their application to the biomedical field. However, one notable limitation of these methods is their complexity in interpreting the encoded biomedical knowledge within the obtained latent variables. As an alternative to latent representation methods, models based on graphs offer an intuitive way of depicting the relationships underlying biological systems (133,134). **Network biology** aims to model biological entities, such as genes and proteins, as nodes linked by edges that signify their relationship in specific biological contexts. Networks have become

extremely popular in biology due to the versatility that they provide for the representation of knowledge, while also facilitating the analysis of biological relationships. Topological analysis of biological networks are characterized by an inherently interpretable way to model and identify relationships between concepts. For example, detection of network communities (i.e. densely connected nodes) can reveal disease-related patterns (27,135).

Despite their inherent interpretability, network biology approaches are computationally demanding compared to other ML approaches. Biological networks often involve a vast number or relationships, leading to increased computational complexity and a major necessity of scalable methods (136,137).

2.2.1. Similarity Network Fusion (SNF)

SNF is an edge-prioritization methodology for data aggregation based on finding a common similarity network from a set of similarity networks representing multiple biological aspects (138). The algorithm starts from calculating pair-wise similarity matrices for each data type, rendering weighted networks where each sample is a node, and the similarities are codified as edge weights (**See section 4, Figure 5C-D**). The following step consists in the fusion of the networks, which occurs iteratively through message passing events. Each network is updated in a stepwise manner to resemble the others, eventually converging into a single ‘fused’ network. This process prioritizes strong node similarities, keeping low-weight edges shared across all graphs.

The similarity matrix serves as input for sample clustering, providing the contribution of each data source in determining patient proximity. SNF was first applied to several TCGA cohorts (138), and recent remarkable applications include predicting clinical outcomes in neuroblastoma (139) and integrating multi-omics data in respiratory conditions (140) such as Bronchiectasis (141) and COPD (10). Furthermore, application of SNF helped in disease subtyping of a group of rare diseases known as idiopathic inflammatory myopathies (142).

SNF-based methods have exhibited a significant potential in recent years (143), but also revealed a common issue in most integrative approaches. The original version of SNF (139) excludes patients and features from the analysis if they have > 20% missing data in each data type. While this is an arguably understandable decision, major questions arise: How can we handle missing data to prevent potential biases? Dealing with missing data is one of the topics explored in the research article presented in Chapter 4.

2.2.2. Graph Embeddings

Section 2 is structured making a distinction between dimensionality reduction methods and network-based techniques as data integration frameworks. However, we should note that both concepts are not mutually exclusive. Graph embedding approaches are widely popular techniques for dimensionality reduction.

Graph embeddings aim to simplify the analysis of large-scale networks, where mathematical magnitude of adjacency matrices (**See Section 4, Figure 5A-B**) becomes high-dimensional. Embeddings are reduced vectorial spaces packing graph features, favouring the scalability of topological network analysis.

This methodology draws inspiration from **Word2vec**, a well-known Natural Language Processing (NLP) approach for word prediction (144). Given a set of phrases, the model predicts the probability of each known word to be next one in the sentence. The underlying neural network architecture, known as **skip-gram neural network**, is formed by three layers: a binarized input layer of the size of all known words, a hidden vector layer with a predefined number of features and an output layer of the same dimensionality as the input layer. This output layer returns the probability of each node being the next in the sentence.

For dimensionality reduction, the skip-gram model can be used for embedding both vertices and entire graphs, with the hidden layer becoming the reduced feature space to be found. A classic example of this is DeepWalk (145), where random walks (**See Section 4.1, Figure 6E**) serve as “sentences” and the nodes act as words.

Recent advances have resulted in the extension of random walk-based embeddings to complex networks (that we will explore in **Section 4.2.**) with applications such as OhmNet (146) and MultiVERSE (147). The latter has been successfully applied for the study of rare disease-gene associations.

3. Network Biology: an interpretable framework for biomedical data representation and integration

Network biology is the research discipline focused on studying complex biological systems by means of graph theory, a branch of mathematics that focuses on the study of relational data structures (**graph** or **network**). A network is formed by a set of entities (**nodes** or **vertices**) and the relationships between them (**edges**). Graph theory is widely applied in various fields, including network analysis (148), computer science and physics (149). Graphs provide a versatile framework for data representation, making them suitable for both simple and complex problems across multiple domains. In the biological context, graph theory is particularly well-suited, as it can effectively represent multiple biological concepts as nodes, and their relationships as edges (150).

In a simple graph two nodes are connected by one edge (**Figure 5A**). A simple graph can also be represented in the form of a binary **adjacency matrix** (**Figure 5B**). Interactome networks, which we briefly mentioned in Section 2.3., **Proteomics**, are generally depicted as simple graphs. These graphs depict sets of proteins (or protein subunits) as nodes, while the edges represent physical interactions between such proteins. Such a representation of an interactome network provide a subtle way to identify promiscuous interacting proteins, but assessing the significance of these interactions becomes unfeasible unless **edge weights** are included, resulting in **weighted networks** (**Figure 5C**).

Weighted networks assign values (**weights**) to each edge, allowing for the prioritization of certain connections over others. A gene expression correlation network is an example of a biological weighted network. By computing the correlation values between pairs of expression arrays for a given set of samples, a non-binary adjacency matrix is obtained, where each matrix entry corresponds to the weight (i.e., correlation) between the nodes (**Figure 5D**).

Directionality can also serve as a valuable means to convey additional information within a network. A **directed network** (**Figure 5E**) includes edges coming from root nodes to target nodes. Ontologies and hierarchical networks are examples of directed graphs, where a *child* concept will be connected to its *parent* with a directed edge, which, may or not be weighted. The corresponding adjacency matrix of directed networks is therefore non-symmetric (**Figure 5F**). In a further complex level, graphs may include multiple edges with the same involved nodes (i.e., **incident nodes**): **multigraphs.** Multigraphs offer the possibility of representing edges of different nature between two nodes (**Figures 5G and 5H**).

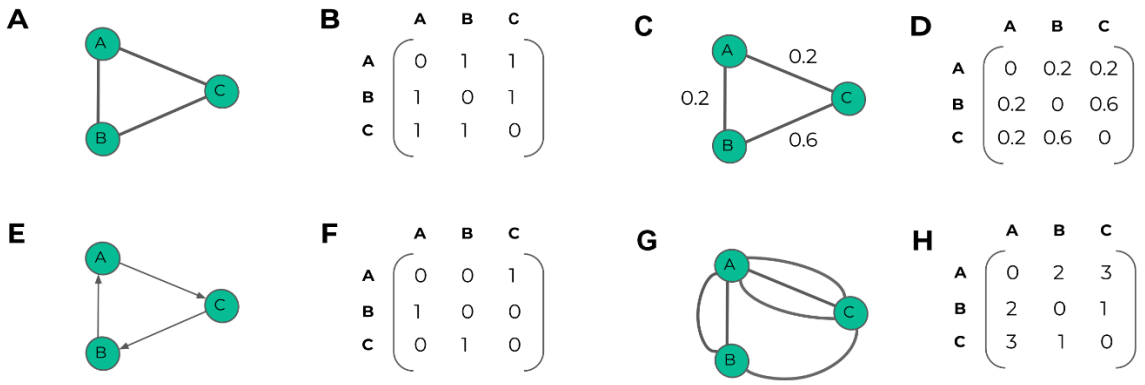


Figure 5. A depiction of common graph representations, and their corresponding adjacency matrices. A) Simple graph. **B)** Adjacency matrix for A. **C)** Weighted network. **D)** Weighted adjacency matrix of C. **E)** Directed graph. **F)** Adjacency matrix for E. **G)** Multigraph. **H)** Adjacency matrix of a multigraph.

3.1. Scale-free networks and topological graph analysis

For now, we have introduced how networks can be used for a reliable representation of relationships between different entities. However, there exists a deeper fundamental axiom underlying that justifies the utilization of networks as means in biomedicine: the topology of real network differs from that of randomly generated networks, exhibiting **scale-free** characteristics (151).

A graph is scale-free depending on a property of its nodes: **degree**. The degree of a node is defined as the number of edges that link the node to other nodes of the network (**Figure 6A**). Nodes presenting high number of edges compared to other neighbour vertices are commonly referred to as **hubs**. For directed networks, we refer to **outdegree** (edges outgoing from the node) and **indegree** (edges incoming to the node) (**Figure 6B**). If the degree distribution of a network does not follow a Poisson distribution (which can be expected for any randomly generated networks), then the graph is considered scale-free.

As topology of networks depends on the underlying relationships represented, several graph properties can be used to understand the modelled system (**Figure 6**). **Centrality measures** allow to rank the importance of nodes and edges in terms of a given graph property. For example, hubs are central in a graph regarding **degree centrality**. Another commonly used centrality measure is **betweenness centrality** (152)**.** Betweenness is based on **shortest paths** (**Figure 6C**) (153). The shortest path between two nodes is the path where the number of nodes traversed (in unweighted networks) or the sum of the edge weights (in weighted networks) is minimal. Computation of betweenness centrality for a given node *v* starts by identifying shortest paths between all pairs of nodes in the graph; then, for each pair of nodes, the fraction of shortest paths traversing *v* is calculated. The betweenness centrality of *v* is the sum of this fraction for all node pairs (**Figure 6D**). This way, *v* is central in terms of betweenness depending on the extent by which it is traversed by shortest paths.

A widely used algorithm for measuring distances in graphs is based on **random walks** (154) (**Figure 6E**). A random walk is a Markov chain (a stochastic model where the probability of each event is dependent on the state acquired during the previous one) where a particle explores the graph starting from a given initial node, named **seed**. At each step, the particle traverses the network, moving through existing edges, following the associated probabilities of each edge.

Recursively, probability of reaching other nodes of the graph is updated until the corresponding probability distribution converges. The probabilities associated with each node can be considered as a measure of connectivity from the seed to other nodes.

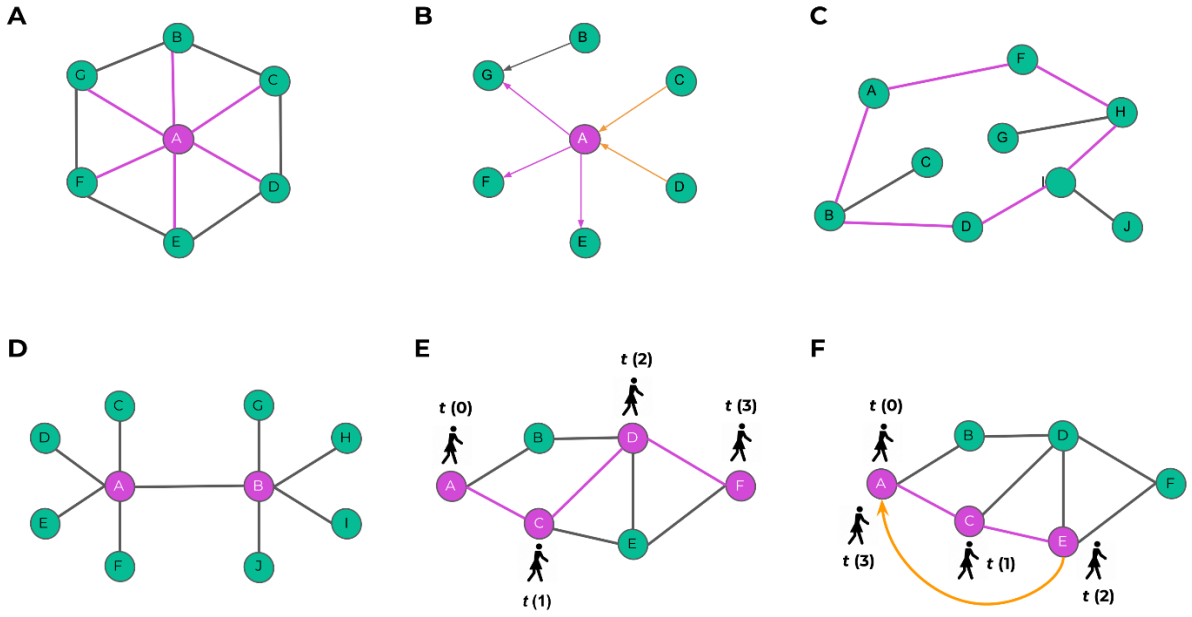


Figure 6. Visual representation of typical analysis of network topology. **A)** A hub node, and its corresponding degree. **B)** A graph highlighting outdegree (pink edges) and indegree (orange edges) of a node. **C)** A graph showcasing the shortest paths between node B and H (pink edges). **D)** A graph displaying the nodes with the highest betweenness centrality. **E**) An example of a random walk process. A 'walker' particle starts from node A at time point 0, to reach vertex F at time 3, traversing C and D in the process (pink path). **F)** Random walk with restart. After reaching node E, a random 'restart' event occurs, forcing the particle to come back to seed node A.

An extension of the procedure is the **random walk with restart** (**Figure 6F**), which provides an elegant solution to avoid walks from getting trapped in dead ends. In this extension, the particle is allowed to randomly restart the process at any step of the walk (with an associated restart probability), at any node of a given set of seeds This approach results in a stationary distribution that represents the distance of the seed node set to all other nodes of the network.

In network medicine, topological analysis of distances, as well as centrality measures allow for a thorough understanding of the importance of encoded biological entities such as genes, proteins, and metabolites. Due to the complexity of biological systems, and the need for a holistic interpretation of disease-related processes, researchers aim to identify meaningfully connected regions of biomedical networks. Identifying such regions (known as **communities**) has a huge relevance because nodes within the same community are expected to have stronger functional relationships (155–157).

One commonly used algorithm for community detection in networks is the **Louvain algorithm**, introduced by Blondel et al. (158) in 2008 (**Figure 7**). This greedy optimization technique maximizes a structural metric from the network: **modularity** (159). Modularity measures the fraction of significantly enriched edges within a set of vertices compared to a randomly generated graph model. As a quality measure for network partitions, recent research pinpointed how the algorithm outperforms similar heuristics (160). Louvain algorithm procedure starts by assigning each node to its own community (**Figure 7**). Then, node-wise, the potential variation of the global modularity is recursively computed when the node is moved to a neighbour community, until no further changes increase modularity. This initial stage is followed by subsequent steps, where the nodes within the detected communities are aggregated into a **super-node**. The resulting super-node network is a weighted network where the weight of the super-node edges is the sum of the weight of the edges existing between the corresponding communities. For the new super-node network, the procedure is iteratively recomputed, until the community structure converges, ultimately yielding a hierarchical community structure.

An important feature of modularity is resolution. The resolution parameter plays a crucial role in modularity as a quality metric for community divisions. It allows for the adjustment of community size and composition, resulting in the emergence of multiple, equally valid partitions at different resolution levels (161).

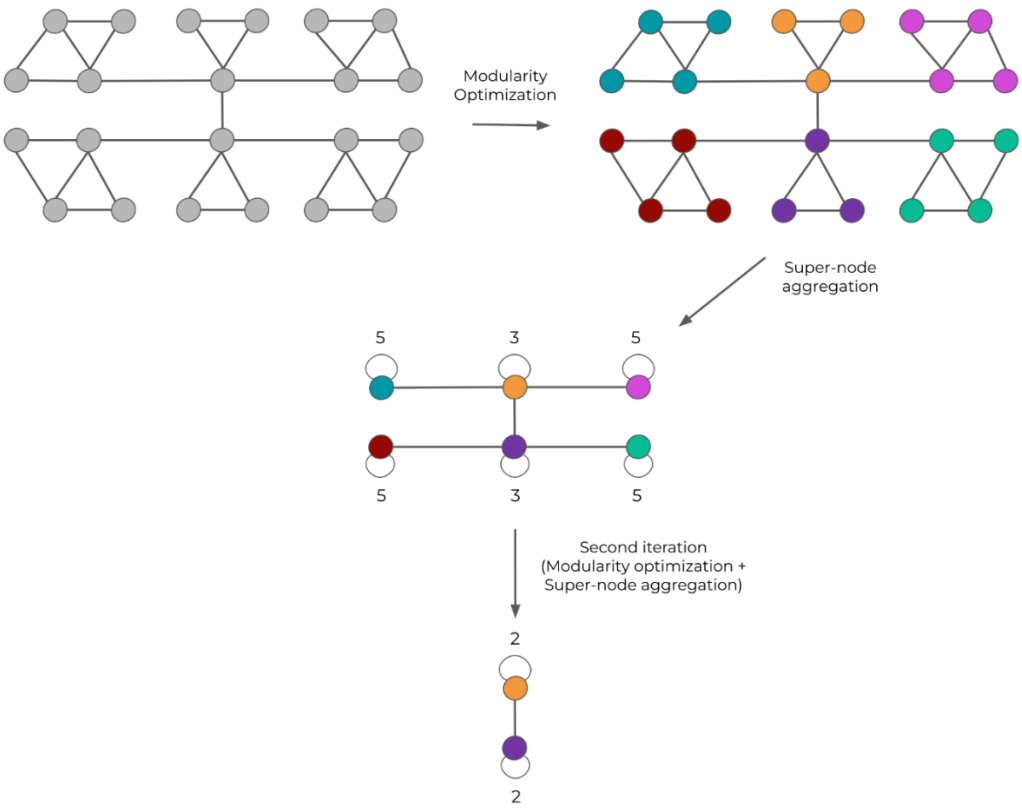


Figure 7. Louvain network community detection algorithm. This recursive heuristic works in iterations of two processes. The first step is the optimization of modularity, a quality measure for the partition. When the optimization is finished, nodes pertaining to the same community are unified into a super-node. The resulting super-node graph is the input for the following iteration, until the process converges.

The resolution parameter poses a significant challenge in community detection, particularly in the context of network biology. As all community structures at any resolution value are equally valid, an arbitrary choice of this parameter can have a great influence on the conclusions drawn from the community analysis. A possible solution to deal with this limitation is to identify persistent community memberships across a range of resolution values of interest. Strong modular structures are indeed characterized by persistent common community identity along resolution ranges (161). We explore the power of such persistency analysis, as well as its potential for the analysis of biomedical data in Chapters 3 and 4.

3.2. Network-based representation of biomedical data

In previous sections, we have highlighted the potential of graphs as a framework for data representation, along with classic topological metrics to analyze them. However, different data types require distinct graph representations. Here, we introduce several examples of network models used in the literature for the different data types used in the research articles presented in **Chapters 3 and 4.**

3.2.1. Proteomic networks

Proteomic data networks can represent various concepts, with the most widely focused type being protein-protein (PPI) networks. These networks offer a mechanistic perspective of the human interactome (47,48), and are a particularly important target of the research presented in Chapters 3 and 4. However, several other network models have proved useful for the analysis of proteomic data. For instance, protein structure networks offer valuable insights into the three-dimensional arrangements of proteins (162), offering detailed understanding on how proteins fold and function within cellular processes. Protein co-expression networks, on the other hand, provide a comparable view of protein co-occurrence within a sample to that of gene co-expression (163). All this diverse protein network models can provide comprehensive understanding of the dynamics of proteomic data across different biological conditions.

3.2.2. Pathway networks

The concept of a biological pathway is probably the most naturally translatable to a network representation, encompassing interrelated biochemical events. For example, signalling cascade processes can be represented as directed networks. In these networks, an edge indicates a chemical reaction catalysed by a source node, thereby influencing the state of a target node (164). The representation of chemical reactions in this manner holds huge potential for disease modelling. This type of networks also allows for the representation and analysis of more complex relationships, such as

feedback loops and other common regulatory features of metabolic pathways, allowing researchers to identify key steps of this processes to be targeted for intervention (165).

Another interesting network representation of pathways information, explored in the research presented in Chapters 3 and 4, is based on share pathway annotations. In such representations, nodes are connected if they are annotated to the same pathway in a given database (e.g., Reactome or KEGG pathways). We will introduce the benefit of this representation in the following section (**Section 3.3., Multilayer and complex networks**).

3.2.3 Metabolomic networks

Although most metabolomic analysis are focused on the knowledge coming from metabolic pathway representations, several aspects of metabolism can be represented as networks (166), including mass spectrometry-data (167,168) and, as explored in Chapters 3 and 4, gene networks of shared interacting metabolites, which we also introduce briefly in the next section. In this sense, a common challenge for metabolic network inference is dealing with metabolite instances marked by significant promiscuity, stemming from their low specificity in reactivity.

3.2.4 Drug-based networks

Drug-based relationships have been commonly represented as networks to facilitate drug repurposing analysis, accelerating the study of the potential effects of drug pairs that may arise from combining different drugs. This approach offers a means to prioritize testing and enhance the efficiency of the study (169).

Interestingly, drug associations to common target genes can be used to represent gene relationships that reflect drug information, a configuration we explore in the article presented in Chapter 4. An example of the recent efforts into providing large-scale network-based resources is DrugMAP, a comprehensive collection of information from multiple databases covering drug data (80).

3.2.5 Chromatin interaction networks

Recently, network approaches have been employed to analyze the 3D organization of chromatin, with the objective of studying the processes underlying chromatin structure dynamics (170,171). Furthermore, significant progress has been made in studying the interactome that mediates promoter interactions, for example, in cell differentiation (172). In **Annex III**, an additional analysis of HiChIP data (which includes DNA-protein interactions and protein-protein interactions) in prostate cancer is provided.

3.2.6 Disease networks

Network representations are also well-suited for the analysis of disease relationships (150). These associations can be based on multiple features, such as bulk or single-cell multi-omic data (173) or phenotypic information (174). Comorbidity dynamics have been well studied by means of network analysis (175,176). Disease comorbidity networks allow to measure both the positive and negative impact of a disease on the likelihood of developing another one, providing additional insights for potential drug targets (177). The network configuration discussed in the article presented in Chapter 4, like the ones mentioned earlier, focuses on genes affected in human diseases as the central elements. Particularly, the relationships identify genes whose variants have been described to be associated to the same disease, connecting genes based on disease knowledge, for instance, genome wide association studies (GWAS).

3.2.7. Transcriptomic networks

When inferring networks from bulk RNA-seq gene expression profiles, the classical methodology relies on one assumption: genes presenting high correlation across a dataset are likely to share similar regulative processes (178). This way, the study of gene co-expression can be performed from multiple perspectives, including correlation of gene profiles across a patient dataset (179) and vice versa. However, when working with single cell RNA-seq data, the network representation becomes more complex. Although it is possible to perform pseudo-bulk RNA-seq analysis from

single cell RNA-seq, the noise coming from cell subpopulations (180) require of the development of network inference tools accounting for such variability. In this sense, notable network inference approaches include GENIE3 (181) and SCENIC (182).

3.3. Multilayer and complex networks

The main motivation behind the research articles presented in this thesis is to propose novel frameworks for precision medicine that use in a jointly way several network networks encoding information from diverse sources. Particularly, we focus on the potential of **multilayer networks,** systems formed by collections of interconnected networks (hereby called **layers**) (**Figure 8B and Figure 9**) (184,185), for the analysis of disease information.

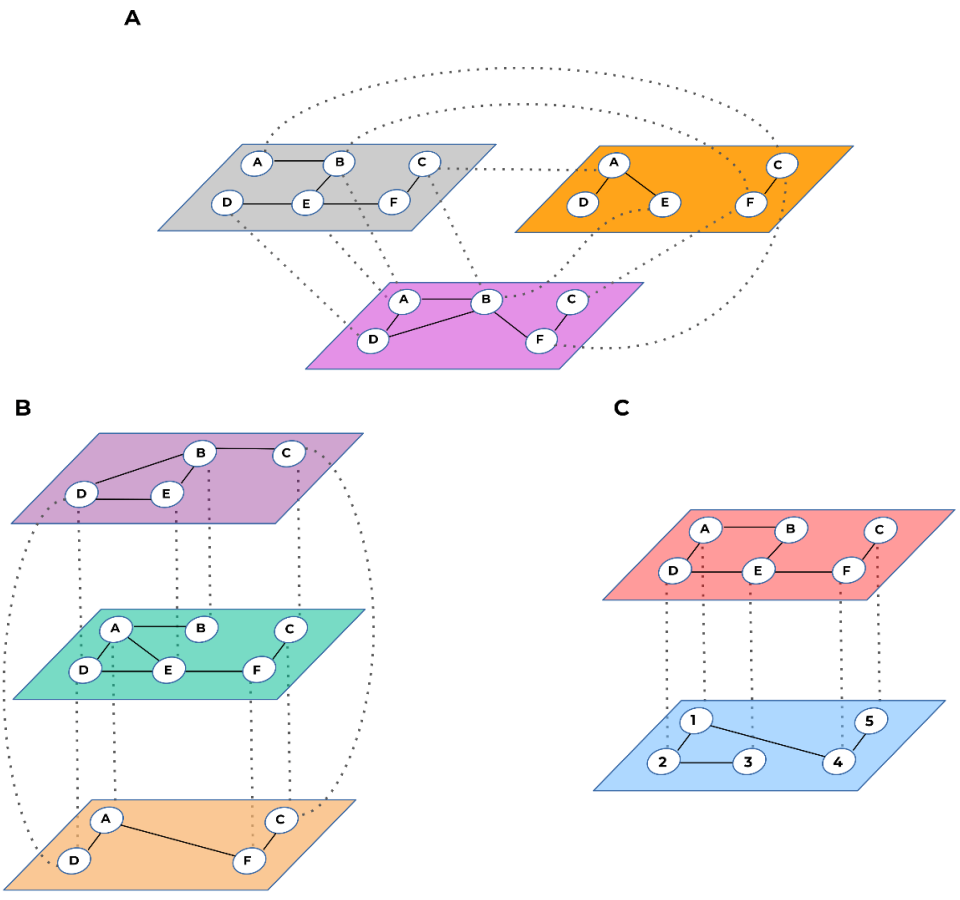


Figure 8. **Visual representation of complex graph constructions**. **A**) A multiplex network. **B**) Multilayer network. **C**) Heterogeneous network.

A multilayer network is a complex graph presenting intralayer edges within each layer, and interlayer edges connecting nodes across different layers. This representation has proven valuable for biomedicine in multiple scenarios, enabling to study biomolecular interactions (146) and diseases (183) and facilitating the integration and interpretation of heterogeneous data resources. Several established tools for network analysis have been recently adapted for multiplex networks.

These include classical graph theory approaches such as random walk with restart (154) and community detection algorithms (184), which we will introduce and discuss in subsequent sections. However, due to the relatively recent introduction of complex networks in the biomedical domain, the application of machine learning approaches in this area is currently limited to few biomedical (185) contexts, including biomedical graph embedding (146,186), biological association prediction (187), cancer driver gene detection (188) and reconstruction of molecular mechanisms from single cell RNA-seq data (189).

An illustrative example of the concept of a multilayer network is the representation of multiple transport systems in a city (190). Connections between nodes within the same layer (in this example, the stations) correspond to **intralayer edges**, while the connections between the different graph units of the systems (in this example, stations where a change between transport options is available) are the **interlayer edges**. This way, multilayer graphs can be adapted to encode specific knowledge, allowing for a finer analysis of the represented system.

Still, the absence of a standardized nomenclature to denominate the different structures that complex networks may adopt has resulted in multiple designations such as *multilayer networks*, *heterogeneous networks,* and *multiplex networks* populating the literature. Here, we will herein follow an adaptation of the nomenclature used by Valdeolivas et al. (2019) in their work for the extension of the Random Walk with Restart algorithm to high-order graphs (154).

The presented definition of a multilayer network is, in fact, a specific instance of a **multiplex network**. In general, a multiplex network refers to a set of interconnected graphs where the layers share the same set of nodes (**Figure 8A**). A **multilayer network** is a multiplex network instance where interlayer edges exist only between nodes of the same identity through the different layers (**Figure 8B**). A **heterogeneous network**, in contrast, presents graphs with different node sets, as well as bipartite networks corresponding to the interlayer edges between pairs of layers (**Figure 8C**).

The way in which a multilayer network is modelled plays a crucial role on how the underlying biology is studied. **Figure 9** depicts a commonly applied multilayer network structure for the integration of biomedical data, illustrating a three-layer multilayer gene network. These networks connect multiple simple graph layers, where each layer represents genes connected based on data from a specific biomedical source. This model results in the integration of diverse gene associations within each layer.

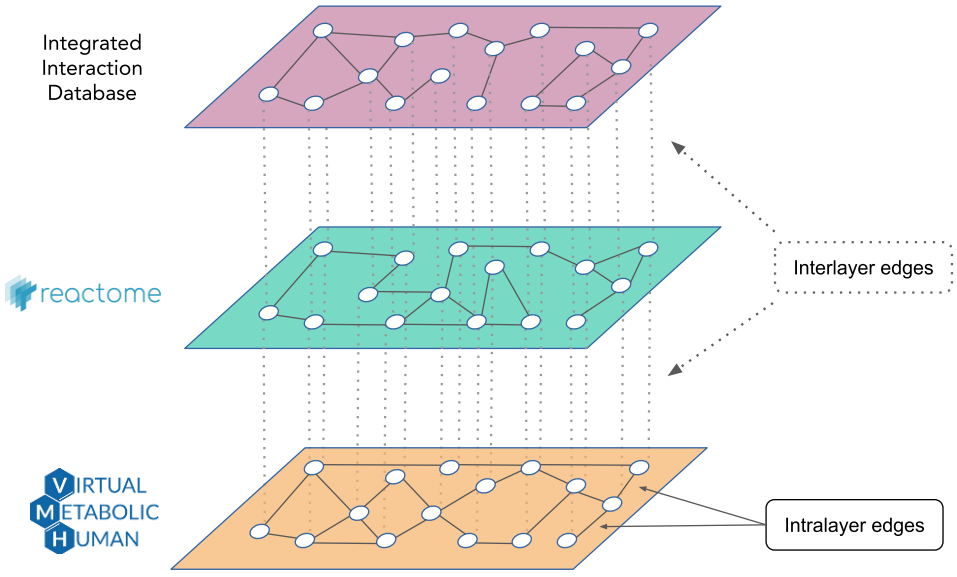


Figure 9. Schematical depiction of a gene multilayer network formed by three layers, from the research presented on Chapter 3. Each layer represents gene associations retrieve from biomedical knowledge databases that provides specific omics information. Intralayer edges describe gene relationships in each database, while interlayer edges exist between nodes sharing the same gene identity in different layers.

It is important to emphasize that reasonable adjustments can (and should) be made to the data extracted from databases to ensure a coherent representation of biological entities across the layers. For instance, within a gene multilayer network's protein interactome layer, nodes conveniently represent genes, while the edges symbolize the physical interactions among gene products.

The research articles presented in Chapters 3 and 4 explore the effectivity of the multilayer network system in detecting genes related to disease processes, providing a valuable tool for studying and understanding diseases in a comprehensive manner.

3.4. Topological analysis of complex networks

The analysis of node relationships within the multilayer network is also performed by means of topological analysis. Extension of topological exploration tools to the multilayer network level has been undertaken recently, including the previously mentioned algorithm for random walk with restart exploration (154) as well as community detection tools (184). The adaptation of these approaches represents a significant stride forward in analyzing biological systems and advancing the field of network science.

3.4.1. Random Walk with Restart on complex graphs

Extending topological analysis to multilayer and heterogeneous graphs imply considering both interlayer and intralayer connections simultaneously, but with different implications. An elegant (and computationally scalable) solution to model such analysis was introduced in 2019 by Valdeolivas et al. (154) as a way to prioritize disease associated genes using biomedical database knowledge.

In the algorithm proposed, the walker is allowed to navigate through the multilayer structure, jumping between layers when encountering a node that exists in other layers. Furthermore, in addition to the probability that controls restart events, a second probability selects the restarting seeds of each layer.

This way, when a restart event occurs and the seed node exists in more than one of the layers, the corresponding probability array assigns specific importance weights to each network. As a result, it is possible to prioritize those layers that may present more meaningful information for the biological problem of analysis. The original publication of the algorithm showcased the effectiveness and applicability of random walks on complex networks in the context of rare disease studies. It demonstrated how random walks can be used to explore the neighbourhood of known causal genes, enabling the identification of novel candidate genes that may have a modifying effect.

3.4.2. Multilayer community detection

Initial approaches to perform community detection in complex graphs primarily focused on network aggregations, treating all interactions as equivalent (191). However, such an approach may overlook valuable information encoded within known interlayer edges.

Detecting communities in multilayer networks implies the challenge of considering the impact of all the layers to the definition of a community. For this task, a number of algorithms have been proposed (192,193). In the context of the research presented in this PhD thesis, we focus on the proposed adaptation of the Louvain algorithm for multilayer networks introduced by Didier et al (184). This adaptation utilizes a new metric called **multiplex-modularity**. The article demonstrates how multiplex-modularity can be defined as the sum of the individual modularities of each graph and serves as an appropriate metric for optimizing the partitions of the multiplex network. Furthermore, the publication presents the higher performances achieved by the algorithm compared to a battery of graph aggregative methods, and most importantly, demonstrates its potential for the analysis of real biological multiplex networks.

By applying community detection based on multiplex-modularity to gene multilayer networks, we could perform several analyses of the persistence of gene community associations demonstrating its relevance for rare disease research. This methodology

is displayed in the research presented on Chapters 3 and 4 for the analysis of gene relationships based on multi-omics data in scenarios characterized by data scarcity.

4. Main Objectives

The **main objectives of the PhD thesis** are the following:

1. Explore the potential of network biology in addressing the challenges of precision medicine, related to limited data in areas such as rare disease research.

2. Develop new methodologies based on multilayer network modelling, for the integration and analysis of biomedical data, with a specific emphasis on bringing personalized medicine to contexts with constraints in patient availability, namely, precision oncology and rare disease scenarios.

3. Create novel approaches based on multilayer networks to enhance the interpretability of biomedical studies, harnessing the potential of the integrated complementary resources to uncover new relationships in precision medicine studies.

4. Establish multilayer network-based models as an effective approach for achieving effective patient stratification, as well as facilitating explainable dimensionality reduction and feature selection.

5. Apply and adapt these new methods for the work with real-world data, aiming to address specific questions in rare disease research, such as the identification of genetic modifiers of disease severity, a largely neglected biomedical challenge.

Chapter 2

Artificial intelligence in cancer research: learning at different levels of data granularity

Publication Record

This chapter introduces the original review article '*Artificial intelligence in cancer research: learning at different levels of data granularity'* published in the Molecular Oncology journal (2021 Journal Impact factor: 7.449; Q1 in the Oncology field. Rank: 51/245). (2022 Journal Impact Factor: 6.6; Q1 in the Oncology field. Rank: 50/241).

Co-authors & affiliations

Davide Cirillo[1,*], Iker Núñez Carpintero[1], Alfonso Valencia[1,2]

1. Barcelona Supercomputing Center (BSC), Barcelona, Spain

2. ICREA, Barcelona, Spain

* Corresponding author: davide.cirillo@bsc.es

Reference

Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. Molecular Oncology. 2021;15(4):817–29.

Contribution of the PhD Candidate

The PhD candidate contribution is on the writing of the review article, as stated in the section '**Author contributions**' (page 46).

Molecular Oncology

FEBS PRESS

REVIEW

Artificial intelligence in cancer research: learning at different levels of data granularity

Davide Cirillo[1], Iker Núñez-Carpintero[1] and Alfonso Valencia[1,2]

1 Barcelona Supercomputing Center (BSC), Barcelona, Spain
2 ICREA, Barcelona, Spain

Keywords
artificial intelligence; cancer research; data granularity; machine learning

Correspondence
D. Cirillo, Barcelona Supercomputing Center (BSC), C/Jordi Girona 29, 08034, Barcelona, Spain
Tel: +34 934137971
Email: davide.cirillo@bsc.es

(Received 30 September 2020, revised 20 December 2020, accepted 10 January 2021, available online 20 February 2021)

doi:10.1002/1878-0261.12920

From genome-scale experimental studies to imaging data, behavioral footprints, and longitudinal healthcare records, the convergence of big data in cancer research and the advances in Artificial Intelligence (AI) is paving the way to develop a systems view of cancer. Nevertheless, this biomedical area is largely characterized by the co-existence of big data and small data resources, highlighting the need for a deeper investigation about the cross-talk between different levels of data granularity, including varied sample sizes, labels, data types, and other data descriptors. This review introduces the current challenges, limitations, and solutions of AI in the heterogeneous landscape of data granularity in cancer research. Such a variety of cancer molecular and clinical data calls for advancing the interoperability among AI approaches, with particular emphasis on the synergy between discriminative and generative models that we discuss in this work with several examples of techniques and applications.

1. Introduction

Data granularity refers to the level of detail observable in the data. The finer the granularity, the more detailed are the observations. In cancer research, data granularity reflects the amount of molecular and clinical information that is collected about a patient or a group of patients, not only in terms of dataset size but also in terms of diversity of measurements, scales, and data types. At present, the available data in cancer research may not always provide the level of granularity required for effective decision-making. For instance, healthcare resources exhibit a shortage of information about specific cancer subtypes, minority groups, and rare cancers, such as the case of pediatric oncology [1]; national cancer registries tend to collect mainly first-line treatments and display reduced accessibility to actionable information [2]; and exigent legal and ethical approvals hurdle the timeliness of cancer data availability [3]. In this scenario, several initiatives devoted to some of these facets have been created, such as the Collaboration for Oncology Data in Europe (CODE; www.code-cancer.com), Rare Cancers Europe (RCE; www.rarecancerseurope.org), and the Cancer Drug Development Forum (CDDF) [4]. Nevertheless, the granularity of oncological data is highly scattered worldwide, resulting in a continuum of scale, quality, and completeness of the available datasets, that we refer to as *data continuum*. This aspect is particularly relevant in the context of the development of Artificial Intelligence (AI) systems, which are largely characterized by data-intensive computational modeling approaches to assist clinical decision-making [5–7].

In this work, we examine how cancer data granularity (from population studies to subgroups stratification) relates to multiple AI approaches (from deep learning to linear regression), and provide possible

Abbreviations
AI, Artificial Intelligence; CEDCD, Cancer Epidemiology Descriptive Cohort Database; EGA, European Genome-phenome Archive; EHR, Electronic Health Record; FDA, Food and Drug Administration; GAN, Generative Adversarial Network; HLA, Human Leukocyte Antigen; HPC, High Performance Computing; MHC, Major Histocompatibility Complex; TCGA, The Cancer Genome Atlas.

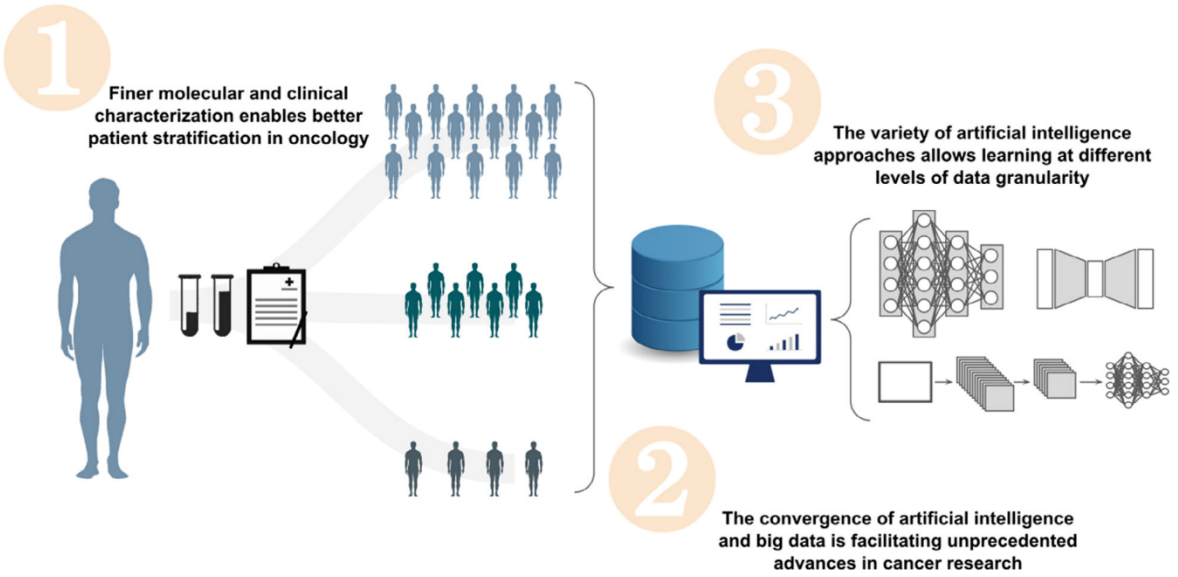


Fig. 1. The interplay between data generated with different levels of granularity and the multiplicity of AI approaches in cancer research.

solutions to reconcile the interoperability between these two components to ensure modeling strategies within the data continuum (Fig. 1). This work brings forward the specific need of developing AI techniques able to transcend the current limitations in their applications to the heterogeneous levels of granularity typical of cancer datasets.

The article is structured in three parts. In the first part, we analyze the ongoing process of confluence of big data and AI in cancer research ('Big data in cancer research' and 'The role of AI in cancer research'), and report on the main data types and areas of application ('Main areas of application and data types of AI in cancer research'). In the second part, we challenge the current focus on big data by examining two large-scale projects, namely the Cancer Genome Atlas (TCGA) and the Cancer Epidemiology Descriptive Cohort Database (CEDCD), under the lens of data granularity ('Heterogeneous levels of data granularity in cancer research'), and provide an overview on multiple AI approaches that allow learning at different levels of data granularity as well as discuss challenges and limitations ('Sample size and label availability: limitations and solutions'). In the third part, we deliver the conclusions to the article and a perspective view on the future of AI in cancer research ('Conclusions and Perspectives').

2. Big data in cancer research

Cancer research has been witnessing unprecedented innovations in recent years, including a major paradigm shift from histological level to molecular level characterization of cancers with a strong impact on treatment and medical practice [8,9]. An illustrative example of this change is the current, finer categorization of blood cancers into multiple subtypes based on the patient's genetic information [10]. Moreover, new technologies, such as CRISPR gene editing [11] and CAR T-cell therapy [12], are pushing the frontiers of clinical intervention and research. Additionally, single-cell multi-omics and imaging of preclinical personalized cancer models, such as organoids [13], are proving extremely valuable in dissecting key aspects of tumor evolution, as demonstrated by the research activities of initiatives such as LifeTime [14].

Such variety of data, including structure and unstructured clinical and molecular information (e.g., genetic tests, medical records, imaging data), outlines a horizon of possibilities for advancing oncology. Efforts to fill the gap between molecular and clinical information have been proposed, such as the concept of the Patient Dossier [15], which aims to facilitate the information flow between complex genomic pipelines and basic queries involving several aspects of the patient's health. Nevertheless, the progress in our understanding of cancer is not dependent on the sole availability of large amounts of high-quality and diversified data. The ongoing accumulation of records on a large number of patients is reinforcing the pressing need of cancer research and clinical care to embrace computational solutions to effectively utilize all this information. The effective utilization of cancer big data entails all the

Molecular Oncology **15** (2021) 817–829 © 2021 The Authors. *Molecular Oncology* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

steps from data processing and storage to data mining, analysis, and final applications, such as the identification of patient-specific oncogenic processes [16] and biomarkers [17]. Moreover, the continuous improvement of data quality through standardization procedures that ensure responsible molecular and clinical data sharing, interoperability, and security is a key aspect for cancer research that is strongly catalyzed by initiatives such as the Global Alliance for Genomics and Health (GA4GH; https://www.ga4gh.org).

As traditional data management methods cannot handle the scale and variety of cancer data acquired and generated daily, advanced infrastructures for permanent archiving and sharing are presently flourishing. An example of an extensive repository of data resulting from biomedical research projects is the European Genome-phenome Archive (EGA; https://ega-archive.org/). EGA collects various data types, including public access data (e.g., somatic mutation, gene expression, anonymized clinical data, protein expression) and controlled access data (e.g., germline genetic variants). EGA stores data from cancer-centric data sources, including TCGA, the International Cancer Genome Consortium (ICGC), the Clinical Proteomic Tumor Analysis Consortium (CPTAC), and the OncoArray Consortium.

3. The role of AI in cancer research

Although advanced solutions for big data management are facilitating the handling of biomedical information, the road to clinical success (e.g., better prevention and diagnosis, improved treatment decisions, effective patient-clinical trial matching) must involve ways to leverage the data and to be able to gain actionable insights from it [18,19]. Predictive analytics and machine learning are thriving areas of research and application in cancer research, characterized by interdisciplinarity and diversity of approaches, which henceforth we collectively refer to as AI. At present, 6 Food and Drug Administration (FDA)-approved AI-based radiological devices with applications in oncology are available for mammography analyses and computer tomography (CT)-based lesion detection [20], and 74 AI algorithms for digital pathology have received FDA clearance [21]. Moreover, more than 300 AI-related clinical trials have been registered at ClinicalTrial.gov [22] and seven randomized trials assessing AI in medicine have been published [23]. These examples are some of the many AI systems that stem from research and development advances in real-time decision-making for health care, which are systematically surveyed and compared [24].

Biomedical big data coupled with the ability of machines to learn and find solutions to problems have ensured that AI is currently playing a major role in the progress of biomedicine [25–27] and particularly cancer research [28,29]. Indeed, big data and AI complement each other, as AI feeds off of big data, from which it can learn how to carry out tasks such as classifying groups of patients, forecasting disease progression, and delivering adaptive treatment recommendations. AI and big data have the potential to fathom and overcome issues such as the reliability of biomarkers and genetic information [30,31], the potential disparities in patient populations [32,33], and the limited understanding of side effects [34] despite the growing promise of combination therapy [35,36] and drug repurposing [37].

The convergence of AI and big data can help interlace the threads of the complex landscape of oncological medicine resources, which is currently pervaded by a high level of heterogeneity and lack of standards [38]. In this regard, international efforts, such as the European-Canadian Cancer Network (EUCANCan; https://eucancan.com/) and individualizedPaediatricCure (https://ipc-project.eu/), are advancing the potential of federated data infrastructures to improve standardized data reporting and the development of cancer-specific AI solutions.

To facilitate this progress, automated strategies for end-to-end AI processes operating on big data, from data governance to deployment of AI applications, have been developed. The intensive workloads of AI operating on big data demand computational resources that must be able to achieve extreme scale and high performance while being cost-effective and environmentally sustainable [39]. High performance computing (HPC), or supercomputing, architectures are facilitating the deployment of pioneering AI applications in biomedicine [40,41]. In this view, HPC represents a critical capacity to gain competitive advantages, including not only faster and more complex computation schemes but also at lower costs and higher impact. Innovative software and hardware solutions, as well as model training implementations that support fine-grained parallelism and restrain memory costs, aim to accelerate the forthcoming convergence of AI and HPC. For this reason, community-driven benchmarking infrastructures for objective and quantitative evaluation of bioinformatics methods and algorithms [42,43] as well as domain-specific evaluation campaigns [44] are acquiring an increasing importance within the cancer research community.

4. Main areas of application and data types of AI in cancer research

The variety of modalities of available data (i.e., molecular profiles, images, texts) enables the full potential of

AI in cancer research. For instance, imaging data has been used to train AI models for skin cancer classification [45] and lymph node metastasis detection [46], while sequencing data has been used for variant functional impact assessment [47] and patient survival prediction [48]. These examples employ artificial neural networks, specifically deep learning, which has marked the biggest trend in AI over the last decade [49]. Deep learning has largely been applied to cancer data integration and modeling, such as the classification of medical images and digital health data, often in combination with processing of electronic health records (EHRs), and included in systems supporting physician–computer interactions [50].

In an ideal scenario, a comprehensive collection of cancer patient data should include both data derived from the patient (e.g., demographic information, familial history, symptoms, comorbidities, histopathological features, immunohistochemistry, nucleic acid sequencing, biochemical analyses, digital images, experience measurements using digital devices) but also results generated from the application of AI. In this regard, the main AI implementations in cancer research encompass (a) statistical and mathematical models of the system under study and (b) simulations of such models aiming to explore the system's properties and behavior in different conditions. The main data types employed in such models and simulations comprise multi-omics and immunogenomics data, longitudinal data (e.g., EHRs), behavioral data (e.g., wearable devices and social media), and imaging data [51].

Multi-omics data play a central role in cancer research. Given the interplay between different biological phenomena (e.g., gene expression, epigenetic modifications, protein–protein interactions), the development of approaches to integrate multiple layers of data has become a subject of profound interest in this area. Harmonizing such heterogeneous sources of information represents a challenge that, in recent years, has led to the development of platforms that leverage data of large-scale pan-cancer initiatives and offer analytical functions, such as LinkedOmics [52] and DriverDBv3 [53].

Recent developments in AI for cancer research are contributing significantly to the field of cancer immunology, in particular neoantigen prediction. Thanks to the predictive power of deep learning, large-scale sequencing data of neoantigens and major histocompatibility complex (MHC) molecules can be used to test possible binding of truncated proteins of a tumor cell and the patient's human leukocyte antigen (HLA) system, enabling the discovery of treatment targets that would be both patient- and tumor-specific. Following this concept, a recent study was able to validate a personalized vaccine for melanoma using candidate neoantigens obtained with a tool using deep learning, NetMHCpan [54,55]. Other recently developed tools using deep learning are devoted to the prediction of antigen presentation in the context of HLA-class II, such as MARIA [56] and NetMHCIIpan [57]. Being promising targets for personalized immunotherapies, neoantigen prediction is a blooming area for which expert recommendations have been recently set out by the European Society for Medical Oncology (ESMO) including optimal selection schemes for candidate prioritization, pipelines for binding affinity prediction and mutated peptide annotation and comparison [58].

Deep learning is widely employed in the processing and analysis of medical imaging data which has resulted in a wide variety of applications, achieving remarkable results in prognosis prediction from routinely obtained tissue slides [59], tumor detection and classification [45,60] and, more recently, real-time tumor diagnosis [61,62].

It is important to note that the collection of EHRs is growing at levels comparable to those of genomic and molecular data. In this regard, EHRs represent a type of data whose processing has proven AI particularly challenging. Indeed, the high variety of clinical terminology, highly specialized words, abbreviations and short notes, makes EHRs content processing through general-purpose Natural Language Processing (NLP) models extremely arduous. Recent efforts focus on the generation of unified semantic systems and the organization of community challenges [63] from which automatically annotated corpora can be derived, which will facilitate the progress in this area [64,65]. One of the main challenges that all these advanced technologies, including modern approaches to digital and systems medicine, are currently facing is their integration and clinical exploitation in the health systems [66]. Indeed, many complex aspects, such as regulation, commercialization, and ethics, are playing a central role in the operational transformation of modern cancer care. For instance, despite the astounding advances in smartphones and Internet of Things (IoT) technologies, which largely facilitate the collection of patient-generated health data, regulatory priorities and positions as well as limitations in device-based data analytics directly affect the slow uptake of such digital medicine solutions in oncology [67].

5. Heterogeneous levels of data granularity in cancer research

Despite the availability of cancer big data, a prominent feature of the current data landscape in oncology

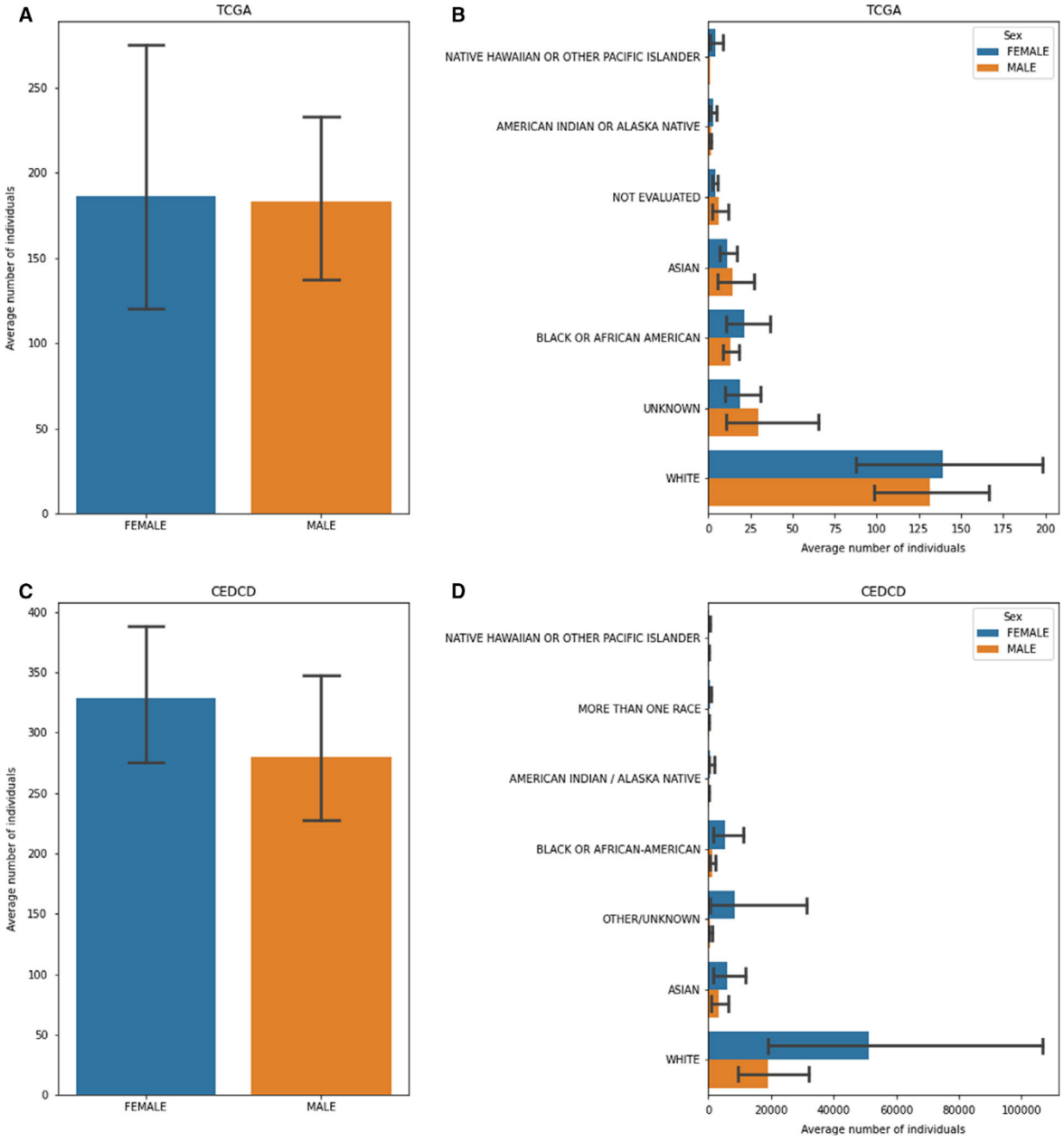


Fig. 2. Demographic features of the individuals represented in TCGA and CEDCD projects. (A) Average number of individuals per cancer type in TCGA disaggregated by sex; (B) average number of individuals per cancer type in TCGA disaggregated by race and sex; (C) average number of individuals per cancer type in CEDCD cohort studies disaggregated by sex; and (D) average number of individuals per CEDCD cohort studies disaggregated by race and sex.

is the imbalance between the amount of data per patient and the cohort size. Indeed, while thousands to millions observables per patient are routinely generated, a typical cohort size of specific groups of patients is relatively small [68].

As an example, we examine the curated clinical data of TCGA project [69] (Fig. 2A,B). The average number of unique patients per cancer type ($N = 33$) is 335.78 (on average, 182.93 male and 186.32 female individuals). As expected, these numbers reduce when

disaggregated by race (on average, 131.62 White, 13.43 Black or African American, 14.64 Asian male individuals, and 139.12 White, 21.17 Black or African American, 10.96 Asian female individuals) and, even more, by tumor stage, if this annotation is available. For instance, the 87 patients with mesothelioma, a rare but fatal cancer causally linked to asbestos exposure [70], distribute unevenly in the six stages (stage I, IA, IB, II, III, IV) by sex and race. White males are the most represented patients (80.4%), mostly appearing in late stages, reflecting both the gradual onset of the disease [71] and its incidence in developing countries that have consumed asbestos over past decades (83% in males and 17% in females as of 2017 in the United Kingdom; source: https://www.cancerresearchuk.org/). This observation highlights not only the overriding importance of early detection and better risk assessment tools based on socio-economic factors but also the need for effective AI-based approaches to learn from the little data that might be available.

A similar trend can be observed in prospective cohort studies, such as those reported in the CEDCD (https://cedcd.nci.nih.gov/), which collects large observational population studies aimed to prospectively investigate the environmental, lifestyle, clinical, and genetic determinants of cancer incidence (Fig. 2C,D). As of September 2020, the average number of participants diagnosed with cancer per cohort ($N = 61$) is 14624.65. However, when disaggregated by sex and cancer type ($N = 25$), this number decreases to an average of 328.50 women and 279.75 men per cancer type in each cohort. Also, the cohort composition is markedly skewed toward specific race categories, with an average of 19 172.77 White, 1330.83 Black or African American, 3420.39 Asian male participants, and 51 347.72 White, 5446.14 Black or African American, 6058.08 Asian female participants per cohort. These observations highlight the need for devising better strategies to improve the low enrollment rates in cohort studies and overcome the obstacles to minority populations engagement [72,73].

6. Sample size and label availability: limitations and solutions

In the area of cancer research, a long-standing challenge is the insufficient availability of massive high-quality labeled datasets coupling exhaustive molecular profiles with matching detailed clinical annotations [18]. In the current scattered scenario, there is a growing need to exploit the multiplicity of AI approaches for the nonexclusive utilization of the available data with different levels of granularity.

Most AI applications in cancer research are mainly based on two types of learning algorithms: supervised and unsupervised learning [74–76]. Supervised learning involves models that map data instances to labels in order to perform tasks such as classification and regression. Unsupervised learning involves models that extract information from data instances without labels to perform tasks such as clustering and dimensionality reduction. Additionally, many hybrid types of learning (e.g., semi-supervised learning) as well as specific learning techniques (e.g., transfer learning) are largely employed. All these approaches can be either discriminative or generative, whether they estimate the conditional probability of a label given an instance or the conditional probability of an instance given a label, respectively [77]. Thus, discriminative models can distinguish between different instances, while generative models can produce new ones.

Label availability and the varied scales of cancer data call for advancing the interoperability among AI approaches, in particular the synergy of discriminative and generative models. These models can be used, in turn, for inference and data augmentation, feeding back a finer characterization and accessibility of data for further training (Fig. 3).

Label availability can guide the choice of an AI approach or another for either discriminative or generative purposes. The dearth of ground-truth labels which are necessary to perform supervised tasks represents one of the main limitations to the use of AI in many areas of cancer research. The collection, curation and validation of labels by experts is an expensive and laborious process resulting in datasets that are too small to estimate complex models required to answer complex questions [78]. Models with low statistical power may lead to nonconvergence as well as biased and inadmissible outcomes, undermining reproducibility and reliability. Beside limited label availability and sample size, other limiting factors for AI can be identified, such as number of features, depth of hyperparameter optimization, and number of cross-validation folds [79].

When informative and defensible background information is available (e.g., previous studies, meta-analyses, expert knowledge), Bayesian statistics may produce reasonable results with small sample sizes [80–82]. Indeed, well-considered decisions are strongly endorsed in the choice of 'thoughtful' priors as opposed to naïvely using Bayesian estimation in small sample contexts. Nevertheless, prior information about the distribution of the parameters cannot be explicitly available and often difficult to derive.

If only a very limited amount of labels is available, AI approaches operating with minimal training data

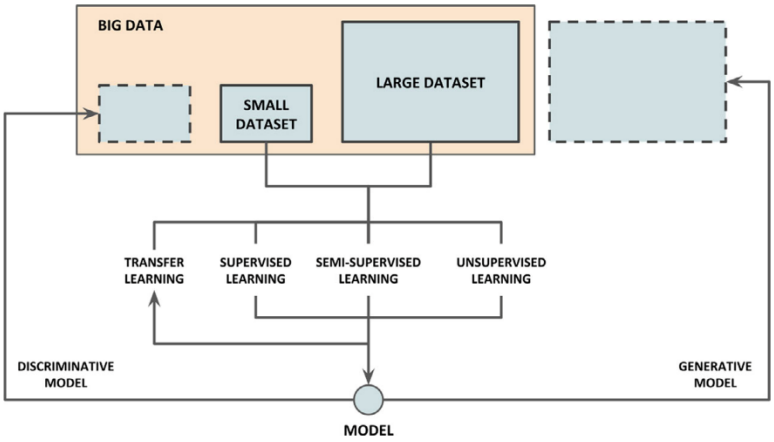


Fig. 3. Synergy of AI solutions for cancer research in the data continuum. Based on label availability of large and small datasets (e.g., over- and under-represented cancer subgroups), several learning approaches (supervised, semi-supervised, unsupervised, transfer learning) can be attained to create both generative and discriminative models. While discriminative models can be used to identify smaller subsets from the totality of big data (represented as small dashed rectangle on the upper left corner), generative models can be used for data augmentation by producing large volumes of synthetic instances (represented as a large dashed rectangle on the upper right corner).

exist, including transfer learning and meta-learning techniques for few-, one-, and zero-shot learning (surveyed in [83,84]). As an example, re-using a model trained on high-resource language pairs, such as French-English, can improve translation on low-resource language pairs, such as Uzbek-English [85]. Due to the ability of learning from minimal data, transfer learning and meta-learning are increasingly gaining momentum having the potential to mitigate many criticisms over deep learning concerning the requisite extensive computational resources and training data [86].

Transfer learning re-uses the weights of pretrained models in a similar learning task [87]. For instance, it has been recently applied to model anticancer drug response in a small dataset transferring the information learnt from large datasets [88]. This study illustrates the potential of transfer learning to improve future drug response prediction performance on patients by transferring information from patient-derived models, such as xenografts and organoids. Nevertheless, although transfer learning is designed to transfer information from a support domain to a target domain, very limited target training data can hamper the efficient adaptation to a new task even with shared features between the support and target data.

Meta-learning is based on the concept of 'learning to learn' consisting of improving performance over multiple learning episodes instead of multiple data instances. Meta-learning learns from the meta-data of previously experienced tasks, including model configurations (e.g., hyperparameter settings), evaluations (e.g., accuracies), and other measurable properties, enabling the search of an optimal model, or combinations of models, for a new task [89]. Recently, meta-learning has been applied to the prediction of cancer survival [90]. Despite the high adaptability of meta-learning, this study shows how the related tasks used for training should contain a reasonable amount of transferable information to achieve a significant improvement in performance compared to other learning strategies. For instance, if the samples of a specific cancer display very unique and distinct features, learning directly from them may represent a more effective strategy than learning from other cancer samples.

If the training data are only partially labeled, semi-supervised learning techniques, such as pseudolabeling and entropy minimization, proved successful and, for this reason, dedicated standard evaluation practices have been recently devised [91]. Semi-supervised learning jointly uses unlabeled and a smaller set of labeled data to improve the performances of one or both unsupervised and supervised tasks using the information learnt from the other or both [92]. Inherent limitations of semi-supervised learning mainly include strong assumptions about the feature space carrying relevant information about the prediction task. In this regard, the assumed dependency between labeled and unlabeled sets is deemed to effectively reveal fitting decision boundaries for predictive models. However, it has been shown that causal tasks, such as semantic segmentation in cancer imaging analysis, do not comply with these assumptions [93] and high-quality supervised baselines are crucial to assess the added value of unlabeled data in semi-supervised learning settings.

If enough labeled data are initially available for training, data augmentation can be achieved using generative models based on neural networks, such as generative adversarial networks (GANs) [94], variational autoencoders [95], and transformer models [96]. These approaches display technical open challenges that need further investigation, for instance the training instability and low mode diversity of GANs [97]. Oversampling datasets can also be achieved by creating synthetic instances to increase the training data and avoid class imbalance [98]. Moreover, similar to image data augmentation techniques and synonym replacement in texts, other methods based on data manipulations and new instances interpolation, such as the Synthetic Minority Oversampling Technique (SMOTE) algorithm [99], have been proposed.

Synthetic data generation represents a promising solution to the ethical and privacy barriers that may prevent in-depth data analysis and modeling of patients' information. For instance, the generation of synthetic data points has been exploited as a privacy-preserving approach to overcome the limitations and difficulties of data anonymization [100]. Indeed, instead of partially de-identifying data or censoring and removing protected variables, synthetic patient records can be fabricated from real-world data and used for model development and healthcare applications testing. Moreover, synthetic data can also be generated to specifically mirror the clinical features of a patient, thus creating a so-called digital twin or avatar for computationally evaluation of personalized drug treatments [101].

7. Conclusions and perspectives

Cancer is a disease that exhibits features of complex systems (e.g., self-organization, emerging patterns, adaptive and collective behavior, nonlinear dynamics). Cancer complexity is exemplified by the definition of the so-called hallmarks of cancer [102], which holds a systems view of the disease to be investigated through computational approaches. Computational cancer research is a multidisciplinary area aimed to advance the biomedical understanding of cancer by harnessing the power of data analytics and AI to advance in both basic and clinical settings [103,104]. With the rapid development of precision medicine and big data applications in cancer research, AI is setting down exceptional opportunities and ambitious challenges in this area [105,106], facilitating the progress toward individually tailored preventive and therapeutic interventions. The acquisition of a deep understanding of such interindividual differences relies on the development of AI systems that enable the identification of biomedically relevant patterns from several data from multiple modalities, spanning a varied range of data types, and displaying heterogeneous levels of granularity. Among the many details defining data granularity in cancer research, such as scales, measurements, and data types, sample size and label availability are the most evident factors that have a direct impact on the application of AI in cancer research. The range of AI modeling approaches that allow learning from both large and small datasets to discriminate or generate observations show the extraordinary potential of operating within a continuum of dataset sizes. This synergy among multiple learning techniques, namely supervised, semisupervised, transfer, and unsupervised learning, encompasses the entire spectrum of data granularity, including both the effective generalization from few examples with applications to multidimensional data, and the effective ability of models trained on big data to uncover small subgroups and subtle details. These AI approaches are not short of limitations and general assumptions that need to be considered before naïvely apply them. In this regard, it is particularly important to develop robust systems for testing and benchmarking AI applications, with adequate data resources and cleaver strategies that can be converted into certifications for the use of AI in real-world medical scenarios, as recently proposed for diagnostic imaging algorithms [107]. We envisage a growing use of such a multiplicity of AI approaches in cancer research that will enable an interconnected integration of automatic learning processes within the data continuum, from big data to small data as well as from small data to big data.

Data accessibility

The code to reproduce the barplots in Fig. 2 is available at: https://github.com/cirillodavide/cancer_data_granularity.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreements 826121 ('iPC—individualizedPaediatricCure: cloud-based virtual-patient models for precision pediatric oncology').

Conflict of interest

The authors declare no conflict of interest.

Author contributions

AV and DC conceived the study. All the authors, AV, DC, and IN-C, contributed to the writing of the article.

References

1 Pui CH, Gajjar AJ, Kane JR, Qaddoumi IA & Pappo AS (2011) Challenging issues in pediatric oncology. *Nat Rev Clin Oncol* **8**, 540–549.

2 Pop B, Fetica B, Blaga ML, Trifa AP, Achimas-Cadariu P, Vlad CI & Achimas-Cadariu A (2019) The role of medical registries, potential applications and limitations. *Med Pharm Rep* **92**, 7–14.

3 Mascalzoni D, Dove ES, Rubinstein Y, Dawkins HJS, Kole A, McCormack P, Woods S, Riess O, Schaefer F, Lochmüller H *et al.* (2015) International charter of principles for sharing bio-specimens and data. *Eur J Hum Genet* **23**, 721–728.

4 Verweij J, Hendriks HR & Zwierzina H (2019) Cancer drug development forum innovation in oncology clinical trial design. *Cancer Treat Rev* **74**, 15–20.

5 van der Ploeg T, Austin PC & Steyerberg EW (2014) Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* **14**, 137.

6 Steyerberg EW, Uno H, Ioannidis JPA & van Calster B (2018) Collaborators poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol* **98**, 133–143.

7 Richter AN & Khoshgoftaar TM (2018) A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artif Intell Med* **90**, 1–14.

8 Ogino S, Fuchs CS & Giovannucci E (2012) How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert Rev Mol Diagn* **12**, 621–628.

9 Loomans-Kropp HA & Umar A (2019) Cancer prevention and screening: the next step in the era of precision medicine. *NPJ Precis Oncol* **3**, 3.

10 Taylor J, Xiao W & Abdel-Wahab O (2017) Diagnosis and classification of hematologic malignancies on the basis of genetics. *Blood* **130**, 410–423.

11 Stadtmauer EA, Fraietta JA, Davis MM, Cohen AD, Weber KL, Lancaster E, Mangan PA, Kulikovskaya I, Gupta M, Chen F *et al.* (2020) CRISPR-engineered T cells in patients with refractory cancer. *Science* **367**, 6481.

12 Mohseni YR, Tung SL, Dudreuilh C, Lechler RI, Fruhwirth GO & Lombardi G (2020) The future of regulatory T cell therapy: promises and challenges of implementing CAR technology. *Front Immunol* **11**, 1608.

13 Kim J, Koo BK & Knoblich JA (2020) Human organoids: model systems for human biology and medicine. *Nat Rev Mol Cell Biol* **21**, 571–584.

14 Rajewsky N, Almouzni G, Gorski SA, Aerts S, Amit I, Bertero MG, Bock C, Bredenoord AL, Cavalli G, Chiocca S *et al.* (2020) LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386.

15 Vazquez M & Valencia A (2019) Patient dossier: healthcare queries over distributed resources. *PLoS Comput Biol* **15**, e1007291.

16 Vasudevan S, Flashner-Abramson E, Remacle F, Levine RD & Kravchenko-Balasha N (2018) Personalized disease signatures through information-theoretic compaction of big cancer data. *Proc Natl Acad Sci USA* **115**, 7694–7699.

17 Crichton DJ, Altinok A, Amos CI, Anton K, Cinquini L, Colbert M, Feng Z, Goel A, Kelly S, Kincaid H *et al.* (2020) Cancer biomarkers and big data: a planetary science approach. *Cancer Cell* **38**, 757–760.

18 Azuaje F (2019) Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol* **3**, 6.

19 Clarke MA & Fisher J (2020) Executable cancer models: successes and challenges. *Nat Rev Cancer* **20**, 343–354.

20 Benjamens S, Dhunnoo P & Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 509.

21 ACR Data Science Institute FDA Cleared AI Algorithms. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

22 CONSORT-AI and SPIRIT-AI Steering Group (2019) Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* **25**, 1467–1468.

23 Topol EJ (2020) Welcoming new guidelines for AI clinical research. *Nat Med* **26**, 1318–1320.

24 Ahmed Z & Mohamed K (2000) Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* **2000**, baa010.

25 Agrawal R & Prabakaran S (2020) Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* **124**, 525–534.

26 Singh O, Singh R & Saxena A (2020) AI and precision medicine for oncology. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*. http://dx.doi.org/10.2139/ssrn.3566788.

27 Goecks J, Jalili V, Heiser LM & Gray JW (2020) How machine learning will transform biomedicine. *Cell* **181**, 92–101.

28 Ho D (2020) Artificial intelligence in cancer therapy. *Science* **367**, 982–983.

29 Liang G, Fan W, Luo H & Zhu X (2020) The emerging roles of artificial intelligence in cancer drug development and precision therapy. *Biomed Pharmacother* **128**, 110255.

30 Shi W, Ng CKY, Lim RS, Jiang T, Kumar S, Li X, Wali VB, Piscuoglio S, Gerstein MB, Chagpar AB

et al. (2018) Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep* **25**, 1446–1457.
31 Reiter JG, Baretti M, Gerold JM, Makohon-Moore AP, Daud A, Iacobuzio-Donahue CA, Azad NS, Kinzler KW, Nowak MA & Vogelstein B (2019) An analysis of genetic heterogeneity in untreated cancers. *Nat Rev Cancer* **19**, 639–650.
32 Zavala VA, Bracci PM, Carethers JM, Carvajal-Carmona L, Coggins NB, Cruz-Correa MR, Davis M, de Smith AJ, Dutil J, Figueiredo JC *et al.* (2021) Cancer health disparities in racial/ethnic minorities in the United States. *Br J Cancer* **124**, 315–332.
33 Li CH, Prokopec SD, Sun RX, Yousif F & Schmitz N (2020) PCAWG tumour subtypes and clinical translation, boutros, P.C., PCAWG Consortium Sex differences in oncogenic mutational processes. *Nat Commun* **11**, 4330.
34 Grassberger C, Ellsworth SG, Wilks MQ, Keane FK & Loeffler JS (2019) Assessing the interactions between radiotherapy and antitumour immunity. *Nat Rev Clin Oncol* **16**, 729–745.
35 Zervantonakis I (2020) Improving cancer combination therapy by timing drug administration. *Sci Transl Med* **12**, eabb5671.
36 Bayat Mokhtari R, Homayouni TS, Baluch N, Morgatskaya E, Kumar S, Das B & Yeger H (2017) Combination therapy in combating cancer. *Oncotarget* **8**, 38022–38043.
37 Zhang Z, Zhou L, Xie N, Nice EC, Zhang T, Cui Y & Huang C (2020) Overcoming cancer therapeutic bottleneck by drug repurposing. *Signal Transduct Target Ther* **5**, 113.
38 Montouchet C, Thomas M, Anderson J & Foster S (2018) The oncology data landscape in Europe: Report. European Federation of Pharmaceutical Industries and Associations. https://www.efpia.eu/media/412192/efpia-onco-data-landscape-1-report.pdf
39 Strubell E, Ganesh A & McCallum A. (2020) Energy and policy considerations for modern deep learning research. In Proceedings of the The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, pp. 13693–13696.
40 Kovatch P, Gai L, Cho HM, Fluder E & Jiang D (2020) Optimizing high-performance computing systems for biomedical workloads. In Proceedings of the 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 183–192.
41 Castrignanò T, Gioiosa S, Flati T, Cestari M, Picardi E, Chiara M, Fratelli M, Amente S, Cirilli M, Tangaro MA *et al.* (2020) ELIXIR-IT HPC@CINECA: high performance computing resources for the bioinformatics community. *BMC Bioinformatics* **21**, 333.
42 Capella-Gutierrez S, de la Iglesia D, Haas J, Lourenco A, Fernández JM, Repchevsky D, Dessimoz C, Schwede T, Notredame C, Gelpi JL *et al.* (2017) Lessons learned: recommendations for establishing critical periodic scientific benchmarking. *bioRxiv* 181677 [PREPRINT]. https://doi.org/10.1101/181677.
43 Rappoport N & Shamir R (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* **46**, 10546–10562.
44 Hirschman L, Yeh A, Blaschke C & Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6** (Suppl 1), S1.
45 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM & Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118.
46 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, the CAMELYON16 Consortium, Hermsen M, Manson QF *et al.* (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210.
47 Zhou J & Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931–934.
48 Poirion OB, Chaudhary K, Huang S & Garmire LX (2020) DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *medRxiv* 19010082. https://doi.org/10.1101/19010082
49 Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S & Dean J (2019) A guide to deep learning in healthcare. *Nat Med* **25**, 24–29.
50 Norgeot B, Glicksberg BS & Butte AJ (2019) A call for deep-learning healthcare. *Nat Med* **25**, 14–15.
51 Troyanskaya O, Trajanoski Z, Carpenter A, Thrun S, Razavian N & Oliver N (2020) Artificial intelligence and cancer. *Nat Cancer* **1**, 149–152.
52 Vasaikar SV, Straub P, Wang J & Zhang B (2018) LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res* **46**, D956–D963.
53 Liu SH, Shen PC, Chen CY, Hsu AN, Cho YC, Lai YL, Chen FH, Li CY, Wang SC, Chen M *et al.* (2020)

DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res* **48**, D863–D870.
54 Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A, Peter L *et al.* (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221.
55 Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S & Nielsen M (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13.
56 Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, Muftuoglu Y, Sworder BJ, Diehn M, Levy R *et al.* (2019) Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol* **37**, 1332–1343.
57 Reynisson B, Alvarez B, Paul S, Peters B & Nielsen M (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449–W454.
58 De Mattos-Arruda L, Vazquez M, Finotello F, Lepore R, Porta E, Hundal J, Amengual-Rigo P, Ng CKY, Valencia A, Carrillo J *et al.* (2020) Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. *Ann Oncol* **31**, 978–990.
59 Jiang D, Liao J, Duan H, Wu Q, Owen G, Shu C, Chen L, He Y, Wu Z, He D *et al.* (2020) A machine learning-based prognostic predictor for stage III colon cancer. *Sci Rep* **10**, 1–9.
60 Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F *et al.* (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9.
61 Yamada M, Saito Y, Imaoka H, Saiko M, Yamada S, Kondo H, Takamaru H, Sakamoto T, Sese J, Kuchiba A *et al.* (2019) Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci Rep* **9**, 14465.
62 Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, Eichberg DG, D'Amico RS, Farooq ZU, Lewis S *et al.* (2020) Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* **26**, 52–58.
63 Miranda-Escalada A, Farré E & Krallinger M (2020) Named entity recognition, concept normalization and clinical coding: overview of the Cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results. In Proceedings of the Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, pp. 303–323.
64 Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M *et al.* (2018) Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* **1**, 18.
65 Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, Cowan J, Weeke P, Mosley JD, Wells QS *et al.* (2014) Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med* **6**, 234cm3.
66 Topol EJ (2019) A decade of digital medicine innovation. *Sci Transl Med* **11**, eaaw7610.
67 Jim HSL, Hoogland AI, Brownstein NC, Barata A, Dicker AP, Knoop H, Gonzalez BD, Perkins R, Rollison D, Gilbert SM *et al.* (2020) Innovations in research and clinical care using patient-generated health data. *CA A Cancer J Clin* **70**, 182–199.
68 Willems SM, Abeln S, Feenstra KA, de Bree R, van der Poel EF, Baatenburg de Jong RJ, Heringa J & van den Brekel MWM (2019) The potential use of big data in oncology. *Oral Oncol* **98**, 8–12.
69 Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV *et al.* (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11.
70 Yap TA, Aerts JG, Popat S & Fennell DA (2017) Novel insights into mesothelioma biology and implications for therapy. *Nat Rev Cancer* **17**, 475–488.
71 Kondola S, Manners D & Nowak AK (2016) Malignant pleural mesothelioma: an update on diagnosis and treatment options. *Ther Adv Respir Dis* **10**, 275–288.
72 Greiner KA, Friedman DB, Adams SA, Gwede CK, Cupertino P, Engelman KK, Meade CD & Hébert JR (2014) Effective recruitment strategies and community-based participatory research: community networks program centers' recruitment in cancer prevention studies. *Cancer Epidemiol Biomarkers Prev* **23**, 416–423.
73 Unger JM, Cook E, Tai E & Bleyer A (2016) The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am Soc Clin Oncol Educ Book* **36**, 185–198.
74 Russell SJ & Norvig P (2010) E. Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, NJ. ISBN 9780136042594.
75 Mohri M, Rostamizadeh A & Talwalkar A (2012) Foundations of Machine Learning. MIT Press, Cambridge, MA, ISBN 9780262018258.
76 Bishop CM (2006) Pattern Recognition and Machine Learning. Springer, New York, NY. ISBN 9780387310732.

77 Ng AY & Jordan MI (2002) On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In Advances in Neural Information Processing Systems 14 (Dietterich TG, Becker S & Ghahramani Z, eds), pp. 841–848.MIT Press, Cambridge, MA.
78 Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* **15**, 20170387.
79 Vabalas A, Gowen E, Poliakoff E & Casson AJ (2019) Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365.
80 Smid SC, McNeish D, Miočević M & van de Schoot R (2020) Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct Equ Modeling* **27**, 131–161.
81 McNeish D (2016) On using Bayesian methods to address small sample problems. *Struct Equ Modeling* **23**, 750–773.
82 Zondervan-Zwijnenburg M, Peeters M, Depaoli S & Van de Schoot R (2017) Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Res Hum Dev* **14**, 305–320.
83 Wang Y, Yao Q, Kwok JT & Ni LM (2020) Generalizing from a few examples. *ACM Comput Surv* **53**, 1–34.
84 Xian Y, Lampert CH, Schiele B & Akata Z (2019) Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans Pattern Anal Mach Intell* **41**, 2251–2265.
85 Zoph B, Yuret D, May J & Knight K (2016) Transfer learning for low-resource neural machine translation. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1568–1575.Association for Computational Linguistics, Austin, TX.
86 Hospedales T, Antoniou A, Micaelli P & Storkey A (2020) Meta-learning in neural networks: a survey. *arXiv*. 2004.05439 [cs.LG].
87 Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H & He Q (2021) A comprehensive survey on transfer learning. *Proc IEEE* **109**, 43–76.
88 Zhu Y, Brettin T, Evrard YA, Partin A, Xia F, Shukla M, Yoo H, Doroshow JH & Stevens RL (2020) Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep* **10**, 18040.
89 Vanschoren J (2019) Meta-Learning. In Automated Machine Learning: Methods, Systems Challenges (Hutter F, Kotthoff L & Vanschoren J, eds), pp. 35–61.Springer International Publishing, Cham. ISBN 9783030053185.
90 Qiu YL, Zheng H, Devos A, Selby H & Gevaert O (2020) A meta-learning approach for genomic survival analysis. *Nat Commun* **11**, 187.
91 Oliver A, Odena A, Raffel C, Cubuk ED & Goodfellow IJ (2018) Realistic evaluation of deep semi-supervised learning algorithms. In Proceedings of the Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 3239–3250.Curran Associates Inc., Red Hook, NY.
92 van Engelen JE & Hoos HH (2020) A survey on semi-supervised learning. *Mach Learn* **109**, 373–440.
93 Castro DC, Walker I & Glocker B (2020) Causality matters in medical imaging. *Nat Commun* **11**, 3673.
94 Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A & Bengio Y (2014) Generative adversarial nets. In Proceedings of the Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pp. 2672–2680.MIT Press, Cambridge, MA.
95 Kingma DP & Welling M (2019) An introduction to variational autoencoders. *FNT in Machine Learning* **12**, 307–392.
96 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L & Polosukhin I (2017) Attention is all you need. *arXiv*. 1706.03762 [cs.CL].
97 Alqahtani H, Kavakli-Thorne M & Kumar G (2019) Applications of generative adversarial networks (GANs): an updated review. *Arch Comput Methods Eng* **9**, 147.
98 Shorten C & Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *J Big Data* **6**, 1106.
99 Fernandez A, Garcia S, Herrera F & Chawla NV (2018) SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J Artif Intell Res* **61**, 863–905.
100 Goncalves A, Ray P, Soper B, Stevens J, Coyle L & Sales AP (2020) Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* **20**, 108.
101 Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, Jörnsten R, Lee EJ, Li X, Lilja S *et al.* (2019) Digital twins to personalize medicine. *Genome Med* **12**, 4.
102 Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.
103 de Anda-Jáuregui G & Hernández-Lemus E (2020) Computational oncology in the multi-omics era: state of the art. *Front Oncol* **10**, 423.
104 Tan A, Huang H, Zhang P & Li S (2019) Network-based cancer precision medicine: a new emerging paradigm. *Cancer Lett* **458**, 39–45.

105 Filipp FV (2019) Opportunities for artificial intelligence in advancing precision medicine. *Curr Genet Med Rep* **7**, 208–213.
106 Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56.
107 Larson DB, Harvey H, Rubin DL, Irani N, Tse JR & Langlotz CP (2020) Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. *J Am Coll Radiol* **S1546-1440**, 31020–6.

Chapter results summary

The main concepts introduced in the review publication presented in this chapter are the following:

1. Finer investigation is required to understand the inherent biomedical data granularity, including sample size and label availability.

2. Lack of proper considerations regarding data granularity reflecting the variety and size of measurements in cancer research can have a deep impact in clinical decision-making.

3. AI has the potential to transcend the current limitations of Big Data analysis with respect to data granularity.

4. Traditional data management may be overwhelmed by the scale size of the current available biomedical information, making computational efforts crucial for its efficient handling.

5. Application of AI-based approaches is already making a huge impact in areas such as prediction of clinical outcome, subtyping and disease diagnosis.

6. High Performance Computing (**HPC**) is key for efficient AI development, not only in terms of computational performance scalability, but also for reducing research costs.

7. Heterogeneous levels of data granularity difficult Big Data analysis, provided the imbalance between cohort size and feature cardinality. Moreover, disaggregation of cohorts by demographic factors can reveal imbalances and potential dataset biases.

8. In small cohorts, an intrinsic property of rare disease research, imbalance impact becomes even more apparent, and may lead to biased outcomes.

9. Multi-omics data modelling, transfer learning, meta-learning and data augmentation come as promising options to overcome limitations from imbalanced data analysis, while accounting for the underlying granularity of oncological data.

Chapter 3

Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes

Publication Record

This chapter presents the preprint of the original research article '*Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes*' currently under revision at the Nature Communications journal (2021 Journal Impact factor: 17.694; Q1 in the Multidisciplinary sciences field. Rank: 6/74). (2022 Journal Impact factor: 16.6; Q1 in the Multidisciplinary sciences field. Rank: 6/73).

Co-authors & affiliations

Iker Núñez-Carpintero [1,*], Emily O'Connor [2,4,*], Maria Rigau [1,5,6], Mattia Bosio [1,7], Sally Spendiff [2], Yoshiteru Azuma [8,9], Ana Topf [10,11], Rachel Thompson [2] , Peter A.C. 't Hoen [12], Teodora Chamova [13] , Ivailo Tournev [13,14], Velina Guergueltcheva [15], Steven Laurie [16], Sergi Beltran [16,17,18], Salvador Capella [1,7], Davide Cirillo [1,#], Hanns Lochmüller [2,3,4,16,19], Alfonso Valencia [1,20]

1. Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain

2. Children's Hospital of Eastern Ontario Research Institute; Ottawa, Canada

3. Division of Neurology, Department of Medicine, The Ottawa Hospital; Ottawa, Canada

4. Brain and Mind Research Institute, University of Ottawa, Ottawa, Canada

5. MRC London Institute of Medical Sciences, Du Cane Road, London, W12 0NN, United Kingdom

6. Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN, United Kingdom

7. Spanish National Bioinformatics Institute Unit, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

8. Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan

9. Department of Pediatrics, Aichi Medical University, Nagakute, Japan

10. John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

11. Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

12. Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands

13. Department of Neurology, Expert Centre for Hereditary Neurologic and Metabolic Disorders, Alexandrovska University Hospital, Medical University-Sofia, Sofia, Bulgaria

14. Department of Cognitive Science and Psychology, New Bulgarian University, Sofia 1618, Bulgaria

15. Clinic of Neurology, University Hospital Sofiamed, Sofia University St. Kliment Ohridski, Sofia, Bulgaria.

16. Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain

17. Universitat Pompeu Fabra (UPF), Barcelona, Spain

18. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain.

19. Department of Neuropediatrics and Muscle Disorders, Medical Center – University of Freiburg, Faculty of Medicine, Freiburg, Germany

20. ICREA, Pg. Lluís Companys 23, 08010, Barcelona, Spain

* These authors contributed equally

Corresponding author: davide.cirillo@bsc.es

Current reference

Núñez-Carpintero I, O'Connor E, Rigau M, Bosio M, Spendiff S, Azuma Y, Topf A, et al. Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes *bioRxiv*; (2023). p. 2023.01.19.524736.

Available from: https://www.biorxiv.org/content/10.1101/2023.01.19.524736v1

Contribution of the PhD candidate

As first co-main author of this research publication, the PhD candidate developed the multilayer network-based pipeline presented in the article, reconstructing the mechanistic process underlying phenotypic severity of the disease. The PhD candidate also analyzed the relationship of the candidate variants as well as suggested the new candidate gene for experimental validation, which was undertaken by the other co-main author, Dr. Emily O'Connor. See '**Author contributions**' (page 91).

Article abstract

Exploring the molecular basis of disease severity in rare disease scenarios is a challenging task provided the limitations on data availability. Causative genes have been described for Congenital Myasthenic Syndromes (CMS), a group of diverse minority neuromuscular junction (NMJ) disorders; yet a molecular explanation for the phenotypic severity differences remains unclear. Here, we present a workflow to explore the functional relationships between CMS causal genes and altered genes from each patient, based on multilayer network analysis of protein-protein interactions, pathways, and metabolomics.

Our results show that CMS severity can be ascribed to the personalized impairment of extracellular matrix components and postsynaptic modulators of acetylcholine receptor (AChR) clustering. We explore this in more detail for one of the proteins not previously associated with the NMJ, USH2A. Loss of the zebrafish USH2A ortholog revealed some effects on early movement and gross NMJ morphology.

This work showcases how coupling multilayer network analysis with personalized -omics information provides molecular explanations to the varying severity of rare diseases, paving the way for sorting out similar cases in other rare diseases.

Keywords: multi-omics data, network biology, multilayer networks, personalized medicine, applied network science, network community analysis, rare diseases, congenital myasthenic syndromes.

1. Introduction

Understanding phenotypic severity is crucial for prediction of disease outcomes, as well as for administration of personalized treatments. Different severity levels among patients presenting the same medical condition could be explained by characteristic relationships between diverse molecular entities (i.e., gene products, metabolites, etc.) in each individual. In this setting, multi-omics data integration is becoming a promising tool for research, as it has the potential to gain complex insights of the molecular determinants underlying disease heterogeneity. However, even in a scenario where the level of biomedical detail available to study is growing in an exponential manner (Karczewski and Snyder, 2018), the analysis of the molecular determinants of disease severity is not typically addressed in rare disease research literature (Boycott et al., 2013), despite its obvious relevance at the medical and clinical level. Rare diseases represent a challenging setting for the application of precision medicine because, by definition, they affect a small number of patients, and therefore the data available for study is considerably limited in comparison to other conditions. Accordingly, leveraging the wealth of biomedical knowledge of diverse nature coming from publicly available databases has the potential to address data limitations in rare diseases (Buphamalai et al., 2021; Mitani and Haneuse, 2020). In this sense, multilayer networks can offer a holistic representation of biomedical data resources (Gosak et al., 2018; Halu et al., 2019), which may allow exploration of the biology related to a given disease independently of cohort sizes and their available omics data.

Here, in order to evaluate and demonstrate the potential of multilayer networks as means of assessing severity in rare disease scenarios, we provide an illustrative case where we develop a framework for analyzing a patient cohort affected by Congenital Myasthenic Syndromes (CMS), a group of inherited rare disorders of the neuromuscular junction (NMJ). Fatigable weakness is a common hallmark of these syndromes, that affects approximately 1 patient in 150,000 people worldwide.

Location	Phenotype	Inheritance	Gene
2q31.1	CMS1A, slow-channel	AD	*CHRNA1*
2q31.1	CMS1B, fast-channel	AR, AD	
17p13.1	CMS2A, slow-channel	AD	*CHRNB1*
17p13.1	CMS2C, associated with acetylcholine receptor deficiency	AR	
2q37.1	CMS3 A, slow-channel	AD	*CHRND*
2q37.1	CMS3 B, fast-channel	AR	
2q37.1	CMS3 C, associated with acetylcholine receptor deficiency	AR	
17p13.2	CMS4 A, slow-channel	AR, AD	*CHRNE*
17p13.2	CMS4 B, fast-channel	AR	
17p13.2	CMS4 C, associated with acetylcholine receptor deficiency	AR	
3p25.1	CMS5	AR	*COLQ*
10q11.23	CMS6, presynaptic	AR	*CHAT*
1q32.1	CMS7, presynaptic	AD	*SYT2*
1p36.33	CMS8, with pre- and postsynaptic defects	AR	*AGRN*
9q31.3	CMS9, associated with acetylcholine receptor deficiency	AR	*MUSK*
4p16.3	CMS10	AR	*DOK7*
11p11.2	CMS11, associated with acetylcholine receptor deficiency	AR	*RAPSN*
2p13.3	CMS12, with tubular aggregates	AR	*GFPT1*
11q23.3	CMS13, with tubular aggregates	AR	*DPAGT1*
9q22.33	CMS14, with tubular aggregates	AR	*ALG2*
1p21.3	CMS15, without tubular aggregates	AR	*ALG14*
17q23.3	CMS16	AR	*SCN4A*
11p11.2	CMS17	AR	*LRP4*
20p12.2	CMS18	AD	*SNAP25*
10q22.1	CMS19	AR	*COL13A1*
2q12.3	CMS20, presynaptic	AR	*SLC5A7*
10q11.23	CMS21, presynaptic	AR	*SLC18A3*
2p21	CMS22	AR	*PREPL*
22q11.21	CMS23, presynaptic	AR	*SLC25A1*
15q23	CMS24, presynaptic	AR	*MYO9A*
12p13.31	CMS25, presynaptic	AR	*VAMP1*

3p21.31	CMS, related to GMPPB	AR	*GMPBB*
20q13.33	CMS, presynaptic	AR	*LAMA5*
3p21.31	CMS, with nephrotic syndrome	AR	*LAMB2*
8q24.3	CMS, with plectin defect	AR	*PLEC*
12q24.13	CMS, related to RPH3A	AR	*RPH3A*
9p13.3	CMS, presynaptic, related to MUNC13-1	AR	*UNC13B*
2q37.1	Escobar syndrome	AR	*CHRNG*

Table 1. Location, phenotype, inheritance, and genes involved in CMS (adapted from https://omim.org/phenotypicSeries/PS601462 and http://www.musclegenetable.fr). AR: autosomal recessive; AD: autosomal dominant.

The inheritance of CMS is autosomal recessive in the majority of patients. CMS can be considered a relevant use case because, while patients share similar clinical and genetic features (Finsterer, 2019), phenotypic severity of CMS varies greatly, with patients experiencing a range of muscle weakness and movement impairment. While over 30 genes are known to be monogenic causes of different forms of CMS (**Table 1**), these genes do not fully explain the ample range of observed severities, which has been suggested to be determined by additional factors involved in neuromuscular function (Thompson et al. 2019). Examples of CMS-related genes are AGRN, LRP4 and MUSK which code for proteins that mediate communication between the nerve ending and the muscle, which is crucial for formation and maintenance of the NMJ (**Figure 1**).

In particular, the AGRN-LRP4 receptor complex activates MUSK by phosphorylation, inducing clustering of the acetylcholine receptor (AChR) in the postsynaptic membrane allowing the presynaptic release of acetylcholine (ACh) to trigger muscle contraction (Burden et al., 2013; Li et al., 2018) . Additional evidence of CMS severity heterogeneity emerged within the NeurOmics and RD-Connect projects (Lochmüller et al., 2018) studying a small population (about 100 individuals) of gypsy ethnic origin from Bulgaria.

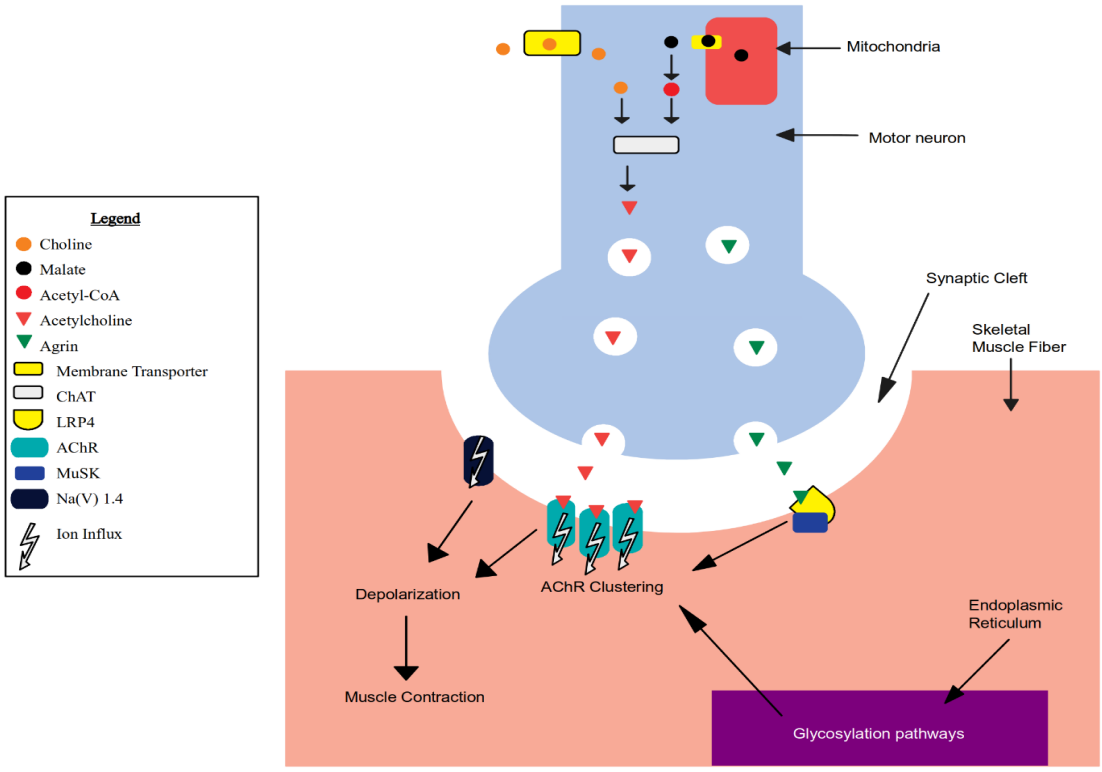


Figure 1. A schematic depiction of the main molecular activities of known CMS causal genes (Methods) taking place at the neuromuscular junction (NMJ) in the presynaptic terminal (in blue), synaptic cleft (in white), and skeletal muscle fiber (in yellow) (for a detailed description of this system see **Supplementary Information**).

All affected individuals shared the same causal homozygous mutation (a deletion within the AChR ε subunit, *CHRNE* c.1327delG (A. Abicht et al. 1999)), however, the severity of symptoms across this cohort varies considerably regardless of age, gender, and initiated therapy, suggesting the existence of additional genetic causes for the diversity of disease phenotypes. By analyzing multi-omics data, we performed an in-depth characterization of 20 CMS patients, representing the two opposite ends of the spectrum observed in the wider cohort, aiming to investigate the molecular basis of the observed differences in the individual severity of the disease. Clinically, CMS severity ranges from minor symptoms (e.g., exercise intolerance) to more severe CMS forms and is dependent on the causal genetic impairments (Abicht et al., 1993; Della Marina et al., 2020). Severe CMS is typically presented with reduced

Forced Vital Capacity (FVC), severe generalized muscle fatigue and weakness, proximal and bulbar muscle fatigue and weakness, impaired myopathic gait and hyperlordosis. Two CMS severity levels have been identified through extensive phenotyping, namely a severe disease phenotype (8 patients) and a not-severe disease phenotype (2 intermediate and 10 mild patients) (**Suppl. Table 1**). Out of the tested demographic factors (age, sex) and clinical tests (speech, mobility, respiratory dysfunctions, among others), FVC and shoulder lifting ability show a significant association with the severity classes (two-tailed Fisher's exact test p-values of 0.0128 and 0.0418, respectively**; Suppl. Figure 1**). We sought to interrogate whether severity was determined by additional genetic variations impacting neuromuscular activity, on top of the causative *CHRNE* mutation.

We analyzed three main types of genetic variations: single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and compound heterozygous variants (two recessive alleles located at different loci within the same gene in a given individual). The extensive analysis of the genomic information did not render any SNPs that could be considered a unique cause of disease severity by being common to all the cases. Nevertheless, a number of CNVs and compound heterozygous variants were found to appear exclusively in the different severity groups, in one or more patients. Moreover, the compound heterozygous variants of the severe group are enriched in pathways related to the extracellular matrix (ECM) receptors, which have been proposed as a target for CMS therapy (Ito and Ohno 2018). To investigate the functional relationship between these variants and CMS severity, we designed an analytical workflow based on multilayer networks (**Figure 2**), allowing the integration of external biological knowledge to acquire deeper functional insights. A multilayer network consists of several layers of nodes and edges describing different aspects of a system (Kivelä et al., 2014). In biomedicine, this data representation has been used to study biomolecular interactions (Zitnik and Leskovec, 2017) and diseases (Halu et al., 2019), facilitating integration and interpretation of heterogeneous sources of data.

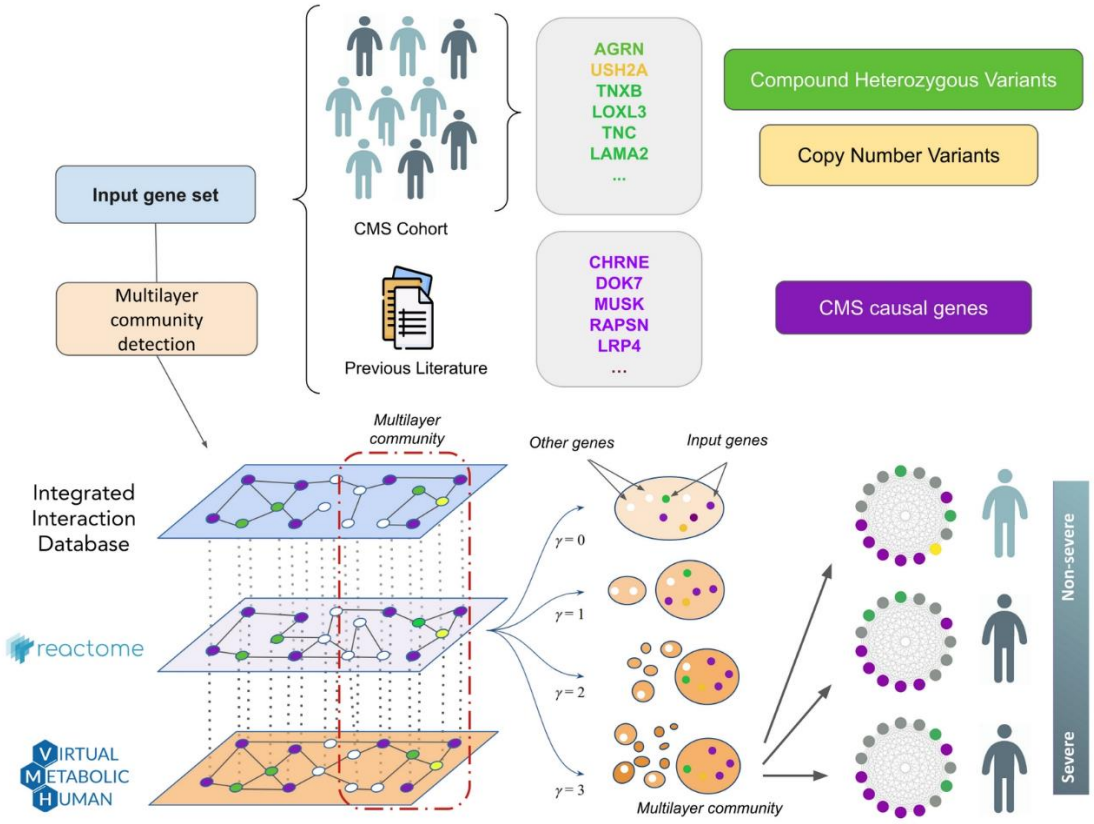


Figure 2. Analytical workflow employed to address the severity of a cohort of patients affected by Congenital Myasthenic Syndromes (CMS). A multi-scale functional analysis approach, based on multilayer networks, was used to identify the functional relationships between genetic alterations obtained from omics data (Whole Genome Sequencing, WGS; RNA-sequencing, RNAseq) with known CMS causal genes. Modules of CMS linked genes are detected using graph community detection at a resolution range (γ) (Methods) where the most prominent changes in community structure occur. Modules that emerged from this analysis were characterized at single individual level.

Several established tools for network analysis have been recently adapted for multilayer networks, such as random walk with restart (Edler et al., 2017; Valdeolivas et al., 2019), community detection algorithms (Didier et al., 2015) and node embeddings (Pio-Lopez et al., 2021). By crossing patient genomic data with the information provided by a biomedical knowledge multilayer network, we are able to describe the functional relationships of new genetic modifiers responsible for the different phenotypic severity levels, showcasing the potential of multilayer networks to provide support on the analysis of rare disease patients.

2. Results

2.1. Variants do not segregate with patient severity

We first searched for variants able to segregate the disease phenotypes (severe and not-severe) by analyzing a large panel of mutational events (mutations in isoforms, splicing sites, small and long noncoding genes, promoters, TSS, predicted pathogenic mutations, loss of function mutations, among others). We could not find one single mutation or combinations of mutations that were able to completely segregate the two groups (Supplementary Information) although partial segregation can be observed (**Suppl. Table 2**).

As already described for monogenic diseases (Kousi and Katsanis, 2015) and cancer (Castro-Giner et al., 2015), we hypothesized that distinct weak disease-promoting effects may represent patient-specific causes to CMS severity, which bring damage to sets of genes that are functionally related. To find these causes, we sought to search for variants with the potential to alter gene functions, such as CNVs and compound heterozygous variants, which have been previously reported to be key to CMS (Abicht et al., 1993; Bevilacqua et al., 2017; Richard et al., 2003; Yang et al., 2018).

2.2. Compound heterozygous variants are functionally related

In order to explore the hypothesis that disease severity in this cohort is due to variants in patient-specific critical elements, we sought to identify potentially damaging compound heterozygous variants and CNVs. We analyzed the gene lists associated with these mutations to search for evidence of alterations in relevant pathways for the severe (n=8) and not-severe cases (n=12). We first performed a functional enrichment analysis (**Methods**) of the genes with CNVs found in the two groups. The set of affected genes in the severe group is composed of 26 unique genes (10 private to the severe group), while the not-severe group presented 86 unique genes (**Suppl. Table 3**).

None of these gene sets showed any functional enrichment. Moreover, none of these genes had been described as causal for CMS, and none carried compound heterozygous variants. (Suppl. Figure 2). As for compound heterozygous variants, the set of affected genes in the severe group is composed of 112 unique genes (89 private to the severe group), while the not-severe group resulted in 152 unique genes (**Suppl. Table 3**).

We found that the severe group shows significant enrichment in genes belonging to extracellular matrix (ECM) pathways, in particular “ECM receptor interactions” (KEGG hsa04512, adjusted p-value 0.002337) and “ECM proteoglycans” (Reactome R-HSA30001787, adjusted p-value 0.001237), which are the top-hit pathways when the 89 genes appearing only in the severe group are considered. Both these pathways share common genes, namely TNXB, LAMA2, TNC, and AGRN. The role of extracellular matrix proteins for the formation and maintenance of the NMJ has recently drawn attention to the study of CMS (Beeson, 2016; Rodríguez Cruz et al., 2018).

In particular, within the genes linked with ECM pathways, AGRN and LAMA2 stand out for their implication in CMS and other rare neuromuscular diseases (Bertini et al., 2011; Bönnemann et al., 2014; Nicole et al., 2014). ECM-related pathways are not enriched in the not-severe set of genes (KEGG hsa04512, adjusted p-value 0.6170). Moreover, top-hit pathways of the not-severe set of genes are not explicitly related to ECM and not consistent between Reactome and KEGG (Reactome “Susceptibility to colorectal cancer” R-HSA-5083636, adjusted p-value 4.131e-7, genes MUC3A/5B/12/16/17/19; KEGG “Huntington's disease” hsa05016, adjusted p-value 0.07103, genes REST, CREB3L4, CLTCL1, DNAH2/8/10/11).

These findings support our hypothesis that severe patients might present disruptions in NMJ functionally related genes that, combined with CHRNE causative alteration, may be responsible for the worsening of symptoms.

2.3. CMS-specific monolayer and multilayer community detection

As disease-related genes tend to be interconnected (Menche et al. 2015), we sought to analyze the relationships among the CMS linked genes (i.e., known CMS causal genes, and severe and not-severe compound heterozygous variants and CNVs; **Methods**) using network community clustering analysis.

We employed the Louvain algorithm (**Methods**) to find groups of interrelated genes in three monolayer networks that represents biological knowledge contained in databases, separately: the Reactome database (Fabregat et al. 2018), the Recon3D Virtual Metabolic Human database (Brunk et al. 2018) (both downloaded in May 2018), and from the Integrated Interaction Database (IID) (Kotlyar et al. 2016) (downloaded in October 2018) (**Suppl. Figure 3**). The first network consists of 10,618 nodes (genes) and 875,436 edges, representing shared pathways between genes. The second network consists of 1,863 nodes (genes) and 902,188 edges, representing shared reaction metabolites between genes. The third network consists of 18,018 nodes (genes) and 947,606 edges, representing aggregated protein-protein interactions from all tissues (**Methods: Monolayer community detection**). The last two networks represent the 'metabolome' and the 'interactome' data, respectively. By measuring community similarity (**Methods**), we observed that the same CMS linked genes did not form the same communities across the different networks (**Suppl. Figure 4**).

These results show that, although disease related genes are prone to form well-defined communities in distinct networks (Goh et al. 2007; Cantini et al. 2015), different facets of biological information (i.e., reactome, metabolome, interactome) reflect diverse participation modalities of such genes into communities. In order to deliver an integrated analysis of such heterogeneous information, we further consider them as a multilayer network (Gosak et al. 2018). (**Methods: Monolayer community detection and Multilayer community detection**).

2.4. Large-scale multilayer community detection of disease associated genes

We first sought to test the hypothesis that disease-related genes tend to be part of the same communities also in a multilayer network setting. We used the curated gene-disease associations database DisGeNET (Piñero et al., 2017), showing that disease-associated genes are significantly found to be members of the same multilayer communities (**Wilcoxon test p-value < 0.001 in a range of resolution parameters described in the Methods**). We preprocessed DisGeNET database by filtering out diseases and disease groups with only one associated gene (6,352 diseases), and those whose number of associated genes was more than 1.5 * interquartile range (IQR) of the gene associated per disease distribution (823 diseases with more than 33 associated genes) (**Suppl. Figure 5A-B**). This procedure prevents a possible analytical bias due to the higher amounts of genes annotated to specific disease groups (e.g., entry C4020899, “Autosomal recessive predisposition”, annotates 1445 genes). We then retrieved the communities of each associated gene, excluding 428 genes not present in our multilayer network and the diseases left with only one associated gene. The final analysis comprised a total of 5,892 diseases with an average number of 7.38 genes per disease. For each disease, we counted the number of times that disease-associated genes are found in the same multilayer communities and compared the distribution of such frequencies with that of balanced random associations (1000 randomizations). Results show that disease-associated genes are significantly found in the same multilayer communities across the resolution interval (**Suppl. Figure 5C**).

2.5. Modules within the CMS multilayer communities

We define a module as a group of CMS linked genes that are systematically found to be part of the same multilayer community while increasing the multilayer network community resolution parameter (**Methods; Supplementary Information; Figures 3-4**). Within each of these communities, we identified smaller modules of CMS linked

genes that are specific to the severe and not-severe groups. We tested the significance of obtaining these exact genes in the severe and not-severe largest modules upon severity class label shuffling among all individuals (1000 randomizations). We found that 13 (p-value 0.022) and 14 (p-value 0.027) are the minimum number of genes composing the modules that are not expected to be found at random in the severe and not-severe largest components, respectively (**Suppl. Figure 6**).

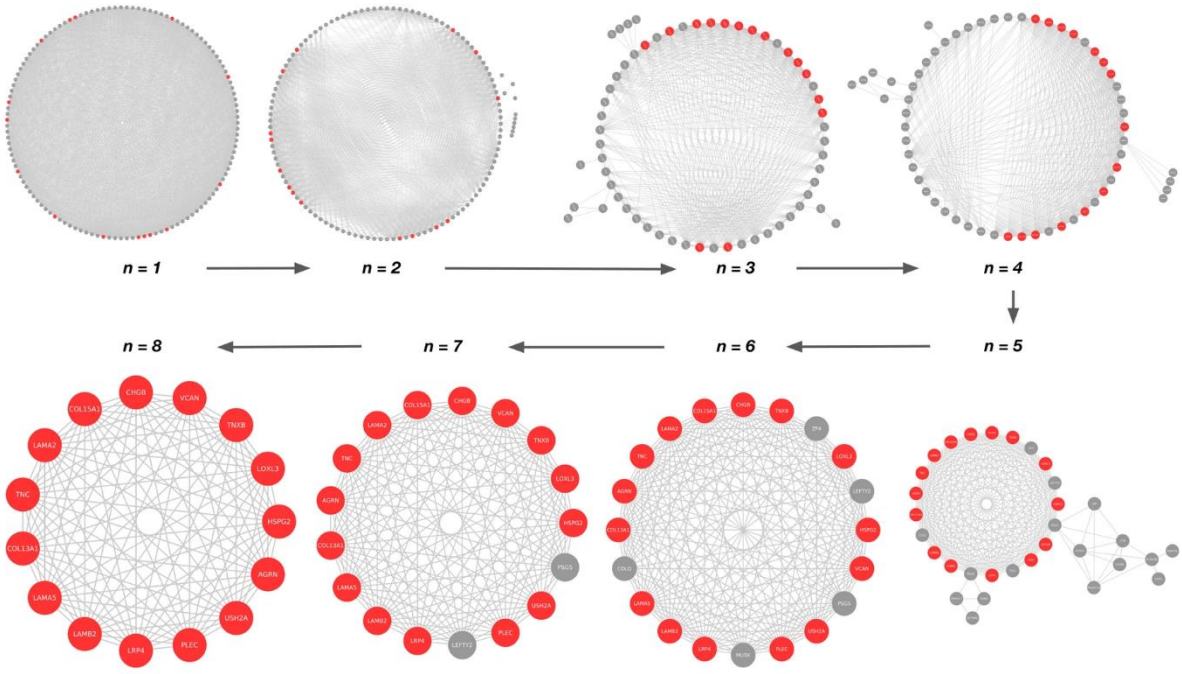


Figure 3. Identification of the largest module containing genes that are found in the same community in the entire range of resolution parameters (Methods). In each module, genes are connected if they are found in the same multilayer communities at n values of the resolution parameter γ within the range under consideration (γ (0,4]). The arrows indicate the systematic increase of ∈ n. At n = 8, the module contains genes that are always found in the same community in the entire range of resolution (see Supplementary Information "Multilayer community detection analysis"). The largest modules containing the CMS linked gene set (highlighted in red), which includes known CMS causal genes, severe-specific heterozygous compound variants and CNVs, are shown.

In the two groups, the significantly largest module that contains known CMS causal genes is composed of 15 genes (**Figure 4**). 6 out of these 15 are previously described CMS causal genes (**Methods**), namely the ECM heparan sulfate proteoglycan agrin (*AGRN*); the cytoskeleton component plectin (*PLEC*), causative of myasthenic disease (Forrest et al. 2010); the agrin receptor *LRP4*, key for AChR clustering at NMJ (Barik et al. 2014) and causative of CMS by compound heterozygous variants (Ohkawara et al. 2014); the ECM components *LAMA5* and *LAMB2* laminins, and *COL13A1* collagen. Considering all nodes (not only CMS linked) the number of nodes in the module is 482.

All the other genes of the two modules are involved in a varied spectrum of muscular dysfunctions, discussed in the following sections. As the location of the causal gene products determine the most common classification of the disease (i.e., presynaptic, synaptic, and postsynaptic CMS) (Rodríguez Cruz et al., 2018), we determined class and localization of the members of the found modules (**Table 2**).

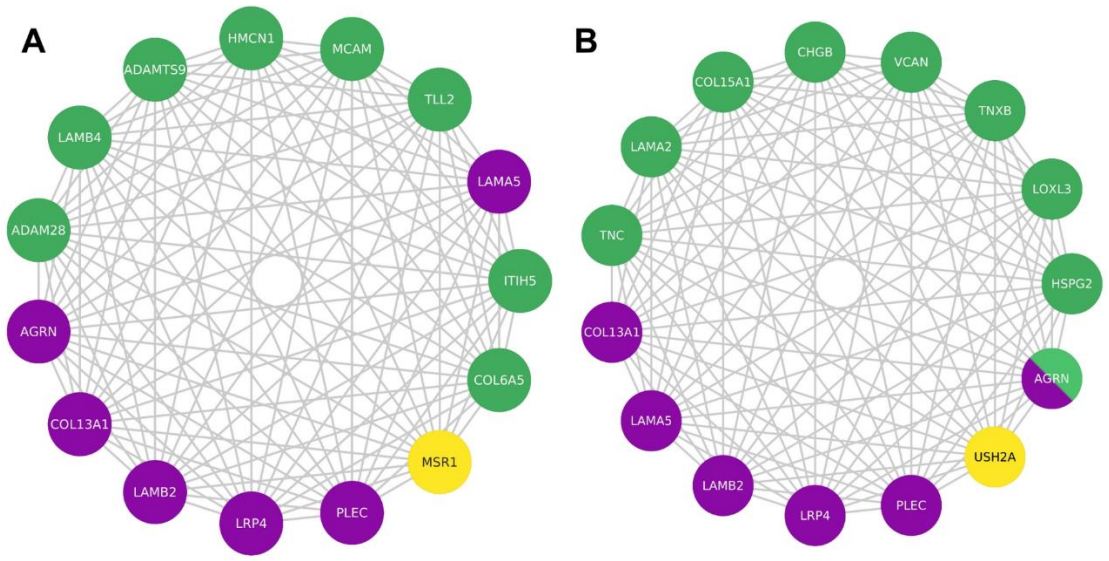


Figure 4. Largest module, containing known CMS causal genes, within the multilayer communities of CMS linked genes that are specific to the not-severe (A) and severe (B) groups. In green, compound heterozygous variants; in yellow, CNVs; in purple, known CMS causal genes. Being a CMS causal gene bearing compound heterozygous variants, AGRN is depicted using both green and purple.

Laminins, well-known CMS glycoproteins, are affected in both severe (LAMA2, USH2A) and not-severe (LAMB4) groups, and are bound by specific receptors that are damaged in the not-severe group (MCAM) (Dagur and McCoy, 2015). Collagens, known CMS-related factors, are associated with the not-severe group (COL6A5), and bound by specific receptors that are damaged in the not-severe group (MSR1) (Di Martino et al., 2023). However, collagen biosynthesis is affected in both severe and not-severe groups. Indeed, metalloproteinases, damaged in the not-severe group, are responsible for the proteolytic processing of lysyl oxidases (*LOXL3*), which are implicated in collagen biosynthesis (Panchenko et al. 1996) and damaged in the severe group. Alterations in proteoglycans (*AGRN, HSPG2, VCAN, COL15A1*) (Iozzo and Schaefer, 2015) , tenascins (*TNC, TNXB*) (Flück et al., 2008; van Dijk et al., 1993), and chromogranins (*CHGB*) (Andreose et al., 1994) are specific of the severe group. We observed no genes associated with proteoglycan damage in the not-severe group, suggesting a direct involvement of ECM in CMS severity.

2.6. Personalized analysis of the severe cases

We sought to analyze the 15 genes of the largest module of the severe group in each one of the 8 patients, hereafter referred to using the WGS sample labels (**Suppl. Table 1**). At the topological level, all incident interactions existing between the genes of the severe module (**Figure 4B**) are related to the protein-protein interaction and pathway layers (**Supplementary Figure 7**). Overall, these genes have a varied range of expression levels in tissues of interest (**Suppl. Figure 8**), for instance in skeletal muscle *HSPG2, LAMA2, PLEC* and *LAMB2* show medium expression levels (9 to 107 TPM) while the others show low expression levels (0.6 to 9 TPM) (**Methods**).

Patient 2, a 15 years old male, presents compound heterozygous variants in tenascin C (*TNC*), mediating acute ECM response in muscle damage (Flück et al., 2008; Sorensen et al., 2018), and CNVs (specifically, a partial heterozygous copy number loss) in usherin (*USH2A*), which have been associated with hearing and vision loss (Austin-Tse et al., 2018).

Patient 16, a 25 years old female, presents compound variants in tenascin XB (*TNXB*), which is mutated in Ehlers-Danlos syndrome, a disease that has already been reported to have phenotypic overlap with muscle weakness (Kirschner et al., 2005; Matsumoto and Aoki, 2020; Okuda-Ashitaka and Matsumoto, 2023; Voermans and Engelen, 2008) and whose compound heterozygous variants have been reported for a primary myopathy case (Pénisson-Besnier et al., 2013; Voermans et al., 2014), and versican (*VCAN*), which has been suggested to modify tenascin C expression (Keller et al., 2012) and is upregulated in Duchenne muscular dystrophy mouse models (McRae et al., 2017, 2020).

Patient 13, a 26 years old male, presents compound mutations in laminin α2 chain (LAMA2), a previously reported gene related to various muscle disorders (AMIN et al., 2019; Dimova and Kremensky, 2018; Løkken et al., 2015) whose mutations cause reduction of neuromuscular junction folds (Rogers and Nishimune, 2017), and collagen type XV α chain (COL15A1), which is involved in guiding motor axon development (Guillon et al., 2016) and functionally linked to a skeletal muscle myopathy (Eklund et al., 2001; Muona et al., 2002).

Patient 12, a 49 years old female, presents compound mutations in chromogranin B4 (CHGB), potentially associated with amyotrophic lateral sclerosis early onset (Gros-Louis et al., 2009; Pampalakis et al., 2019). Patient 18, a 51 years old man, presents compound mutations in agrin (AGRN), a CMS causal gene that mediates AChR clustering in the skeletal fiber membrane (Huzé et al., 2009) (Jacquier et al., 2022).

Patient 20, a 57 years old male, presents compound mutations in lysyl oxidase-like 3 (LOXL3), involved in myofiber extracellular matrix development by improving integrin signaling through fibronectin oxidation and interaction with laminins (Kraft-Sheleg et al., 2016), and perlecan (HSPG2) (Zoeller et al., 2008), a protein present on skeletal muscle basal lamina (Carmen et al., 2019; Larraín et al., 1997), whose deficiency leads to muscular hypertrophy (Xu et al., 2010), that is also mutated in Schwartz-Jampel syndrome (Stum et al., 2006), Dyssegmental dysplasia Silverman-

Handmaker type (DDSH) (Arikawa-Hirasawa et al., 2001) and fibrosis (Lord et al., 2018), such as Patient 19, a 62 years old female. Furthermore, based on the estimated familial relatedness (**Methods**) and personal communication (**February 2018, Teodora Chamova),** patients 19 and 20 are siblings (**Suppl. Table 4**).

2.7. Functional consequences of variants in the severe-specific module

Studying the functional impact of the compound heterozygous variants in the severe-specific genes of the module, we observed that in 6 of the 8 patients at least one of the variants is predicted to be deleterious by the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) (**Methods; Suppl. Table 5**). For example, as for Patient 18, who presents 3 different variants in AGRN gene, only rs200607541 is predicted to be deleterious by VEP's Condel (score = 0.756), SIFT (score = 0.02), and PolyPhen (score = 0.925). In particular, the variant (a C>T transition) presents an allele frequency (AF) of 4.56E-03 (gnomAD exomes) (Karczewski et al., 2020) and affects a region encoding a position related to an EGF-like domain (SMART:SM00181) and a Follistatin-N-terminal like domain (SMART:SM00274). Both of these domains are part of the Kazal domain superfamily which are specially found in the extracellular part of agrins (PFAM: CL0005) (Laskowski and Kato, 1980; Porten et al., 2010).

On the other hand, Patient 16 presents a total of 38 TNXB transcripts affected by three gene variants (rs201510617, rs144415985, rs367685759) that are all predicted to be deleterious by the three scoring systems, have allele frequencies of 3.17E-02, 4.83E-02 and 5.90E-03, respectively; and in overall, are affecting two conserved domains.

The first consists of a fibrinogen related domain that is present in most types of tenascins (SMART:SM00186), while the second is a fibronectin type 3 domain (SMART:SM00060) that is found in various animal protein families such as muscle proteins and extracellular-matrix molecules (Bork and Doolittle, 1992).

Activity localization	Class	CMS causal gene	Phenotype group		Function	Synaptic localization (Manual curation)	Localization (UniProt)
			Not-severe	Severe			
ECM (ECM)	Proteoglycans	AGRN	-	AGRN	Cell hydration and growth factor trapping	Pre- and postsynaptic (PMID:29462312)	Synaptic basal lamina / ECM
		-	-	HSPG2		Basement membrane (PMID:30453502)	Basement membrane / ECM
		-	-	VCAN		ECM (PMID:29211034)	ECM
		-	-	COL15A1		Basement membrane (PMID:26937007)	ECM
	Collagens	COL13A1	-	-	Structural support	Basement membrane, post-synaptic (PMID:30768864)	Post-synaptic cell membrane
		-	COL6A5	-		Basement membrane (PMID:23869615)	Extracellular matrix
	Laminins	LAMA5	-	-	Web-like structures	Pre-synaptic (PMID:28544784)	Basement membrane / ECM
		LAMB2	-	-		Basement membrane (PMID:27614294)	Basement membrane / ECM / Synaptic cleft
		-	LAMB4	-		Myenteric plexus basement membrane (PMID:28595269)	Basement membrane / ECM
		-	-	LAMA2		Pre-synaptic (PMID:9396756)	Basement membrane / ECM
		-	-	USH2A		Neuronal projection of stereocilia (PMID:19023448)	Stereocilia membrane / Secreted (Extracellular region)

	Fibulins	-	HMCN1	-	Scaffolding	Glomerular Extracellular matrix (PMID:29488390)	Basement membrane / ECM
	Tenascins	-	-	TNC	Anti-adhesion	Basement membrane (PMID:29466693)	ECM / Perisynaptic ECM (Ensembl)
				TNXB		Basement membrane (PMID:23768946)	ECM
	Enzymes	-	-	LOXL3	Collagen assembly	Basement membrane (PMID:26954549)	Secreted (extracellular region)
			ADAMTS9	-	Proteoglycan cleavage	Secreted to ECM (PMID:30626608)	ECM
			ADAM28			ECM (PMID:24613731)	Cell membrane / Secreted (extracellular region)
	Neuropeptides	-	-	CHGB	Regulatory peptides precursor	Pre- and postsynaptic (PMID:7526287)	Secreted (extracellular region)
	Others	-	ITIH5	-	Hyaluronic acid binding	ECM (PMID:27143355)	Secreted (extracellular region)
Cell surface	Receptors	-	MSR1	-	Proteoglycan and collagen binding	Macrophage surface Scavenger Receptor (PMID:12488451)	Plasma membrane
			MCAM			Plasma membrane (PMID:28923978)	Plasma membrane
		LRP4	-	-	Laminin binding	Post-synaptic (PMID:25319686)	Post-synaptic cell membrane
Cytoplasm	Cytoskeleton	PLEC	-	-	Structural support	Post-synaptic (PMID:20624679)	Post-synaptic cytoskeleton

Table 2. Localization and functions of proteins encoded by the genes found in the largest modules of the multilayer communities of severe and not-severe groups. In green, compound heterozygous variants; in yellow, CNVs; in purple, known CMS causal genes. Synaptic localization was retrieved from manual curation and Uniprot database (Methods).

Two of the severe patients (Patients 12 and 19) present severe-only specific compound heterozygous variants that are not predicted to be deleterious. However, one variant in the *CHGB* gene (rs742710, AF=1.07E-01), present in patient 12, has been previously reported to be potentially causative for amyotrophic lateral sclerosis early onset (Gros-Louis et al., 2009; Pampalakis et al., 2019). This gene has also been strongly suggested in literature as a possible marker for onset prediction in multiple sclerosis (Mo et al., 2013), and other related neural diseases like Parkinson's (Nilsson et al., 2009) and Alzheimer's disease (Chen et al., 2019).

As for patient 19, the variant rs146309392 (AF=8.40E-04) in the gene HSPG2 has been previously referred to be causal of Dyssegmental dysplasia as a compound heterozygous mutation (Arikawa-Hirasawa et al., 2001). This variant, as pointed out before, is shared with sibling patient 20.

One severe individual (Patient 3), a 37 years old female, does not carry compound heterozygous variants included in this module but others at a lower resolution parameter value (Suppl. Figure 9; Suppl. Table 6). Interestingly, most of the genes carrying severe-specific deleterious compound heterozygous variants in this patient (*CDH3, FAAP100, FCGBP, GFY, RPTN*) are not related to processes at the NMJ level (Hull et al., 2016; Johansson et al., 2009; Kaneko-Goto et al., 2013; Ramanagoudr-Bhojappa et al., 2018; Swuec et al., 2017). Nevertheless, three of these variants occur in genes potentially involved in NMJ functionality. In particular, variants rs111709242 (AF=2.64E-03) and rs77975665 (AF=3.03E-02) affect gene *PPFIBP2*, which encodes a member of the liprin family (liprin-β) that has been described to control synapse formation and postsynaptic element development (Astigarraga et al., 2010; Bernadzki et al., 2017).

Furthermore, the variant rs111709242 is predicted to be deleterious by the SIFT algorithm (see Suppl. Table 6). Interestingly, *PPFIBP2* appears in modules at lower resolution parameter values associated with known CMS causal genes (e.g., *DOK7, RPSN, RPH3A, VAMP1, UNC13B*) (**Supplementary Figure 9**). In addition, variant

rs151154986 (AF=2.18E-02) affects the acyl-CoA thioesterase *ACOT2*, which generate CoA and free fatty acids from acyl-CoA esters in peroxisomes (Grevengoed, et al., 2014). While *ACOT2* is lost early during the module detection process, community detection at the individual layer level (i.e., Louvain community detection for each network) revealed relationships with causal CMS genes throughout all layers of the multilayer network system (**Supplementary Figure 3**). Namely, *ACOT2* shares community membership with *ALG14, DPAGT1, GFPT1, GMPPB* and *SLC25A1A* at the protein-protein interaction network; with *CHAT* and *SLC5A7* at the pathways level, and with *GMPBB, SLC25A1* and *CHAT* at the metabolomic layer.

A role for CoA levels in skeletal muscle for this enzyme class has been previously described (Li et al. 2015). Moreover, this patient presents high relatedness with three not-severe patients (Patients 8, 9, and 10) who in turn display a very high relatedness among them (**Suppl. Table 4**).

2.8. Potential pharmacological implications

Finding a genetic diagnosis might help select the appropriate medication for each patient. For instance, fluoxetine and quinine are used for treating the slow-channel syndrome, an autosomal dominant type of CMS caused by mutations affecting the ligand binding or pore domains of AChR, but this treatment should be avoided in patients with fast-channel CMS (Engel et al. 2015). Within our cohort, 13 (7 mild, 2 moderate and 4 severe) out of 20 individuals from our CMS cohort are receiving a pharmacological treatment consisting of pyridostigmine, an acetylcholinesterase inhibitor used to treat muscle weakness in myasthenia gravis and CMS (Lee, Beeson, and Palace 2018). This treatment slows down acetylcholine hydrolysis, elevating acetylcholine levels at the NMJ, which eventually extends the synaptic process duration when the AChR subunits are mutated. Although the severity could potentially be related to how well a patient responds to the standard treatment with the AChE inhibitors, we could not find a clear correlation between severity and pyridostigmine treatment (two-tailed Fisher's exact test p-value 0.356; **Suppl. Figure 1**).

In addition to the causal mutation in *CHRNE*, our results indicate that severity is related to AChR clustering at the Agrin-Plectin-*LRP4*-Laminins axis level, suggesting the potential benefit of pharmaceutical intervention enhancing the downstream process of AChR clustering. For example, beta-2 adrenergic receptor agonists like ephedrine and salbutamol have been documented as capable of enhancing AChR clustering (Clausen et al., 2018) and proved to be successful in the treatment for severe AChR deficiency syndromes (Cruz et al., 2015; Garg and Goyal, 2022). Furthermore, the addition of salbutamol in pyridostigmine treatments has been described as being able to ameliorate the possible secondary effects of pyridostigmine in the postsynaptic structure (Vanhaesebrouck et al., 2019).

2.9. Experimental validations of *USH2A* involvement at the NMJ

To determine the potential relevance of one of our identified potential modifiers with no previously published relationship to the NMJ, we analyzed its function using zebrafish. For this we chose USH2A, a gene associated with Usher syndrome and Retinitis pigmentosa in humans (OMIM ID 608400, https://omim.org/), which was identified as a copy number loss in patient 2. While we expect the phenotypic outcome (more severe disease) of this genetic difference to manifest when expressed in conjunction with the *CHRNE* mutation causing this patients' CMS, we hypothesized that knockdown of *USH2A* expression alone may cause detectable NMJ impairments.

Therefore, we used a MO to knockdown the expression of the zebrafish orthologue; *ush2a*, and studied the effects on survival, development and NMJ function. Zebrafish *ush2a* is expressed from 1 to 5 dpf, as shown in **Suppl. Figure 10A**. Using a MO targeting the exon 3/intron 3 splice donor site we were able to decrease expression of *ush2a* with a 6 ng to 18 ng MO injection (**Suppl. Figure 10B**).

Survival of control and *ush2a*-MO zebrafish was not significantly affected as compared to wildtype (WT) fish over 5 dpf (log-rank test, WT n = 574, control MO 4 ng n = 46, 6 ng n = 75, 18 ng n = 34, ush2a-MO 2 ng n = 72, 4 ng n = 68, 6 ng n =

360, 12 ng n = 288, 18 ng n = 139, **Suppl. Figure 10C**). There were no obvious gross morphological differences between control MO and *ush2a*-MO fish up to 5 dpf (representative images of 2 dpf fish shown in **Suppl. Figure 10D**).

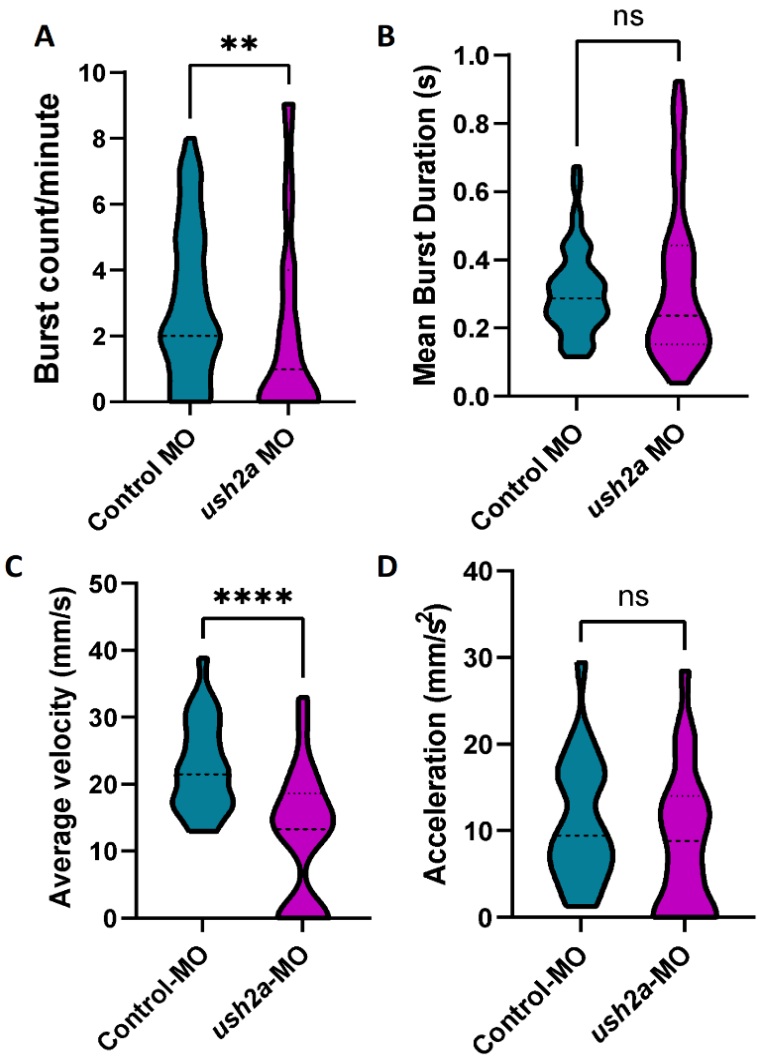


Figure 5. Early movement behaviors in ush2a-MO zebrafish. (A) Chorion rotations per minute (burst count), and **(B)** mean chorion rotation duration in seconds for control and ush2a-MO-injected zebrafish at 1 days post fertilization (dpf). **(C)** Average velocity and **(D)** initial acceleration of control and ush2a-MO zebrafish at 2 dpf in response to touch. Dashed line shows the median, dotted lines show the quartiles, **p < 0.01, ****p < 0.0001, ns = not significant, Mann Whitney test (A and B), unpaired t-test (C and D).

As length is an indicator of developmental stage, we measured the length of 18 ng injected *ush2a*-MO fish at 2 dpf and found a significant reduction in length as compared to controls (p = 0.013, t = 2.59, df = 38, unpaired t-test, control MO n = 20, ush2a-MO n = 20, **Suppl. Figure 10E**).

Eye area can be reduced in zebrafish models of retinitis pigmentosa, the condition that *USH2A* mutations are associated with in humans. We measured eye area in 2 dpf fish and found it to be significantly reduced in 18 ng-injected *ush2a*-MO fish as compared to controls (p = 0.0006, t = 3.73 df = 38, unpaired t-test, control MO n = 20, ush2a-MO n = 20, **Supplementary Figure 10F**). Eye area remains significantly different after normalizing for body length (data not shown).

CMS manifests as fatigable muscle weakness in patients and in developing zebrafish we can study the ability of fish to perform repetitive, well-characterized movements during development to determine whether impairments to the functioning of the neuromuscular system may be present. We quantified the number and duration of chorion movements in 1 dpf fish following administration of a control or 18 ng *ush2a*-MO. This revealed a significant decrease in the number of burst events performed per minute in knockdown fish as compared to controls (p = 0.003, Mann Whitney test, control MO n = 84, ush2a-MO n = 74, **Figure 5A**).

The average duration of each burst event was not significantly affected by loss of *ush2a* (p = 0.467, Mann Whitney test, control MO n = 72, ush2a-MO n = 49, **Figure 5B**). To ascertain whether impairments to movement are present in the knockdown fish while swimming free of the chorion, we also performed a touch response assay at 2 dpf. We observed a significant decrease in average velocity of the fish injected with *ush2a*-MO as compared to control MO in response to a touch stimulus (p < 0.0001, t = 4.42, df = 48, unpaired t-test; n = 25, **Figure 5C**). There was no significant difference in acceleration of *ush2a*-MO fish as compared to controls (p = 0.263, t = 1.13 df = 47, unpaired t-test; control MO n = 24, ush2a-MO n = 25, **Figure 5D**).

To determine whether changes in movement are reflected at the level of gross NMJ structure, analysis of NMJ morphology was performed on 2 dpf zebrafish (**Figure 6A**). A significant decrease in the number of SV2-positive clusters per 100 µm2 (representative of the presynaptic motor neurons) was identified on the fast muscle fibers of *ush2a*-MO fish as compared to controls ($p = 0.0004$, Mann Whitney test, control MO n = 11, ush2a-MO n = 15, **Figure 6B**). SV2-positive clusters overlie postsynaptic AChRs to form NMJs and these receptors can be detected with fluorophore-labelled α-bungarotoxin. Analysis of AChR clusters revealed no significant differences in number per 100 µm2 between the two conditions ($p = 0.217$, Mann Whitney test, control MO n = 11, ush2a-MO n = 15, **Figure 6C**). Colocalization analysis revealed no significant differences in co-occurrence of SV2 and AChR on fast muscle fibers (SV2 colocalization with AChRs: $p = 0.371$, $t = 0.911$, $df = 24$, nested t-test, **Figure 6D** and AChR colocalization with SV2: $p = 0.372$, $t = 0.909$, $df = 24$, control MO n = 11, ush2a-MO n = 15, nested t-test, **Figure 6E**).

There was also no significant difference in colocalization of SV2 with AChRs on slow muscle, however, a significant reduction in co-occurrence of AChRs with SV2 is present on *ush2a*-MO slow muscle (SV2 colocalization with AChRs: $p = 0.516$, $t = 0.660$, $df = 24$, nested t-test, **Figure 6F** and AChR colocalization with SV2: $p = 0.002$, $t = 3.41$, $df = 24$, control MO n = 11, ush2a-MO n = 15, nested t-test, **Figure 6G**). Movement differences in zebrafish may also be caused by changes in muscle growth and development. Therefore, we assessed 2 dpf fish for gross phenotypic differences in muscle fiber orientation and structure using a phalloidin stain to detect actin in muscles (**Suppl. Figure 11A**). We identified no significant differences in muscle fiber dispersion (organization) or myotome size between *ush2a*-MO and control-MO zebrafish ($p = 0.922$, $t = 0.099$, $df = 24$ unpaired t-test and $p = 985$, $t = 0.019$, $df = 24$ nested t-test, respectively. Control MO n = 11 and *ush2a*-MO n = 15, **Suppl. Figure 11B, C**).

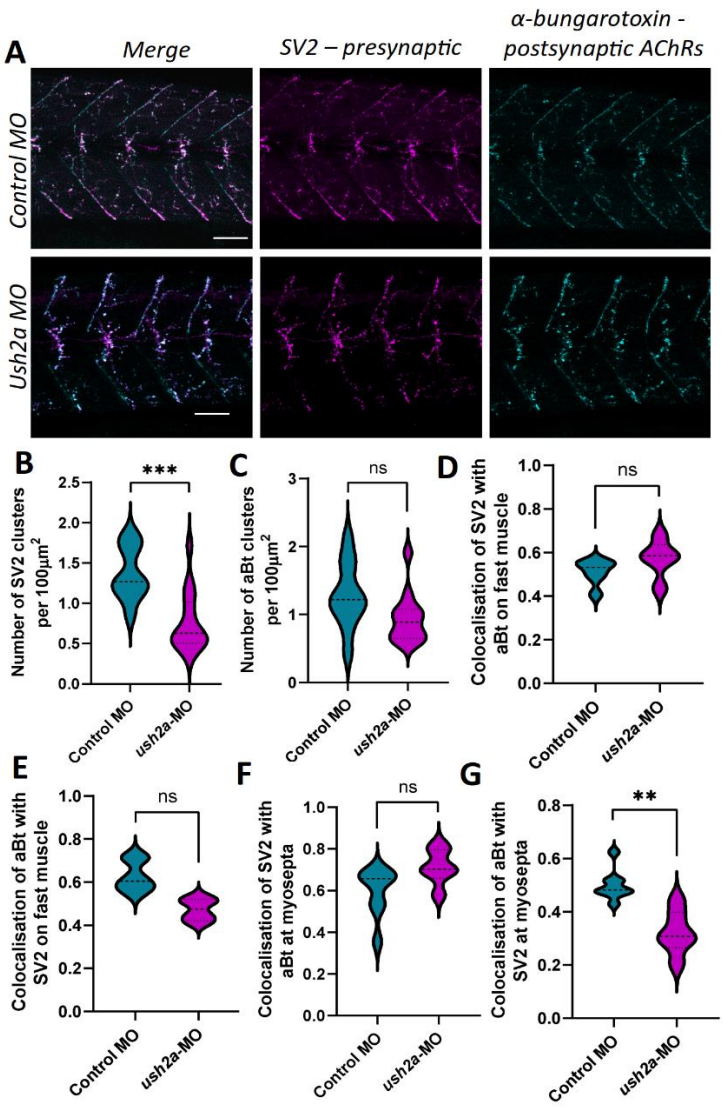


Figure 6. Neuromuscular junction morphology in *ush2a*-MO zebrafish. (A) Representative images of neuromuscular junctions from control and *ush2a*-MO zebrafish at 2 days post fertilization (dpf). Acetylcholine receptors (AChRs) are stained with fluorophore bound α-bungarotoxin (aBt, cyan), and motor neurons detected with an antibody against synaptic vesicle protein 2 (SV2, magenta). Scale bar = 50 µm. (B) Number of SV2-positive clusters and (C) number of aBt-positive clusters per 100 µm^2. (D) Colocalization of SV2 with aBt and (E) colocalization of αBT with SV2 on fast muscle cells, using Mander's correlation coefficient (0 = no colocalization, 1 = full colocalization). (F) Colocalization of SV2 with aBt and (G) colocalization of aBt with SV2 on slow muscle cells at the myosepta, using Mander's correlation coefficient. Dashed line shows the median, dotted lines show the quartiles, **p < 0.05, ***p < 0.001, ns = not significant, nested t-test.

3. Discussion

In this work, we have developed a framework for the analysis of disease severity in scenarios heavily impacted by sample size. Presenting limited numbers of cases is one of the main obstacles for the application of precision medicine methods in rare disease research, as it critically affects the level of expected statistical power, a common hallmark in the analysis of minority conditions (Whicher et al., 2018). This fact makes it difficult to explore the molecular relationships that define the inherently heterogeneous levels of disease severity observed in rare disease populations, making it an atypically addressed biomedical problem (Boycott et al., 2013). Our approach, based on the application of multilayer networks, enable the user to account for the many interdependencies that are not properly captured by a single source of information, effectively combining the available patient genomic information with general biomedical knowledge from relevant databases representing different aspects of molecular biology. The application to a relevant clinical case, where we tested the hypothesis that the severity of CMS is determined by patient-specific alterations that impact NMJ functionality, provided evidence on how the methodology is able to recover the molecular relationships between the candidate patient-specific genomic variants, the observed causal AChR mutation and previously described CMS causal genes (**Table 1**).

Our in-depth functional analysis focused on a cohort of 20 CMS patients, from a narrow, geographically isolated and ethnically homogenous population, who share the same causative mutation in the AChR ε subunit (CHRNE) but present with different levels of severity.

The isolation and endogamy that characterize the population from which these patients come from might have favored the accumulation of damaging variants (Fareed and Afzal, 2017; Petukhova et al., 2009), giving rise to the emergence of compound effects on relevant genes for CMS. This observation has previously been made in similar syndromes (Müller et al., 2004; Ohno et al., 2003) and in a number of

other neuromuscular diseases (Wang et al., 2018; Zhong et al., 2017). Compound heterozygosity is known to happen in CMS (Hantaï et al., 2013) (Thompson et al., 2019). The initial analysis of compound heterozygous variants revealed a significant enrichment of functional categories that are specific to the severe cases, namely ECM functions. This suggests the existence of functional relationships between major actors of the NMJ that are affected by severity-associated damaging mutations. Such interactors include already known CMS causal genes (e.g., *AGRN, LRP4, PLEC*) as well as genes known to interact with them. While severity-specific compound heterozygous variants and CNVs are observed, demographic factors (e.g., sex, age), pharmacological treatment, and personalized omics data (e.g., variant calling, differential gene expression, allele specific expression, splicing isoforms) do not segregate with patient severity.

Therefore, this motivated the developing of our multilayer network community analysis to investigate the relationship between known CMS causal genes and severity-associated variants (compound heterozygous variants and CNVs), integrating pathways, metabolic reactions, and protein-protein interactions. Recently, we used a multilayer network as a means to perform dimensionality reduction tasks for patient stratification in medulloblastoma, a childhood brain tumor (Núñez-Carpintero et al. 2021) (**See Chapter 4**).

Here, we started by analyzing DisGeNET data in order to verify that disease-associated genes tend to belong to the same multilayer communities. We then identified stable and significantly large gene modules within our CMS cohort's multilayer communities and mapped the corresponding damaging mutations back to the single patients, providing a personalized mechanistic explanation of severity differences. Given the difficulties of cohort recruitment for rare diseases, this approach could be used to investigate forms of CMS and other phenotypically variable rare diseases caused by a common mutation.

Overall, our approach revealed major relationships at the protein-protein and pathway layers, with the personalized analysis of these mutations suggesting that CMS severity can be ascribed to the damage of specific molecular functions of the NMJ which, despite affecting individuals in a personalized manner, involve genes belonging to distinct classes and localizations, namely ECM components (proteoglycans, tenascins, chromogranins) and postsynaptic modulators of AChR clustering (*LRP4, PLEC*) (**Table 2**). Alterations of other genes related to the production of ECM components, such as laminins and collagen, are observed but are not specific to the severity levels.

Although at first the usage of metabolomic knowledge as an additional level of the multilayer network system did not seem to provide highly relevant information for the cohort, it provided major information for the personalized analysis of patient 3, whose mutations presented functional relationships with other CMS causal genes outside of the presented severe-specific module (**Supplementary Figure 3**). Finding a personalized genetic diagnosis for phenotypic severity might help select the appropriate medication for each patient. For instance, fluoxetine and quinidine are used for treating the slow-channel syndrome, an autosomal dominant type of CMS caused by mutations affecting the ligand binding or pore domains of AChR, but this treatment should be avoided in patients with fast-channel CMS (Engel et al., 2015). Within our cohort, 13 out of 20 individuals from our CMS cohort are receiving a pharmacological treatment consisting of pyridostigmine, an acetylcholinesterase inhibitor used to treat muscle weakness in myasthenia gravis and CMS (Lee et al., 2018). Although the severity could potentially be related to how well a patient responds to the standard treatment with the AchE inhibitors, we could not find a clear correlation between severity and pyridostigmine treatment (two-tailed Fisher's exact test p-value 0.356; **Suppl. Figure 1**).

Our results indicate that severity is related to AChR clustering at the Agrin-Plectin-LRP4-Laminins axis level, suggesting the potential benefit of pharmaceutical intervention enhancing the downstream process of AChR clustering. Strikingly, beta-

2 adrenergic receptor agonists like ephedrine and salbutamol have been documented as capable of enhancing AChR clustering (Clausen et al., 2018) and proved to be successful in the treatment for severe AChR deficiency syndromes (Rodríguez Cruz et al., 2015; Garg and Goyal, 2022; Sadeh et al., 2011; Vanhaesebrouck et al., 2019) , but a strong molecular explanation for the observed favorable effects was still missing. This study reinforces explainability for the described successful usage of such treatments by relating CMS phenotypic severity with the normal development of AChR clusters at the motor neuron membrane. Several of the genes identified in this analysis do not have previous associations with the NMJ, such as the Usher syndrome and Retinitis pigmentosa associated gene; *USH2A*, identified as a copy number loss in patient 2. To provide proof of principal for this gene acting as a potential modifier of CMS severity, we investigated whether knockdown of *ush2a*, the zebrafish orthologue, could result in NMJ defects. Both CRISPR and TALEN-mediated knockout of *ush2a* in zebrafish have previously revealed phenotypes consistent with Usher syndrome and Retinitis pigmentosa such as hearing loss and progressive visual impairments (Han et al. 2018).

However, neither study assessed impacts on muscle structure or movement of the fish. Zebrafish perform well-characterized movements throughout development, starting with spontaneous chorion rotations from approximately 17 hours post fertilization (hpf, the time at which primary motor axons start extending into the muscle) to 30 hp (Saint-Amant & Drapeau, 1998). We treated 1-cell-stage embryos with a high dose of MO to reduce expression of *ush2a* (or equivalent dose of a control MO) and found a decrease in the number of chorion rotations performed at 24 hpf. These movements are mediated at the level of the spinal cord and are independent of supraspinal inputs (Downes & Granato. 2006), thus implying an early defect in NMJ or muscle development, or in signal transduction in the spinal cord/peripheral nervous system. By 2 dpf zebrafish can respond to touch and do so by rapidly swimming at least 1 body-length away from the stimulus (Saint-Amant & Drapeau. 1998). In *ush2a*-MO fish the average swimming velocity was significantly slower than in

controls, whereas the initial acceleration (proportional to the force of muscle contraction) was unaffected (Sztal et al., 2016). This implies that the initial fast muscle response is not significantly affected at this time-point, but that loss of *ush2a* at the NMJs of slow muscle may be impacting swimming. Defects in movement are reported in many other zebrafish models of CMS, such as those lacking *dok7* (Müller et al. 2010), *gfpt1* (Senderek et al., 2011) and *syt2* (Wen et al., 2010). Our motility findings are supported by the identification of a reduction in colocalization of AChRs with SV2-positive clusters on slow muscle fibers in 2 dpf fish, thus showing an increase in the number of AChRs that have not been contacted by a motor axon. We also identified an overall reduction in the number of SV2-positive clusters, which may be indicative of a defect or delay in development of the motor nervous system.

Previous studies have commented on *USH2A* presence on the basement membranes of perineurium nerve fibers (Pearsall et al., 2002) (Schwaller et al., 2021), however, further functional studies will be required to determine the precise localization of the defect and whether loss of USH2A alone can impact NMJ signaling or whether co-occurrence with *CHRNE* CMS is required. Additional functional work is also required to ascertain the importance of other potential modifiers identified in this study. Particularly, a prospective analysis on the potential NMJ involvement of the unique variants detected for the non-severe group could be of special interest for the study of CMS, potentially discerning their functional relationship to causal CMS genes.

Our work represents a thorough study of a narrow population showing a differential accumulation of damaging mutations in patients with CMS who have varying phenotypic severities, building on the initial impact of *CHRNE* mutations on the NMJ. It is important to remark that CMS is of particular interest among rare diseases, since drugs that influence neuromuscular transmission can produce clear improvements in the affected patients (Engel 2007). In this sense, identifying meaningful molecular relationships between gene variants allow us to gain insight into the disease mechanisms through a multiplex biomedical framework, paving the way for a whole new set of computational approximations for rare disease research.

Acknowledgements

The authors acknowledge the donors and families, Daniel Rico (Newcastle University) for his contribution in early stages of the project, Anaïs Baudot (Aix Marseille Université and Barcelona Supercomputing Center) for her careful revision of the manuscript, Miguel Vázquez (Barcelona Supercomputing Center) for advising about Rbbt analysis, Jon Sánchez Valle (Barcelona Supercomputing Center) and Núria Olvera (Barcelona Supercomputing Center and IDIBAPS) for the insightful discussions.

Funding

The NeurOmics and RD-Connect projects have been funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreements no 2012-305121 and 2012-305444.

I.N.C. was supported by a grant for pre-doctoral contracts for the training of doctors (Project ID: SEV-2015-0493-18-2) (Grant ID: PRE2018-083662) from the Spanish Ministry for Science, Innovation and Universities.

E.O. was supported by an AFM-Téléthon postdoctoral fellowship for the duration of this work.

H.L. receives support from the Canadian Institutes of Health Research (Foundation Grant FDN-167281), the Canadian Institutes of Health Research and Muscular Dystrophy Canada (Network Catalyst Grant for NMD4C), the Canada Foundation for Innovation (CFI-JELF 38412), and the Canada Research Chairs program (Canada Research Chair in Neuromuscular Genomics and Health, 950-232279).

V.G. was a research fellow of the Alexander von Humboldt Foundation.

D.C. was supported by the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018- 1, "iPC - individualizedPaediatricCure" (ref. 826121).

Author contributions

T.C., I.T. and V.G. collected and processed the biopsies; H.L. and R.T. coordinated data sharing; A.T., P.A.C.T., S.B. and S.C. coordinated and performed the omics data analysis with Y.A., S.L., M.R. and M.B.; E.O. performed the experiments in zebrafish; D.C. and A.V. coordinated the multilayer network analysis performed by I.N.C. All authors contributed to the writing and revising of the manuscript.

Ethics approval

This study was approved by the Ethics committee of Sofia Medical University (protocol 4/15-April-2013).

Conflict of interest

None declared.

Methods

WGS and RNA-seq

Whole genome sequencing (WGS) data have been obtained from blood using the Illumina TruSeq PCR-free library preparation kit. Sample sequencing was performed with the HiSeqX sequencing platform (HiseqX v1 or v2 SBS kit, 2x150 cycles), with an average mean depth coverage ≥ 30X. Samples have been analyzed using the RD-Connect specific pipeline: BWA-mem for alignment; Picard for duplicate marking and GATK 3.6.0 for variant calling. RNA sequencing (RNA-seq) data have been obtained from fibroblasts, using Illumina TruSeq RNA Library Preparation Kit v2, sequencing with an average of 60M reads per sample (paired-end 2X125 cycles). Data has been processed with the following pipeline (Laurie et al. 2016): STAR 2.35a for alignment, RSEM 1.3.0 for quantification, and GATK 3.6.0 for variant calling. All analyses have been performed using the human genome GRCh37d5 as reference.

Copy number variants

Copy Number Variants (CNVs) have been extracted using ClinCNV (https://github.com/imgag/ClinCNV) by employing a set of Eastern European samples as a background control group. Out of the 569 autosomal CNVs we selected as potential candidates the CNVs of the following types that overlapped with protein-coding genes: 1) whole gene gains or losses, and 2) partial losses (deletions overlapping with exons but not with the whole gene). The list of potential candidates included 55 CNVs that created a total of 82 whole gene gains or losses and 28 partial losses.

Compound heterozygous variants

Compound heterozygous variants have been obtained by phasing the WGS variant calls with the RNA-seq aligned BAM files using phASER (Castel et al. 2016). At first, variants are imputed using Sanger Imputation Service with EAGLE2 pre-phasing

step (Durbin 2014). PhASER is then applied to extend phased regions to gene-wide haplotypes. By accurately reflecting the muscle transcriptome, fibroblasts have been previously proved to be excellent and minimally invasive diagnostic tools for rare neuromuscular diseases (Gonorazky et al. 2019). We then annotated variants with eDiVA tool (www.ediva.crg.es) (Bosio et al. 2019), and removed all mutations with Genome Aggregation Database (gnomAD) (Lek et al. 2016) that show allele frequency > 3% globally, all variants outside exonic and splicing regions using Ensembl annotation, all synonymous mutations, and all variants with read depth (coverage) smaller than 8. Afterwards we selected all genes with at least two hits on different alleles as genes affected by damaging compound heterozygous variants. Each sample has been processed individually throughout the whole process.

Monolayer community detection

We performed a network community detection analysis using the Louvain clustering algorithm (Blondel et al. 2008) implemented in R package igraph (https://igraph.org/) with default parameters. We carried out the analysis using three (monolayer) networks, obtained from Reactome database (Fabregat et al. 2018), from the Recon3D Virtual Metabolic Human database (Brunk et al. 2018) (both downloaded in May 2018), and from the Integrated Interaction Database (IID) (Kotlyar et al. 2016) (downloaded in October 2018).

Additional information on network connectivity metrics (e.g., node centrality distributions and specific centrality information for severe-specific module genes) is conveniently provided as a jupyter notebook script, accessible from the following link: https://github.com/ikernunezca/CMS/blob/master/Scripts/Multilayer_Network_Information_and_Connectivity_Patterns.ipynb. All gene identifiers of each network were converted to NCBI Entrez gene identifiers using R packages AnnotationDbi v1.44.0 and org.Hs.eg.db v3.7.0 (http://bioconductor.org/). After detecting the community structure from each layer independently, we retrieved the community membership of the genes of interest, henceforth called “CMS linked genes”, i.e., known CMS causal

genes, and severe and not-severe compound heterozygous variants and CNVs. We then defined a community similarity measure as Jaccard Index, i.e., the number of shared genes of interest between the communities divided by the sum of the total number of genes of each community.

Multilayer community detection

We constructed a multilayer gene network composed of the three monolayer networks described in the previous section (Reactome, Virtual Metabolic Human and Integrated Interaction Database). Each of these three networks represents one layer of the multilayer network and, in general, three facets of fundamental molecular processes in the cell (**Suppl. Figure 11**). The multilayer community detection analysis was performed by using MolTi software (Didier, Brun, and Baudot 2015), which adapts the Louvain clustering algorithm with modularity maximization to multilayer networks. The algorithm is parametrized by the resolution (γ): the higher the value of γ, the smaller the size of the detected multilayer communities.

By varying the resolution parameter γ it is possible to uncover the modular structure of network communities (Fortunato and Barthelemy 2007). By exploring a wide range of resolution parameter values, we identified γ=4 (727 communities, each one composed of 26.46 genes on average) as an extreme value before both size and number of the detected multilayer communities stabilize (**Suppl. Figure 12**). The most dramatic changes in number and composition of detected communities are observed in the resolution parameter interval $\gamma \in (0,4]$.

We, therefore, used this parameter interval to test the hypothesis that disease-related genes consistently appear in the same multilayer communities, as well as to identify modules containing CMS linked genes within them. In this analysis, we define a module as a group of CMS linked genes that are systematically found to be part of the same multilayer community while increasing the resolution parameter (see Supplementary Information "Multilayer community detection analysis").

Additional analyses and data availability

We retrieved known CMS causal genes from the GeneTable of Neuromuscular Disorders (http://www.musclegenetable.fr, version November 2018) (Bonne, Rivier, and Hamroun 2017). Segregation analysis of WGS data has been performed using Rbbt (Vázquez et al. 2010). DisGeNET database (Piñero et al. 2017) was downloaded in November 2018. The association between CMS severity, demographic factors and clinical tests was assessed with a two-tailed Fisher's test using R statistical environment (www.R-project.org). Networks were rendered with Cytoscape (Saito et al. 2012). We used VCFtools (Danecek et al. 2011) to compute familial relatedness Ω among patients, scaled to $-\log_2(2\Omega)$. We used Enrichr (E. Y. Chen et al. 2013) for the functional enrichment analysis of the gene lists under study. We used Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016) to assess the impact of the compound heterozygous variants in the genes of the severe-specific largest module. Expression levels in tissues of interest (GTEx and Illumina Body Map) were retrieved from EBI Expression Atlas (www.ebi.ac.uk/) by filtering with the following keywords: 'nerve', 'muscle cell', 'fibroblast' and 'nervous system' (0.5 TPM default cutoff). We used Expression Atlas expression level categories: low (0.5 to 10 TPM), medium (11 to 1000 TPM), and high (more than 1000 TPM). Synaptic localization was retrieved from the UniProt database (https://www.uniprot.org/).

Zebrafish morpholino injections

Zebrafish have one orthologue of human *USH2a: ush2a*, as identified using the UCSC database (http://genome.ucsc.edu/, GRCz11/danRer11 assembly). We confirmed that *ush2a* is expressed throughout the first 5 days post fertilization (dpf). Gene Tools LLC (USA) then designed and synthesized an antisense morpholino oligonucleotide (MO) targeting the splice donor site of exon 3/intron 3 of *ush2a* (5'-3' GAGAAATGCTGCTCACCTGTAGAGC, ENSDART00000086201.5). We also obtained a control MO that targets a human beta-globin mutation (5'-3' CCTCTTACCTCAGTTACAATTTATA). MOs were diluted to 2 ng/nl in Danieau buffer

(58 mM NaCl, 5 mM HEPES, 0.7 mM KCl, 0.6 mM $Ca(NO_3)_2$, 0.4 mM $MgSO_4$; pH 7.6) and supplemented with 1% phenol red, before being injected into the yolk-sac of 1-cell stage embryos. A range of doses between 6 and 18 ng per 1-cell stage embryo were trialed for success in reducing *ush2a* expression and producing a measurable phenotypic change. A dose of 18 ng per 1-cell stage embryo was selected for behavioral and morphological analysis, as survival was not significantly affected for any dose tested. Embryos were maintained at 28.5°C in blue water (system water with 0.1 µg/ml Methylene Blue) for up to 5 dpf and survival recorded daily. At 2 dpf zebrafish were imaged using a Leica EZ4 W stereomicroscope and eye size and length measured using Fiji (ImageJ).

Chorion movement analysis in zebrafish

At 1 dpf (24 hours post fertilization), zebrafish were recorded in their chorions for 1 minute at 30 frames per second using a Leica EZ4 W stereomicroscope. Videos were analyzed using DanioScope software (Noldus Information Technology Inc., Leesburg, VA) to automatically assess duration of bursts and burst count/minute (bursts are full rotations performed by the zebrafish within the chorion).

Touch response analysis

At 2 dpf, a touch response assay was performed as previously described (O'Connor et al. 2018). Only fish with a normal phenotype were used for movement analysis. Briefly, fish that had not hatched from the chorion were enzymatically dechorionated with pronase (1 mg/ml, Sigma) for 10 min in blue water, followed by 3x washes in blue water. An individual fish was placed in a petri dish containing blue water and a Sony RX0 II (DSC-RX0M2) camera was placed 20 cm above the petri dish. A ruler with 1 mm markings was used as a scale for recordings. A gel loading pipette tip was used to touch the zebrafish on the back of the head and the response recorded. Videos were imported into Fiji ImageJ (Schindelin et al. 2012) as FFmpeg movies and movements analyzed using the Trackmate plugin (Tinevez et al. 2017). Values for average speed were exported and used to derive initial acceleration.

RNA isolation, cDNA synthesis and RT-PCR in zebrafish

RNA was isolated from pools of around 20 2 dpf zebrafish (control MO and *ush2a* MO-injected) following removal of chorions with pronase (Streptomyces griseus, Roche,1 mg/ml in blue water). Zebrafish were washed 3 times with blue water, euthanized with a 1:1 ratio of fresh system water:4 mg/ml tricaine methanesulfonate (Sigma). Fish were homogenized in RLT buffer (RNeasy mini kit, Qiagen) using 5 mm stainless steel beads with a TissueLyser II (Qiagen) at 25 Hz for 2 mins. RNA was then isolated following the RNeasy kit manufacturer's instructions, including on-column DNase digestion. RNA was measured using a Nanodrop ND-1000 and 1 µg used for cDNA synthesis according to manufacturer's instructions (5X All-In-One RT MasterMix, abm). Reverse-transcriptase PCR (RT-PCR) was performed to check for *ush2a* gene expression and knockdown success in MO-treated embryos, using MyTaq™ DNA Polymerase (Meridian Bioscience) and primers as follows: *eef1a1l1* forward 5'-CTGGAGGCCAGCTCAAACATGG-3', reverse 5'-CTTGCTGTCTCCAGCCACATTAC-3' and *ush2a* forward 5'-CTGGGCACACTTGGCTCTAC -3', reverse 5'-TTCTTCAATCTCCCTGTTGGTT-3'.

Immunofluorescent staining, imaging and analysis of zebrafish neuromuscular junctions and muscle fibers

Whole mount staining of 2 dpf zebrafish NMJs was performed as previously described (O'Connor et al. 2019). Briefly, a mouse anti-synaptic vesicle protein 2 (SV2) antibody was used to visualize motor neurons (1:200, AB2315387, Developmental Studies Hybridoma Bank) and Alexa Fluor 488-α-bungarotoxin conjugate (1:1000, B13422, Invitrogen) was used for visualizing acetylcholine receptors (AChRs). Phalloidin-iFluor 594 was used to visualize filamentous actin within muscle fibers (1:1000, ab176757). Z-stack images encompassing the depth of the midsection of the zebrafish tail were obtained using a 20× air objective on an LSM800 confocal microscope. Analysis of NMJ structure was performed as previously described (O'Connor et al. 2019), using Fiji (ImageJ, Madison, WI, USA). The number of SV2-

positive and α-bungarotoxin-positive clusters per 100 μm^2 were measured. Co-localization analysis between SV2 and α-bungarotoxin was performed on maximum intensity projections using the 'JACoP' Fiji plugin (Bolte & Cordelières, 2006). Briefly, each fluorophore was subject to manual thresholding to remove background, and the Mander's correlation coefficient calculated to give a value between 0 and 1, reflecting the degree of co-occurrence of signals between both SV2 and α-bungarotoxin, and α-bungarotoxin with SV2. For phalloidin-stained fish, average myotome size was measured, and degree of fiber dispersion quantified using the directionality plugin. Data was collected from at least 4 myotomes per fish.

Statistics for zebrafish experiments

Statistical analysis was performed using GraphPad Prism software (v9.3.0). Outliers were removed from data using the ROUT method (Q = 1 %). Cleaned data was tested for normal distribution then depending on outcome either a nonparametric Mann-Whitney test or parametric unpaired t-test were applied for behavioral studies and degree of dispersion. For NMJ morphology experiments in which 4+ myotomes (technical replicates) per fish (biological replicates) were analyzed, data was assessed for significance using a nested t-test to avoid pseudo-replication. Statistical significance was taken as $p < 0.05$, degrees of freedom (df) and t-value are given for all parametric tests, and n numbers listed in the results section. Survival analysis was performed using the log-rank test comparing WT to each other condition, and threshold for significance was corrected for multiple comparisons using the Bonferroni method ($p < 0.006$). Zebrafish studies were blinded before image/video acquisition and unblinded following analysis.

Data availability

The datasets generated and analyzed in this study are not publicly available due to sensible content (genomics information in a rare disease). Reasonable requests for further information will be carefully evaluated by the corresponding author and co-authors.

Code availability

All code and the Cytoscape session rendering Figures 3 and 4, as well as Supplementary Figures 3, 6 and 9 are available for reproducibility purposes at: https://github.com/ikernunezca/CMS. The analysis of multilayer community communities can also be performed using CmmD (Núñez-Carpintero et al., 2021) (https://github.com/ikernunezca/CmmD) with parameters: resolution_start: 0, resolution_end: 4, interval: 0.5 and the CMS linked genes as nodelist.

Chapter results summary

The main concepts introduced in the research article presented in this chapter are the following:

1. The study of phenotypic severity in rare diseases, despite being a huge clinically relevant problem, is still a typically neglected scenario provided the challenging setting that minor conditions represent.

2. Multilayer networks provide an integrative framework for the exploration of relevant biomedical data resources, that is independent of cohort size limitations.

3. Detection of multilayer communities at multiple levels of modularity resolution allows for an evaluation of the robustness of the functional relationships of genes affected by patient-specific damaging mutations.

4. The detected gene modules, provide a thorough understanding of the specific damaged NMJ processes in the patients presenting severe phenotypic affectations, as well as their functional connection to already known causative processes of Congenital Myasthenic Syndromes.

5. Identified severe-specific compound heterozygous variants affect key mediators for the presentation of AChR at the post-synaptic level, a crucial process for normal muscle contraction, in a patient-wise manner. This article additionally provides explainability on the potential of AChR clustering as a therapeutic target.

6. One of the studied patients additionally presents a partial heterozygous copy number loss affecting the gene *USH2A*, previously unknown to play functional roles at the NMJ level. We provide extensive experimental demonstration of its importance by studying a zebrafish morpholino model affecting the orthologous gene: *ush2a*.

7. Overall, this chapter presents an important example of the potential of multilayer network analysis for the detection of severity related genes, efficiently overcoming inherent limitations of rare clinical settings.

References

Abicht, Angela, Juliane Müller, and Hanns Lochmüller. 1993. “Congenital Myasthenic Syndromes.” In *GeneReviews®*, edited by Margaret P. Adam, Holly H. Ardinger, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen Stephens, and Anne Amemiya. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1168/.

Abicht, A., R. Stucka, V. Karcagi, A. Herczegfalvi, R. Horváth, W. Mortier, U. Schara, et al. 1999. “A Common Mutation (Epsilon1267delG) in Congenital Myasthenic Patients of Gypsy Ethnic Origin.” *Neurology* 53 (7): 1564–69. https://doi.org/10.1212/wnl.53.7.1564.

Amin, Mutaz, Yousuf Bakhit, Mahmoud Koko, Mohamed Osama Mirgahni Ibrahim, M. A. Salih Muntaser Ibrahim, and Osheik A. Seidi. 2019. “Rare Variant in LAMA2 Gene Causing Congenital Muscular Dystrophy in a Sudanese Family. A Case Report.” *Acta Myologica: Myopathies and Cardiomyopathies: Official Journal of the Mediterranean Society of Myology* 38 (1): 21–24.

Andreose, J. S., C. Sala, and G. Fumagalli. 1994. “Immunolocalization of Chromogranin B, Secretogranin II, Calcitonin Gene-Related Peptide and Substance P at Developing and Adult Neuromuscular Synapses.” *Neuroscience Letters* 174 (2): 177–80.

Arikawa-Hirasawa, Eri, William R. Wilcox, Alexander H. Le, Neil Silverman, Prasanthi Govindraj, John R. Hassell, and Yoshihiko Yamada. “Dyssegmental Dysplasia, Silverman-Handmaker Type, Is Caused by Functional Null Mutations of the Perlecan Gene.” *Nature Genetics* 27, no. 4 (April 2001): 431–34. https://doi.org/10.1038/86941.

Astigarraga, Sergio, Kerstin Hofmeyer, Reza Farajian, and Jessica E. Treisman. 2010. “Three Drosophila Liprins Interact to Control Synapse Formation.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (46): 15358–68. https://doi.org/10.1523/JNEUROSCI.1862-10.2010.

Austin-Tse, Christina A., Diana L. Mandelker, Andrea M. Oza, Heather Mason-Suares, Heidi L. Rehm, and Sami S. Amr. 2018. “Analysis of Intragenic USH2A Copy Number Variation Unveils Broad Spectrum of Unique and Recurrent Variants.” *European Journal of Medical Genetics* 61 (10): 621–26. https://doi.org/10.1016/j.ejmg.2018.04.006.

Barik, Arnab, Yisheng Lu, Anupama Sathyamurthy, Andrew Bowman, Chengyong Shen, Lei Li, Wen-cheng Xiong, and Lin Mei. 2014. “LRP4 Is Critical for Neuromuscular Junction Maintenance.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (42): 13892–905. https://doi.org/10.1523/JNEUROSCI.1733-14.2014.

Beeson, David. 2016. “Congenital Myasthenic Syndromes: Recent Advances.” *Current Opinion in Neurology* 29 (5): 565–71. https://doi.org/10.1097/WCO.0000000000000370.

Bernadzki, Krzysztof M., Marta Gawor, Marcin Pęziński, Paula Mazurek, Paweł Niewiadomski, Maria J. Rędowicz, and Tomasz J. Prószyński. 2017. “Liprin-α-1 Is a Novel Component of the Murine Neuromuscular Junction and Is Involved in the Organization of the Postsynaptic Machinery.” *Scientific Reports* 7 (1): 9116. https://doi.org/10.1038/s41598-017-09590-7.

Bertini, Enrico, Adele D’Amico, Francesca Gualandi, and Stefania Petrini. 2011. “Congenital Muscular Dystrophies: A Brief Review.” *Seminars in Pediatric Neurology* 18 (4): 277–88. https://doi.org/10.1016/j.spen.2011.10.010.

Bevilacqua, Jorge A., Marian Lara, Jorge Díaz, Mario Campero, Jessica Vázquez, and Ricardo A. Maselli. 2017. “Congenital Myasthenic Syndrome Due to DOK7 Mutations in a Family from Chile.” *European Journal of Translational Myology* 27 (3):6832. https://doi.org/10.4081/ejtm.2017.6832.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/17425468/2008/10/P10008.

Bolte, Sussane, and Fabrice P. Cordelières. “A Guided Tour into Subcellular Colocalization Analysis in Light Microscopy.” *Journal of Microscopy* 224, no. 3 (2006): 213–32. https://doi.org/10.1111/j.1365-2818.2006.01706.x.

Bonne, Gisèle, François Rivier, and Dalil Hamroun. 2017. “The 2018 Version of the Gene Table of Monogenic Neuromuscular Disorders (Nuclear Genome).” *Neuromuscular Disorders: NMD* 27 (12): 1152–83. https://doi.org/10.1016/j.nmd.2017.10.005.

Bönnemann, Carsten G., Ching H. Wang, Susana Quijano-Roy, Nicolas Deconinck, Enrico Bertini, Ana Ferreiro, Francesco Muntoni, et al. 2014. “Diagnostic Approach to the Congenital Muscular Dystrophies.” *Neuromuscular Disorders: NMD* 24 (4): 289–311. https://doi.org/10.1016/j.nmd.2013.12.011.

Bork, P., and R. F. Doolittle. 1992. “Proposed Acquisition of an Animal Protein Domain by Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (19): 8990–94.

Bosio, Mattia, Oliver Drechsel, Rubayte Rahman, Francesc Muyas, Raquel Rabionet, Daniela Bezdan, Laura Domenech Salgado, et al. 2019. “EDiVA-Classification and Prioritization of Pathogenic Variants for Clinical Diagnostics.” *Human Mutation*, April. https://doi.org/10.1002/humu.23772.

Boycott, Kym M., Megan R. Vanstone, Dennis E. Bulman, and Alex E. MacKenzie. 2013. “Rare-Disease Genetics in the Era of next-Generation Sequencing: Discovery to Translation.” Nature Reviews Genetics 14 (10): 681–91. https://doi.org/10.1038/nrg3555.

Brunk, Elizabeth, Swagatika Sahoo, Daniel C. Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, et al. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3): 272–81. https://doi.org/10.1038/nbt.4072.

Buphamalai, Pisanu, Tomislav Kokotovic, Vanja Nagy, and Jörg Menche. 2021. ‘Network Analysis Reveals Rare Disease Signatures across Multiple Levels of Biological Organization’. *Nature Communications* 12 (1): 6306. https://doi.org/10.1038/s41467-021-26674-1.

Burden, Steven J., Norihiro Yumoto, and Wei Zhang. 2013. “The Role of MuSK in Synapse Formation and Neuromuscular Disease.” *Cold Spring Harbor Perspectives in Biology* 5 (5): a009167. https://doi.org/10.1101/cshperspect.a009167.

Cantini, Laura, Enzo Medico, Santo Fortunato, and Michele Caselle. 2015. “Detection of Gene Communities in Multi-Networks Reveals Cancer Drivers.” *Scientific Reports* 5 (December): 17386. https://doi.org/10.1038/srep17386.

Carmen, Laurino, Vadala' Maria, Julio Cesar Morales-Medina, Annamaria Vallelunga, Beniamino Palmieri, and Tommaso Iannitti. 2019. "Role of Proteoglycans and Glycosaminoglycans in Duchenne Muscular Dystrophy." *Glycobiology* 29 (2): 110–23. https://doi.org/10.1093/glycob/cwy058.

Castel, Stephane E., Pejman Mohammadi, Wendy K. Chung, Yufeng Shen, and Tuuli Lappalainen. 2016. "Rare Variant Phasing and Haplotypic Expression from RNA Sequencing with PhASER." *Nature Communications* 7: 12817. https://doi.org/10.1038/ncomms12817.

Castro-Giner, Francesc, Peter Ratcliffe, and Ian Tomlinson. 2015. "The Mini-Driver Model of Polygenic Cancer Evolution." *Nature Reviews. Cancer* 15 (11): 680–85. https://doi.org/10.1038/nrc3999.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. "Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool." *BMC Bioinformatics* 14 (April): 128. https://doi.org/10.1186/1471-2105-14-128.

Chen, Yuewen, Jinying Xu, Xiaopu Zhou, Saijuan Liu, Yulin Zhang, Shuangshuang Ma, Amy K. Y. Fu, Nancy Y. Ip, and Yu Chen. 2019. "Changes of Protein Phosphorylation Are Associated with Synaptic Functions during the Early Stage of Alzheimer's Disease." *ACS Chemical Neuroscience* 10 (9): 3986–96. https://doi.org/10.1021/acschemneuro.9b00190.

Clausen, Lisa, Judith Cossins, and David Beeson. 2018. "Beta-2 Adrenergic Receptor Agonists Enhance AChR Clustering in C2C12 Myotubes: Implications for Therapy of Myasthenic Disorders." *Journal of Neuromuscular Diseases* 5 (2): 231–40. https://doi.org/10.3233/JND-170293.

Dagur, Pradeep K., and J. Philip McCoy. 2015. "Endothelial-Binding, Proinflammatory T Cells Identified by MCAM (CD146) Expression: Characterization and Role in Human Autoimmune Diseases." *Autoimmunity Reviews* 14 (5): 415–22. https://doi.org/10.1016/j.autrev.2015.01.003.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Della Marina, Adela, Eva Wibbeler, Angela Abicht, Heike Kölbel, Hanns Lochmüller, Andreas Roos, and Ulrike Schara. 2020. “Long Term Follow-Up on Pediatric Cases With Congenital Myasthenic Syndromes—A Retrospective Single Centre Cohort Study.” *Frontiers in Human Neuroscience* 14. https://www.frontiersin.org/articles/10.3389/fnhum.2020.560860.

Di Martino, Alberto, Matilde Cescon, Claudio D’Agostino, Francesco Schilardi, Patrizia Sabatelli, Luciano Merlini, and Cesare Faldini. 2023. “Collagen VI in the Musculoskeletal System.” *International Journal of Molecular Sciences* 24 (6): 5095. https://doi.org/10.3390/ijms24065095.

Didier, Gilles, Christine Brun, and Anaïs Baudot. 2015. “Identifying Communities from Multiplex Biological Networks.” *PeerJ* 3: e1525. https://doi.org/10.7717/peerj.1525.

Dimova, Ivanka, and Ivo Kremensky. 2018. “LAMA2 Congenital Muscle Dystrophy: A Novel Pathogenic Mutation in Bulgarian Patient.” *Case Reports in Genetics* 2018: 3028145. https://doi.org/10.1155/2018/3028145.

Downes, Gerald B., and Michael Granato. 2006. “Supraspinal Input Is Dispensable to Generate Glycine-Mediated Locomotive Behaviors in the Zebrafish Embryo.” *Journal of Neurobiology* 66, no. 5: 437–51. https://doi.org/10.1002/neu.20226.

Durbin, Richard. 2014. “Efficient Haplotype Matching and Storage Using the Positional Burrows-Wheeler Transform (PBWT).” *Bioinformatics (Oxford, England)* 30 (9): 1266–72. https://doi.org/10.1093/bioinformatics/btu014.

Edler, Daniel, Ludvig Bohlin, and Martin Rosvall. 2017. “Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap.” *Algorithms* 10 (4): 112. https://doi.org/10.3390/a10040112.

Eklund, L., J. Piuhola, J. Komulainen, R. Sormunen, C. Ongvarrasopone, R. Fässler, A. Muona, et al. 2001. “Lack of Type XV Collagen Causes a Skeletal Myopathy and Cardiovascular Defects in Mice.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (3): 1194–99. https://doi.org/10.1073/pnas.031444798.

Engel, Andrew G. 2007. “The Therapy of Congenital Myasthenic Syndromes.” *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics* 4 (2): 252–57. https://doi.org/10.1016/j.nurt.2007.01.001.

Engel, Andrew G., Xin-Ming Shen, Duygu Selcen, and Steven M. Sine. 2015. “Congenital Myasthenic Syndromes: Pathogenesis, Diagnosis, and Treatment.” *The Lancet. Neurology* 14 (4): 420–34. https://doi.org/10.1016/S1474-4422(14)70201-7.

Fabregat, Antonio, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, et al. 2018. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 46 (D1): D649–55. https://doi.org/10.1093/nar/gkx1132.

Fareed, Mohd, and Mohammad Afzal. 2017. “Genetics of Consanguinity and Inbreeding in Health and Disease.” *Annals of Human Biology* 44 (2): 99–107. https://doi.org/10.1080/03014460.2016.1265148.

Finsterer, Josef. 2019. “Congenital Myasthenic Syndromes.” *Orphanet Journal of Rare Diseases* 14 (1): 57. https://doi.org/10.1186/s13023-019-1025-5.

Forrest, Katharine, Jemima E. Mellerio, Stephanie Robb, Patricia J. C. Dopping-Hepenstal, John A. McGrath, Lu Liu, Stefan J. A. Buk, Safa Al-Sarraj, Elizabeth Wraige, and Heinz Jungbluth. 2010. “Congenital Muscular Dystrophy, Myasthenic Symptoms and Epidermolysis Bullosa Simplex (EBS) Associated with Mutations in the PLEC1 Gene Encoding Plectin.” *Neuromuscular Disorders: NMD* 20 (11): 709–11. https://doi.org/10.1016/j.nmd.2010.06.003.

Fortunato, S., and M. Barthelemy. 2007. “Resolution Limit in Community Detection.” *Proceedings of the National Academy of Sciences* 104 (1): 36–41. https://doi.org/10.1073/pnas.0605965104.

Garg, Divyani, and Vinay Goyal. “Positive Response to Inhaled Salbutamol in Congenital Myasthenic Syndrome Due to CHRNE Mutation.” Muscle & Nerve n/a, no. n/a. Accessed June 2, 2022. https://doi.org/10.1002/mus.27563.

Goh, K.-I., M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. 2007. “The Human Disease Network.” *Proceedings of the National Academy of Sciences* 104 (21): 8685–90. https://doi.org/10.1073/pnas.0701361104.

Gonorazky, Hernan D., Sergey Naumenko, Arun K. Ramani, Viswateja Nelakuditi, Pouria Mashouri, Peiqui Wang, Dennis Kao, et al. 2019. “Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease.” *American Journal of Human Genetics* 104 (3): 466–83. https://doi.org/10.1016/j.ajhg.2019.01.012.

Gosak, Marko, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. 2018. “Network Science of Biological Systems at Different Scales: A Review.” *Physics of Life Reviews* 24: 118–35. https://doi.org/10.1016/j.plrev.2017.11.003.

Grevengoed, Trisha J., Eric L. Klett, and Rosalind A. Coleman. 2014. “Acyl-CoA Metabolism and Partitioning.” *Annual Review of Nutrition* 34: 1–30. https://doi.org/10.1146/annurev-nutr-071813-105541.

Gros-Louis, Francois, Peter M. Andersen, Nicolas Dupre, Makoto Urushitani, Patrick Dion, Frederique Souchon, Monique D’Amour, et al. 2009. “Chromogranin B P413L Variant as Risk Factor and Modifier of Disease Onset for Amyotrophic Lateral Sclerosis.” *Proceedings of the National Academy of Sciences* 106 (51): 21777–82. https://doi.org/10.1073/pnas.0902174106.

Guillon, Emilie, Sandrine Bretaud, and Florence Ruggiero. 2016. “Slow Muscle Precursors Lay Down a Collagen XV Matrix Fingerprint to Guide Motor Axon Navigation.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 36 (9): 2663–76. https://doi.org/10.1523/JNEUROSCI.2847-15.2016.

Halu, Arda, Manlio De Domenico, Alex Arenas, and Amitabh Sharma. 2017. “The Multiplex Network of Human Diseases.” *BioRxiv*. https://doi.org/10.1101/100370.

Han, Shanshan, Xiliang Liu, Shanglun Xie, Meng Gao, Fei Liu, Shanshan Yu, Peng Sun, et al. 2018. “Knockout of Ush2a Gene in Zebrafish Causes Hearing Impairment and Late Onset Rod-Cone Dystrophy.” *Human Genetics* 137, no. 10: 779–94. https://doi.org/10.1007/s00439-018-1936-6.

Hantaï, Daniel, Sophie Nicole, and Bruno Eymard. 2013. “Congenital Myasthenic Syndromes: An Update.” *Current Opinion in Neurology* 26 (5): 561. https://doi.org/10.1097/WCO.0b013e328364dc0f.

Hull, Sarah, Gavin Arno, Anthony G. Robson, Suzanne Broadgate, Vincent Plagnol, Martin McKibbin, Stephanie Halford, et al. 2016. “Characterization of CDH3-Related Congenital Hypotrichosis With Juvenile Macular Dystrophy.” *JAMA Ophthalmology* 134 (9): 992–1000. https://doi.org/10.1001/jamaophthalmol.2016.2089.

Huzé, Caroline, Stéphanie Bauché, Pascale Richard, Frédéric Chevessier, Evelyne Goillot, Karen Gaudon, Asma Ben Ammar, et al. 2009. “Identification of an Agrin Mutation That Causes Congenital Myasthenia and Affects Synapse Function.” *American Journal of Human Genetics* 85 (2): 155–67. https://doi.org/10.1016/j.ajhg.2009.06.015.

Iozzo, Renato V., and Liliana Schaefer. 2015. “Proteoglycan Form and Function: A Comprehensive Nomenclature of Proteoglycans.” *Matrix Biology: Journal of the International Society for Matrix Biology* 42 (March): 11–55. https://doi.org/10.1016/j.matbio.2015.02.003.

Ito, Mikako, and Kinji Ohno. 2018. “Protein-Anchoring Therapy to Target Extracellular Matrix Proteins to Their Physiological Destinations.” *Matrix Biology: Journal of the International Society for Matrix Biology* 68–69: 628–36. https://doi.org/10.1016/j.matbio.2018.02.014.

Jacquier, Arnaud, Valérie Risson, Thomas Simonet, Florine Roussange, Nicolas Lacoste, Shams Ribault, Julien Carras, et al. 2022. “Severe Congenital Myasthenic Syndromes Caused by Agrin Mutations Affecting Secretion by Motoneurons.” *Acta Neuropathologica* 144 (4): 707–31. https://doi.org/10.1007/s00401-022-02475-8.

Johansson, Malin E. V., Kristina A. Thomsson, and Gunnar C. Hansson. 2009. “Proteomic Analyses of the Two Mucus Layers of the Colon Barrier Reveal That Their Main Component, the Muc2 Mucin, Is Strongly Bound to the Fcgbp Protein.” *Journal of Proteome Research* 8 (7): 3549–57. https://doi.org/10.1021/pr9002504.

Kaneko-Goto, Tomomi, Yuki Sato, Sayako Katada, Emi Kinameri, Sei-ichi Yoshihara, Atsushi Nishiyori, Mitsuhiro Kimura, et al. 2013. “Goofy Coordinates the Acuity of Olfactory Signaling.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33 (32): 12987–96, 12996a. https://doi.org/10.1523/JNEUROSCI.4948-12.2013.

Karczewski, Konrad J., and Michael P. Snyder. 2018. “Integrative Omics for Health and Disease.” Nature Reviews Genetics 19 (5): 299–310. https://doi.org/10.1038/nrg.2018.4.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581, no. 7809 (May 2020): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Keller, Kate E., Ying Ying Sun, Janice A. Vranka, Lauren Hayashi, and Ted S. Acott. 2012. “Inhibition of Hyaluronan Synthesis Reduces Versican and Fibronectin Levels in Trabecular Meshwork Cells.” *PloS One* 7 (11): e48523. https://doi.org/10.1371/journal.pone.0048523.

Kirschner, Janbernd, Ingrid Hausser, Yaqun Zou, Gudrun Schreiber, Hans-Jürgen Christen, Susan C. Brown, Ingrun Anton-Lamprecht, Francesco Muntoni, Folker Hanefeld, and Carsten G. Bönnemann. 2005. “Ullrich Congenital Muscular Dystrophy: Connective Tissue Abnormalities in the Skin Support Overlap with Ehlers-Danlos Syndromes.” *American Journal of Medical Genetics. Part A* 132A (3): 296–301. https://doi.org/10.1002/ajmg.a.30443.

Kivelä, Mikko, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. 2014. “Multilayer Networks.” *Journal of Complex Networks* 2 (3): 203–71. https://doi.org/10.1093/comnet/cnu016.

Kotlyar, Max, Chiara Pastrello, Nicholas Sheahan, and Igor Jurisica. 2016. “Integrated Interactions Database: Tissue-Specific View of the Human and Model Organism Interactomes.” *Nucleic Acids Research* 44 (D1): D536-541. https://doi.org/10.1093/nar/gkv1115.

Kousi, M., and N. Katsanis. 2015. “Genetic Modifiers and Oligogenic Inheritance.” *Cold Spring Harbor Perspectives in Medicine* 5 (6): a017145–a017145. https://doi.org/10.1101/cshperspect.a017145.

Kraft-Sheleg, Ortal, Shelly Zaffryar-Eilot, Olga Genin, Wesal Yaseen, Sharon Soueid-Baumgarten, Ofra Kessler, Tatyana Smolkin, et al. 2016. “Localized LoxL3-Dependent Fibronectin Oxidation Regulates Myofiber Stretch and Integrin-Mediated Adhesion.” *Developmental Cell* 36 (5): 550–61. https://doi.org/10.1016/j.devcel.2016.02.009.

Larraín, Juan, Jaime Alvarez, John R. Hassell, and Enrique Brandan. 1997. “Expression of Perlecan, a Proteoglycan That Binds Myogenic Inhibitory Basic Fibroblast Growth Factor, Is Down Regulated during Skeletal Muscle Differentiation.” *Experimental Cell Research* 234 (2): 405–12. https://doi.org/10.1006/excr.1997.3648.

Laskowski, M., and I. Kato. 1980. “Protein Inhibitors of Proteinases.” *Annual Review of Biochemistry* 49: 593–626. https://doi.org/10.1146/annurev.bi.49.070180.003113.

Laurie, Steve, Marcos Fernandez-Callejo, Santiago Marco-Sola, Jean-Remi Trotta, Jordi Camps, Alejandro Chacón, Antonio Espinosa, et al. 2016. “From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing.” *Human Mutation* 37, (12): 1263–71. https://doi.org/10.1002/humu.23114.

Lee, Manon, David Beeson, and Jacqueline Palace. 2018. “Therapeutic Strategies for Congenital Myasthenic Syndromes.” *Annals of the New York Academy of Sciences* 1412 (1): 129–36. https://doi.org/10.1111/nyas.13538.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91. https://doi.org/10.1038/nature19057.

Li, Lei O., Trisha J. Grevengoed, David S. Paul, Olga Ilkayeva, Timothy R. Koves, Florencia Pascual, Christopher B. Newgard, Deborah M. Muoio, and Rosalind A. Coleman. 2015. “Compartmentalized Acyl-CoA Metabolism in Skeletal Muscle Regulates Systemic Glucose Homeostasis.” *Diabetes* 64 (1): 23–35. https://doi.org/10.2337/db13-1070.

Li, Lei, Wen-Cheng Xiong, and Lin Mei. 2018. “Neuromuscular Junction Formation, Aging, and Disorders.” *Annual Review of Physiology* 80 (1): 159–88. https://doi.org/10.1146/annurev-physiol-022516-034255.

Lochmüller, Hanns, Dorota M. Badowska, Rachel Thompson, Nine V. Knoers, Annemieke Aartsma-Rus, Ivo Gut, Libby Wood, et al. 2018. “RD-Connect, NeurOmics and EURenOmics: Collaborative European Initiative for Rare Diseases.” *European Journal of Human Genetics: EJHG* 26 (6): 778–85. https://doi.org/10.1038/s41431018-0115-5.

Løkken, Nicoline, Alfred Peter Born, Morten Duno, and John Vissing. 2015. “LAMA2-Related Myopathy: Frequency among Congenital and Limb-Girdle Muscular Dystrophies.” *Muscle & Nerve* 52 (4): 547–53. https://doi.org/10.1002/mus.24588.

Lord, Megan S., Fengying Tang, Jelena Rnjak-Kovacina, James G. W. Smith, James Melrose, and John M. Whitelock. 2018. “The Multifaceted Roles of Perlecan in Fibrosis.” *Matrix Biology: Journal of the International Society for Matrix Biology* 68–69: 150–66. https://doi.org/10.1016/j.matbio.2018.02.013.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4.

McRae, Natasha, Leonard Forgan, Bryony McNeill, Alex Addinsall, Daniel McCulloch, Chris Van der Poel, and Nicole Stupka. 2017. “Glucocorticoids Improve Myogenic Differentiation In Vitro by Suppressing the Synthesis of Versican, a Transitional Matrix Protein Overexpressed in Dystrophic Skeletal Muscles.” *International Journal of Molecular Sciences* 18 (12). https://doi.org/10.3390/ijms18122629.

Menche, Jörg, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. 2015. “Disease Networks. Uncovering Disease-Disease Relationships through the Incomplete Interactome.” *Science (New York, N.Y.)* 347 (6224): 1257601. https://doi.org/10.1126/science.1257601.

Mitani, Aya A., and Sebastien Haneuse. 2020. “Small Data Challenges of Studying Rare Diseases.” JAMA Network Open 3 (3): e201965. https://doi.org/10.1001/jamanetworkopen.2020.1965.

Mo, Michelle, Ha Thi Hoang, Stefan Schmidt, Robert B. Clark, and Barbara E. Ehrlich. 2013. “The Role of Chromogranin B in an Animal Model of Multiple Sclerosis.” *Molecular and Cellular Neurosciences* 56 (September): 102–14. https://doi.org/10.1016/j.mcn.2013.04.003.

Müller, Juliane S., Angela Abicht, Hans-Jürgen Christen, Rolf Stucka, Ulrike Schara, Wilhelm Mortier, Angela Huebner, and Hanns Lochmüller. 2004. “A Newly Identified Chromosomal Microdeletion of the Rapsyn Gene Causes a Congenital Myasthenic Syndrome.” *Neuromuscular Disorders: NMD* 14 (11): 744–49. https://doi.org/10.1016/j.nmd.2004.06.010.

Müller, Juliane S., Catherine D. Jepson, Steven H. Laval, Kate Bushby, Volker Straub, and Hanns Lochmüller. 2010. “Dok-7 Promotes Slow Muscle Integrity as Well as Neuromuscular Junction Formation in a Zebrafish Model of Congenital Myasthenic Syndromes.” *Human Molecular Genetics* 19, no. 9: 1726–40. https://doi.org/10.1093/hmg/ddq049.

Muona, Anu, Lauri Eklund, Timo Väisänen, and Taina Pihlajaniemi. 2002. “Developmentally Regulated Expression of Type XV Collagen Correlates with Abnormalities in Col15a1(-/-) Mice.” *Matrix Biology: Journal of the International Society for Matrix Biology* 21 (1): 89–102.

Nicole, Sophie, Amina Chaouch, Torberg Torbergsen, Stéphanie Bauché, Elodie de Bruyckere, Marie-Joséphine Fontenille, Morten A. Horn, et al. 2014. “Agrin Mutations Lead to a Congenital Myasthenic Syndrome with Distal Muscle Weakness and Atrophy.” *Brain: A Journal of Neurology* 137 (Pt 9): 2429–43. https://doi.org/10.1093/brain/awu160.

Nilsson, Anna, Maria Fälth, Xiaoqun Zhang, Kim Kultima, Karl Sköld, Per Svenningsson, and Per E. Andrén. 2009. “Striatal Alterations of Secretogranin-1, Somatostatin, Prodynorphin, and Cholecystokinin Peptides in an Experimental Mouse Model of Parkinson Disease.” *Molecular & Cellular Proteomics: MCP* 8 (5): 1094–1104. https://doi.org/10.1074/mcp.M800454-MCP200.

Núñez-Carpintero, Iker, Marianyela Petrizzelli, Andrei Zinovyev, Davide Cirillo and Alfonso Valencia. 2021. “The multilayer community structure of medulloblastoma.” *iScience* 24. https://doi.org/10.1016/j.isci.2021.102365

O’Connor, Emily, Ana Töpf, René P Zahedi, Sally Spendiff, Daniel Cox, Andreas Roos and Hanns Lochmüller. 2018. “Clinical and research strategies for limb-girdle congenital myasthenic syndromes.” *Annals of the New York Academy of Sciences* 1412, 102–112. https://doi.org/10.1111/nyas.13520

O’Connor, Emily, George Cairns, Sally Spendiff, David Burns, Stefan Hettwer, Armin Mäder, Juliane Müller, Rita Horvath, Clarke Slater, Andreas Roos, Hanns Lochmüller. 2019. “Modulation of Agrin and RhoA Pathways Ameliorates Movement Defects and Synapse Morphology in MYO9A-Depleted Zebrafish.” *Cells* 8, no. 8, 848. https://doi.org/10.3390/cells8080848.

Ohkawara, Bisei, Macarena Cabrera-Serrano, Tomohiko Nakata, Margherita Milone, Nobuyuki Asai, Kenyu Ito, Mikako Ito, et al. 2014. “LRP4 Third β-Propeller Domain Mutations Cause Novel Congenital Myasthenia by Compromising Agrin-Mediated MuSK Signaling in a Position-Specific Manner.” *Human Molecular Genetics* 23 (7): 1856–68. https://doi.org/10.1093/hmg/ddt578.

Ohno, K. 2003. “E-Box Mutations in the RAPSN Promoter Region in Eight Cases with Congenital Myasthenic Syndrome.” *Human Molecular Genetics* 12 (7): 739–48.

Okuda-Ashitaka, Emiko, and Ken-Ichi Matsumoto. 2023. “Tenascin-X as a Causal Gene for Classical-like Ehlers-Danlos Syndrome.” *Frontiers in Genetics* 14: 1107787. https://doi.org/10.3389/fgene.2023.1107787.

Pampalakis, Georgios, Konstantinos Mitropoulos, Georgia Xeromerisiou, Efthymios Dardiotis, Georgia Deretzi, Maria Anagnostouli, Theodora Katsila, Michail Rentzos, and George P. Patrinos. 2019. “New Molecular Diagnostic Trends and Biomarkers for Amyotrophic Lateral Sclerosis.” *Human Mutation* 40 (4): 361–73. https://doi.org/10.1002/humu.23697.

Panchenko, M. V., W. G. Stetler-Stevenson, O. V. Trubetskoy, S. N. Gacheru, and H. M. Kagan. 1996. “Metalloproteinase Activity Secreted by Fibrogenic Cells in the Processing of Prolysyl Oxidase. Potential Role of Procollagen C-Proteinase.” *The Journal of Biological Chemistry* 271 (12): 7113–19.

Pearsall, Nicole, Gautam Bhattacharya, Jim Wisecarver, Joe Adams, Dominic Cosgrove, and William Kimberling. 2002. “Usherin Expression Is Highly Conserved in Mouse and Human Tissues.” Hearing Research 174, no. 1: 55–63. https://doi.org/10.1016/S0378-5955(02)00635-4.

Pénisson-Besnier, Isabelle, Valérie Allamand, Philippe Beurrier, Ludovic Martin, Joost Schalkwijk, Ivonne van Vlijmen-Willems, Corine Gartioux, et al. 2013. “Compound Heterozygous Mutations of the TNXB Gene Cause Primary Myopathy.” *Neuromuscular Disorders: NMD* 23 (8): 664–69. https://doi.org/10.1016/j.nmd.2013.04.009.

Petukhova, Lynn, Yutaka Shimomura, Muhammad Wajid, Prakash Gorroochurn, Susan E. Hodge, and Angela M. Christiano. 2009. “The Effect of Inbreeding on the Distribution of Compound Heterozygotes: A Lesson from Lipase H Mutations in Autosomal Recessive Woolly Hair/Hypotrichosis.” *Human Heredity* 68 (2): 117–30. https://doi.org/10.1159/000212504.

Piñero, Janet, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. 2017. “DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants.” *Nucleic Acids Research* 45 (D1): D833–39. https://doi.org/10.1093/nar/gkw943.

Pio-Lopez, Léo, Alberto Valdeolivas, Laurent Tichit, Élisabeth Remy, Anaïs Baudot. 2021. "MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach." *Scientific reports* 11(1): 8794.

Porten, Elmar, Beate Seliger, Verena A. Schneider, Stefan Wöll, Daniela Stangel, Rene Ramseger, and Stephan Kröger. 2010. "The Process-Inducing Activity of Transmembrane Agrin Requires Follistatin-like Domains." *The Journal of Biological Chemistry* 285 (5): 3114–25. https://doi.org/10.1074/jbc.M109.039420.

Ramanagoudr-Bhojappa, Ramanagouda, Blake Carrington, Mukundhan Ramaswami, Kevin Bishop, Gabrielle M. Robbins, MaryPat Jones, Ursula Harper, et al. 2018. "Multiplexed CRISPR/Cas9-Mediated Knockout of 19 Fanconi Anemia Pathway Genes in Zebrafish Revealed Their Roles in Growth, Sexual Development and Fertility." *PLoS Genetics* 14 (12): e1007821. https://doi.org/10.1371/journal.pgen.1007821.

Richard, P., K. Gaudon, F. Andreux, E. Yasaki, C. Prioleau, S. Bauché, A. Barois, et al. 2003. "Possible Founder Effect of Rapsyn N88K Mutation and Identification of Novel Rapsyn Mutations in Congenital Myasthenic Syndromes." *Journal of Medical Genetics* 40 (6): e81.

Rodríguez Cruz, Pedro M., Jacqueline Palace, Hayley Ramjattan, Sandeep Jayawant, Stephanie A. Robb, and David Beeson. 2015. "Salbutamol and Ephedrine in the Treatment of SevereAChR Deficiency Syndromes." *Neurology* 85 (12): 1043–47. https://doi.org/10.1212/WNL.0000000000001952.

Rodríguez Cruz, Pedro M., Jacqueline Palace, and David Beeson. 2018. "The Neuromuscular Junction and Wide Heterogeneity of Congenital Myasthenic Syndromes." *International Journal of Molecular Sciences* 19 (6). https://doi.org/10.3390/ijms19061677.

Rogers, Robert S., and Hiroshi Nishimune. 2017. "The Role of Laminins in the Organization and Function of Neuromuscular Junctions." *Matrix Biology: Journal of the International Society for Matrix Biology* 57–58: 86–105. https://doi.org/10.1016/j.matbio.2016.08.008.

Sadeh, Menachem, Xin-Ming Shen, and Andrew G. Engel. 2011. "Beneficial Effect of Albuterol in Congenital Myasthenic Syndrome with Epsilon-Subunit Mutations." *Muscle & Nerve* 44, no. 2: 289–91. https://doi.org/10.1002/mus.22153.

Saint-Amant, Louis and Pierre Drapeau. “Time Course of the Development of Motor Behaviors in the Zebrafish Embryo.” 1998. *Journal of Neurobiology* 37, no. 4: 622–32. https://doi.org/10.1002/(sici)1097-4695(199812)37:4<622::aid-neu10>3.0.co;2-s.

Saito, Rintaro, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. 2012. “A Travel Guide to Cytoscape Plugins.” *Nature Methods* 9 (11): 1069–76. https://doi.org/10.1038/nmeth.2212.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak and Albert Cardona. 2012. “Fiji - an Open Source platform for biological image analysis.” *Nature Methods* 9, 10.1038/nmeth.2019. https://doi.org/10.1038/nmeth.2019

Schwaller, Fred, Valérie Bégay, Gema García-García, Francisco J. Taberner, Rabih Moshourab, Brennan McDonald, Trevor Docter, Johannes Kühnemund, Julia Ojeda-Alonso, Ricardo Paricio-Montesinos, Stefan G. Lechner, James F. A. Poulet, Jose M. Millan and Gary R. Lewin. 2021. “USH2A is a Meissner’s Corpuscle Protein Necessary for Normal Vibration Sensing in Mice and Humans.” *Nature Neuroscience* 24, no. 1: 74–81. https://doi.org/10.1038/s41593-020-00751-y.

Senderek, Jan, Juliane S Müller, Marina Dusl, Tim M. Strom, Velina Guergueltcheva, Irmgard Diepolder, Steven H. Laval, Susan Maxwell, Judy Cossins, Sabine Krause, Nuria Muelas, Juan J. Vilchez, Jaume Colomer, Cecilia Jimenez Mallebrera, Andres Nascimento, Shahriar Nafissi, Ariana Kariminejad, Yalda Nilipour, Bita Bozorgmehr, Hossein Najmabadi, Carmelo Rodolico, Jörn P Sieb, Ortrud K. Steinlein, Beate Schlotter, Benedikt Schoser, Janbernd Kirschner, Ralf Herrmann, Thomas Voit, Anders Oldfors, Christopher Lindbergh, Andoni Urtizberea, Maja von der Hagen, Angela Hübner, Jacqueline Palace, Kate Bushby, Volker Straub, David Beeson, Angela Abicht and Hanns Lochmüller 2011. “Hexosamine Biosynthetic Pathway Mutations Cause Neuromuscular Transmission Defect.” *American Journal of Human Genetics* 88, no. 2: 162–72. https://doi.org/10.1016/j.ajhg.2011.01.008.

Sorensen, Jacob R., Caitlin Skousen, Alex Holland, Kyle Williams, and Robert D. Hyldahl. 2018. “Acute Extracellular Matrix, Inflammatory and MAPK Response to Lengthening Contractions in Elderly Human Skeletal Muscle.” *Experimental Gerontology* 106: 28–38. https://doi.org/10.1016/j.exger.2018.02.013.

Stum, Morgane, Claire-Sophie Davoine, Savine Vicart, Léna Guillot-Noël, Haluk Topaloglu, Francisco Javier Carod-Artal, Hülya Kayserili, et al. 2006. “Spectrum of HSPG2 (Perlecan) Mutations in Patients with Schwartz-Jampel Syndrome.” *Human Mutation* 27 (11): 1082–91. https://doi.org/10.1002/humu.20388.

Swuec, Paolo, Ludovic Renault, Aaron Borg, Fenil Shah, Vincent J. Murphy, Sylvie van Twest, Ambrosius P. Snijders, Andrew J. Deans, and Alessandro Costa. 2017. “The FA Core Complex Contains a Homo-Dimeric Catalytic Module for the Symmetric Mono-Ubiquitination of FANCI-FANCD2.” *Cell Reports* 18 (3): 611–23. https://doi.org/10.1016/j.celrep.2016.11.013.

Sztal, Tamar E., Avnika A. Ruparelia, Caitlin Williams, and Robert J. Bryson-Richardson. 2016. “Using Touch-Evoked Response and Locomotion Assays to Assess Muscle Performance and Function in Zebrafish.” *Journal of Visualized Experiments* : JoVE, no. 116: 54431. https://doi.org/10.3791/54431.

Thompson, Rachel, Anastasios Papakonstantinou Ntalis, Sergi Beltran, Ana Töpf, Eduardo de Paula Estephan, Kiran Polavarapu, Peter A. C. ’t Hoen, Paolo Missier, and Hanns Lochmüller. 2019. “Increasing Phenotypic Annotation Improves the Diagnostic Rate of Exome Sequencing in a Rare Neuromuscular Disorder” *Human Mutation*, June. https://doi.org/10.1002/humu.23792.

Tinevez, Jean-Yves, Nick Perry, Johannes Schindelin, Genevieve M. Hoopes, Gregory D. Reynolds, Emmanuel Laplantine, Sebastian Y. Bednarek, Spencer L. Shorte, and Kevin W. Eliceiri. “TrackMate: An Open and Extensible Platform for Single-Particle Tracking.” *Methods (San Diego, Calif.)* 115 (February 15, 2017): 80–90. https://doi.org/10.1016/j.ymeth.2016.09.016.

Valdeolivas, Alberto, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaëlle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. 2019. “Random Walk with Restart on Multiplex and Heterogeneous Biological Networks.” *Bioinformatics (Oxford, England)* 35 (3): 497–505. https://doi.org/10.1093/bioinformatics/bty637.

Dijk, Fleur S. van, Neeti Ghali, Serwet Demirdas, and Duncan Baker. 1993. “TNXB-Related Classical-Like Ehlers-Danlos Syndrome.” In *GeneReviews*®, edited by Margaret P. Adam, Ghayda M. Mirzaa, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen W. Gripp, and Anne Amemiya. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK584019/.

Vanhaesebrouck, An E, Richard Webster, Susan Maxwell, Pedro M Rodriguez Cruz, Judith Cossins, James Wickens, Wei-wei Liu, et al. 2019. “B2-Adrenergic Receptor Agonists Ameliorate the Adverse Effect of Long-Term Pyridostigmine on Neuromuscular Junction Structure.” *Brain* 142, no. 12: 3713–27. https://doi.org/10.1093/brain/awz322.

Vázquez, Miguel, Rubén Nogales, Pedro Carmona, Alberto Pascual, and Juan Pavón. 2010. “Rbbt: A Framework for Fast Bioinformatics Development with Ruby.” In *Advances in Bioinformatics*, edited by Miguel P. Rocha, Florentino Fernández Riverola, Hagit Shatkay, and Juan Manuel Corchado, 74:201–8. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13214-8_26.

Voermans, N. C., and B. G. van Engelen. 2008. “Differential Diagnosis of Muscular Hypotonia in Infants: The Kyphoscoliotic Type of Ehlers-Danlos Syndrome (EDS VI).” *Neuromuscular Disorders: NMD* 18 (11): 906; author reply 907. https://doi.org/10.1016/j.nmd.2008.05.016.

Voermans, N. C., Karin Gerrits, Baziel G. van Engelen, and Arnold de Haan. 2014. “Compound Heterozygous Mutations of the TNXB Gene Cause Primary Myopathy.” *Neuromuscular Disorders: NMD* 24 (1): 88–89. https://doi.org/10.1016/j.nmd.2013.10.007.

Wang, Dan-Ni, Zhi-Qiang Wang, Yu-Qing Chen, Guo-Rong Xu, Min-Ting Lin, and Ning Wang. 2018. “Limb-Girdle Muscular Dystrophy Type 2I: Two Chinese Families and a Review in Asian Patients.” *The International Journal of Neuroscience* 128 (3): 199–207. https://doi.org/10.1080/00207454.2017.1380640.

Wen, Hua, Michael W. Linhoff, Matthew J. McGinley, Geng-Lin Li, Glen M. Corson, Gail Mandel, and Paul Brehm. 2010. “Distinct Roles for Two Synaptotagmin Isoforms in Synchronous and Asynchronous Transmitter Release at Zebrafish Neuromuscular Junction.” *Proceedings of the National Academy of Sciences of the United States of America* 107, no. 31: 13906–11. https://doi.org/10.1073/pnas.1008598107.

Whicher, Danielle, Sarah Philbin, and Naomi Aronson. 2018. “An Overview of the Impact of Rare Disease Characteristics on Research Methodology.” *Orphanet Journal of Rare Diseases* 13 (1): 14. https://doi.org/10.1186/s13023-017-0755-5.

Xu, Zhuo, Naoki Ichikawa, Keisuke Kosaki, Yoshihiko Yamada, Takako Sasaki, Lynn Y. Sakai, Hisashi Kurosawa, Nobutaka Hattori, and Eri Arikawa-Hirasawa. 2010. “Perlecan Deficiency Causes Muscle Hypertrophy, a Decrease in Myostatin Expression, and Changes in Muscle Fiber Composition.” *Matrix Biology: Journal of the International Society for Matrix Biology* 29 (6): 461–70. https://doi.org/10.1016/j.matbio.2010.06.001.

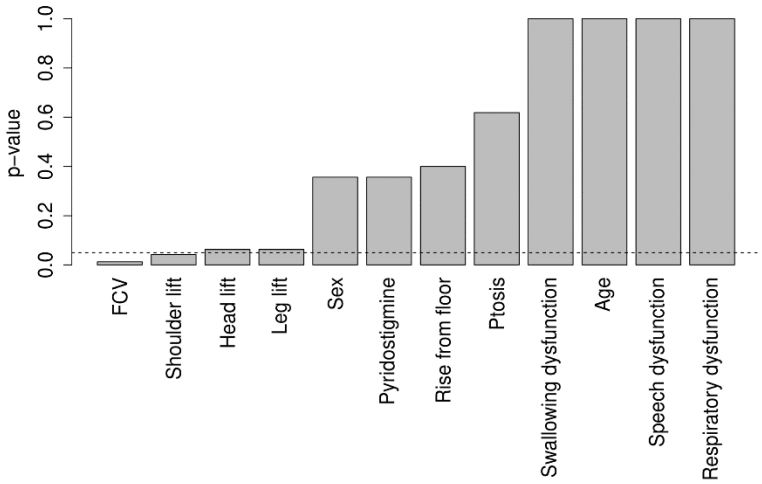
Yang, Kunfang, Hongyi Cheng, Fang Yuan, Linyi Meng, Rongrong Yin, Yuanfeng Zhang, Simei Wang, et al. 2018. “CHRNE Compound Heterozygous Mutations in Congenital Myasthenic Syndrome: A Case Report.” *Medicine* 97 (17): e0347. https://doi.org/10.1097/MD.0000000000010347.

Zhong, Jingzi, Gang Chen, Yiwu Dang, Haixia Liao, Jiapeng Zhang, and Dan Lan. 2017. “Novel Compound Heterozygous PLEC Mutations Lead to Early-onset Limb-girdle Muscular Dystrophy 2Q.” *Molecular Medicine Reports* 15 (5): 2760–64. https://doi.org/10.3892/mmr.2017.6309.

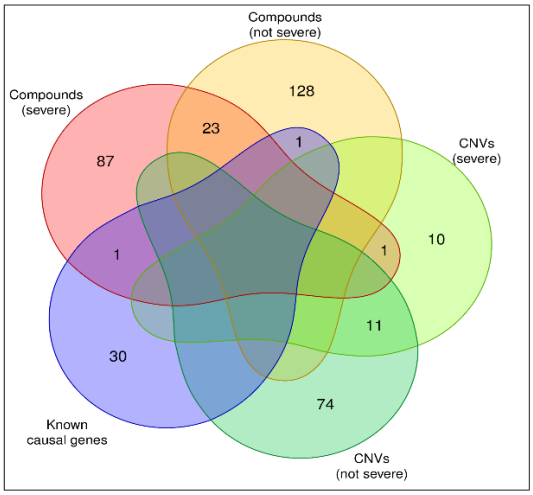
Zitnik, Marinka, and Jure Leskovec. 2017. “Predicting Multicellular Function through Multi-Layer Tissue Networks.” *Bioinformatics (Oxford, England)* 33 (14): i190–98. https://doi.org/10.1093/bioinformatics/btx25

Zoeller, Jason J., Angela McQuillan, John Whitelock, Shiu-Ying Ho, and Renato V. Iozzo. 2008. “A Central Function for Perlecan in Skeletal Muscle and Cardiovascular Development.” *The Journal of Cell Biology* 181 (2): 381–94. https://doi.org/10.1083/jcb.200708022.

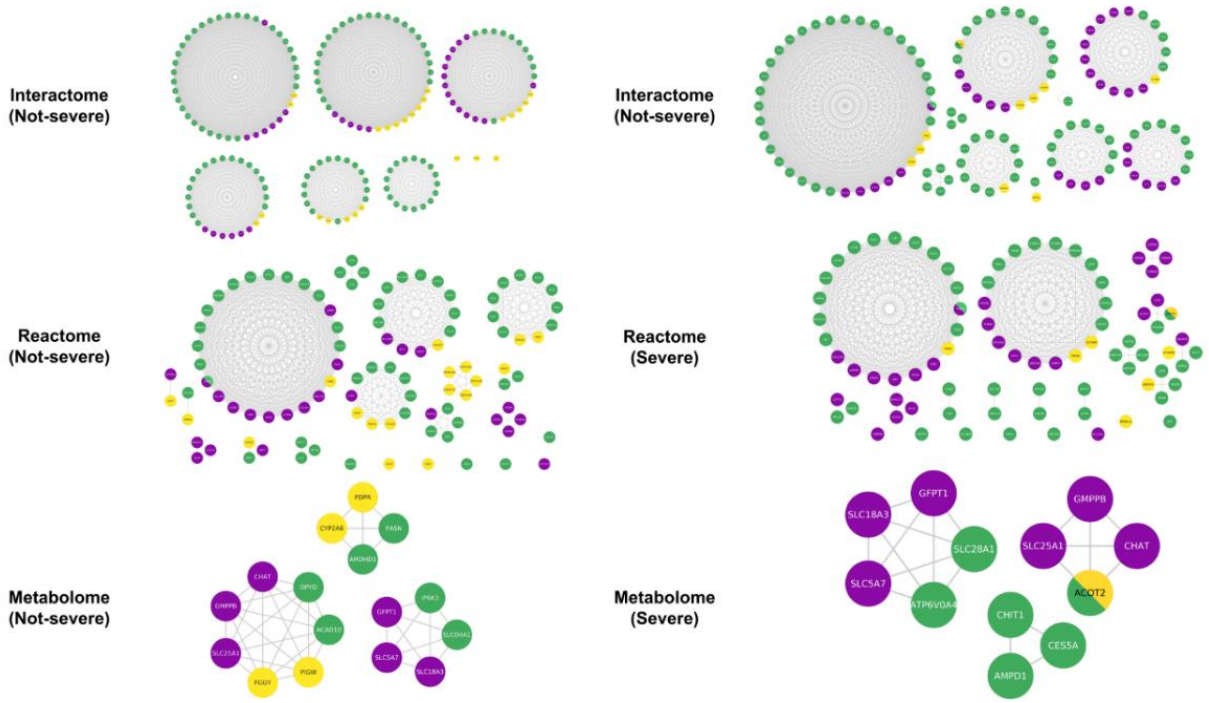
Supplementary Figures



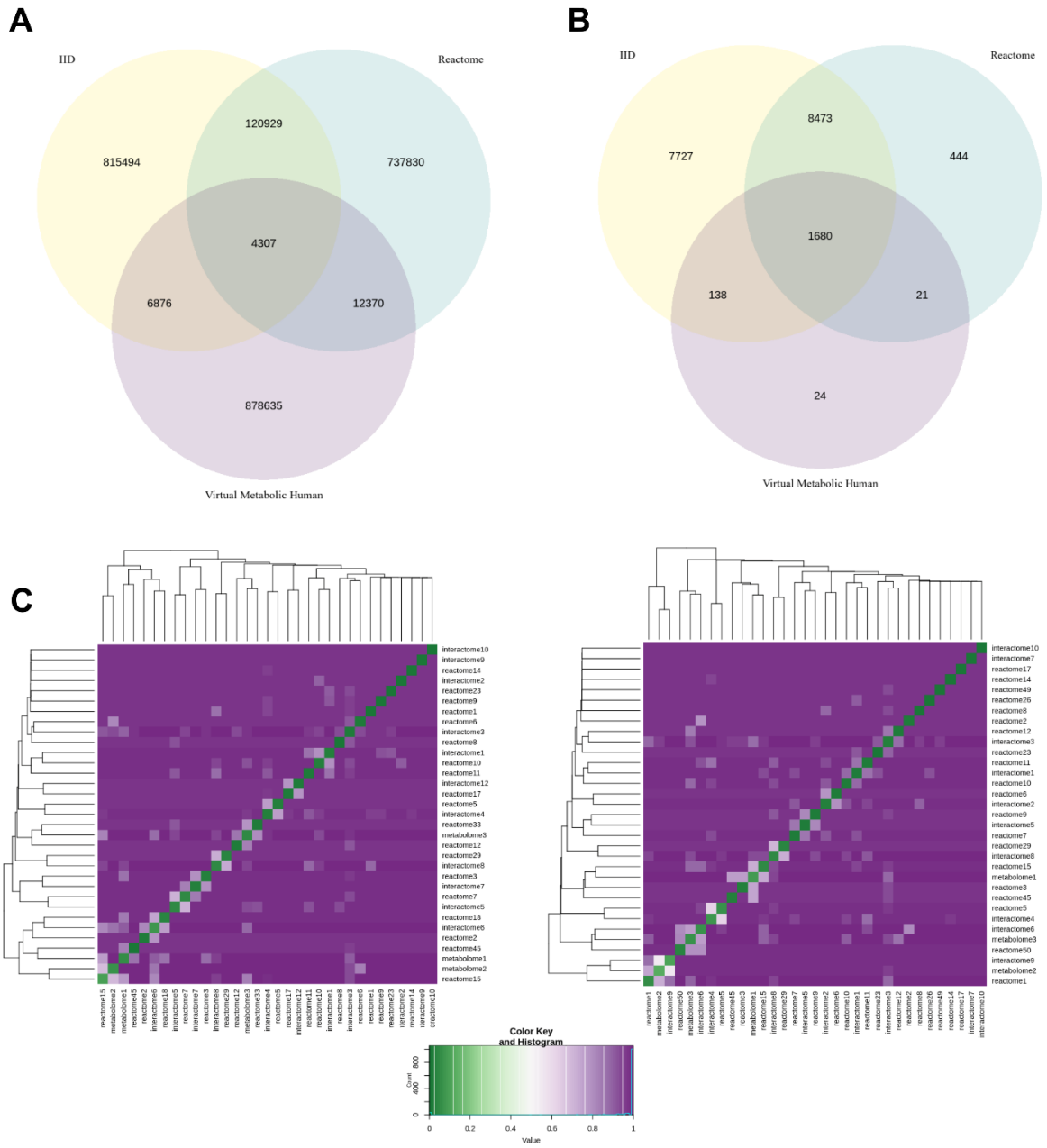
Suppl. Figure 1. Association between CMS severity (severe and not-severe phenotypes) and demographic factors (age, sex), pharmacological treatment (pyridostigmine), and clinical tests (speech, respiratory, swallowing functionality, ability to list shoulder, head, leg, eyelids (ptosis), and to rise from the floor, and Forced Vital Capacity (FVC)) (**Suppl. Table 1**). Classes were defined based on **Suppl. Table 1**. Age was discretized into two classes ('young' and 'old') based on the average age of all the patients (40 years). Bar plot reports the p-values of a two-tailed Fisher's exact test (Methods). The dotted line indicates a p-value of 0.05.



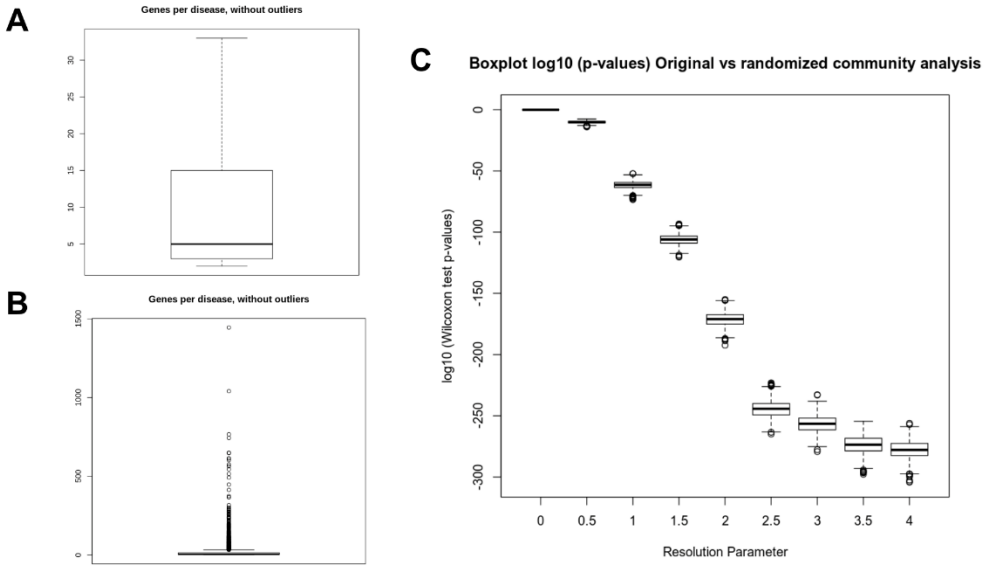
Suppl. Figure 2. Venn diagram of the genes associated with CNVs and compound heterozygous variants in not-severe and severe phenotypes as well as known CMS causal genes



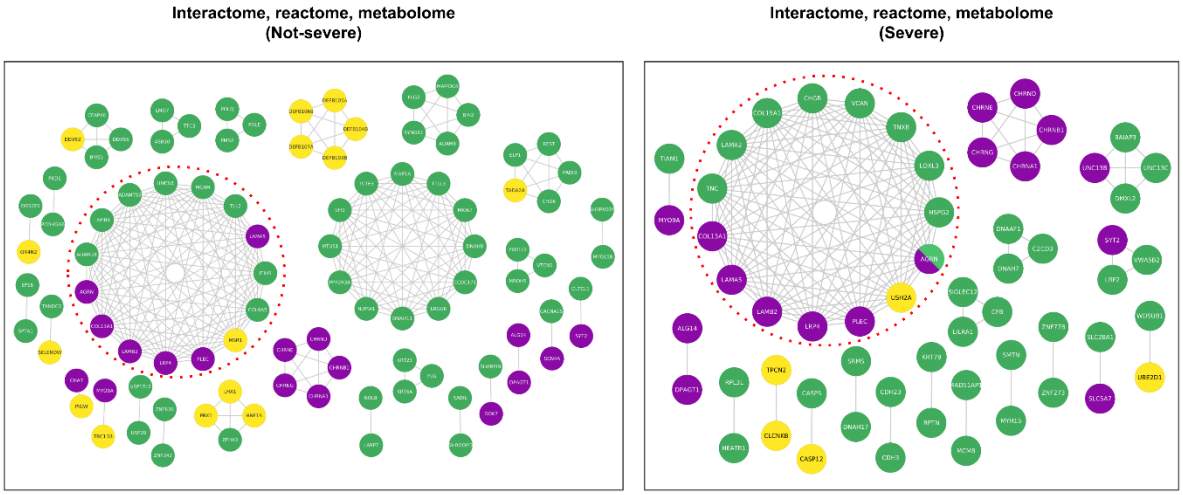
Suppl. Figure 3. Communities of CMS linked genes in the monolayer networks. Nodes are connected if they share membership to the same community from the clustering obtained using the Louvain algorithm. In green compound heterozygous variants; in yellow, CNVs; in purple, known CMS causal genes. Being a causal gene bearing compound heterozygous variants, AGRN is depicted in both purple and green. Being a gene presenting both compound heterozygous mutations and copy number variations, ACOT2 is depicted in both green and yellow.



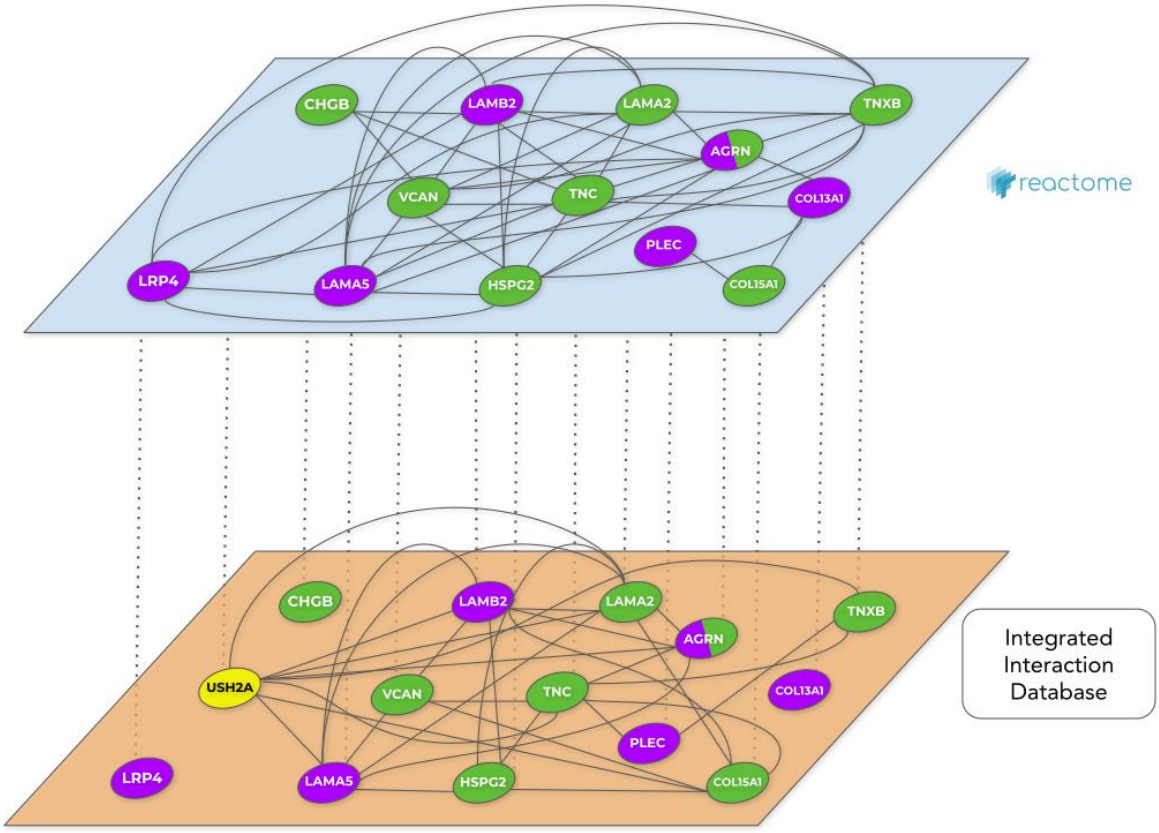
Suppl. Figure 4. (**A**) Edge overlap among the layers of the multilayer network. Each layer is identified by the name of the database from which the information was retrieved. (**B**) Node overlap among the layers of the multilayer network. (**C**) Heatmap of the Jaccard index dissimilarity among communities of CMS linked genes in the monolayer networks. Left, not-severe group, right, severe group.



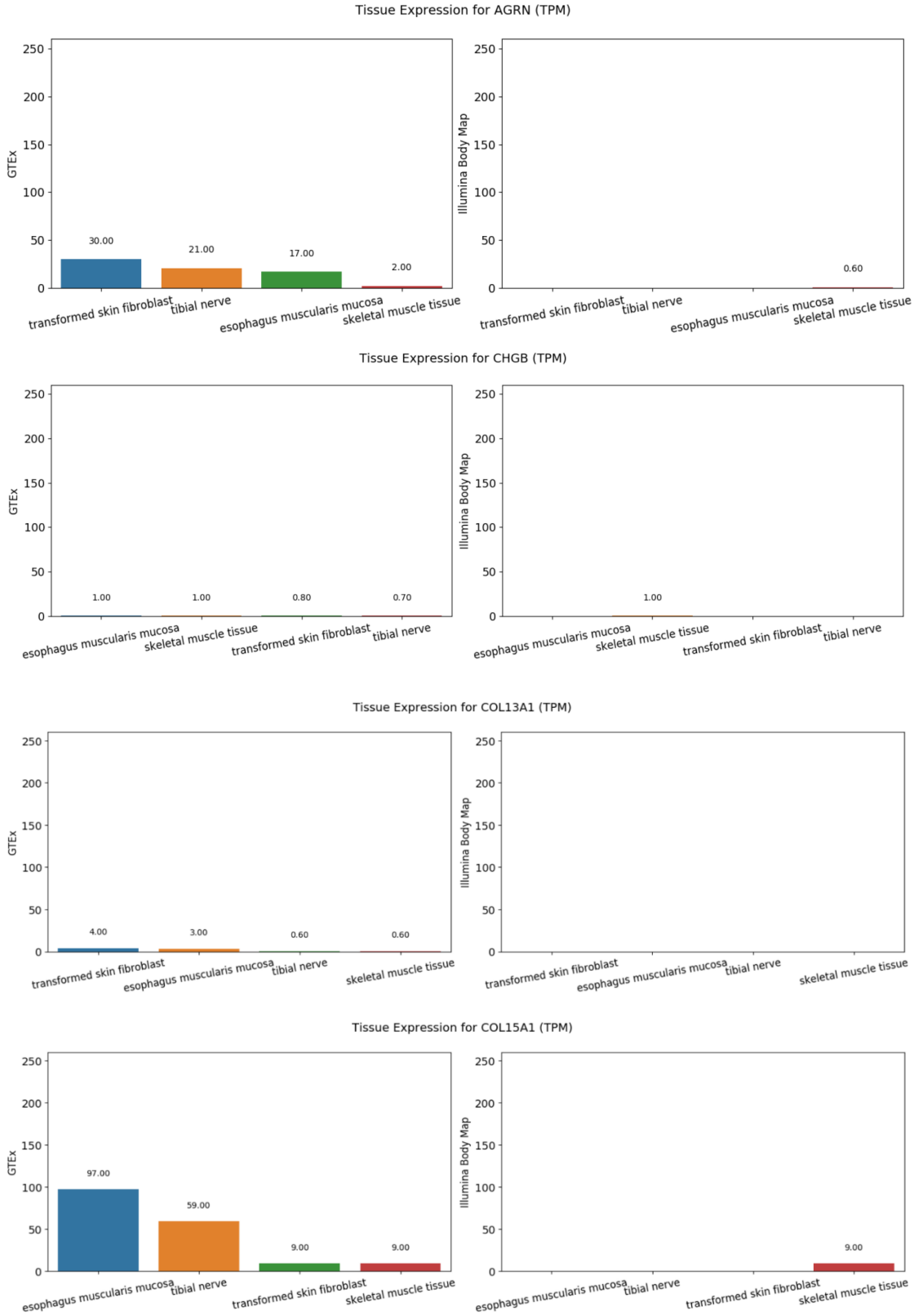
Suppl. Figure 5. Distribution of the number of genes per disease in the DisGeNET database, not showing (**A**) and showing (**B**) outliers (Methods). Distribution of p-values (two-sided Wilcoxon test; logarithmic scale) associated with DisGeNET multilayer communities along the range of the MolTi resolution parameter under evaluation (**C**; Methods).

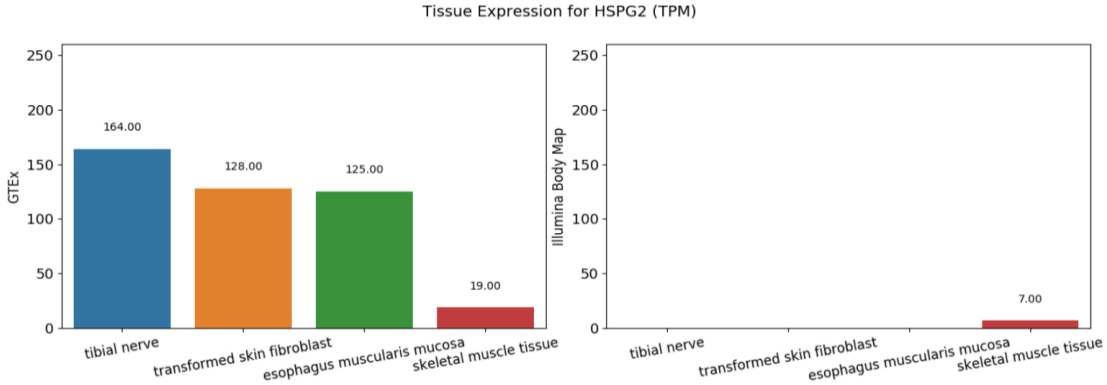


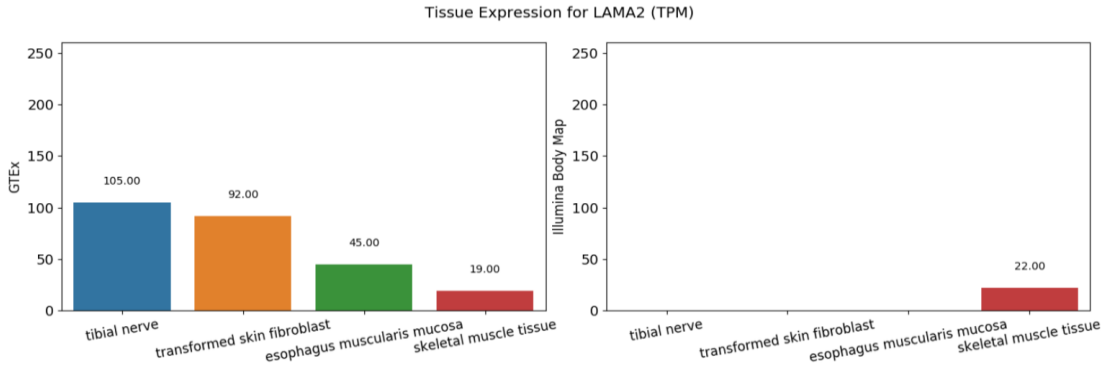
Suppl. Figure 6. Multilayer modules containing CMS linked genes of the severe and not-severe groups. Nodes are connected if the genes share membership to the same multilayer community across the range of MolTi resolution parameter (Methods). Modules with a size not expected to be found by chance are highlighted with a dotted circle (p-value < 0.05). In turquoise, compound heterozygous variants; in yellow, CNVs; in pink, known CMS causal genes.

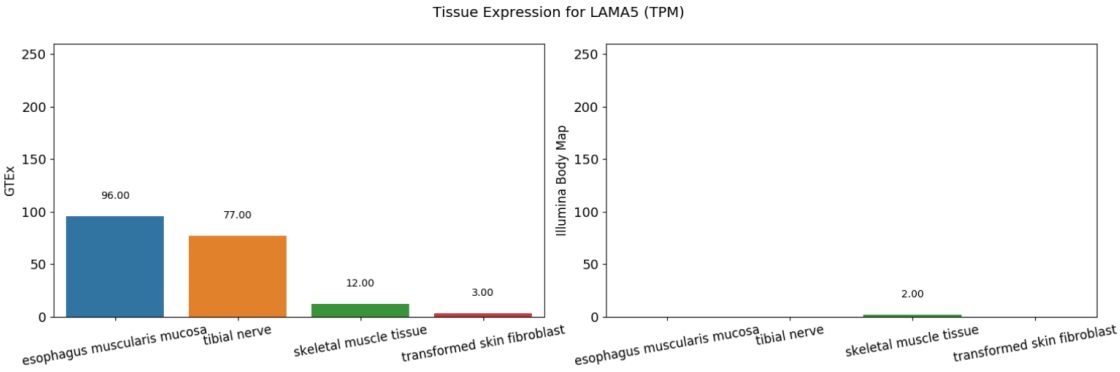


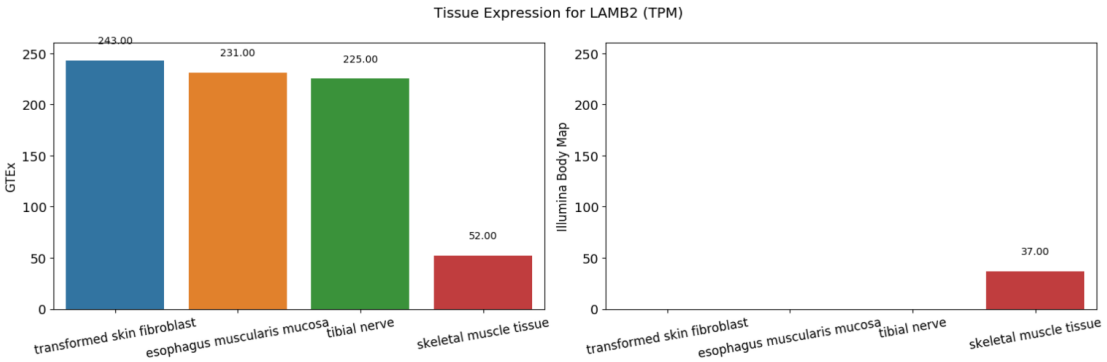
Suppl. Figure 7. Nature of the existing incident interactions between the genes identified in the severe-specific module. In green compound heterozygous variants; in yellow, CNVs; in purple, known CMS causal genes. As *LOXL3* does not present incident interactions in any of the two layers, but with other module genes that are not represented for not being a CMS linked gene, *LOXL3* is not depicted. Provided that *USH2A* does not exist on the pathways layer, it is only depicted on the protein-protein interaction layer.

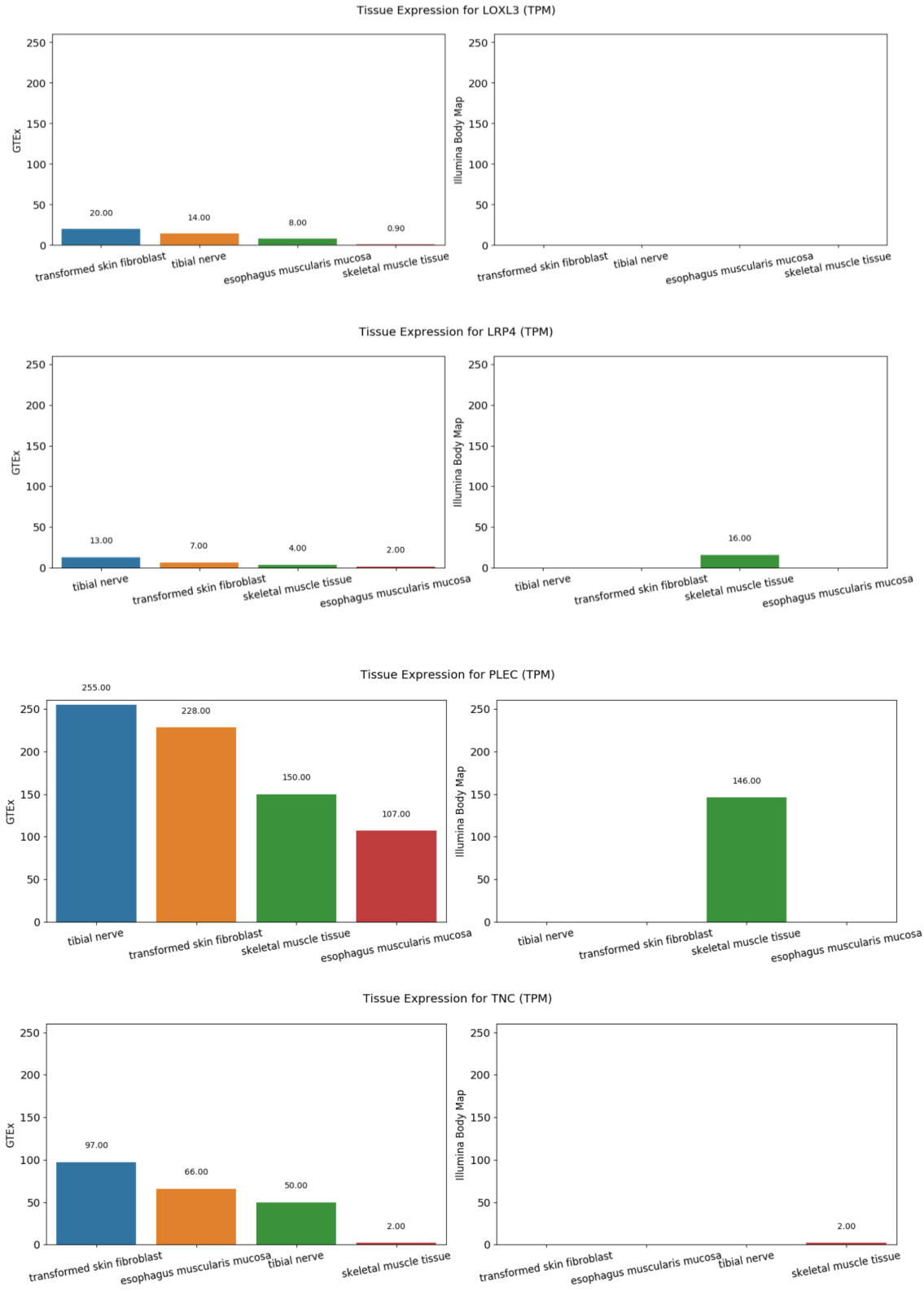

Tissue Expression for AGRN (TPM)
GTEx
Illumina Body Map
30.00
21.00
17.00
2.00
0.60
transformed skin fibroblast
tibial nerve
esophagus muscularis mucosa
skeletal muscle tissue
Tissue Expression for CHGB (TPM)
1.00
1.00
0.80
0.70
Tissue Expression for COL13A1 (TPM)
4.00
3.00
0.60
0.60
Tissue Expression for COL15A1 (TPM)
97.00
59.00
9.00
9.00

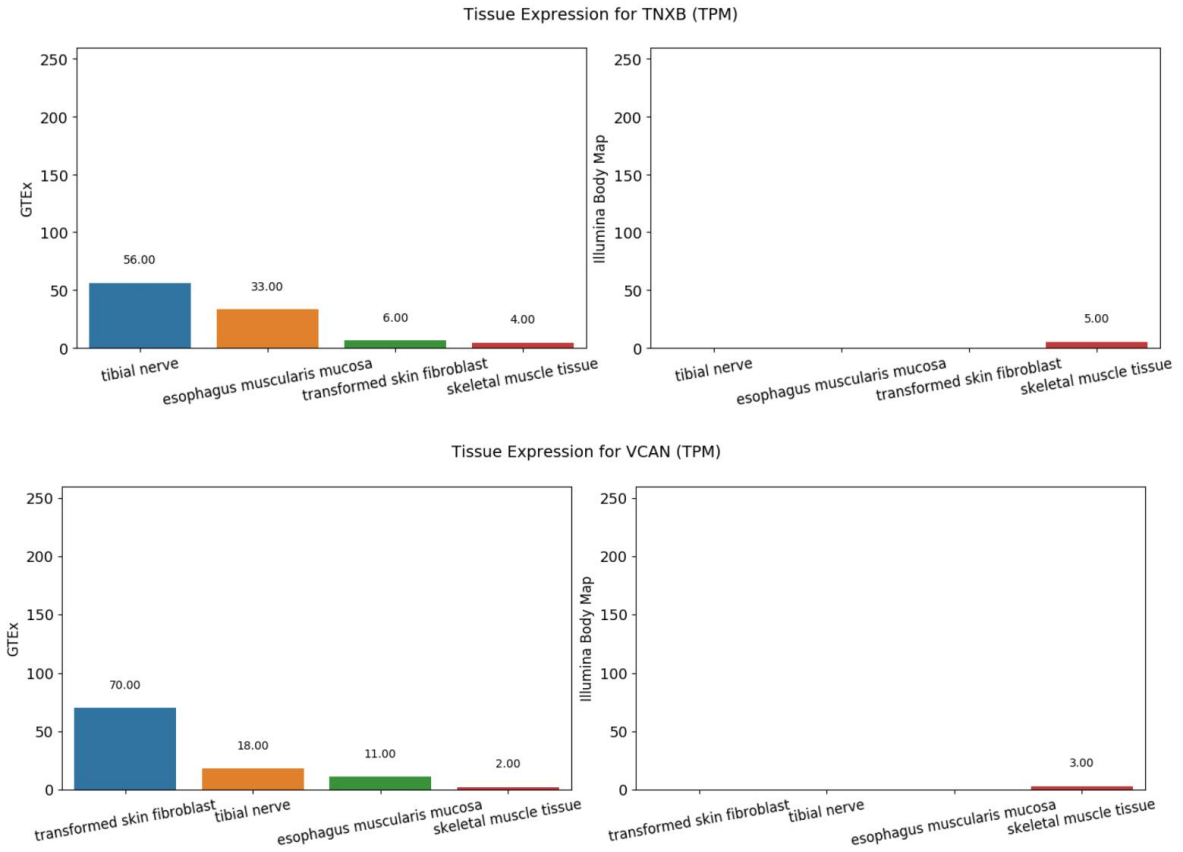

Tissue Expression for HSPG2 (TPM)
GTEx
Illumina Body Map
164.00
128.00
125.00
19.00
7.00
tibial nerve
transformed skin fibroblast
esophagus muscularis mucosa
skeletal muscle tissue


Tissue Expression for LAMA2 (TPM)
GTEx
Illumina Body Map
105.00
92.00
45.00
19.00
22.00
tibial nerve
transformed skin fibroblast
esophagus muscularis mucosa
skeletal muscle tissue

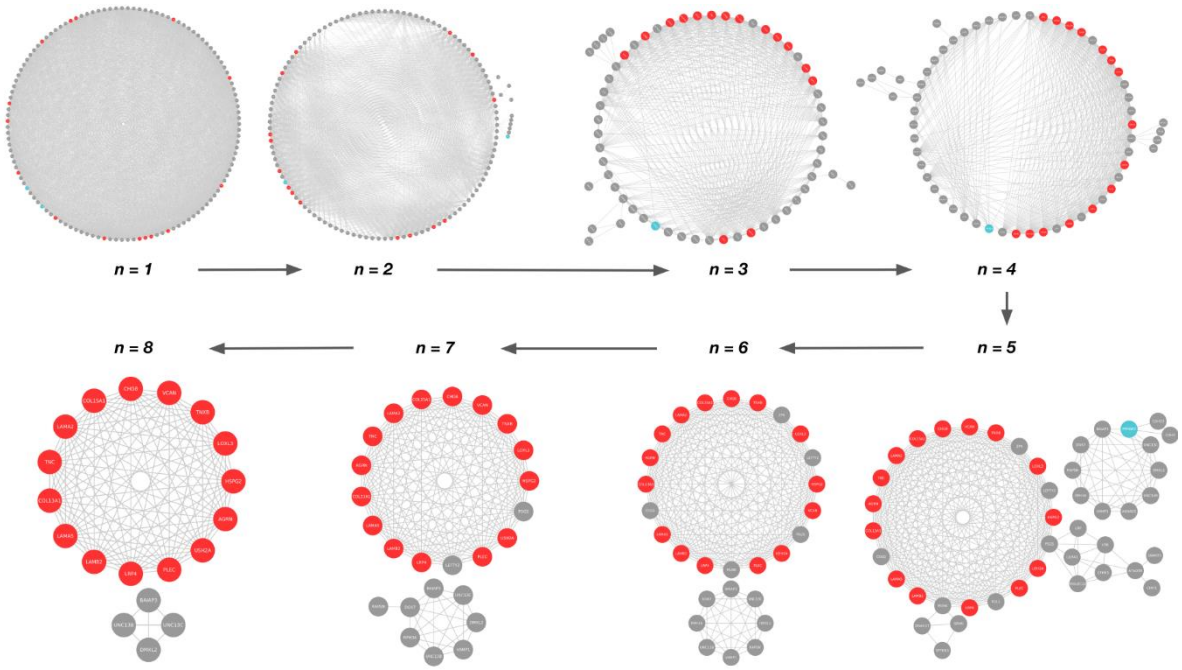

Tissue Expression for LAMA5 (TPM)
GTEx
Illumina Body Map
96.00
77.00
12.00
3.00
2.00
esophagus muscularis mucosa
tibial nerve
skeletal muscle tissue
transformed skin fibroblast


Tissue Expression for LAMB2 (TPM)
GTEx
Illumina Body Map
243.00
231.00
225.00
52.00
37.00
transformed skin fibroblast
esophagus muscularis mucosa
tibial nerve
skeletal muscle tissue

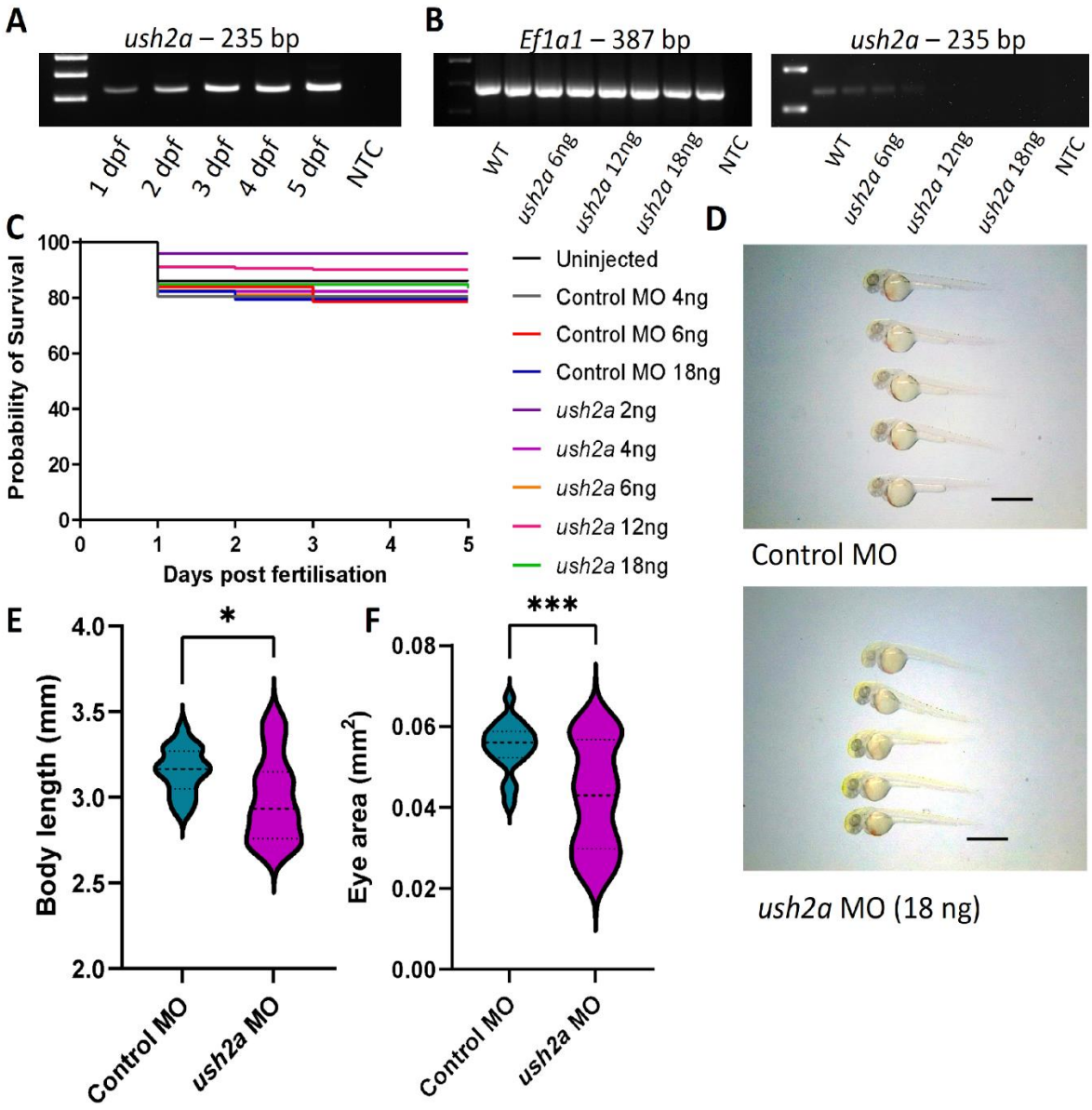

Tissue Expression for LOXL3 (TPM)
GTEx
Illumina Body Map
20.00
14.00
8.00
0.90
transformed skin fibroblast
tibial nerve
esophagus muscularis mucosa
skeletal muscle tissue
Tissue Expression for LRP4 (TPM)
13.00
7.00
4.00
2.00
16.00
tibial nerve
transformed skin fibroblast
skeletal muscle tissue
esophagus muscularis mucosa
Tissue Expression for PLEC (TPM)
255.00
228.00
150.00
107.00
146.00
Tissue Expression for TNC (TPM)
97.00
66.00
50.00
2.00



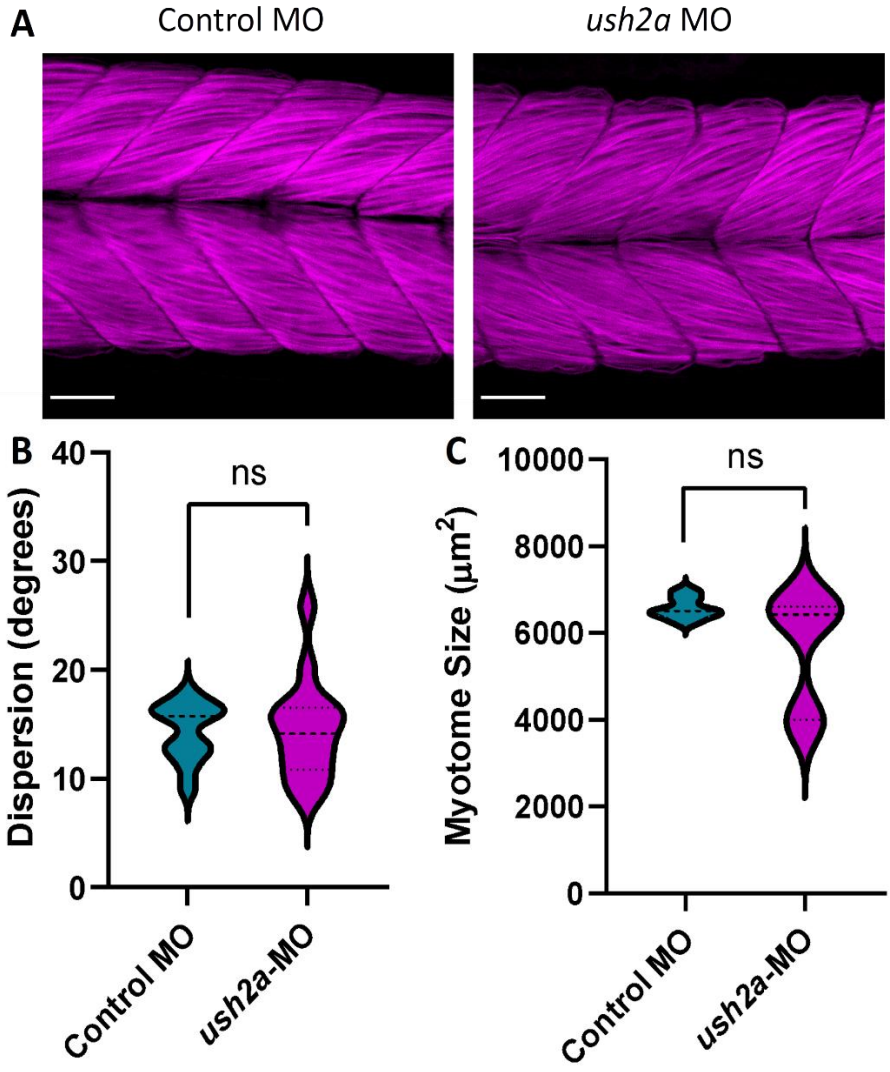
Suppl. Figure 8. Tissue-specific expression levels (Transcripts Per Million, TPM) of the genes contained in the largest module within the multilayer communities of the severe group (Methods). Expression levels are reported for GTEx (left panels) and Illumina Body Map (right panels) using EBI Expression Atlas default cutoff (0.5 TPM). Missing bars indicate no data availability (e.g., COL13A1, LOXL3). As its expression is below the cutoff for the tissues of interest in both GTEx and Illumina Body Map (Methods), USH2A is not reported. Expression level categories based on Expression Atlas: low (0.5 to 10 TPM), medium (11 to 1000 TPM), and high (more than 1000 TPM).



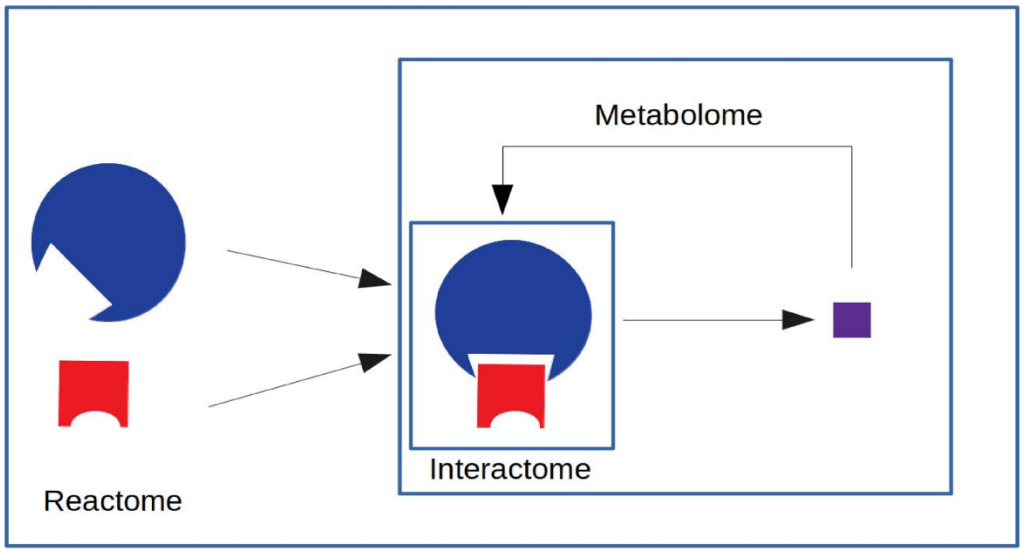
Suppl. Figure 9. Presence of *PPFIBP2* and *ACOT2* in the multilayer communities across the range of resolution parameter values (**Methods**). Genes of the severe-specific module are highlighted in red. *PPFIBP2* (present from n=1 to n=5) and *ACOT2* (present from n=1 to n=2) are depicted in blue.



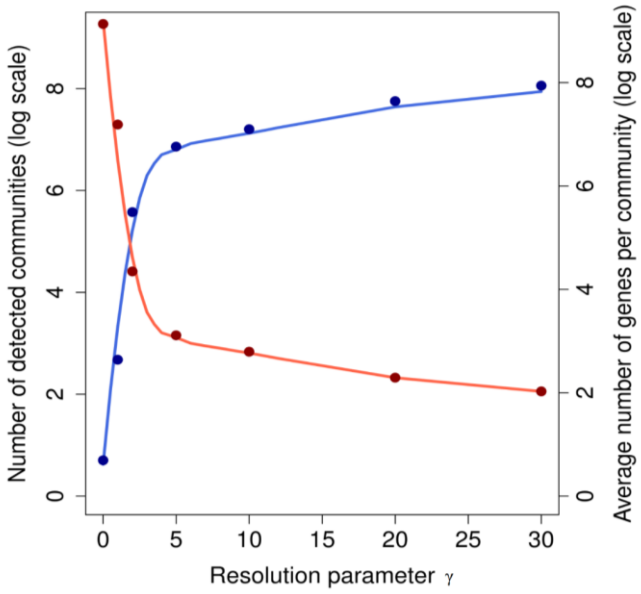
Suppl. Figure 10. Survival and phenotype of Ush2a-MO zebrafish. **(A)** Reverse-transcriptase (RT)-PCR of wildtype (WT) zebrafish at 1-5 days post fertilization (dpf) showing expression of *ush2a* throughout early development. **(B)** RT-PCR of control *ush2a*-MO fish at 2 dpf showing consistent expression of eef1a1l1 and a loss of expression of *ush2a* in *ush2a*-MO fish when injected with 6, 12 and 18 ng of MO. NTC = no template control. **(C)** Survival of WT, control MO and *ush2a*-MO injected zebrafish over 5 dpf. **(D)** Example light microscope images of control MO and 18 ng *ush2a*-MO-injected zebrafish at 2 dpf. Scale bar = 2mm. **(E)** Length of 2 dpf control and *ush2a*-MO zebrafish from the tip of the head to the tail. **(F)** Eye area of control and *ush2a*-MO zebrafish at 2 dpf. Dashed line shows the median, dotted lines show the quartiles, *p < 0.05, ***p < 0.001, unpaired t-test.



Suppl. Figure 10. Muscle morphology in *ush2a*-MO zebrafish. **(A)** Representative images of phalloidin-stained muscle fibers (detects filamentous-actin) in control and *ush2a* MO 2 dpf fish. Regularly arranged muscle fibers can be observed, with no obvious indications of disorganization, missing fibers, or fiber size changes. Scale bar = 50 µm. **(B)** Dispersion, as a measure of muscle fiber orientation and arrangement, showed no significant differences between the two groups. **(C)** Myotome size was also similar in control and *ush2a*-MO injected zebrafish. Dashed line shows the median, dotted lines show the quartiles, ns = not significant, nested t-test/unpaired t-test.



Suppl. Figure 11. Distinct layers of biological information covered in the analysis. In this example, enzyme A physically interacts (**interactome)** with enzyme B for the production of a metabolite that is further processed (**metabolome**). Moreover, all these molecules are part of the same pathway (**reactome**).



Suppl. Figure 12. Variation of number (blue line) and size (red line) of detected multilayer communities as a function of the MolTi resolution parameter γ. Curves are fitted with LOESS (locally estimated scatterplot smoothing) regression with span 0.6. Sampled points along the explored interval are shown for ease of visualization.

Supplementary Table availability and legends

Supplementary tables are available at the following link: https://doi.org/10.1101/2023.01.19.524736

Supplementary Table legends:

Suppl. Table 1. Clinical characterization of 20 CMS patients with distinct severity levels, namely severe and not-severe (mild and moderate). The annotation of clinical test responses with Human Phenotype Ontology (HPO) (https://hpo.jax.org/) terms has been manually curated. FCV: Forced Vital Capacity, i.e., volume of air that can forcibly be blown out after full inspiration. Y: yes. N: no. NI: no information.

Suppl. Table 2. Partially segregating mutations. In the table, mutations segregating at least 50% of one group (i.e., 5 out of 8 severe and 6 out of 10 mild patients) are reported (the mutation categories are described in Supplementary Information).

Suppl. Table 3. Genes associated with CNVs and compound heterozygous variants in not-severe and severe phenotypes. Severe-specific genes and known CMS causal genes are reported.

Suppl. Table 4. Estimated familiar relatedness between the analyzed patients. Only patients presenting positive relatedness (**Methods**) are shown.

Suppl. Table 5. Functional effect prediction of Ensembl VEP (**Methods**) for the compound heterozygous variants found in the largest module within the multilayer communities of the severe group. Deleterious variants are highlighted in bold.

Suppl. Table 6. Functional effect prediction of Ensembl VEP (**Methods**) for the compound heterozygous variants found in patient 3. Deleterious variants are highlighted in bold.

Supplementary Information

Functions of CMS-associated genes in the neuromuscular junction

Acetylcholine biosynthesis and release

Acetylcholine, the main neurotransmitter involved in skeletal muscle contraction, is synthesized in the presynaptic neuron, by the choline acetyltransferase (enzyme encoded by *CHAT* gene), using Acetyl-CoA and choline as substrate in the reaction (Nachmansohn and Machado 1943). Compound heterozygous mutations in this gene were identified by Ohno et al. (K. Ohno et al. 2001) causing CMS in 5 patients.

Solute carriers are critical for this process. Three genes encoding this class of transporters have been previously related to CMS and neuromuscular transmission defects, namely *SLC5A7* (Bauché et al. 2016), *SLC25A* (Chaouch et al. 2014), and *SLC18A3* (O'Grady et al. 2016). *SLC5A7* encodes the membrane choline transporter (Okuda and Haga 2000; Apparsundaram, Ferguson, and Blakely 2001). Acetyl-CoA presence is in part dependent on malate exported from mitochondria, by the action of *SLC25A1* transporter (Kaplan, Mayor, and Wood 1993). Finally, after *CHAT* generates the acetylcholine, this is carried into synaptic vesicles by *SLC18A3* gene product, the VAChT transporter (Eiden et al. 2004).

Another CMS causal gene that might have a detrimental effect at presynaptic level is *PREPL*. This gene encodes a serine oligopeptidase essential for the activation of clathrin associated adaptor protein 1 (AP1), which is needed by VAChTr to fill the synaptic vesicles with acetylcholine (Radhakrishnan et al. 2013). Régal et al. (2014) described a CMS case caused by a heterozygous deletion. Rabphilin 3a (*RPH3A*) is also involved in vesicle trafficking in the presynaptic element (Guillén et al. 2013; Shirataki et al. 1993) and has recently been described as causative of a specific form of CMS (Ricardo A. Maselli et al. 2018).

Other genes described as causal of CMS related to the vesicle generation and exocytosis are *SNAP25* (Shen et al. 2014), *VAMP1* (Salpietro et al. 2017; Shen et

al. 2017), *SYT2* (Herrmann et al. 2014; Whittaker et al. 2015) and *UNC13B* (Andrew G. Engel et al. 2016). *SNAP25* encodes synaptosomal-associated protein 25 (Sørensen et al. 2003), which is a part of the SNARE complex, where also synaptobrevin 1 (VAMP1) is allocated (Liu, Sugiura, and Lin 2011). This SNARE complex is key for the Ca^{2+}-induced exocytosis of synaptic vesicles, a process in which Synaptotagmin 2 (*SYT2*), the Ca2+ sensor, is also critical (Pang et al. 2006). *UNC13B* encodes a homolog protein to rat Munc13-1. This protein has a calmodulin site and also regulates synaptic vesicles by mediating in the SNARE complex conformation (Ma et al. 2011, 13).

Acetylcholine Receptor clustering

While acetylcholine is the main neurotransmitter in the neuromuscular junction contraction process, another important molecule, the proteoglycan agrin (*AGRN*), is released by exocytosis from the motor neuron into the synaptic cleft, where it binds the LRP4 receptor. A special type of myosin, MYO9, is known to affect agrin exocytosis upon depletion, causing a characteristic type of CMS (O'Connor et al. 2016; 2018). Agrin binding to *LRP4* leads to MuSK self-phosphorylation. Activated MuSK recruits Dok-7, which in the end stimulates Rapsyn for AChRs (acetylcholine receptors) clustering at the skeletal muscle fiber membrane (Burden, Yumoto, and Zhang 2013). MuSK, Dok-7 and Rapsyn (*RAPSN*) absence have been previously reported to result in AChR deficiency, poor neuromuscular junction development and causal of some CMS cases (Chevessier et al. 2004; Azulay et al. 1994; Kinji Ohno et al. 2002; Kumar et al. 2018). Interestingly, promoting MuSK activity has been described as capable of preserving neuromuscular synapses in Amyotrophic Lateral Sclerosis mice models (Cantor et al. 2018).

Plectin, encoded by the gene *PLEC*, is essential in the AChR clustering process as it bridges AChRs to the postsynaptic intermediate filament network (IF) via interaction with rapsyn (Mihailovska et al. 2014). Mutations in this gene are also described to cause CMS (Banwell et al. 1999; Selcen et al. 2011). MuSK is also required for the

anchoring of endplate acetylcholinesterase (AChE) at the NMJ extracellular matrix (ECM), via a collagenic-like peptide encoded by *COLQ* gene (Cartaud et al. 2004). AChE is involved in terminating impulse transmission, by hydrolysis of acetylcholine. Mutations in *COLQ* have been reported as causative for a specific form of CMS (K. Ohno et al. 1998; Donger et al. 1998).

The acetylcholine receptor itself is the main source of CMS-related mutations. In adult individuals, the receptor acts as a cation ligand-gated ion channel formed by 5 homologous subunits, being α2βδε its stoichiometry, with ε subunit replacing embryonic γ. The channel is mainly permeable to Na^+ and K^+, and to Ca^{2+} in a lesser way. When acetylcholine binds to AChR, the channel opens triggering the membrane depolarization (Brisson and Unwin 1985). All the genes encoding the receptor subunits (*CHRNA, CHRNB, CHRND, CHRNE* and *CHRNG*) have been described as causal for different CMS types (A. G. Engel et al. 1982; Quiram et al. 1999; Brownlow et al. 2001; K. Ohno et al. 1995; Morgan et al. 2006). CHRNE, which encodes the ε subunit of the AChR receptor, accounts as causative for ~50% of all reported CMS cases, although frequencies might vary depending on ethnicity (Abicht et al., 1993; Finsterer, 2019). The high prevalence of ε subunit mutations may be the result of partial compensation of its functionality by the embryonic γ (encoded by CHRNG), which is substituted after birth given its lower conductance levels. Mutations in other subunits reduce patient survival as no compensation mechanism occurs (Engel et al., 1996). Both Fast-Channel CMS (abnormally short AChR opening time) and Slow-Channel (abnormally long AChR opening time) CMS have been reported for mutations on CHRNE.

The *SCN4A* gene encodes the α subunit of the voltage-gated sodium channel (Nav1.4), which is key for the generation and propagation of action potentials through the skeletal muscle fiber, what causes Ca^{2+} release and fiber contraction. Many *SCN4A* mutations have been associated with different muscle channelopathies (Wu et al. 2016; Zaharieva et al. 2016; Tsujino et al. 2003), including CMS.

Another process related to AChR clustering is the Endoplasmic Reticulum glycosylation pathway. Normally, mutations in genes that are part of these processes cause congenital disorders of glycosylation (CDG) (Jaeken and Matthijs 2009). However, mutations in some of the pathway components (*DPAGT1, ALG2, ALG14, GFPT1 and GMPPB*) have been described as causal of some CMS variants (Belaya et al. 2012; Cossins et al. 2013, 2; Senderek et al. 2011; Belaya et al. 2015).

As for the ECM, collagens are also involved in the receptor clustering process. Collagen XIII (encoded by *COL13A1* gene) is known to be a key regulator of NMJ maturation process and AChR clustering (Latvanlehto et al. 2010). Logan et al. (Logan et al. 2015, 19) reported a specific form of CMS being caused by mutations on this gene. Laminins α5 and β2 are also involved molecules in AChR clustering (Rogers and Nishimune 2017). Each one of the different laminins have its own role during NMJ maturation and development, with mutations in *LAMA5* (Ricardo A. Maselli et al. 2017) and *LAMB2* (R. A. Maselli et al. 2009, 2) being causative of the CMS disease.

Segregation analyses

We employed Rbbt (Vázquez et al. 2010) framework to stratify CMS patients based on mutations, aiming to assess whether non-severe (n=12) and severe (n=8) patients segregate any of the following mutation types (**Figure S1**):

'**overlapping**' = the mutation overlaps the span of the gene, from first exon to last, including introns

'**mutated_isoform**' = the mutation produces a mutated isoform, i.e., an AA change (on one isoform or just the principal isoform, depending on the options used)

'**splicing**' = the mutation falls within a splicing site, they are deemed to break the protein function

'**affected**' = the mutations affects the encoded protein, by introducing a mutated isoform or a splice site mutation

'**damaged_mutated_isoform**' = the mutation makes a specific protein isoform damaged as predicted by damage or pathogenicity predictions

'broken' = the mutation seems to break the protein function, due it introducing a damaging mutation or a splice site mutation

'**TSS**' = the mutation falls within a transcription starting site (1000 bases from TSS)

'**compound**' = the gene has at least two mutations that affect it

'**homozygous**' = the gene is affected by a homozygous mutation

'**missing**' = the genes function may be entirely missing due to a homozygous or a compound mutation possibly affecting both alleles

'**gc19_pc**' **.promCore**' = core promoter of protein coding gene (hg19)

'**gc19_pc.promDomain**' = promoter domain of protein coding gene (hg19)

'**gc19_pc.5utr**' = 5'UTR of a protein coding gene (hg19)

'**gc19_pc.3utr**' = 3'UTR of a protein coding gene (hg19)

'**gc19_pc.ss**' = splicing site of a protein coding gene (hg19)

'**lncrna.promDomain**' = core promoter of a long noncoding RNA with coding potential

'**lncrna.promCore**' = core promoter of a long noncoding RNA with coding potential

'**lncrna.ss**' = splicing site of a long noncoding RNA with coding potential

'**lncrna.ncrna**' = long noncoding RNA

'**smallrna.ncrna**' = small RNA

We define complete segregating mutations that are present in one group and not in the other. No complete segregating mutations were observed (**Figure S1**), while partial segregation mutations (i.e., present in at least 50% of the patients of one group and not in the other) can be appreciated (**Suppl. Table 2**).

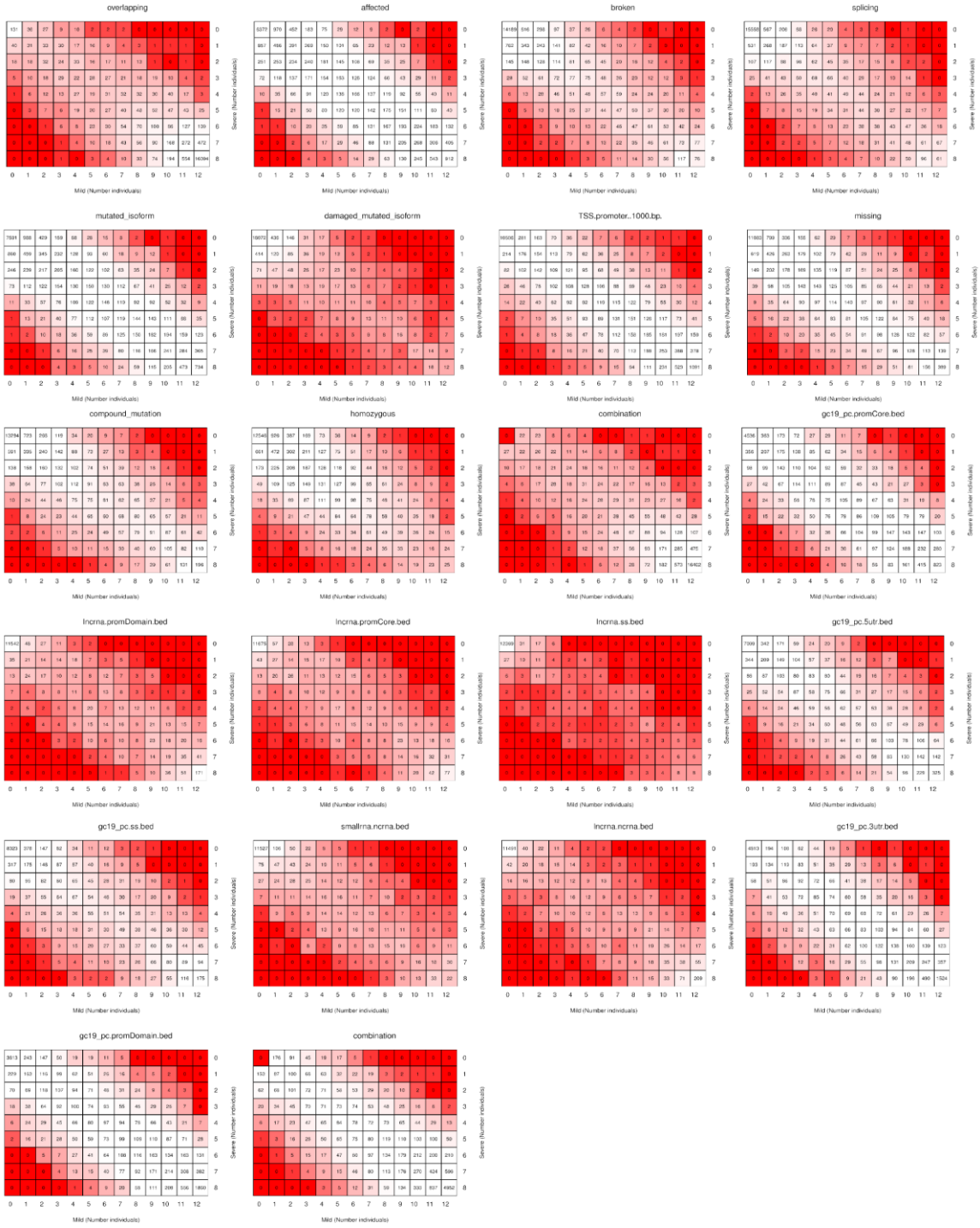


Figure S1. Segregation analysis of several mutation types (described in the text). The number of mutations that overlap in the two groups (Non-severe and Severe) are reported for sets of individuals (0 to 12 for non-severe individuals, 0 to 8 for severe individuals).

Multilayer community detection analysis

In this work, we performed a multilayer community detection analysis using MolTi software (Didier, Brun, and Baudot 2015), which adapts the Louvain clustering algorithm with modularity maximization to multilayer networks. The algorithm is parametrized by the resolution parameter γ: the higher the value of γ, the smaller the size of the detected multilayer communities. Given the intrinsic resolution limit of modularity, the reliability of community detection should be assessed ad hoc using quality functions that are able to capture the actual community structure of a network (Fortunato and Barthelemy 2007). In this work, we were interested in the identification of communities that robustly express functional relationships among the CMS linked genes (i.e., known CMS causal genes, and severe and non-severe compound heterozygous variants and CNVs). Accordingly, we sought to determine the largest module of CMS linked genes that are found in the same multilayer communities at any value of resolution within the parameter range in which the community structure is more variable (see **Supplementary Figure 12**). The adopted procedure is illustrated in **Figure S2**.

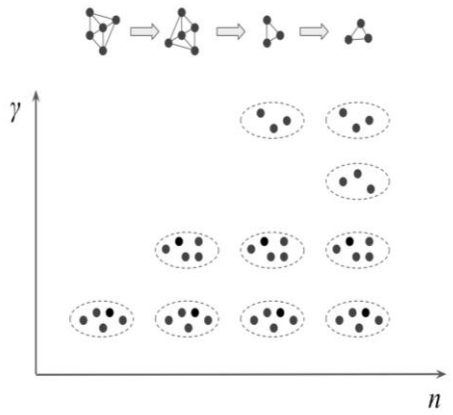


Figure S2. Module identification based on detected multilayer communities. Genes that are found in the same community at *n* values of the resolution parameter γ are represented as fully connected modules (upper panel). The resolution range considered is γ∈(0,4] with intervals of 0.5 (Methods). The module corresponding to the highest *n* contains genes that are systematically found in the same community across the entire range of resolution.

Supplementary Information references

Abicht, A., Müller, J.S., Lochmüller, H., 1993. Congenital Myasthenic Syndromes Overview, in: Adam, M.P., Mirzaa, G.M., Pagon, R.A., Wallace, S.E., Bean, L.J., Gripp, K.W., Amemiya, A. (Eds.), GeneReviews®. University of Washington, Seattle, Seattle (WA).

Apparsundaram, S., S. M. Ferguson, and R. D. Blakely. 2001. 'Molecular Cloning and Characterization of a Murine Hemicholinium-3-Sensitive Choline Transporter'. *Biochemical Society Transactions* 29 (Pt 6): 711–16.

Azulay, J. P., J. Pouget, D. Figarella-Branger, R. Colamarino, J. F. Pellissier, and G. Serratrice. 1994. '[Isolated proximal muscular weakness disclosing myasthenic syndrome]'. *Revue Neurologique* 150 (5): 377–81.

Banwell, B. L., J. Russel, T. Fukudome, X. M. Shen, G. Stilling, and A. G. Engel. 1999. 'Myopathy, Myasthenic Syndrome, and Epidermolysis Bullosa Simplex Due to Plectin Deficiency'. Journal of Neuropathology and Experimental Neurology 58 (8): 832–46.

Bauché, Stéphanie, Seana O'Regan, Yoshiteru Azuma, Fanny Laffargue, Grace McMacken, Damien Sternberg, Guy Brochier, et al. 2016. 'Impaired Presynaptic High-Affinity Choline Transporter Causes a Congenital Myasthenic Syndrome with Episodic Apnea'. *American Journal of Human Genetics* 99 (3): 753–61. https://doi.org/10.1016/j.ajhg.2016.06.033.

Belaya, Katsiaryna, Sarah Finlayson, Clarke R. Slater, Judith Cossins, Wei Wei Liu, Susan Maxwell, Simon J. McGowan, et al. 2012. 'Mutations in DPAGT1 Cause a Limb-Girdle Congenital Myasthenic Syndrome with Tubular Aggregates'. *American Journal of Human Genetics* 91 (1): 193–201. https://doi.org/10.1016/j.ajhg.2012.05.022.

Belaya, Katsiaryna, Pedro M. Rodríguez Cruz, Wei Wei Liu, Susan Maxwell, Simon McGowan, Maria E. Farrugia, Richard Petty, et al. 2015. 'Mutations in GMPPB Cause Congenital Myasthenic Syndrome and Bridge Myasthenic Disorders with Dystroglycanopathies'. *Brain: A Journal of Neurology* 138 (Pt 9): 2493–2504. https://doi.org/10.1093/brain/awv185.

Brisson, A., and P. N. Unwin. 1985. 'Quaternary Structure of the Acetylcholine Receptor'. *Nature*:315 (6019): 474–77.

Brownlow, S., R. Webster, R. Croxen, M. Brydson, B. Neville, J. P. Lin, A. Vincent, J. Newsom-Davis, and D. Beeson. 2001. 'Acetylcholine Receptor Delta Subunit Mutations Underlie a Fast-Channel Myasthenic Syndrome and Arthrogryposis Multiplex Congenita'. *The Journal of Clinical Investigation* 108 (1): 125–30. https://doi.org/10.1172/JCI1293

Burden, Steven J., Norihiro Yumoto, and Wei Zhang. 2013. "The Role of MuSK in Synapse Formation and Neuromuscular Disease." *Cold Spring Harbor Perspectives in Biology* 5 (5): a009167. https://doi.org/10.1101/cshperspect.a009167.

Cantor, Sarah, Wei Zhang, Nicolas Delestrée, Leonor Remédio, George Z. Mentis, and Steven J. Burden. 2018. "Preserving Neuromuscular Synapses in ALS by Stimulating MuSK with a Therapeutic Agonist Antibody." *ELife* 7 (February): e34375. https://doi.org/10.7554/eLife.34375.

Cartaud, Annie, Laure Strochlic, Manuel Guerra, Benoît Blanchard, Monique Lambergeon, Eric Krejci, Jean Cartaud, and Claire Legay. 2004. "MuSK Is Required for Anchoring Acetylcholinesterase at the Neuromuscular Junction." *The Journal of Cell Biology* 165 (4): 505–15. https://doi.org/10.1083/jcb.200307164.

Chaouch, Amina, Vito Porcelli, Daniel Cox, Shimon Edvardson, Pasquale Scarcia, Anna De Grassi, Ciro L. Pierri, et al. 2014. "Mutations in the Mitochondrial Citrate Carrier SLC25A1 Are Associated with Impaired Neuromuscular Transmission." *Journal of Neuromuscular Diseases* 1 (1): 75–90. https://doi.org/10.3233/JND-140021.

Chevessier, Frédéric, Brice Faraut, Aymeric Ravel-Chapuis, Pascale Richard, Karen Gaudon, Stéphanie Bauché, Cassandra Prioleau, et al. 2004. "MUSK, a New Target for Mutations Causing Congenital Myasthenic Syndrome." *Human Molecular Genetics* 13 (24): 3229–40. https://doi.org/10.1093/hmg/ddh333.

Cossins, Judith, Katsiaryna Belaya, Debbie Hicks, Mustafa A. Salih, Sarah Finlayson, Nicola Carboni, Wei Wei Liu, et al. 2013. "Congenital Myasthenic Syndromes Due to Mutations in ALG2 and ALG14." *Brain: A Journal of Neurology* 136 (Pt 3): 944–56. https://doi.org/10.1093/brain/awt010.

Didier, Gilles, Christine Brun, and Anaïs Baudot. 2015. "Identifying Communities from Multiplex Biological Networks." *PeerJ* 3 (December): e1525. https://doi.org/10.7717/peerj.1525.

Donger, C., E. Krejci, A. P. Serradell, B. Eymard, S. Bon, S. Nicole, D. Chateau, et al. 1998. “Mutation in the Human Acetylcholinesterase-Associated Collagen Gene, COLQ, Is Responsible for Congenital Myasthenic Syndrome with End-Plate Acetylcholinesterase Deficiency (Type Ic).” *American Journal of Human Genetics* 63 (4): 967–75. https://doi.org/10.1086/302059.

Eiden, Lee E., Martin K.-H. Schäfer, Eberhard Weihe, and Burkhard Schütz. 2004. “The Vesicular Amine Transporter Family (SLC18): Amine/Proton Antiporters Required for Vesicular Accumulation and Regulated Exocytotic Secretion of Monoamines and Acetylcholine.” *Pflugers Archiv: European Journal of Physiology* 447 (5): 636–40. https://doi.org/10.1007/s00424-003-1100-5.

Engel, A. G., E. H. Lambert, D. M. Mulder, C. F. Torres, K. Sahashi, T. E. Bertorini, and J. N. Whitaker. 1982. ‘A Newly Recognized Congenital Myasthenic Syndrome Attributed to a Prolonged Open Time of the Acetylcholine-Induced Ion Channel’. *Annals of Neurology* 11 (6):553–69. https://doi.org/10.1002/ana.410110603.

Engel, A. G., K. Ohno, C. Bouzat, S. M. Sine, and R. C. Griggs. 1996. “End-Plate Acetylcholine Receptor Deficiency Due to Nonsense Mutations in the Epsilon Subunit.” *Annals of Neurology* 40 (5): 810–17. https://doi.org/10.1002/ana.410400521.

Engel, Andrew G., Duygu Selcen, Xin-Ming Shen, Margherita Milone, and C. Michel Harper. 2016. ‘Loss of MUNC13-1 Function Causes Microcephaly, Cortical Hyperexcitability, and Fatal Myasthenia’. *Neurology.Genetics* 2 (5): e105. https://doi.org/10.1212/NXG.0000000000000105.

Finsterer, Josef. 2019. “Congenital Myasthenic Syndromes.” *Orphanet Journal of Rare Diseases* 14 (1): 57. https://doi.org/10.1186/s13023-019-1025-5.

Fortunato, S., and M. Barthelemy. 2007. ‘Resolution Limit in Community Detection’. Proceedings of the National Academy of Sciences 104 (1): 36–41. https://doi.org/10.1073/pnas.0605965104.

Guillén, Jaime, Cristina Ferrer-Orta, Mònica Buxaderas, Dolores Pérez-Sánchez, Marta Guerrero-Valero, Ginés Luengo-Gil, Joan Pous, et al. 2013. ‘Structural Insights into the Ca2+ and PI(4,5)P2 Binding Modes of the C2 Domains of Rabphilin 3A and Synaptotagmin 1’. *Proceedings of the National Academy of Sciences of the United States of America* 110 (51): 20503–8. https://doi.org/10.1073/pnas.1316179110.

Herrmann, David N., Rita Horvath, Janet E. Sowden, Michael Gonzalez, Michael Gonzales, Avencia Sanchez-Mejias, Zhuo Guan, et al. 2014. 'Synaptotagmin 2 Mutations Cause an Autosomal-Dominant Form of Lambert-Eaton Myasthenic Syndrome and Nonprogressive Motor Neuropathy'. *American Journal of Human Genetics* 95 (3): 332–39. https://doi.org/10.1016/j.ajhg.2014.08.007.

Jaeken, Jaak, and Gert Matthijs. 2009. 'From Glycosylation to Glycosylation Diseases'. *Biochimica Et Biophysica Acta* 1792 (9): 823. https://doi.org/10.1016/j.bbadis.2009.08.003.

Kaplan, R. S., J. A. Mayor, and D. O. Wood. 1993. 'The Mitochondrial Tricarboxylate Transport Protein. CDNA Cloning, Primary Structure, and Comparison with Other Mitochondrial Transport Proteins'. *The Journal of Biological Chemistry* 268 (18): 13682–90.

Kumar, Ashutosh, Sheila Asghar, Robert Kavanagh, and Matthew P. Wicklund. 2018. 'Unique Presentation of Rapidly Fluctuating Symptoms in a Child with Congenital Myasthenic Syndrome Due to RAPSN Mutation'. *Muscle & Nerve* 58 (4): E23–24. https://doi.org/10.1002/mus.26200.

Latvanlehto, Anne, Michael A. Fox, Raija Sormunen, Hongmin Tu, Tuomo Oikarainen, Anu Koski, Nikolay Naumenko, et al. 2010. 'Muscle-Derived Collagen XIII Regulates Maturation of the Skeletal Neuromuscular Junction'. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (37): 12230–41. https://doi.org/10.1523/JNEUROSCI.5518-09.2010.

Liu, Yun, Yoshie Sugiura, and Weichun Lin. 2011. 'The Role of Synaptobrevin1/VAMP1 in Ca2+-Triggered Neurotransmitter Release at the Mouse Neuromuscular Junction'. *The Journal of Physiology* 589 (Pt 7): 1603–18. https://doi.org/10.1113/jphysiol.2010.201939.

Logan, Clare V., Judith Cossins, Pedro M. Rodríguez Cruz, David A. Parry, Susan Maxwell, Pilar Martínez-Martínez, Joey Riepsaame, et al. 2015. 'Congenital Myasthenic Syndrome Type 19 Is Caused by Mutations in COL13A1, Encoding the Atypical Non-Fibrillar Collagen Type XIII A1 Chain'. American *Journal of Human Genetics* 97 (6): 878–85. https://doi.org/10.1016/j.ajhg.2015.10.017.

Ma, Cong, Wei Li, Yibin Xu, and Josep Rizo. 2011. 'Munc13 Mediates the Transition from the Closed Syntaxin-Munc18 Complex to the SNARE Complex'. *Nature Structural & Molecular Biology* 18 (5): 542–49. https://doi.org/10.1038/nsmb.2047.

Maselli, R. A., J. J. Ng, J. A. Anderson, O. Cagney, J. Arredondo, C. Williams, H. B. Wessel, H. Abdel-Hamid, and R. L. Wollmann. 2009. 'Mutations in LAMB2 Causing a Severe Form of Synaptic Congenital Myasthenic Syndrome'. *Journal of Medical Genetics* 46 (3): 203–8. https://doi.org/10.1136/jmg.2008.063693.

Maselli, Ricardo A., Juan Arredondo, Jessica Vázquez, Jessica X. Chong, University of Washington Center for Mendelian Genomics, Michael J. Bamshad, Deborah A. Nickerson, et al. 2017. 'Presynaptic Congenital Myasthenic Syndrome with a Homozygous Sequence Variant in LAMA5 Combines Myopia, Facial Tics, and Failure of Neuromuscular Transmission'. *American Journal of Medical Genetics*. Part A 173 (8): 2240–45. https://doi.org/10.1002/ajmg.a.38291.

Maselli, Ricardo A., Jessica Vázquez, Leah Schrumpf, Juan Arredondo, Marian Lara, Jonathan B. Strober, Peter Pytel, Robert L. Wollmann, and Michael Ferns. 2018. 'Presynaptic Congenital Myasthenic Syndrome with Altered Synaptic Vesicle Homeostasis Linked to Compound Heterozygous Sequence Variants in RPH3A'. *Molecular Genetics & Genomic Medicine* 6 (3): 434–40. https://doi.org/10.1002/mgg3.370.

Mihailovska, Eva, Marianne Raith, Rocio G. Valencia, Irmgard Fischer, Mumna Al Banchaabouchi, Ruth Herbst, and Gerhard Wiche. 2014. 'Neuromuscular Synapse Integrity Requires Linkage of Acetylcholine Receptors to Postsynaptic Intermediate Filament Networks via Rapsyn-Plectin 1f Complexes'. *Molecular Biology of the Cell* 25 (25): 4130–49. https://doi.org/10.1091/mbc.E14-06-1174.

Morgan, Neil V., Louise A. Brueton, Phillip Cox, Marie T. Greally, John Tolmie, Shanaz Pasha, Irene A. Aligianis, et al. 2006. 'Mutations in the Embryonal Subunit of the Acetylcholine Receptor (CHRNG) Cause Lethal and Escobar Variants of Multiple Pterygium Syndrome'. *American Journal of Human Genetics* 79 (2): 390–95. https://doi.org/10.1086/506256.

Nachmansohn, D., and A. L. Machado. 1943. 'THE FORMATION OF ACETYLCHOLINE. A NEW ENZYME: "CHOLINE ACETYLASE"'. *Journal of Neurophysiology* 6 (5): 397–403. https://doi.org/10.1152/jn.1943.6.5.397.

O'Connor, Emily, Ana Töpf, Juliane S. Müller, Daniel Cox, Teresinha Evangelista, Jaume Colomer, Angela Abicht, et al. 2016. 'Identification of Mutations in the MYO9A Gene in Patients with Congenital Myasthenic Syndrome'. *Brain: A Journal of Neurology* 139 (Pt 8): 2143–53. https://doi.org/10.1093/brain/aww130.

O'Connor, Emily, Vietxuan Phan, Isabell Cordts, George Cairns, Stefan Hettwer, Daniel Cox, Hanns Lochmüller, and Andreas Roos. 2018. 'MYO9A Deficiency in Motor Neurons Is Associated with Reduced Neuromuscular Agrin Secretion'. *Human Molecular Genetics* 27 (8): 1434–46. https://doi.org/10.1093/hmg/ddy054.

O'Grady, Gina L., Corien Verschuuren, Michaela Yuen, Richard Webster, Manoj Menezes, Johanna M. Fock, Natalie Pride, et al. 2016. 'Variants in SLC18A3, Vesicular Acetylcholine Transporter, Cause Congenital Myasthenic Syndrome'. *Neurology* 87 (14): 1442–48. https://doi.org/10.1212/WNL.0000000000003179.

Ohno, K., J. Brengman, A. Tsujino, and A. G. Engel. 1998. 'Human Endplate Acetylcholinesterase Deficiency Caused by Mutations in the Collagen-like Tail Subunit (ColQ) of the Asymmetric Enzyme'. *Proceedings of the National Academy of Sciences of the United States of America* 95 (16): 9654–59.

Ohno, K., D. O. Hutchinson, M. Milone, J. M. Brengman, C. Bouzat, S. M. Sine, and A. G. Engel. 1995. 'Congenital Myasthenic Syndrome Caused by Prolonged Acetylcholine Receptor Channel Openings Due to a Mutation in the M2 Domain of the Epsilon Subunit'. *Proceedings of the National Academy of Sciences of the United States of America* 92 (3): 758–62.

Ohno, K., A. Tsujino, J. M. Brengman, C. M. Harper, Z. Bajzer, B. Udd, R. Beyring, S. Robb, F. J. Kirkham, and A. G. Engel. 2001. 'Choline Acetyltransferase Mutations Cause Myasthenic Syndrome Associated with Episodic Apnea in Humans'. *Proceedings of the National Academy of Sciences of the United States of America* 98 (4):2017–22. https://doi.org/10.1073/pnas.98.4.2017.

Ohno, Kinji, Andrew G. Engel, Xin-Ming Shen, Duygu Selcen, Joan Brengman, C. Michel Harper, Akira-Tsujino, and Margherita Milone. 2002. 'Rapsyn Mutations in Humans Cause Endplate Acetylcholine-Receptor Deficiency and Myasthenic Syndrome'*. American Journal of Human Genetics* 70 (4): 875–85. https://doi.org/10.1086/339465.

Okuda, T., and T. Haga. 2000. 'Functional Characterization of the Human High-Affinity Choline Transporter'. FEBS Letters 484 (2): 92–97.

Pang, Zhiping P., Jianyuan Sun, Josep Rizo, Anton Maximov, and Thomas C. Südhof. 2006. 'Genetic Analysis of Synaptotagmin 2 in Spontaneous and Ca2+-Triggered Neurotransmitter Release'. *The EMBO Journal* 25 (10): 2039–50. https://doi.org/10.1038/sj.emboj.7601103.

Quiram, P. A., K. Ohno, M. Milone, M. C. Patterson, N. J. Pruitt, J. M. Brengman, S. M. Sine, and A. G. Engel. 1999. 'Mutation Causing Congenital Myasthenia Reveals Acetylcholine Receptor Beta/Delta Subunit Interaction Essential for Assembly'*. The Journal of Clinical Investigation* 104 (10): 1403–10. https://doi.org/10.1172/JCI8179.

Radhakrishnan, Karthikeyan, Jennifer Baltes, John W. M. Creemers, and Peter Schu. 2013. 'Trans-Golgi Network Morphology and Sorting Is Regulated by Prolyl-Oligopeptidase-like Protein PREPL and the AP-1 Complex Subunit M1A'. *Journal of Cell Science* 126 (Pt 5): 1155–63. https://doi.org/10.1242/jcs.116079.

Régal, Luc, Xin-Ming Shen, Duygu Selcen, Chantal Verhille, Sandra Meulemans, John W. M. Creemers, and Andrew G. Engel. 2014. 'PREPL Deficiency with or without Cystinuria Causes a Novel Myasthenic Syndrome'. *Neurology* 82 (14): 1254–60. https://doi.org/10.1212/WNL.0000000000000295.

Rogers, Robert S., and Hiroshi Nishimune. 2017. 'The Role of Laminins in the Organization and Function of Neuromuscular Junctions'. *Matrix Biology: Journal of the International Society for Matrix Biology* 57–58: 86–105. https://doi.org/10.1016/j.matbio.2016.08.008.

Salpietro, Vincenzo, Weichun Lin, Andrea Delle Vedove, Markus Storbeck, Yun Liu, Stephanie Efthymiou, Andreea Manole, et al. 2017. 'Homozygous Mutations in VAMP1 Cause a Presynaptic Congenital Myasthenic Syndrome'. *Annals of Neurology* 81 (4): 597–603. https://doi.org/10.1002/ana.24905.

Selcen, D., V. C. Juel, L. D. Hobson-Webb, E. C. Smith, D. E. Stickler, A. V. Bite, K. Ohno, and A. G. Engel. 2011. 'Myasthenic Syndrome Caused by Plectinopathy'. *Neurology* 76 (4): 327–36. https://doi.org/10.1212/WNL.0b013e31820882bd.

Senderek, Jan, Juliane S. Müller, Marina Dusl, Tim M. Strom, Velina Guergueltcheva, Irmgard Diepolder, Steven H. Laval, et al. 2011. 'Hexosamine Biosynthetic Pathway Mutations Cause Neuromuscular Transmission Defect'. *American Journal of Human Genetics* 88 (2): 162–72. https://doi.org/10.1016/j.ajhg.2011.01.008.

Shen, Xin-Ming, Rosana H. Scola, Paulo J. Lorenzoni, Cláudia S. K. Kay, Lineu C. Werneck, Joan Brengman, Duygu Selcen, and Andrew G. Engel. 2017. 'Novel Synaptobrevin-1 Mutation Causes Fatal Congenital Myasthenic Syndrome'. *Annals of Clinical and Translational Neurology* 4 (2): 130–38. https://doi.org/10.1002/acn3.387.

Shen, Xin-Ming, Duygu Selcen, Joan Brengman, and Andrew G. Engel. 2014. 'Mutant SNAP25B Causes Myasthenia, Cortical Hyperexcitability, Ataxia, and Intellectual Disability'. *Neurology* 83 (24): 2247–55. https://doi.org/10.1212/WNL.0000000000001079.

Shirataki, H., K. Kaibuchi, T. Sakoda, S. Kishida, T. Yamaguchi, K. Wada, M. Miyazaki, and Y. Takai. 1993. 'Rabphilin-3A, a Putative Target Protein for Smg P25A/Rab3A P25 Small GTP-Binding Protein Related to Synaptotagmin'. *Molecular and Cellular Biology* 13 (4): 2061–68.

Sørensen, Jakob B., Gábor Nagy, Frederique Varoqueaux, Ralf B. Nehring, Nils Brose, Michael C. Wilson, and Erwin Neher. 2003. 'Differential Control of the Releasable Vesicle Pools by SNAP-25 Splice Variants and SNAP-23'. *Cell* 114 (1): 75–86.

Tsujino, Akira, Chantal Maertens, Kinji Ohno, Xin-Ming Shen, Taku Fukuda, C. Michael Harper, Stephen C. Cannon, and Andrew G. Engel. 2003. 'Myasthenic Syndrome Caused by Mutation of the SCN4A Sodium Channel'. *Proceedings of the National Academy of Sciences of the United States of America* 100 (12): 7377–82. https://doi.org/10.1073/pnas.1230273100.

Vázquez, Miguel, Rubén Nogales, Pedro Carmona, Alberto Pascual, and Juan Pavón. 2010. 'Rbbt: A Framework for Fast Bioinformatics Development with Ruby'. In *Advances in Bioinformatics*, edited by Miguel P. Rocha, Florentino Fernández Riverola, Hagit Shatkay, and Juan Manuel Corchado, 74:201–8. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13214-8_26.

Whittaker, Roger G., David N. Herrmann, Boglarka Bansagi, Bashar Awwad Shiekh Hasan, Robert Muni Lofra, Eric L. Logigian, Janet E. Sowden, et al. 2015. 'Electrophysiologic Features of

SYT2 Mutations Causing a Treatable Neuromuscular Syndrome'. *Neurology* 85 (22): 1964–71. https://doi.org/10.1212/WNL.0000000000002185.

Wu, Fenfen, Wentao Mi, Yu Fu, Arie Struyk, and Stephen C. Cannon. 2016. 'Mice with an NaV1.4 Sodium Channel Null Allele Have Latent Myasthenia, without Susceptibility to Periodic Paralysis'. *Brain: A Journal of Neurology* 139 (Pt 6): 1688–99. https://doi.org/10.1093/brain/aww070.

Zaharieva, Irina T., Michael G. Thor, Emily C. Oates, Clara van Karnebeek, Glenda Hendson, Eveline Blom, Nanna Witting, et al. 2016. 'Loss-of-Function Mutations in SCN4A Cause Severe Foetal Hypokinesia or "classical" Congenital Myopathy'. *Brain: A Journal of Neurology* 139 (Pt 3): 674–91. https://doi.org/10.1093/brain/awv352.

Chapter 4

The multilayer community structure of medulloblastoma

Publication Record

This chapter introduces the original research article '*The multilayer community structure of medulloblastoma*' published in the CellPress journal iScience (2021 Journal Impact factor: 6.107; Q1 in the Multidisciplinary Sciences field. Rank: 15/74). (2022 Journal Impact factor: 5.8; Q1 in the Multidisciplinary Sciences field. Rank: 15/73).

Co-authors & affiliations

Iker Nuñez-Carpintero[1], Marianyela Petrizzelli[2,3,4], Andrei Zinovyev[2,3,4,5], Davide Cirillo[1,*], and Alfonso Valencia[1,6]

1. Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain

2. Institut Curie, PSL Research University, 75005 Paris, France

3. INSERM, U900, 75005 Paris, France

4. MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006 Paris, France

5. Lobachevsky University, 603000 Nizhny Novgorod, Russia

6. ICREA - Institució Catalana de Recerca i Estudis Avanc¸ ats, Pg. Lluís Companys 23, 08010, Barcelona, Spain

* Corresponding author: davide.cirillo@bsc.es

Reference

Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. *iScience*. 2021 Apr 23; 24(4). Available from: https://www.cell.com/iscience/abstract/S2589-0042(21)00333-3

Contribution of the PhD Candidate

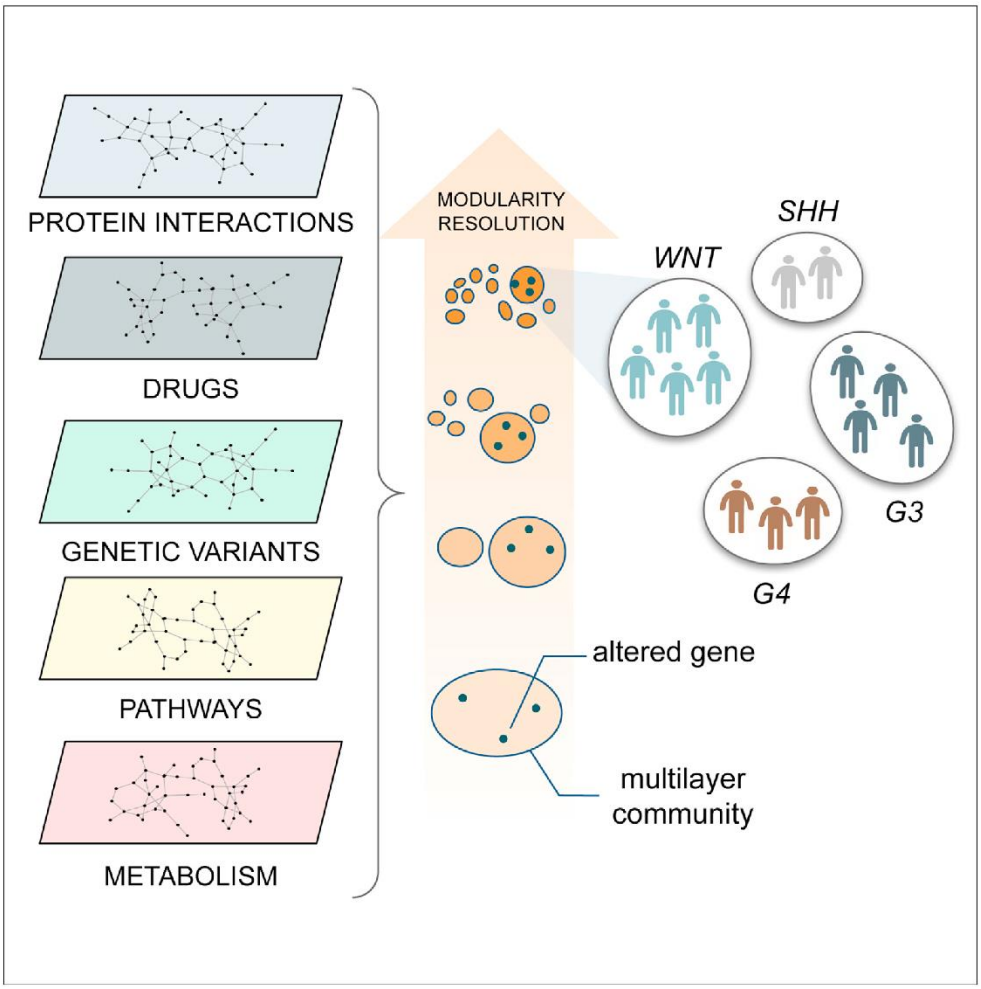
As main author, the PhD candidate developed the presented pipeline, tested its functionality in the introduced cohorts, and designed and implemented the different sensitivity analysis, additionally performing the network enrichment analysis (‘**Author contributions**’, page 169).

iScience

CellPress
OPEN ACCESS

Article

The multilayer community structure of medulloblastoma



Iker Núñez-Carpintero, Marianyela Petrizzelli, Andrei Zinovyev, Davide Cirillo, Alfonso Valencia

davide.cirillo@bsc.es

Highlights

The molecular interpretation of rare diseases is a challenging task

Multilayer networks allow patient stratification and explainability

We identify subgroup-specific genes and multilayer associations in medulloblastoma

Multilayer community analysis enables the molecular interpretation of rare diseases

Núñez-Carpintero et al., iScience 24, 102365 April 23, 2021
https://doi.org/10.1016/j.isci.2021.102365

iScience

CellPress
OPEN ACCESS

Article

The multilayer community structure of medulloblastoma

Iker Núñez-Carpintero,[1] Marianyela Petrizzelli,[2,3,4] Andrei Zinovyev,[2,3,4,5] Davide Cirillo,[1,7,*] and Alfonso Valencia[1,6]

SUMMARY

Multilayer networks allow interpreting the molecular basis of diseases, which is particularly challenging in rare diseases where the number of cases is small compared with the size of the associated multi-omics datasets. In this work, we develop a dimensionality reduction methodology to identify the minimal set of genes that characterize disease subgroups based on their persistent association in multilayer network communities. We use this approach to the study of medulloblastoma, a childhood brain tumor, using proteogenomic data. Our approach is able to recapitulate known medulloblastoma subgroups (accuracy >94%) and provide a clear characterization of gene associations, with the downstream implications for diagnosis and therapeutic interventions. We verified the general applicability of our method on an independent medulloblastoma dataset (accuracy >98%). This approach opens the door to a new generation of multilayer network-based methods able to overcome the specific dimensionality limitations of rare disease datasets.

INTRODUCTION

To improve our understanding of complex systems, it is crucial to take into account the multiple types of relationships that inherently define natural systems. The study of the so-called multilayer networks (alternatively multiplex networks) has recently become one of the most important directions in network science (Kivela et al., 2014; Aleta and Moreno 2019). A multilayer network is a network organized into multiple layers representing different types of nodes and edges (Figure S1). Despite offering the means to achieve a comprehensive view of human diseases by accounting for the complexity of accumulated biomedical data, biological multilayer networks exhibit a range of research challenges that still require substantial investigation (Kristensen et al., 2014). Among them, community detection in multilayer networks is an area of investigation that is particularly promising for biomedicine, facilitating the evaluation of relevant associations among genes and the identification of candidate targets for drug development and repurposing (Halu et al., 2019; Valdeolivas et al., 2019).

Popular strategies for community detection in networks include the Louvain algorithm (Blondel et al., 2008), a greedy optimization technique, to maximize a network structural metric that is called modularity (Newman and Girvan 2004). Modularity is defined as the fraction of edges within a group of nodes that is significantly enriched when compared with a random model. It measures the strength of a given partition of the network (Reichardt and Bornholdt 2006). The Louvain algorithm is one of the most widely used meta-heuristics for community detection in large networks. It outperforms other community detection algorithms in accuracy, scalability, and computing time (Yang et al., 2016). Moreover, the algorithm is implemented in a number of network analysis software, and it has been recently adapted to multilayer networks (Didier et al. 2015; Didier et al. 2018).

Nevertheless, community structure determination in networks remains an open problem to such an extent that the preferred formulation of communities is often domain specific (Porter et al. 2009). One major conundrum of modularity-based approaches to community detection is the intrinsic limit of resolution, by which it is *a priori* impossible to rule out that a community defined at a certain level of resolution may be composed of a cluster of smaller communities (Fortunato and Barthélemy 2007; Lancichinetti and Fortunato 2011). In other words, multiple topological descriptions, each one with its own importance, coexist at different scales that are detected at alternative values of resolution (Arenas et al. 2008). As a

[1]Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain

[2]Institut Curie, PSL Research University, 75005 Paris, France

[3]INSERM, U900, 75005 Paris, France

[4]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006 Paris, France

[5]Lobachevsky University, 603000 Nizhny Novgorod, Russia

[6]ICREA - Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010, Barcelona, Spain

[7]Lead contact

*Correspondence: davide.cirillo@bsc.es

https://doi.org/10.1016/j.isci.2021.102365

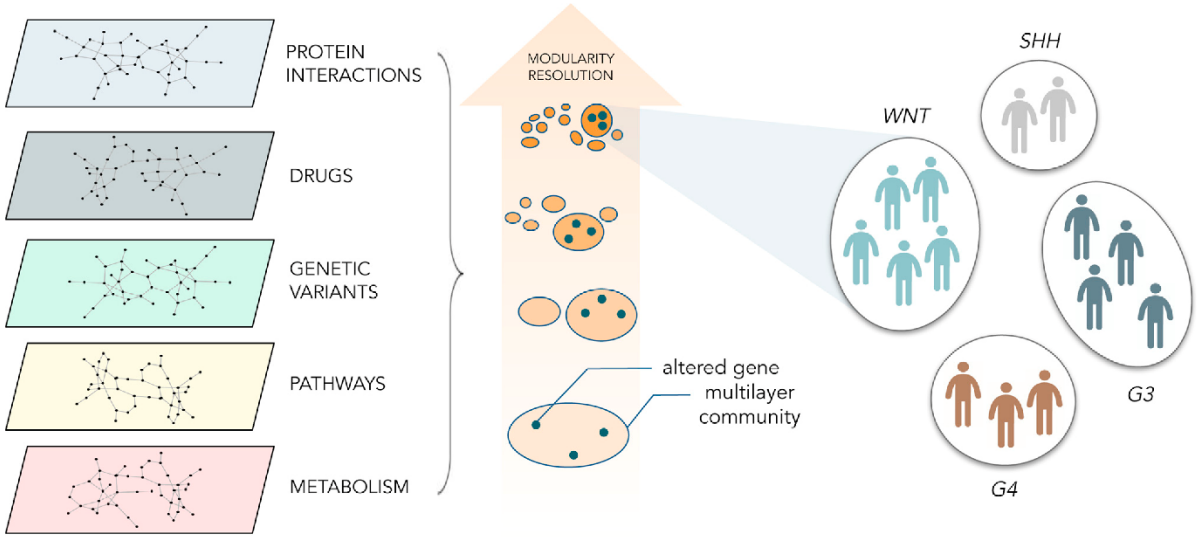


Figure 1. Multilayer community structure analysis of medulloblastoma subgroups

Using multilayer community structure analysis on a network describing gene-gene associations based on protein interactions, drug targets, genetic variants, pathways, and metabolic reactions, we identified the minimum sets of altered genes that optimally cluster the patients with medulloblastoma into previously described subgroups. See also Figures S1 and S2.

consequence, the identification of meaningful network communities, such as groups of genes of interest that robustly express strong associations, heavily depends on the choice of the resolution value to be used. This limitation can be overcome through the identification of stable partitions at different values of resolution. Indeed, the detection of persistent partitions when changing the resolution is indicative of strong modular structures (Arenas et al. 2008).

Here we implemented a methodology to identify groups of genes that are systematically found to belong to the same communities across a range of different resolution values. In this view, two or more genes of interest that are consistently found in the same communities at different values of resolution will be deemed strongly associated based on the multiple biological evidence from the multilayer network. We applied this concept to the analysis of the multilayer community structure of genes altered in a cohort of patients with medulloblastoma (MB) who were previously stratified based on proteogenomic data (Forget et al., 2018) (Figure 1). To this aim, we implemented a dimensionality reduction methodology based on the persistent association of genes in the multilayer network communities (see methods: "multilayer community structure analysis" and Figure 2).

MB is a malignant and fast-growing primary central nervous system tumor, which originates from embryonic cells of the brain or spinal cord with no known causes and a preferential manifestation in children (aged 1–9 years). Despite being rare, MB is the most common cancerous brain tumor in children. Four molecular disease subgroups of pediatric MB with distinct clinicopathological features have been identified: WNT, SHH, Group 3 (G3), and Group 4 (G4) (Taylor et al., 2012; Northcott et al., 2011). WNT is associated with the most favorable prognosis, whereas SHH and G4 are associated with intermediate-level prognosis and G3 with the worst outcome. Seven genes exhibit recurrent genetic alterations in the four subgroups (*SHH* in SHH group, *CTNNB1* in WNT group, *MYC* and MYCN in G3 and G4, *ERBB4*, *SRC*, and *CDK6* in G4 (Kool et al., 2012; Ramaswamy et al., 2016; Taylor et al., 2012; Northcott et al., 2014; Robinson et al., 2012; Northcott et al., 2017; Kool et al., 2014; Clifford et al., 2006; Forget et al., 2018). Each subgroup presents substantial biological heterogeneity and survival differences (Jones et al., 2012) so much so that the identification of more than four subgroups has been recently proposed, in particular as concerns the heterogeneity of G3 and G4 (Schwalbe et al., 2017).

Our results show that our multilayer community structure analysis is able to recapitulate the four MB subgroups (accuracy 94.94%), as well as better characterize them by identifying distinct minimal sets of genes

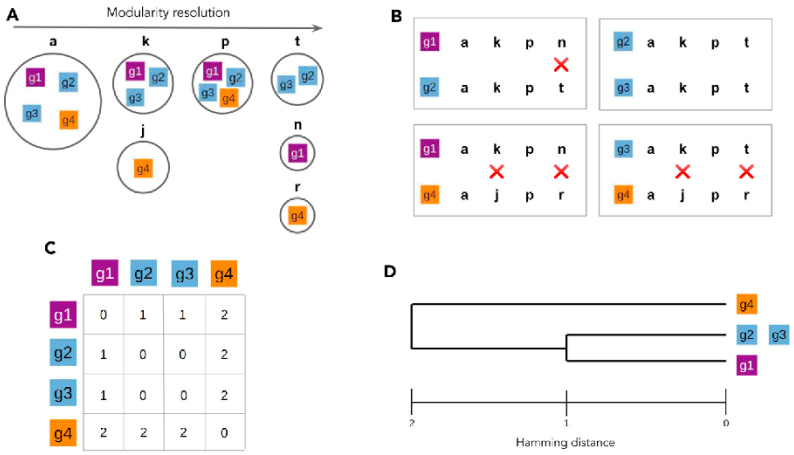


Figure 2. Identification of multilayer community trajectories

(A–D) For a given set of genes, we identified the multilayer communities to which they belong in a range of modularity resolution (A). We then computed the pairwise Hamming distances of the trajectories of communities visited by each gene (B). The corresponding distance matrix (C) was represented in the form of a dendrogram (D) used for clustering analysis. See also Figure S3.

with strong associations based on multiple layers of evidence (Figure S2). We further verify the applicability of our method using an independent MB multi-omics dataset, achieving a very high performance also in this case (accuracy 98.29%). This work represents an important step forward not only in the characterization of MB subgroups but also, in general, in rare tumor research, where the absence of large patient sample cohorts makes the identification of supporting evidence for candidate genes an extremely challenging task.

RESULTS

Multilayer community trajectories

To implement a way to monitor the behavior of multilayer communities containing MB genes upon changes of the modularity resolution, we initially sought to take into account gene mentions in abstracts of scientific publications about MB (see methods: "data sources of medulloblastoma genes"). By interrogating PubTator Central (PTC) (Wei et al., 2019), we retrieved a list of 1,941 multi-species genes, consisting of 1,475 human genes (76%), 389 murine genes (20%), and 77 genes of other species (4%). We identified the multilayer communities to which the human genes (1,387 out of 1,475, represented in the multilayer network) belong in a range of modularity resolution (see methods: "multilayer community structure analysis" and Figure S3). We conceived this particular analysis as a proof of concept for the multilayer community structure analysis.

As shown in Figure 3, there are plain differences in the trajectories of the communities that are visited by each gene. Interestingly, the trajectories of seven genes, whose recurrent genetic alterations are well-known hallmark features of the four molecular disease subgroups (see introduction), branch off from well-separated communities, with the exception of *SRC* and *CTNNB1*, which are physical interactors (IntAct: EBI-15951997).

The landscape of these multilayer community trajectories can be further explored to investigate the so-called operations on dynamic communities (Cazabet et al. 2017), such as birth (a new community appears), death (a community vanishes), and resurgence (a community disappears and appears again later on). Along the explored range of modularity resolution, the 2,186 unique multilayer communities of the text-mined MB genes experience a total of 2,517 death events and 673 resurgence events (Figure S4), indicating not only a high level of instability (all communities disappear at least once) but also a high level of commutability (some communities reappear several times with the same exact composition). These observations led us to realize that each gene is characterized by its own journey throughout the communities found at different levels of resolution. For this reason, we further tested the hypothesis that tracing such trajectories for a set

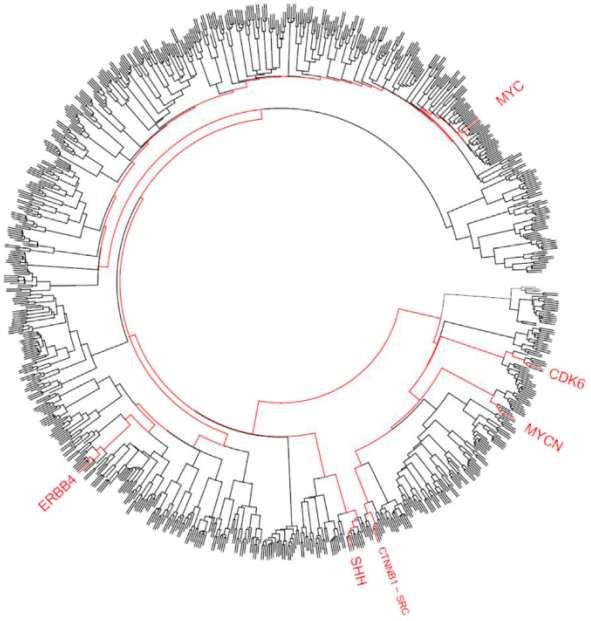


Figure 3. Dendrogram of multilayer community trajectories

The dendrogram represents the Hamming distance among the trajectories of the communities visited by each gene associated to medulloblastoma by text mining in a range of modularity resolution (see methods: "multilayer community structure analysis"). Trajectories of seven genes that are known to characterize medulloblastoma subgroups (see introduction) are highlighted in red. See also Figure S4.

of disease-related genes could be exploited for patient clustering purposes (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups").

Medulloblastoma patient stratification through multilayer structure analysis

We sought to use the trajectories of the multilayer communities visited by the genes altered in MB to achieve patient stratification. Our reference (ground truth) consists of the four classical subgroups (WNT, SHH, G3, G4), which represent a standard categorization of MB despite substantial heterogeneity and the possibility of a more granular stratification have been reported (see introduction). The four subgroups have been recently investigated via network fusion using a cohort of patients with proteogenomic information (Forget et al., 2018). We reanalyzed this cohort to optimally recapitulate the four subgroups, while aiming to reduce the number of critical genes required for this stratification.

We retrieved lists of genes altered in 35 patients who display complete datasets (DNA methylation, RNA sequencing, proteomics, and phosphoproteomics) (see methods: "data sources of medulloblastoma genes"). Partial datasets are available for three additional patients (MB10, MB21, and MB33) that we retained as a validation set (see results: "sensitivity analyses"). We performed a hierarchical clustering based on the multilayer community trajectories of an optimal selection of minimal sets of genes. Optimality means that the features of these selected genes, in terms of their representation in the multilayer communities (parameter λ) and the similarity of their trajectories (parameter θ), allow clustering patients with the maximum accuracy and Matthews correlation coefficient (MCC) to the four subgroups of reference (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups").

We achieved the highest accuracy (94.94%) and MCC (87%) with five clusters (WNT, SHH, G4, G3, and G3-G4), by selecting for each patient those genes that are represented in the communities in sets of at most 6 ($\lambda = 6$) and that are always part of the same communities along their trajectories ($\theta = 0$) (Figures 4 and 5, Tables S1–S3). Strikingly, such high accuracy corresponds to a strict selection of genes, indicating that only a small portion of the genes altered in a patient is sufficient to accomplish an accurate patient

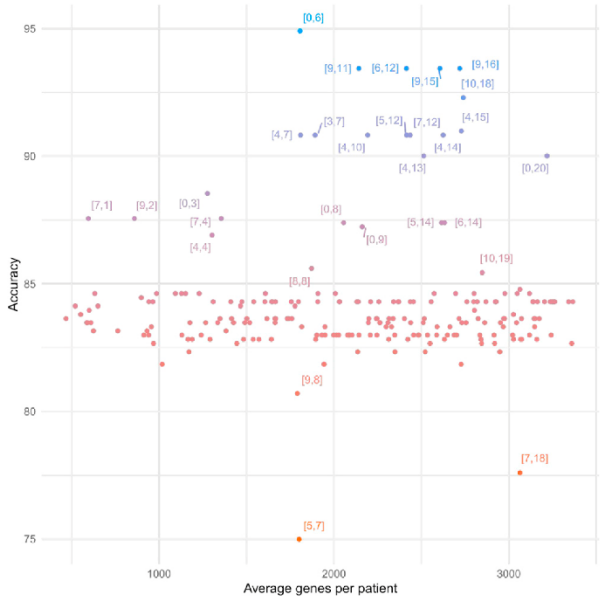


Figure 4. Parameters optimization

Scatterplot comparing the average genes per patient obtained by each iteration of the optimization procedure (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups") and its corresponding accuracy. Values next to each point highlight the corresponding [θ,λ] combination. See also Figures S5–S7 and Tables S1–S3.

segregation. This observation implies that the selected genes are tightly associated and never leave the communities they belong to along their trajectories. An important aspect of this result is that, despite our reference being of four subgroups, we identified five clusters, indicating that only few patients escape the classical categorization and subtler stratas may exist, as suggested in recent studies (Schwalbe et al., 2017; Archer et al., 2018).

Classification of patients with partial molecular information

As the datasets of three patients consist of partial molecular information (see methods: "data sources of medulloblastoma genes"), we excluded these samples from the parameter optimization procedure and used them as a validation set. The three patients belong to subgroups G4 (patient MB10) and WNT (patients MB21 and MB33) (Forget et al., 2018). We assigned each one of the three patients to the cluster of the most similar among the remaining 35 patients based on the Jaccard Index (J) parametrized by the optimal θ and λ (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups"). Patient MB10 shows the highest similarity to patient MB22 (J = 0.263), who belongs to G4 subgroup likewise eight patients in the following ranking positions (Table S4). Patient MB21 shows the highest similarity to three patients of the WNT subgroup (MB31 J = 0.2653; MB34 J = 0.2631; MB30 J = 0.2601). Finally, patient MB33 shows high similarity to two patients of WNT subgroup (MB30 J = 0.2168; MB34 J = 0.2106). Of note, patient MB31 of the WNT subgroup is the fourth most similar patient to MB33 (J = 0.2080), MB16 of the G4 subgroup being the third (J = 0.2081). These results show that the parameters for gene selection optimized based on patients with complete molecular information allow classifying the patients who have only partial molecular information with high accuracy (all three patients are correctly classified).

Robustness analyses

The identified values of θ and λ, optimized on 35 patients, correspond to an average of 1,812.74 genes per patient (SD = 106.97) (i.e., an average dimensionality reduction of 87.56% (SD = 0.44) per patient) (Table S5). Moreover, some of these genes are uniquely found among all patients of distinct clusters (148 genes in G3

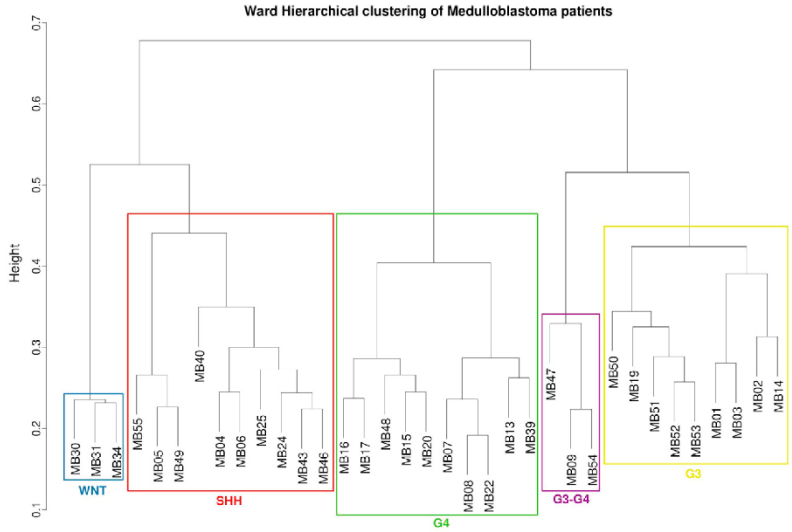


Figure 5. Clustering of medulloblastoma patients

Ward's linkage hierarchical clustering obtained at $\lambda = 6$ and $\theta = 0$. The rectangles indicate the five clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the four medulloblastoma subgroups (Forget et al., 2018): WNT (blue), SHH (red), Group 4 (G4, green), Group 3 (G3, yellow). A fifth cluster is depicted in purple, including three patients originally assigned to subgroups G3 (MB47) and G4 (MB09 and MB54). See also Figures S8 and S9 and Tables S4, S5, and S6.

patients; 83 genes in SHH patients; 115 genes in G4 patients; 46 genes in G3-G4 patients; 260 genes in WNT patients).

We evaluated the robustness of our results with two types of robustness analyses. In the first analysis, we shuffled the altered genes across the cohort 10,000 times, maintaining the same number of genes for each patient as in the original data. This procedure yielded an average accuracy of 54.76% (SD = 0.11) with $\theta = 0$ and $\lambda = 6$ (Figure S5). The distribution of the average optimization accuracies of the randomized sets shows dramatically lower values than those of the original data, indicating that our optimization procedure, when based on a meaningful clinical stratification, is able to identify non-random and very specific gene-subgroup associations (Figure S6).

In the second analysis, we recursively performed the optimization procedure after excluding the identified minimal set of genes at each iteration. We observed a progressive decrease in accuracy and, as expected, higher values of optimal θ and λ in later iterations, indicating less effective gene selection and dimensionality reduction (Figure S7). Overall, we observed that this decay in accuracy upon iterative removal of selected genes can be divided into three phases: a short initial phase (accuracies between 94.94 and 88.57) in which large sets of genes are removed at each iteration (1027.72 on average), a long intermediate phase (accuracies between 79.76 and 69.96) in which less genes are removed (23.31 on average), and a short final phase (between 57.06 and 31.43) in which an average of 1.08 gene is removed at each iteration before the accuracy drops to 0. At the end of this procedure, the cumulative number of removed genes is 5,950.63 (average per patient; 38 patients). These results show the effectiveness of the greedy nature of our optimization algorithm, which is able to achieve high accuracies even when the pool of genes it operates upon is largely reduced.

Sensitivity analyses

To test if our clusters are a good representation of the similarities among patients, we performed a sensitivity analysis with two approaches for clustering significance assessment. The first, based on multiscale bootstrap resampling (Suzuki and Shimodaira 2006), assigns a confidence value, known as approximately unbiased probability value (pvAU), to each cluster. High pvAU indicates high confidence in the clusters. The second, based on a Monte Carlo procedure (Kimes et al., 2017), assigns an empirical p value and a Gaussian

approximate p value to each cluster. An important difference between the two approaches is that the multi-scale bootstrap resampling approach tends to be less conservative than the Monte Carlo-based procedure, which outperformed the first with simulated and real-world data (Kimes et al., 2017).

At the root node, WNT and SHH subgroups are significantly separated from G3 and G4 subgroups with empirical p value of 1.08e-02 (Gaussian approximate p value of 2.59e-03) (Figures S8 and S9). Such two large partitions are poorly supported by the data (pvAU 49.23% and 63.05%, respectively), indicating the possibility of a finer subdivision. Indeed, WNT subgroup significantly separates from SSH subgroup with empirical p value of 5.45e-02 (Gaussian approximate p value of 3.55e-02), whereas G4 subgroup significantly separates from G3 subgroup with empirical p value of 1.85e-02 (Gaussian approximate p value of 6.75e-03).

Unlike the three main subgroups WNT (pvAU 100%), G4 (pvAU 99.97%), and G3 (pvAU 92.79%), SHH appears to be poorly supported by the data as a unique cluster (pvAU = 38.41%), whereas two SHH sub-clusters might exist (pvAU 99.88% and pvAU 99.55%, respectively), although their separation is not statistically significant (empirical p value 1.02e-01; Gaussian approximate p-value 9.74e-02). Of note, a finer partition of SHH subgroup into multiple sub-clusters has been reported by recent studies (Schwalbe et al., 2017; Archer et al., 2018).

The fifth cluster (G3-G4), despite being composed of two patients previously described as G4 (MB09 and MB54), and one as G3 (MB47), is supported by the data (pvAU 83.98%), but its separation from the G3 subgroup is not statistically significant (empirical p value 1.01e-01; Gaussian approximate p value 9.18e-02). Interestingly, patients of this cluster were all assigned to G4 via network fusion and to G3 only using methylation data (Forget et al., 2018). Indeed, an overlap of genetic features between G3 and G4 has also been reported by a study on risk stratification (Schwalbe et al., 2017).

Overall, these sensitivity analyses indicate that (1) 4 of 5 clusters found in our optimization procedure are statistically significant based on a Monte Carlo approach (Kimes et al., 2017) and recapitulate the classical MB molecular subgroups and (2) the small fifth cluster (G3-G4) shares similarities with G3 whose heterogeneity was previously observed (Schwalbe et al., 2017; Forget et al., 2018).

Provenance analysis of the identified gene communities

By performing a network enrichment analysis test (Signorelli et al. 2016), we identified the most significantly overrepresented intra-layer edges among the genes of the minimal sets identified for each patient in each cluster (see methods: "multilayer network enrichment analysis"). In the following, we analyze those associations that are unique of the five clusters and enriched in all patients of each cluster (Table S6). Beside this strict requirement, several other enriched associations are shared among clusters and can be further explored (see resource availability: "data and code availability"). Overall, we found that the minimal set of genes found in all patients of WNT, SHH, and G4 clusters are uniquely enriched in very specific associations in each layer, whereas G3-G4 and G3 clusters tend to display less specific enrichments (i.e., either several or no enriched associations). This reduction of enrichment specificity from WNT to G3 suggests an interesting parallel with the prognosis spectrum of the four classical subgroups, from best (WNT) to worst (G3) outcomes.

- **Molecular associations.** As for the molecular interaction layer, WNT cluster presents enrichment in four proteins: ACVR2A, a receptor involved in the activin signaling pathway (Chen et al., 2006), which is also enriched in this cluster in the pathways layer; ATP4A, a subunit of the ATPase H^+/K^+, a membrane transporter that is target of the Hedgehog signaling pathway, whose low levels of β1 subunit have been related to cell proliferation in MB models (Lee et al., 2015); POU2F2, which has been recently found to play a role in spinal cord development in a mouse model (Masgutova et al., 2019) and suspected to be regulated by miRNAs in MB (Venkataraman et al., 2013); and RBM48, a protein found to be amplified across several cancer tissues and cell lines and that may have a role in apoptotic processes (Hart et al., 2015). SHH cluster is uniquely enriched in molecular interactions of various gene products, including two proto-oncogenes (*ETS1* [Cao et al., 2015] and *JUND* [Elliott et al., 2019]), a calcium voltage-gated channel (CACNA1A) significantly downregulated in MB and other brain tumors (Phan et al., 2017), and interestingly a long noncoding RNA (*LINC00461*), expressed predominantly in the brain and involved in tumorigenesis (Yang et al., 2017). G4 cluster

only presents enrichment in the interactions of ARID4A, a member of the ARID family such as ARID1B, a repressor of Wnt/β-catenin signaling (Vasileiou et al., 2015). G3 cluster is enriched in interactions of the ABC transporter, ABCA3, suspected to be involved in chemoresistance in brain tumor progression (Hadjipanayis and Van Meir 2009); the dystrophin-glycoprotein SGCB; the SUMO ligase PIAS1, which increases the activity of Gli proteins on the Hedgehog pathway (Niewiadomski et al., 2019); and the heat shock protein DNAJB5, which regulates histone deacetylase (HDAC) nuclear shuttling, whose inhibition is considered to be a promising therapy in MB (Becher 2019).

- **Drug-target associations.** As for the drug layer, G3 cluster is the only one showing a unique enrichment in all patients, namely, in lubeluzole, an inhibitor of nitric oxide (NO) synthesis (Maiese et al. 1997). This observation points toward the role of oxidative stress in MB under the light of results from NO synthesis inhibition in experimental models (Haag et al., 2012) and clinical trials in G3 subgroup (Bakhshinyan et al., 2019).
- **Variant-disease associations.** As for the disease layer, the enriched associations may indicate overlapping features between MB and molecular processes underlying other pathologies. WNT cluster is uniquely associated with alveolar rhabdomyosarcoma, a common soft tissue sarcoma in children (Barr 2011), and familial prostate carcinoma. The implication of the overactivation of the Hedgehog signaling pathway in both MB and rhabdomyosarcoma (Azatyan et al., 2019) as well as in prostate cancer (Amakye et al. 2013; Ng and Curran 2011) has been extensively reported. SHH cluster is uniquely associated with macular degeneration and syndromic craniosynostosis, also characterized by ocular abnormalities, suggesting a link with the ophthalmic complications of MB, which occur as a result of the disease and its treatments (Cassidy et al., 2000). Polydactyly (Crane et al., 2018) appears to be uniquely enriched in G4 cluster, MB being a feature of several disorders of infants often characterized by akin skeletal abnormality such as Gorlin syndrome (Lo Muzio 2008) and others (Osterling et al., 2011). Cluster G3-G4 shows many enriched diseases, including several cancers, forms of hypogonadism, and interestingly Dravet syndrome, a genetic disorder that causes severe epilepsy in infants. Of note, MB is among the most frequent tumors of cerebellum presenting with seizures (5%) (Sánchez Fernández and Loddenkemper, 2017). G3 does not display unique enrichments in the disease layer.
- **Pathway associations.** As for the pathway layer, the WNT cluster is uniquely enriched in cell differentiation in early embryogenesis, such as Nodal (Brown et al., 2011) and Activin (Chen et al., 2006) signaling; immune response, such as Dendritic cell-associated C-type lectin-2 (Dectin-2) carbohydrates receptor activity (Graham and Brown 2009); protein metabolism, such as insulin-like growth factor (IGF) regulation (Holly and Perks 2006); and defects in the mismatch repair (MMR) system (Chao and Lipkin 2006). SHH cluster is uniquely enriched in potassium channels of the neuronal system, such as the Kir channel (Radeke et al. 1999), and signal transduction, such as calcitonin (Sexton et al. 1999) and Hedgehog (Briscoe and Thérond 2013) signaling. G4 is associated with fusion events in the *FGFR1* gene (Braun and Shannon 2004) and neuronal system transmission, such as excitatory synaptic transmission by glutamate receptors (Kessels and Malinow 2009). The great majority of these pathways have been directly or indirectly related to MB in the literature, such as the interplay between the embryonic morphogens Nodal and Hedgehog in brain development (Rohr et al., 2001), the activation of Activin signaling in a subset of G3 subgroup (Morabito et al., 2019), the role of *FGFR1* in gliomas (Egbivwie et al., 2019), and the importance of carbohydrate antigen recognition in MB (Read et al., 2009). Clusters G3-G4 and G3 show a varied landscape of enriched pathways.
- **Metabolic reaction associations.** As for the metabolome layer, uniquely enriched metabolites of the WNT cluster are ferricytochrome C (part of mitochondrial respiratory electron transport chain), nicotinamide nucleotide (a derivative of niacin, a form of vitamin B3), superoxide anion (a reactive oxygen species), and ribose 5-phosphate (a precursor to many biomolecules, including DNA and RNA). SHH shows unique enrichments for nicotinate D-ribonucleotide (part of cofactor biosynthesis) and pantothenate (vitamin B5), whereas G4 is uniquely enriched in sulfate (the major sulfur source in humans). Cluster G3 does not present uniquely enriched metabolites, whereas G3-G4 shows several.

Method verification on an independent cohort

To further verify the applicability of our methodology, we performed the same analytical procedure on an independent, non-overlapping, multi-omics MB cohort (Archer et al., 2018) (see methods: "data sources of medulloblastoma genes"). This cohort study collects proteogenomics data from 45 patients and proposes

a finer categorization of SHH and G3 subgroups (SHHa, SHHb, G3a, G3b). A total of 39 patients display complete multi-omics information, whereas 6 lack RNA sequencing, including all 3 patients of the WNT subgroup.

In a first analysis, we were able to recapitulate the 5 clusters (SHHa, SHHb, G3a, G3b, G4) of the 39 patients with complete multi-omics information, achieving the highest accuracy (98.29%, MCC = 0.95) with optimized parameters $\lambda = 3$ and $\theta = 0$ (Figure S10), which corresponds to an average of 842.2 genes per patient (SD = 145.12) and average dimensionality reduction of 92.83% (SD = 0.578). All patients are correctly assigned to their subgroups, whereas only MB136, labeled as a SHHb member, clusters with the SHHa subgroup.

As for the previous analysis, the patients with incomplete multi-omics information were used as validation set and assigned individually to subgroups based on the Jaccard Index (J) (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups"). Patients MB037, MB018, and MB282 are correctly classified as SHHa, G3a, and G4, the most similar patients being MB239 (J = 0.177), MB226 (J = 0.136), and MB091 (J = 0.166), respectively.

In a second analysis, we included all 45 patients achieving the highest accuracy (95.56%, MCC = 0.85) with 7 clusters and $\lambda = 5$ and $\theta = 1$ (Figure S11), which corresponds to an average of 1,073.58 genes per patient (SD = 161.94) and average dimensionality reduction of 90.59% (SD = 1.06). The performance reduction suggests that the addition of patients presenting missing data in the parameters optimization procedure can decrease its performance.

DISCUSSION

Molecular disease subtyping is a fundamental tool to achieve an effective patient stratification for clinical trials and preventive and therapeutic interventions. In some cancers, such as breast cancer and blood cancers, subtyping has been very successful thanks to the statistical power brought by cohorts composed of large numbers of patients. Rare diseases represent a more challenging setting because, by definition, they affect a small number of patients with studies that, in most cases, are in the order of tens of subjects. MB, such as other pediatric cancers, is an illustrative example, two MB subgroups being very well distinguishable (SHH and WNT) and two others being far less characterized (G3 and G4).

In our vision, a meaningful molecular subtyping of rare diseases can be achieved by leveraging the wealth of biomedical information that is available in public knowledge bases and that can be integrated in the form of multilayer networks. In particular, achieving patient stratification by means of structural features (multilayer community trajectories) extracted from a general-purpose multilayer network represents a way to both identify the minimal set of genes that characterize the subgroups and, most importantly, to obtain information about the types of relations that define the associations of such genes (e.g., targeting drugs, pathways, molecular interactions). This way of accomplishing two objectives with one action constitutes the main achievement of our methodology.

In this regard, this work is additionally motivated by the relevance and urgency of implementing computational solutions based on biological multilayer networks. Borrowing from social network science, we use multiplexity as a way to evaluate intimacy of gene associations in MB: the more tightly a group of genes is connected through multiple types of features, the more clearly defined and explainable that community will be (Dickison et al. 2016).

Our results show that we can accurately recapitulate the four established MB subgroups using proteogenomic data and correctly classify the patients with partial molecular profiles. The approach enables an effective dimensionality reduction leading to the identification of a minimal set of altered genes that are sufficient to define MB subgroups. Moreover, the use of a multilayer network in this context allows the retrieval and analysis of the multiple associations among the identified genes, enabling a high level of interpretation of the patient subgroups and the spectrum of prognosis that characterize them, from best (WNT) to worst (G3) outcomes. Analyzing the provenance of the associations that determine the detected communities is extremely beneficial to better characterize the molecular determinants of the patient subgroups and, in turn, achieve a high level of explainability, a matter of considerable debate in computational biology lately (Adadi and Berrada 2020).

An additional important aspect that emerges from our results is that the precise clinical stratification of patients and the completeness of multi-omics information can lead to a better optimization and finer molecular characterization. Indeed, the overall performances of our optimization approach, in terms of both clustering accuracy and dimensionality reduction, are higher using a patient stratification of reference of six subgroups (Archer et al., 2018) compared with the traditional four subgroups (Forget et al., 2018). This indicates that precise clinical hypotheses can lead to precise molecular characterization of patient subgroups, making multilayer networks a powerful and unique tool especially for the study of rare diseases.

Limitations of the study

The main limitations of the study include (1) the scope of the multilayer network, (2) the reliability of the patient stratification of reference, and (3) the suitability of modularity as a quality function for community detection. As for the multilayer network, we distilled high-quality information from reputable and widely used knowledge bases (see methods: "data sources for the construction of the multilayer network"). Our multilayer network encapsulates a comprehensive view of fundamental aspects of human biology, but it can be further expanded to layers with a different content. As for the patient stratification of reference, the categorization of the cohort under study, based on network fusion (Forget et al., 2018), is one of the most recent and highly accurate attempts to cluster patients with MB using multi-omics information. Our analysis can be repurposed for different MB cohorts, available at data sharing platforms such as R2 (http://r2.amc.nl) and Cavatica (www.cavatica.org), among others. As for modularity, it is one of the most well-known quality functions for community detection (Chen et al., 2018). Moreover, the Louvain algorithm has been adapted for multilayer networks (Didier et al. 2018; Didier et al. 2015). Nevertheless, our approach can be applied to other quality functions (e.g., Hamiltonians, partition density) and more recent algorithms, such as the Leiden algorithm (Traag et al. 2019), which, to our knowledge, has currently not been adapted to multilayer networks.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Davide Cirillo (davide.cirillo@bsc.es).

Materials availability

This study did not generate reagents, cell lines, or any biological material.

Data and code availability

The data and code generated during this study is available at dedicated GitHub repositories. The developed CmmD package is available at https://github.com/ikernunezca/CmmD. The code to reproduce all the figures and tables is available at https://github.com/ikernunezca/Medulloblastoma, where the complete lists of network enrichments and the processing of MB gene lists from the cohorts under study are also available. The text mining process is automated in the workflow available at https://github.com/cirillodavide/ipc_textmining. The procedure to generate the multilayer network used in this work is available at https://github.com/cirillodavide/gene_multilayer_network.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102365.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche in the program Investissements d'Avenir (project No. ANR-19-P3IA-0001; PRAIRIE 3IA Institute), the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, "iPC - individualizedPaediatricCure" (ref. 826121), and the Ministry of Science and Higher Education of the Russian Federation (Project No. 14.Y26.31.0022).

The authors would like to thank Anaïs Baudot and Léo Pio-Lopez (Marseille Medical Genetics, Inserm) for advising about multilayer community structure analysis, María Rodríguez Martínez and Matteo Manica (IBM Research, Zurich) for text mining support, and François Serra and Miguel Ponce de León (Barcelona Supercomputing Center) for the insightful discussions.

AUTHOR CONTRIBUTIONS

A.V. and D.C. conceived the study; I.N.C. designed and implemented the computational analyses in consultation with D.C.; M.P. and A.Z. processed the medulloblastoma proteogenomic data. A.V. supervised the project. All the authors contributed to the writing of the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 13, 2020
Revised: March 17, 2021
Accepted: March 24, 2021
Published: April 23, 2021

REFERENCES

Adadi, A., and Berrada, M. (2020). Explainable AI for healthcare: from black box to interpretable models. In Embedded Systems and Artificial Intelligence, *1076*, V. Bhateja, S. Satapathy, and H. Satori, eds (Singapore: Springer), pp. 327–337.

Aleta, A., and Moreno, Y. (2019). Multilayer networks in a nutshell. Annu. Rev. Condens. Matter Phys. https://doi.org/10.1146/annurev-conmatphys-031218-013259.

Amakye, D., Jagani, Z., and Dorsch, M. (2013). Unraveling the therapeutic potential of the hedgehog pathway in cancer. Nat. Med. *19*, 1410–1422.

Archer, T.C., Ehrenberger, T., Mundt, F., Gold, M.P., Krug, K., Mah, C.K., Mahoney, E.L., Daniel, C.J., LeNail, A., Ramamoorthy, D., et al. (2018). Proteomics, post-translational modifications, and integrative analyses reveal molecular heterogeneity within medulloblastoma subgroups. Cancer Cell *34*, 396–410.e8.

Arenas, A., Fernández, A., and Gómez, S. (2008). Analysis of the structure of complex networks at different resolution levels. New J. Phys. https://doi.org/10.1088/1367-2630/10/5/053039.

Azatyan, A., Gallo-Oller, G., Diao, Y., Selivanova, G., Johnsen, J.I., and Zaphiropoulos, P.G. (2019). RITA downregulates hedgehog-GLI in medulloblastoma and rhabdomyosarcoma via JNK-dependent but p53-independent mechanism. Cancer Lett. *442*, 341–350.

Bakhshinyan, D., Adile, A., Venugopal, C., Singh, M., Qazi, M., Kameda-Smith, M., and Singh, S. (2019). MEDU-25. genes preserving stem cell state in group 3 MB BTICs contribute to therapy evasion and relapse. Neuro-Oncology *21*, ii108.

Barr, F.G. (2011). Soft tissue tumors: alveolar rhabdomyosarcoma. Atlas Genet. Cytogenet. Oncol. Haematol. *12*, https://doi.org/10.4267/2042/44650.

Becher, O.J. (2019). HDAC inhibitors to the rescue in sonic hedgehog medulloblastoma. Neuro Oncol. https://doi.org/10.1093/neuonc/noz115.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. https://doi.org/10.1088/1742-5468/2008/10/p10008.

Braun, B.S., and Shannon, K. (2004). The sum is greater than the FGFR1 partner. Cancer Cell *5*, 203–204.

Briscoe, J., and Thérond, P.P. (2013). The mechanisms of hedgehog signalling and its roles in development and disease. Nat. Rev. *14*, 416–429.

Brown, S., Teo, A., Pauklin, S., Hannan, N., Cho, C.H.-H., Lim, B., Vardy, L., Dunn, N.R., Trotter, M., Pedersen, R., et al. (2011). Activin/nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. Stem Cells *29*, 1176–1185.

Cao, P., Fan, F., Dong, G., Yu, C., Feng, S., Song, E., Shi, G., Liang, Y., and Liang, G. (2015). Estrogen receptor α enhances the transcriptional activity of ETS-1 and promotes the proliferation, migration and invasion of neuroblastoma cell in a ligand dependent manner. BMC Cancer *15*, 491.

Cassidy, L., Stirling, R., May, K., Picton, S., and Doran, R. (2000). Ophthalmic complications of childhood medulloblastoma. Med. Pediatr. Oncol. *34*, 43–47.

Cazabet, R., Rossetti, G., and Amblard, F. (2017). Dynamic community detection. In Encyclopedia of Social Network Analysis and Mining, *2*, R. Alhajj and J. Rokne, eds. (Springer), pp. 1–10.

Chao, E.C., and Lipkin, S.M. (2006). Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. Nucleic Acids Res. *34*, 840–852.

Chen, S., Wang, Z.Z., Bao, M.H., Tang, L., Zhou, J., Xiang, J., Li, J.M., and Yi, C.H. (2018). Adaptive multi-resolution modularity for detecting communities in networks. Physica A Stat. Mech. Appl. *491*, 591–603.

Chen, Y.G., Wang, Q., Lin, S.L., Donald Chang, C., Chuang, J., and Ying, S.Y. (2006). Activin signaling and its role in regulation of cell proliferation, apoptosis, and carcinogenesis. Exp. Biol. Med. *231*, 534–544.

Clifford, S.C., Lusher, M.E., Lindsey, J.C., Langdon, J.A., Gilbertson, R.J., Straughton, D., and Ellison, D.W. (2006). Wnt/Wingless Pathway Activation and Chromosome 6 Loss Characterize a Distinct Molecular Sub-Group of Medulloblastomas Associated with a Favorable Prognosis. Cell Cycle *5*, 2666–2670.

Crane, J., Chang, V., Lee, H., Yong, W., Salamon, N., Kianmahd, J., Dorrani, N., Martinez-Agosto, J., and Davidson, T. (2018). PATH-23. germline gnas mutation in an 18-month-old with medulloblastoma. Neuro Oncol. *20*, vi163.

Dickison, M.E., Magnani, M., and Rossi, L. (2016). Multilayer Social Networks (Cambridge University Press).

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. PeerJ *3*, e1525.

Didier, G., Valdeolivas, A., and Baudot, A. (2018). Identifying communities from multiplex biological networks by randomized optimization of modularity. F1000Res. *7*, 1042.

Egbivwie, N., Cockle, J.V., Humphries, M., Ismail, A., Esteves, F., Taylor, C., Karakoula, K., Morton, R., Warr, T., Short, S.C., and Brüning-Richardson, A. (2019). FGFR1 expression and role in migration in low and high grade pediatric gliomas. Front. Oncol. *9*, 103.

Elliott, B., Millena, A.C., Matyunina, L., Zhang, M., Zou, J., Wang, G., Zhang, Q., Bowen, N., Eaton, V., Webb, G., et al. (2019). Essential role of JunD in cell proliferation is mediated via MYC signaling in prostate cancer cells. Cancer Lett. *448*, 155–167.

Forget, A., Martignetti, L., Puget, S., Calzone, L., Brabetz, S., Picard, D., Montagud, A., Liva, S., Sta, A., Dingli, F., et al. (2018). Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. Cancer Cell *34*, 379–395.e7.

Fortunato, S., and Barthélemy, M. (2007). Resolution limit in community detection. Proc. Natl. Acad. Sci. U S A *104*, 36–41.

Graham, L.M., and Brown, G.D. (2009). The dectin-2 family of C-type lectins in immunity and homeostasis. Cytokine *48*, 148–155.

Haag, D., Zipper, P., Westrich, V., Karra, D., Pfleger, K., Toedt, G., Blond, F., Delhomme, N., Hahn, M., Reifenberger, J., et al. (2012). Nos2 inactivation promotes the development of medulloblastoma in Ptch1(+/-) mice by deregulation of gap43-dependent granule cell precursor migration. PLoS Genet. *8*, e1002572.

Hadjipanayis, C.G., and Van Meir, E.G. (2009). Brain cancer propagating cells: biology, genetics and targeted therapies. Trends Mol. Med. *15*, 519–530.

Halu, A., De Domenico, M., Arenas, A., and Sharma, A. (2019). The multiplex network of human diseases. NPJ Syst Biol Appl *5*, 15.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526.

Holly, J., and Perks, C. (2006). The role of insulin-like Growth factor binding proteins. Neuroendocrinology *83*, 154–160.

Jones, D.T.W., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.J., Pugh, T.J., Hovestadt, V., Stütz, A.M., et al. (2012). Dissecting the genomic complexity underlying medulloblastoma. Nature *488*, 100–105.

Kessels, H.W., and Malinow, R. (2009). Synaptic AMPA receptor plasticity and behavior. Neuron *61*, 340–350.

Kimes, P.K., Liu, Y., Hayes, D.N., and Marron, J.S. (2017). Statistical significance for hierarchical clustering. Biometrics *73*, 811–821.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., and Porter, M.A. (2014). Multilayer networks. J. Complex Netw. *2*, 203–271.

Kool, M., Jones, D.T.W., Jäger, N., Northcott, P.A., Pugh, T.J., Hovestadt, V., Piro, R.M., Esparza, L.A., Markant, S.L., Remke, M., et al. (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell *25*, 393–405.

Kool, M., Korshunov, A., Remke, M., David, T., Jones, W., Schlanstein, M., Northcott, P.A., Cho, Y.J., Koster, J., Schouten-van Meeteren, A., van Vuurden, D., et al. (2012). Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, group 3, and group 4 medulloblastomas. Acta Neuropathol. *123*, 473–484.

Kristensen, V.N., Lingjærde, O.C., Russnes, H.G., Vollan, H.K.M., Frigessi, A., and Børresen-Dale, A.L. (2014). Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer *14*, 299–313.

Lancichinetti, A., and Fortunato, S. (2011). Limits of modularity maximization in community detection. Phys. Rev. E *84*, 066122.

Lee, S.J., Litan, A., Li, Z., Graves, B., Lindsey, S., Barwe, S.P., and Langhans, S.A. (2015). Na,K-ATPase β1-subunit is a target of sonic hedgehog signaling and enhances medulloblastoma tumorigenicity. Mol. Cancer *14*, 159.

Lo Muzio, L. (2008). Nevoid basal cell carcinoma syndrome (Gorlin syndrome). Orphanet J. Rare Dis. *3*, 32.

Maiese, K., TenBroeke, M., and Kue, I. (1997). Neuroprotection of lubeluzole is mediated through the signal transduction pathways of nitric oxide. J. Neurochem. *68*, 710–714.

Masgutova, G., Harris, A., Jacob, B., Corcoran, L.M., and Clotman, F. (2019). Pou2f2 regulates the distribution of dorsal interneurons in the mouse developing spinal cord. Front. Mol. Neurosci. *12*, 263.

Morabito, M., Larcher, M., Cavalli, F.M., Foray, C., Antoine, F., Mirabal-Ortega, L., Andrianteranagna, M., Druillennec, S., Garancher, A., Masliah-Planchon, J., et al. (2019). An autocrine ActivinB mechanism drives TGFβ/activin signaling in group 3 medulloblastoma. EMBO Mol. Med. *11*, e9830.

Newman, M.E.J., and Girvan, M. (2004). "Finding and evaluating community structure in networks. Phys. Rev. E *69*, 026113.

Ng, J.M., and Curran, T. (2011). "The hedgehog's tale: developing strategies for targeting cancer. Nat. Rev. *11*, 493–501.

Niewiadomski, P., Niedziółka, S.M., Markiewicz, Ł., Uśpieński, T., Baran, B., and Chojnowska, K. (2019). Gli proteins: regulation in development and cancer. Cells *8*, https://doi.org/10.3390/cells8020147.

Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. Nature *547*, 311–317.

Northcott, P.A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C.G., Mack, S., Bouffet, E., Clifford, S.C., Hawkins, C.E., French, P., et al. (2011). Medulloblastoma comprises four distinct molecular variants. J. Clin. Oncol. *29*, 1408–1414.

Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., David, J., Shih, H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature *511*, 428–434.

Osterling, W.L., Boyer, R.S., Hedlund, G.L., and Bale, J.F., Jr. (2011). MPPH syndrome: two new cases. Pediatr. Neurol. *44*, 370–373.

Phan, N.N., Wang, C.Y., Chen, C.F., Sun, Z., Lai, M.-D., and Lin, Y.-C. (2017). Voltage-gated calcium channels: novel targets for cancer therapy. Oncol. Lett. *14*, 2059–2074.

Porter, M.A., Onnela, J.P., and Mucha, P.J. (2009). Communities in networks. arXiv. http://arxiv.org/abs/0902.3788.

Radeke, C.M., Conti, L.R., and Vandenberg, C.A. (1999). Inward rectifier potassium channel Kir 2.3 is inhibited by internal sulfhydryl modification. Neuroreport *10*, 3277–3282.

Ramaswamy, V., Remke, M., Bouffet, E., Bailey, S., Steven, C.C., Doz, F., Kool, M., Dufour, C., Vassal V, Milde, T., et al. (2016). Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. Acta Neuropathol. *131*, 821–831.

Read, T.-A., Fogarty, M.P., Markant, S.L., McLendon, R.E., Wei, Z., Ellison, D.W., Febbo, P.G., and Wechsler-Reya, R.J. (2009). Identification of CD15 as a marker for tumor-propagating cells in a mouse model of medulloblastoma. Cancer Cell *15*, 135–147.

Reichardt, J., and Bornholdt, S. (2006). Statistical mechanics of community detection. Phys. Rev. E *74*, 016110.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43–48.

Rohr, K.B., Anukampa Barth, K., Varga, Z.M., and Wilson, S.W. (2001). The nodal pathway acts upstream of hedgehog signaling to specify ventral telencephalic identity. Neuron *29*, 341–351.

Sánchez Fernández, I., and Loddenkemper, T. (2017). Seizures caused by brain tumors in children. Seizure *44*, 98–107.

Schwalbe, E.C., Lindsey, J.C., Nakjang, S., Crosier, S., Smith, A.J., Hicks, D., Rafiee, G., Hill, R.M., Iliasova, A., Stone, T., et al. (2017). Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. Lancet Oncol. *18*, 958–971.

Sexton, P.M., Findlay, D.M., and Martin, T.J. (1999). Calcitonin. Curr. Med. Chem. *6*, 1067–1093.

Signorelli, M., Vinciotti, V., and Wit, E.C. (2016). NEAT: an efficient network enrichment analysis test. BMC Bioinformatics *17*, 352.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542.

Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.-J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol. *123*, 465–472.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics *35*, 497–505.

Vasileiou, G., Ekici, A.B., Uebe, S., Zweier, C., Hoyer, J., Engels, H., Behrens, J., Reis, A., and Hadjihannas, M.V. (2015). Chromatin-remodeling-factor ARID1B represses wnt/β-catenin signaling. Am. J. Hum. Genet. *97*, 445–456.

Venkataraman, S., Birks, D.K., Balakrishnan, I., Alimova, I., Harris, P.S., Patel, P.R., Handler, M.H., Dubuc, A., Taylor, M.D., Foreman, N.K., et al. (2013). MicroRNA 218 acts as a tumor suppressor by targeting multiple cancer phenotype-associated genes in medulloblastoma. J. Biol. Chem. *288*, 1918–1928.

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. *47*, W587–W593.

Yang, Y., Ren, M., Song, C., Li, D., Hussain Soomro, S., Xiong, Y., Zhang, H., and Fu, H. (2017). LINC00461, a long non-coding RNA, is important for the proliferation and migration of glioma cells. Oncotarget *8*, 84123–84139.

Yang, Z., Algesheimer, R., and Claudio, J.T. (2016). A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. *6*, 1–18.

Chapter results summary

The main concepts introduced in the research article presented in this chapter are the following:

1. Resolution limit is an open problem in community detection. Community detection at different levels of resolution allows for discovery of alternative co-existing modular structures, describing alternative levels of data specificity. Therefore, discovery of meaningful biology is heavily impacted by the initial choice of resolution.

2. Persistent partitions identify strong modular relationships, therefore overcoming such limitation requires of analyzing stable partitions at alternative resolution levels where the most meaningful changes in community composition and size occur.

3. We demonstrate the potential of a new complex graph theory concept, the multilayer community trajectory, as methodology for dimensionality reduction and feature selection in patient classification scenarios, presenting its effectiveness for independent cohorts from a rare brain tumor, medulloblastoma.

4. The concept of multilayer community trajectory reflects on how communities change upon modularity resolution variations, identifying nodes (in this case, the genes) sharing community membership across a given resolution range of interest.

5. Multilayer community trajectories provide a framework to achieve optimal feature selection with limited sample sizes, achieving high performances.

6. Initial exclusion of patients presenting partial molecular information from the optimization procedure enhances the dimensionality reduction process, allowing for an accurate *a posteriori* classification after applying the optimal parameters learnt using patients presenting complete molecular data.

7. Multilayer networks provide a powerful framework for dimensionality reduction and feature selection, while keeping high levels of explainability for biomedical studies, especially for rare disease scenarios.

Transparent methods

Multilayer network definition

A network (i.e. a graph or a monoplex) is defined as a tuple $G = (V, E)$, where denotes the set of nodes (or vertices) in the network and $E \subseteq V \times V$ denotes the set of edges (or links) connecting them (Bollobás 1998). A graph composed of multiple networks, called layers, is referred to as a multilayer network. A multilayer network is defined as a quadruplet $M = (V$, where denotes the set of nodes in the M , E M , V, $L)$ V multilayer network, L denotes the set of layers , V denotes the sets $M \subseteq V \times L$ of nodes $v \in V$ contained in each layer, and E denotes the sets of $M \subseteq V$ $M \times V$ M edges connecting tuples of nodes and layers (v, l), (v', l') $\in V$ (Kivela et al. 2014) M (**Figure S1**). In a multilayer network, an edge can be intra-layer, i.e., it connects nodes in the same layer ($l = l$'), or inter-layer, i.e. it connects nodes from different layers ($l \neq l$'). We built a multilayer network consisting of 5 layers and inter-layer edges imposed only between the same nodes, if any, on different layers.

Multilayer community detection

Communities in the multilayer network have been detected using MolTi software (Didier, Valdeolivas, and Baudot 2018; Didier, Brun, and Baudot 2015), which is available at https://github.com/gilles-didier/MolTi-DREAM. MolTi adapts the Louvain clustering algorithm with modularity maximization to multilayer networks. The Louvain algorithm for community detection consists of two recursive steps. In the first step, nodes are assigned to communities and then moved to others until no increase in modularity is observed. In the second step, the identified communities are aggregated so that a new graph is created, and the entire process starts again and proceeds until convergence.

A community (c) is defined as a group of densely connected nodes in the different layers $l \in L$. The algorithm is parametrized to the resolution parameter γ: the higher

the value of *γ*, the smaller the size of the detected multilayer communities. In MolTi, modularity of a multilayer network X is defined as:

$$Multilayer\ modularity = \sum_{l} \frac{w^{(l)}}{2m^{(l)}} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j}^{(l)} - \gamma \frac{S_i^{(l)} S_j^{(l)}}{2m^{(l)}} \right) \delta_{c_i,c_j}$$

where the first sum runs over all layers of the multilayer network and the second over all edges *{i,j}* of each layer *l*. $X^{(l)}_{i,j}$ is the weight of the edge *{i,j}* in a layer *l*; $S^{(l)}{}_{i}$ is the sum of the weights of all the edges involving vertex *i* in that layer; $m^{(l)}$ is the sum of the weights of all the edges of that layer; $\delta_{ci,cj}$ is equal to 1 if *i* and *j* belong to the same community (*ci* = *cj*) and to 0 otherwise; *γ* is the resolution parameter; $w^{(l)}$ is the user-defined weight associated to the layer l. In our calculations, $w^{(l)}$ and $X^{(l)}_{i,j}$ are both equal to 1, so that $m^{(l)}$ represents the total number of edges in *l* and $S^{(l)}_{i}$ and $S^{(l)}_{j}$ represent the degree of nodes *i* and *j*, respectively.

Data sources for the construction of the multilayer networks

We created a multilayer network consisting of five layers in which nodes represent genes (Entrez identifiers), intra-layer edges represent different types of associations retrieved from publicly available knowledge bases and inter-layer edges exist between the same nodes in the different layers (**Figure S2**). All the data was downloaded on October 19, 2019, and it is available at https://github.com/cirillodavide/gene_multilayer_network.

Molecular associations. In this layer, two genes are connected if a physical or genetic association exists. Molecular associations between human genes were obtained from BioGRID, release 3.5.177. BioGRID (Oughtred et al. 2019) is a multi-species database of interactions, curated from high-throughput datasets and individual studies. Among other prominent primary databases, BioGRID shows the highest coverage for both interactions and proteins (Bajpai et al. 2019).

Drug-target associations. In this layer, two genes are connected if they are both targets of the same drug. Drug-target associations between human genes were obtained from KEGG BRITE "Target-based Classification of Compounds", release br08310. KEGG BRITE (Kanehisa et al. 2019) is a manually curated database of functional hierarchies of various biological objects, such as Drug classifications. The Target-based Classification of Compounds consists of six categories (Protein-coupled receptors, Nuclear receptors, Ion channels, Transportes, Enzymes, Others). One-to-one and unclassified gene-target associations were excluded.

Variant-disease associations. In this layer, two genes are connected if they are both reported to be associated with the same disease in genome-wide association studies (GWAS). Variant-disease associations between human genes were obtained from Monarch Disease Ontology (MonDO), released 2019-09-30. MonDO (Mungall et al. 2017) is a multi-species ontology generated by merging and harmonizing multiple disease resources (ORDO/Orphanet, DO, OMIM, MESH, etc.). In MonDO, gene-disease associations are inferred by integrating gene variants (SNPs, SNVs, QTLs, CNVs, among others) from significant GWAS hits. We retrieved MonDO entries with associated OMIM identifiers from the OWL file, filtering for evidence code ECO:0000220 (sequencing assay evidence) through the Monarch Solr search service.

Pathway associations. In this layer, two genes are connected if they are both annotated to the same pathway. Pathway associations between human genes were obtained from Reactome, release 70. Reactome (Fabregat et al. 2018) is a manually curated pathway database. Associations were retrieved from the lowest level pathway diagram of Reactome hierarchy. We found that all annotations are associated with IEA (inferred from electronic annotations) and TAS (traceable author statement) evidence codes.

Metabolic reaction associations. In this layer, two genes are connected if they are involved in metabolic reactions where product metabolites of one reaction are

reactant metabolites of the other one. Metabolic reaction associations between human genes were obtained from Recon3D (Brunk et al. 2018) through BiGG Models (http://bigg.ucsd.edu), released 2019-09-12. Recon3D is the largest human metabolic network model. Super connected metabolites (e.g., ATP, CO2 , H2O) (Croes et al. 2006) were excluded.

Data sources of medulloblastoma genes

We aim to study the community structures of a multilayer network that contains medulloblastoma-associated genes. We selected genes for our study from two sources: (1) genes mentioned in scientific publications about medulloblastoma identified via text mining; (2) genes that are altered in medulloblastoma patients based on two recent proteogenomic studies (Forget et al. 2018; Archer et al. 2018). The text mined data has been used as a proof-of-concept for the multilayer community structure analysis. The proteogenomic datasets have been used to identify the minimal sets of genes that characterize the medulloblastoma subgroups.

Text mined medulloblastoma genes. PubTator Central (PTC) (Wei et al. 2019) was used to retrieve gene mentions in abstracts of scientific publications indexed in PubMed with the MeSH term “medulloblastoma” (D008527) in February 2020 (see Resource Availability: “Data and Code Availability”).

Medulloblastoma genes from proteogenomic data. Subgroups of 38 medulloblastoma patients (WNT, SHH, G3, G4) were retrieved from (Forget et al. 2018). While 35 patients present DNA methylation, RNA sequencing, proteomic and phosphoproteomic profiles, 3 patients (MB10, MB21, MB33) present only partial molecular information (the three lack RNA sequencing) and were used for validation. Gene methylation levels were mapped from CpG sites using the biomaRt package in R. When multiple CpG sites fell on a gene position, the median value was considered; when it fell on a region that is not annotated, the nearest gene was considered. Based on these pre-processed datasets (Forget et al. 2018), lists of genes, henceforth called “altered genes”, were obtained by selecting the top 30% of the distribution of each

data type. All the items of such lists were converted to Entrez identifiers, resulting in a total of 14039.6 altered genes per patient on average (see Resource Availability: “Data and Code Availability”).

Subgroups of 45 medulloblastoma patients (WNT, SHHa, SHHb, G3a, G3b and G4) were retrieved from (Archer et al. 2018). While 39 patients present DNA acetylation, RNA sequencing, proteomics and phosphoproteomics profiles, 6 patients lack RNA sequencing information, including all 3 patients of the WNT subgroup. When multiple DNA acetylation measurements were linked to the same gene, the median value was considered. Altered genes were obtained with the same criterion as previously described and gene symbols converted to Entrez identifiers, resulting in a total of 11608.6 (SD= 2264.524) altered genes per patient on average.

Multilayer community structure analysis

We analyzed how the multilayer community structure varies within a range of modularity resolution (γ) where the most dramatic changes in size and composition of the communities are observed before both reach a plateau.

We identified the endpoint of this range as the value where the average community size, as a function of the number of communities, establishes a plateau, i.e., where the first derivative equals zero with 0.05 margin of error (**Figure S3**).

The endpoint was found at γ=12 (964 multilayer communities), indicating that $\gamma \in (0,12]$ is the range of interest for our study.

To compare the trajectories of each gene along the communities, we computed the pairwise Hamming distance (Hamming 1950) among the vectors of communities visited by each gene in the range $\gamma \in (0,12]$ with an interval of 0.5. We refer to these vectors as multilayer community trajectories. The higher the distance, the more times two genes belong to different communities within this range (**Figure 2**).

Identification of the minimal set of genes that define medulloblastoma subgroups

The biomedical goal of the study is to identify the minimal number of genes that recapitulate the four biomedically relevant medulloblastoma subgroups (WNT, SHH, G3, and G4) (Forget et al. 2018). Identifying a minimal set of genes is crucial for both the definition of diagnostic signatures and the research on disease mechanisms. To achieve this goal, we performed a series of hierarchical clustering analyses (Ward's linkage method) where the similarity between two patients (A and B) was measured as the Jaccard index (J) of sets of altered genes selected using two parameters, θ and λ:

$$J(A_{\theta,\lambda}, B_{\theta,\lambda}) = \frac{A_{\theta,\lambda} \cap B_{\theta,\lambda}}{A_{\theta,\lambda} \cup B_{\theta,\lambda}}$$

The parameter θ defines the maximum Hamming distance allowed to include genes in the analysis, while the parameter λ defines the maximum number of them that must co-occur in the same communities along their trajectories. For dimensionality reduction purpose, small values of θ and λ guarantee a selection of genes with similar trajectories and in minimal numbers.

For instance, with θ = 2 and λ = 4, patient similarity is computed using sets of at most four genes that did not belong to the same communities at most twice along their trajectories. For each of these clustering analyses, we identified the optimal number of clusters using the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 1987) (**Table S1**).

Based on this approach, we formulated an optimization procedure to systematically evaluate values of θ and λ to identify the ones that maximize the accuracy of recapitualiting patient stratification into the four medulloblastoma subgroups (WNT, SHH, Group 3, and Group 4). We defined accuracy as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where true positives (TP) are patients of the same subgroup who are clustered together, true negatives (TN) are patients of different subgroups who are not clustered together, false positives (FP) are patients of different subgroups who are clustered together, and false negatives (FN) are patients of the same subgroup who are not clustered together. The same optimization procedure can also be formulated to maximize the Matthews Correlation Coefficient (MCC), which is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In both cases, the optimal parameters found are θ = 0 and λ = 6, corresponding to an accuracy of 94.94% (**Figure 4 and Table S2**) and an MCC 87% (**Table S3**). The optimal number of clusters based on PAM is 5, suggesting the existence of subtle differences in a few patients (see Results: “Medulloblastoma patient stratification through multilayer structure analysis”).

Multilayer networks enrichment analysis

To detect overrepresented features (drugs, pathways, etc.) that characterize each cluster, we performed a network enrichment analysis test (NEAT) (Signorelli, Vinciotti, and Wit 2016) in each layer of the multilayer network. NEAT tests whether the number of edges between two groups of nodes is significantly higher (over-enriched) than by chance, assuming a hypergeometric null distribution. In our analyses, the two groups of nodes are (a) the minimal set of genes of a patient that are present in a layer, and (b) the genes annotated to a certain feature of that layer (e.g., the genes annotated to a specific drug in the drug layer). In the specific case of the molecular interaction layer, the annotation feature consists of the neighborhood of each gene of the minimal set of a patient. Once we identify significant hits for each patient using a p-value cutoff

of 0.01 (Benjamini-Hochberg correction for multiple testing), we select those features that are enriched in all the patients of a cluster and unique to each cluster (**Table S6**).

Computational resources

All calculations were performed using the R statistical environment, in particular the packages stats (hierarchical clustering), fpc (k-medoids clustering), pvclust (clustering significance by multiscale bootstrap resampling), sigclust2 (clustering significance by Monte Carlo procedure), and neat (network enrichment analysis). To ease the detection and analysis of the multilayer community trajectories, we developed the R package CmmD, which is openly available at https://github.com/ikernunezca/CmmD.

Supplemental Figures

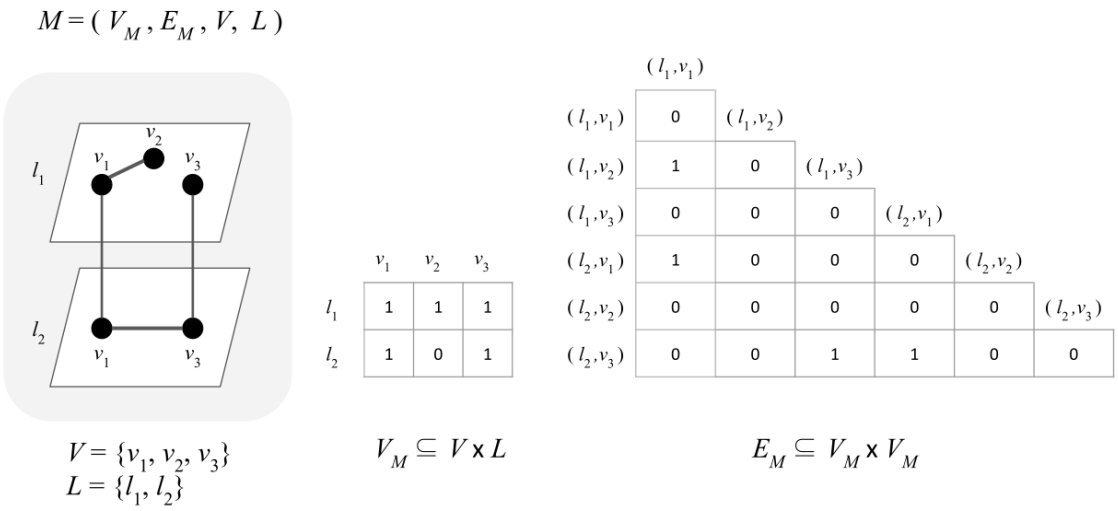


Figure S1. Multilayer network definition, Related to Figure 1 and Figure 2. A multilayer network *M*, such as the one represented inside the grey area, is defined as a quadruplet of four elements (*VM, EM , V, and L*). V and L are the sets of nodes and layers of *M*, respectively. *VM* and *EM* are the sets of nodes contained in each layer and edges connecting them within (intra-layer) and between (inter-layer) layers, respectively. As the one represented here, we build a multilayer network where inter-layer edges only connect the same nodes in each layer.

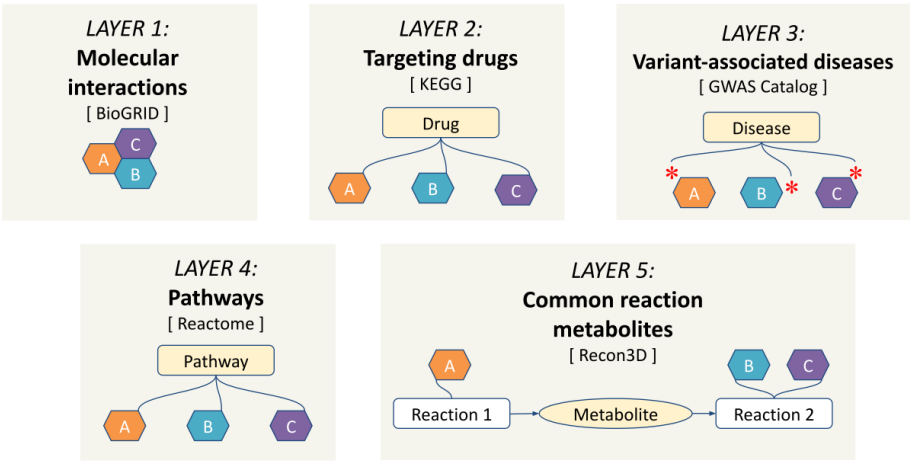


Figure S2. Gene-gene association represented in the fiver layers of the multilayer network, Related to Figure 1 and Figure 2. Gene entities are represented as hexagons. Associations retrieved from the databases in squared parentheses are represented as curved lines. Red asterisks indicate mutations.

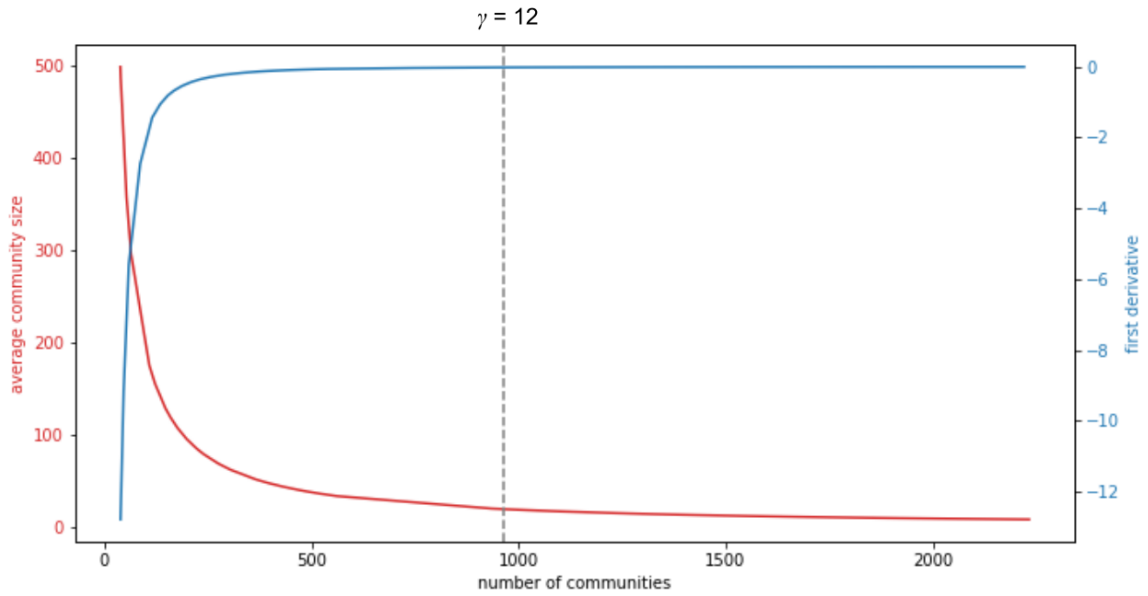


Figure S3. Identification of the resolution range of interest, Related to Figure 2. The modularity resolution parameter (γ) determines the number of communities and their size. The most dramatic changes in both size and number of communities occur in an initial range of resolution, which enables us to detect genes that are strongly associated. We identified the endpoint of this range (γ = 12) as the value where the average community size, as a function of the number of communities, establishes a plateau (i.e., its first derivative equals zero with 0.05 margin of error).

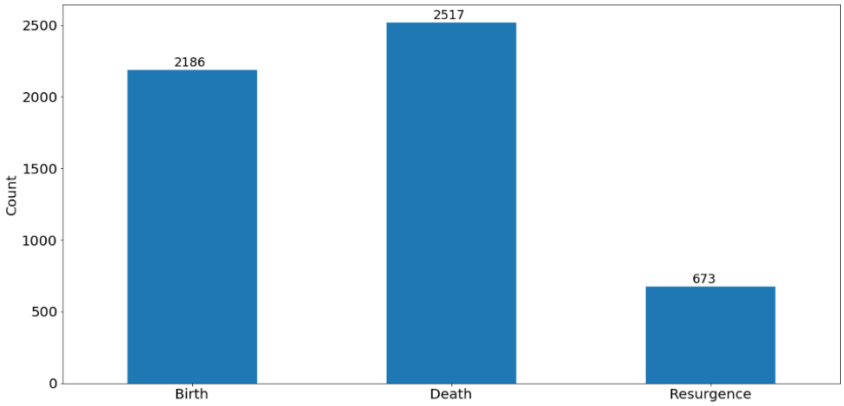


Figure S4. Operations on dynamic communities, Related to Figure 3. Count of dynamic events (birth, death, and resurgence) in the multilayer communities that contain text-mined medulloblastoma genes.

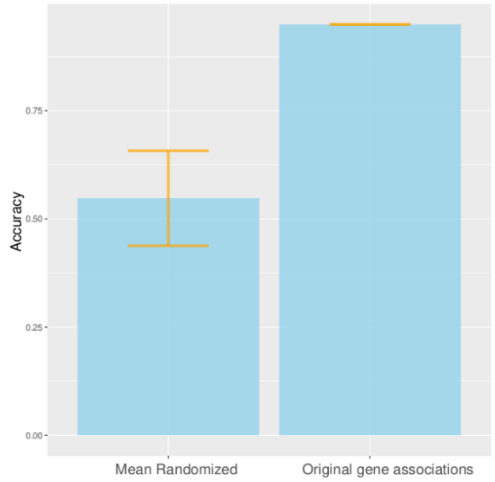


Figure S5. Gene shuffling test, Related to Figure 4 and Figure S6. The bar plots show the comparison between the highest accuracy achieved with the optimization procedure (94.94%, "Original gene associations") and the average accuracy achieved by shuffling the genes in the cohort 10,000 times (54.76%, SD = 0.11, "Mean Randomized"), maintaining the same number of genes for each patient as in the original data and using the optimal parameters θ = 0 and λ = 6.

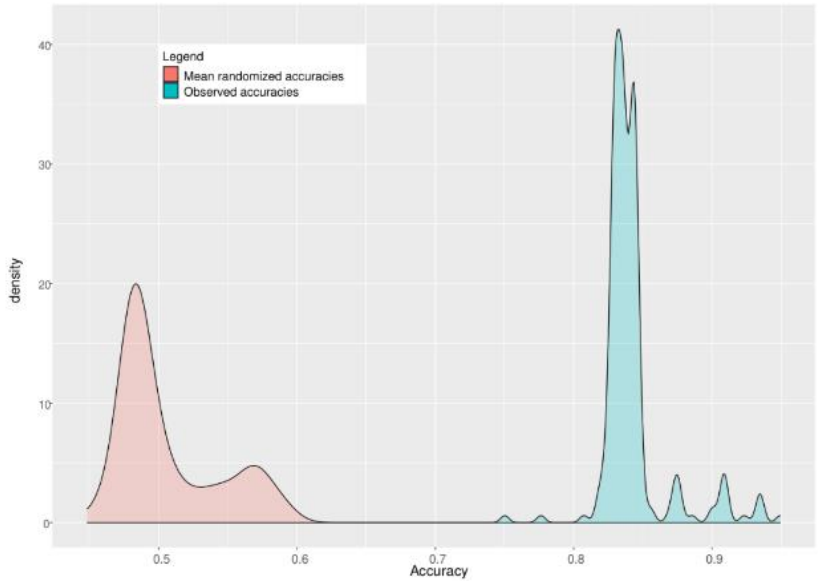


Figure S6. Distributions of optimization accuracies, Related to Figure 4 and Figure S4. The distribution of the optimization accuracies in the original data is reported in green, and the distribution of the average optimization accuracies after shuffling the altered genes across the cohort 10,000 times, maintaining the same number of genes for each patient is reported in red.

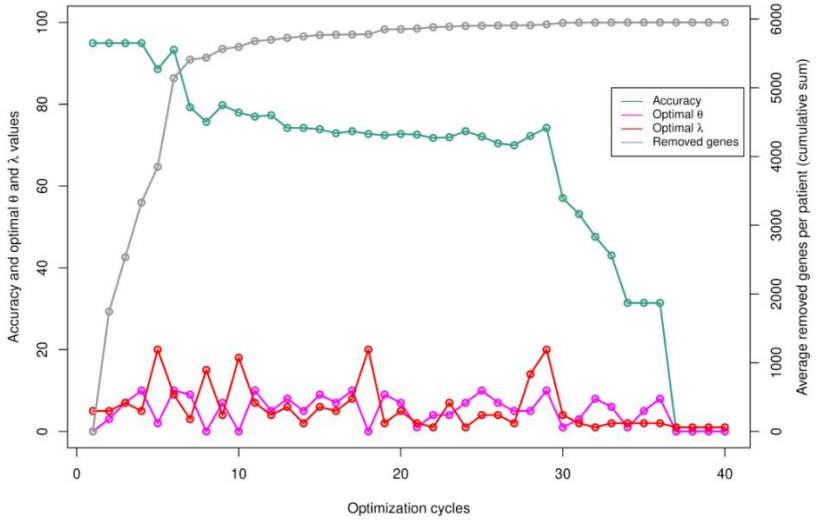


Figure S7. Recursive exclusion test, Related to Figure 4. The plot shows the iterative removal of selected genes in the cohort of 38 medulloblastoma patients. At every iteration, the minimal set of genes, found at optimal values of θ (purple line) and λ (red line) corresponding to highest accuracy (green line), is removed and the optimization procedure is repeated. The cumulative average number of genes per patient that are removed at every iteration is reported (grey line).

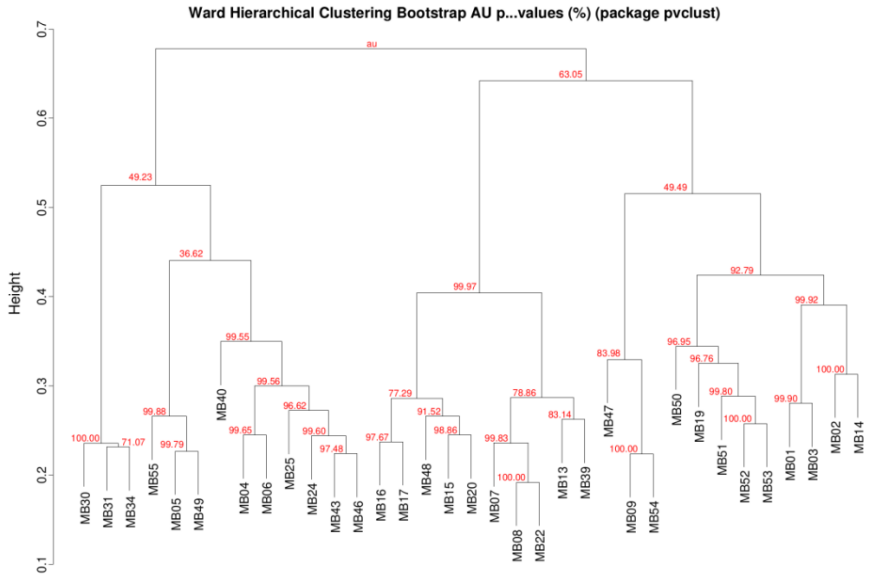


Figure S8. Clustering significance, Related to Figure 5. Significance assessment of hierarchical clustering (Ward method) of medulloblastoma patients using multiscale bootstrap resampling (Suzuki and Shimodaira, 2006). AU p values (%), or approximately unbiased probability value (pvAU), is reported in red on top of each cluster.

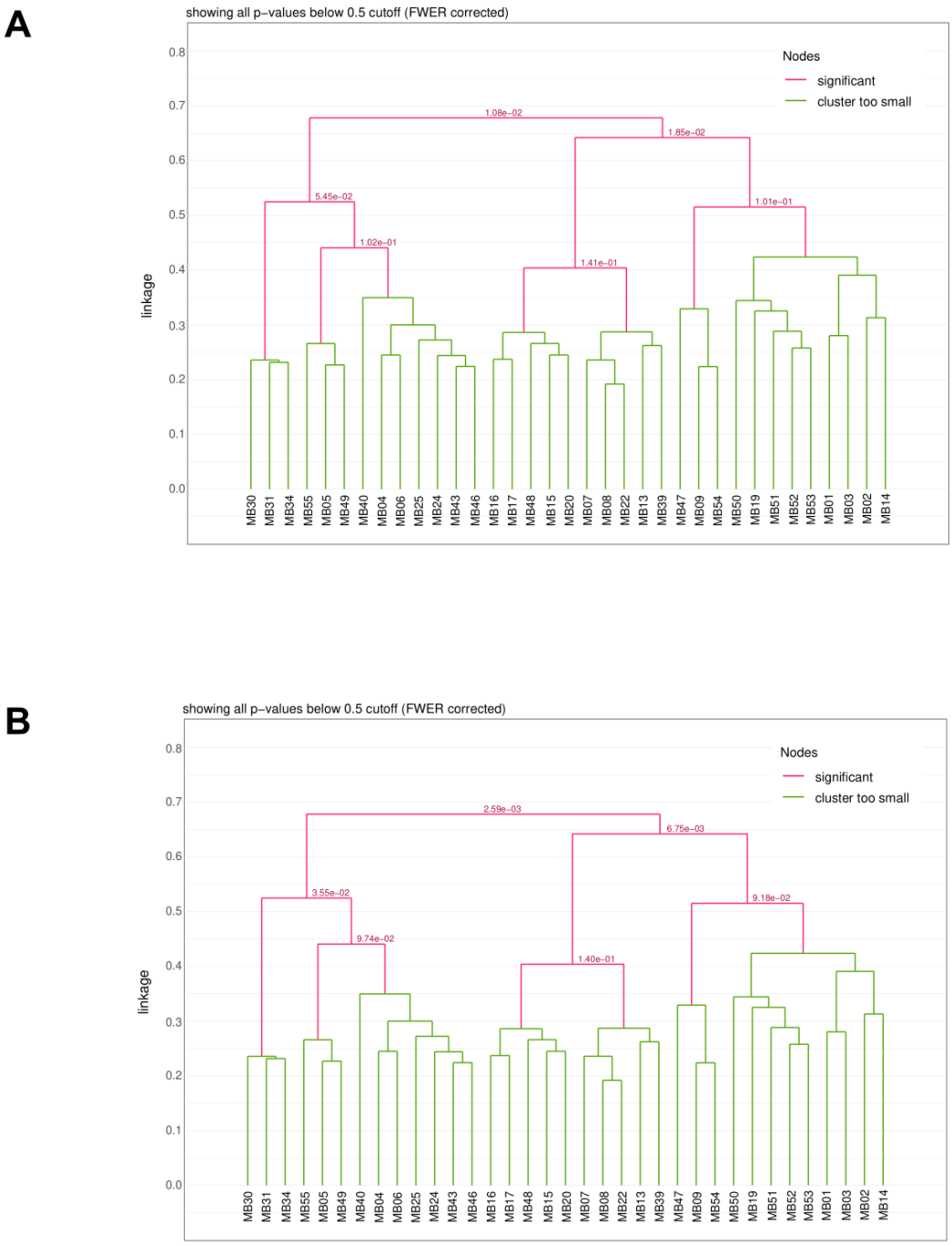


Figure S9. Clustering significance, Related to Figure 5. Significant assessment of hierarchical clustering (Ward method) of medulloblastoma patients using a Monte Carlo procedure (Kimes et al., 2017). (**A**) empirical p-value and (**B**) Gaussian approximate p-value are reported in red on top of each cluster.

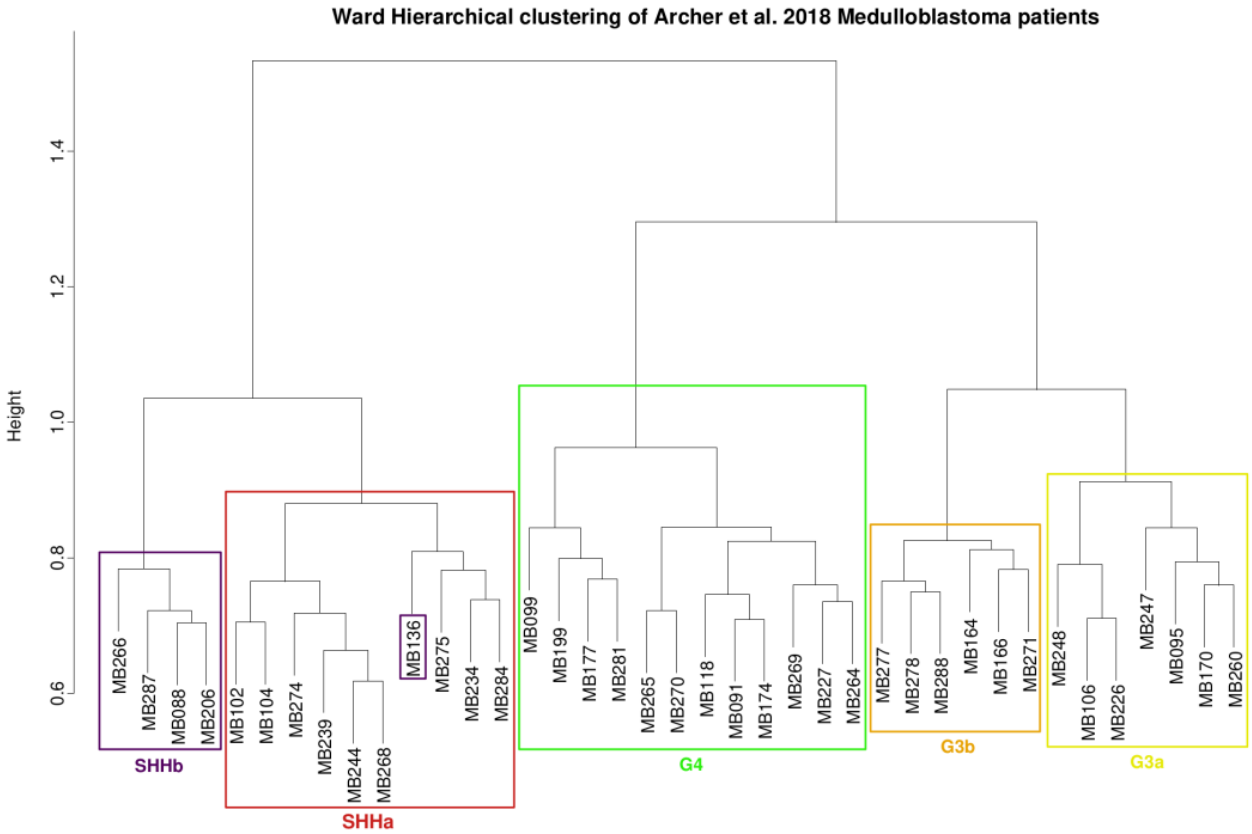


Figure S10. Hierarchical clustering of medulloblastoma patients from Archer et al. 2018, **Related to Figure 2 and Figure S11**. Ward's linkage hierarchical clustering obtained at λ = 3 and θ = 0 for patients with complete multi-omics data (Archer et al., 2018). Rectangles indicate the 5 clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the five medulloblastoma subgroups: SHHa (red), SHHb (purple), Group 4 (G4, green), Group 3a (G3, yellow), Group 3b (G3b, orange). Patient MB136, originally labeled as SHHb subgroup and highlighted with a purple lower level rectangle, clusters within the SHHa subgroup.

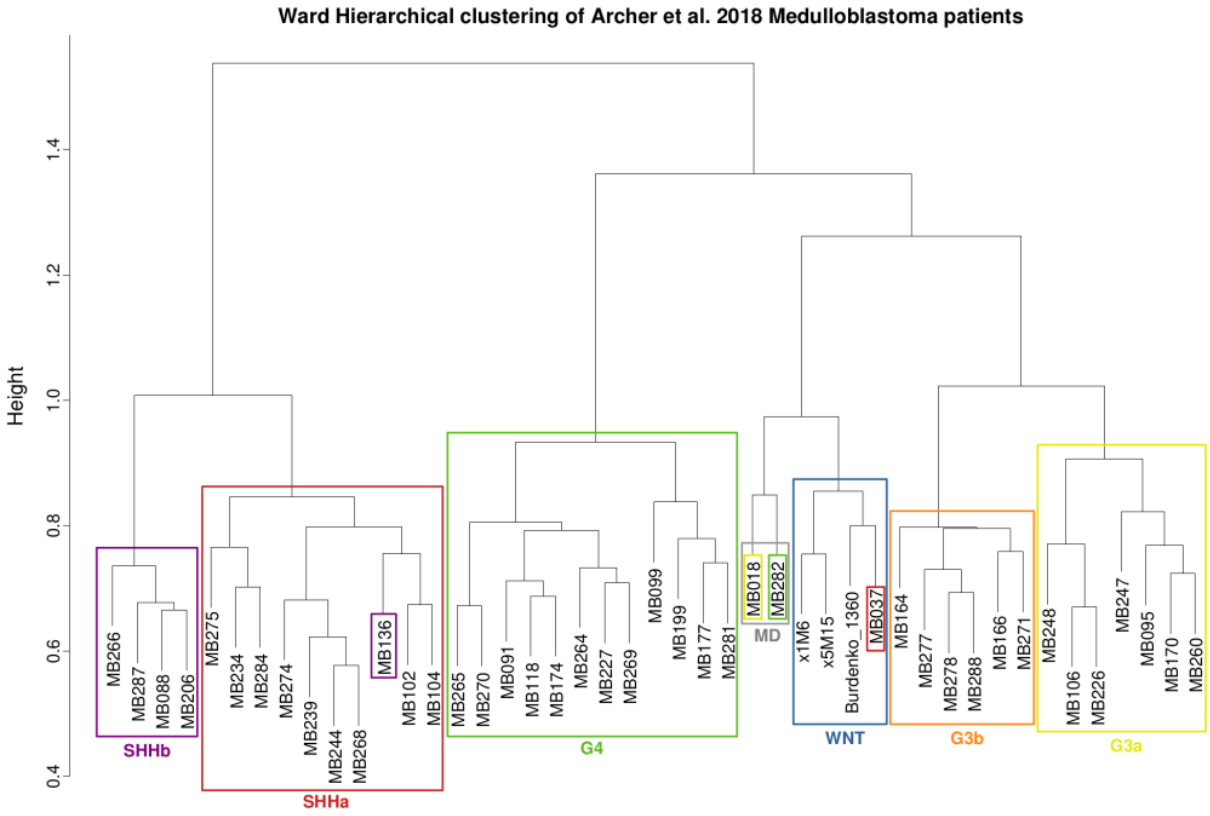


Figure S11. Hierarchical clustering of medulloblastoma patients from Archer et al. 2018, Related to Figure 2 and Figure S10. Ward's linkage hierarchical clustering obtained at $\lambda = 5$ and $\theta = 1$ for patients with complete and incomplete multi-omics data (Archer et al., 2018). Rectangles indicate the 7 clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the six medulloblastoma subgroups: WNT (blue), SHHa (red), SHHb (purple), Group 4 (G4, green), Group 3a (G3, yellow), Group 3b (G3b, orange). Patients with missing data cluster together (MD, Missing Data). Misclassified patients are highlighted with lower-level rectangles indicating their original subgroup.

Supplemental Tables

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	9	10	7	9	9	5	10	6	8	8	10	8	10	9	8	8	9	8	9	6
1	8	10	10	10	9	8	10	8	8	9	8	8	8	8	8	8	9	9	10	8
2	8	8	8	10	8	9	8	10	9	10	8	10	9	9	8	8	8	9	8	10
3	10	10	9	5	9	9	5	10	10	9	10	8	8	10	8	10	10	8	8	8
4	10	10	10	4	8	9	5	10	10	5	9	9	5	5	5	10	8	8	8	8
5	10	8	9	7	8	8	5	10	10	10	9	5	8	6	10	10	8	9	8	8
6	8	9	10	7	9	8	8	10	10	5	9	4	9	6	10	9	9	9	8	8
7	7	9	7	7	7	7	9	8	10	8	9	5	10	10	5	7	9	5	8	8
8	8	8	8	8	8	8	9	4	8	8	8	8	10	10	10	10	9	9	8	8
9	8	7	8	8	8	8	9	4	8	8	4	5	8	10	4	4	9	9	10	10
10	8	9	8	8	9	9	9	8	8	9	10	10	9	9	9	10	9	5	8	10

Table S1. Optimal number of clusters, Related to Figure 4. The matrix shows the optimal number of clusters, based on the partitioning around medoids (PAM) algorithm, for combinations of parameters θ (rows) and λ (columns).

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.842	0.819	0.886	0.829	0.837	**0.949**	0.819	0.874	0.873	0.835	0.824	0.835	0.827	0.833	0.843	0.843	0.837	0.829	0.837	0.9
1	0.84	0.827	0.83	0.827	0.837	0.843	0.83	0.843	0.843	0.837	0.837	0.843	0.847	0.847	0.843	0.843	0.848	0.837	0.83	0.843
2	0.835	0.843	0.847	0.835	0.843	0.837	0.847	0.83	0.83	0.83	0.843	0.83	0.833	0.837	0.843	0.843	0.847	0.833	0.843	0.827
3	0.835	0.83	0.829	0.832	0.829	0.837	0.909	0.83	0.824	0.837	0.83	0.843	0.843	0.83	0.847	0.83	0.83	0.843	0.843	0.843
4	0.835	0.83	0.83	0.869	0.843	0.829	0.909	0.83	0.83	0.909	0.833	0.837	0.9	0.909	0.91	0.83	0.835	0.835	0.843	0.843
5	0.832	0.843	0.829	0.835	0.835	0.843	0.75	0.819	0.83	0.83	0.829	0.909	0.835	0.874	0.819	0.827	0.847	0.837	0.843	0.843
6	0.847	0.833	0.824	0.837	0.837	0.843	0.843	0.83	0.83	0.835	0.837	0.935	0.837	0.874	0.835	0.829	0.833	0.829	0.837	0.843
7	0.876	0.832	0.835	0.876	0.835	0.835	0.833	0.833	0.83	0.843	0.833	0.909	0.83	0.83	0.835	0.835	0.837	0.776	0.829	0.843
8	0.838	0.845	0.847	0.843	0.842	0.847	0.837	0.856	0.847	0.843	0.843	0.843	0.83	0.83	0.83	0.84	0.827	0.838	0.847	0.843
9	0.842	0.876	0.847	0.835	0.843	0.843	0.837	0.807	0.843	0.843	0.935	0.835	0.843	0.83	0.935	0.935	0.837	0.843	0.829	0.83
10	0.837	0.832	0.847	0.847	0.829	0.837	0.829	0.843	0.842	0.829	0.83	0.83	0.837	0.835	0.829	0.824	0.829	0.923	0.855	0.824

Table S2. Optimization accuracies, Related to Figure 4. The matrix shows the accuracies of the optimization procedure (see Methods: “Identification of the minimal set of genes that define medulloblastoma subgroups”) for combinations of parameters θ (rows) and λ (columns). The maximum accuracy achieved is highlighted in bold.

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.603	0.536	0.719	0.564	0.590	**0.876**	0.536	0.682	0.685	0.582	0.554	0.582	0.563	0.581	0.608	0.608	0.590	0.564	0.590	0.754
1	0.596	0.563	0.572	0.563	0.590	0.608	0.572	0.608	0.608	0.590	0.590	0.608	0.617	0.617	0.608	0.608	0.624	0.590	0.572	0.608
2	0.582	0.608	0.617	0.586	0.608	0.590	0.617	0.572	0.572	0.572	0.608	0.572	0.578	0.590	0.608	0.608	0.617	0.581	0.608	0.563
3	0.589	0.572	0.564	0.577	0.564	0.590	0.771	0.572	0.554	0.590	0.572	0.608	0.608	0.572	0.617	0.572	0.572	0.608	0.608	0.608
4	0.589	0.572	0.572	0.678	0.608	0.564	0.771	0.572	0.572	0.771	0.578	0.590	0.751	0.771	0.776	0.572	0.582	0.582	0.608	0.608
5	0.577	0.608	0.564	0.575	0.582	0.608	0.382	0.536	0.572	0.572	0.564	0.771	0.582	0.682	0.536	0.563	0.617	0.590	0.608	0.608
6	0.617	0.581	0.554	0.579	0.590	0.608	0.608	0.572	0.572	0.582	0.590	0.841	0.590	0.682	0.589	0.564	0.581	0.564	0.590	0.608
7	0.694	0.577	0.575	0.694	0.575	0.575	0.581	0.578	0.572	0.608	0.581	0.771	0.572	0.576	0.582	0.575	0.590	0.421	0.564	0.608
8	0.595	0.612	0.617	0.608	0.603	0.617	0.590	0.643	0.617	0.608	0.608	0.608	0.572	0.572	0.572	0.603	0.563	0.595	0.617	0.608
9	0.603	0.694	0.617	0.582	0.608	0.608	0.590	0.521	0.608	0.608	0.841	0.582	0.608	0.572	0.841	0.841	0.590	0.611	0.567	0.572
10	0.587	0.577	0.617	0.617	0.564	0.590	0.564	0.608	0.603	0.564	0.572	0.572	0.590	0.586	0.564	0.554	0.564	0.810	0.642	0.554

Table S3. Optimization MCC, Related to Figure 4. The matrix shows the Matthews Correlation Coefficient (MCC) of the optimization procedure (see Methods: “Identification of the minimal set of genes that define medulloblastoma subgroups”) for combinations of parameters θ (rows) and λ (columns). The maximum MCC value achieved is highlighted in bold.

	"MB10"	"MB21"	"MB33"
"MB01"	0.20855106888361	0.219810040705563	0.199903194578896
"MB02"	0.226933830382106	0.232198142414861	0.202247191011236
"MB03"	0.230385487528345	0.246086956521739	0.203089504770559
"MB04"	0.236396890717878	0.247598253275109	0.193577566711895
"MB05"	0.224057602710716	0.236489232019504	0.185480486781368
"MB06"	0.2255299954894	0.239740820734341	0.190839694656489
"MB07"	0.247404063205418	0.241379310344828	0.191169977924945
"MB08"	0.255896751223854	0.245812395309883	0.191443388072602
"MB09"	0.234858387799564	0.238391376451078	0.19559585492228
"MB10"	1	0.387596899224806	0.417813765182186
"MB13"	0.252108716026242	0.240667545015371	0.196420376319413
"MB14"	0.224178962398858	0.229138475417231	0.200670498084291
"MB15"	0.245346062052506	0.238565022421525	0.199434229137199
"MB16"	0.260057471264368	0.244523915958873	0.208097928436912
"MB17"	0.256641366223909	0.238726790450928	0.205223880597015
"MB19"	0.226726057906459	0.23021582733813	0.196706720071206
"MB20"	0.245344506517691	0.237636761487965	0.19560238204306
"MB21"	0.387596899224806	1	0.47027027027027
"MB22"	0.263229308005427	0.251486830926083	0.197016235190873
"MB24"	0.228245363766049	0.235772357723577	0.198476915754403
"MB25"	0.219874100719424	0.225834046193328	0.190647482014389
"MB30"	0.225081890500702	0.260118235561619	0.216873212583413
"MB31"	0.219325842696629	0.265342163355408	0.208029197080292
"MB33"	0.417813765182186	0.47027027027027	1
"MB34"	0.231185218566922	0.263134851138354	0.210621879255561
"MB39"	0.243792325056433	0.246463780540077	0.193433895297249
"MB40"	0.210699202252464	0.221179624664879	0.191943127962085
"MB43"	0.214088397790055	0.220472440944882	0.188539741219963
"MB46"	0.232285312060066	0.232372505543237	0.200093720712277
"MB47"	0.208278291501541	0.229626485568761	0.183826778612461
"MB48"	0.236533957845433	0.244097995545657	0.194895591647332
"MB49"	0.216193656093489	0.234702093397746	0.172910662824208
"MB50"	0.208942390369733	0.227743271221532	0.176949443016281
"MB51"	0.22202565236621	0.242437153813379	0.195921985815603
"MB52"	0.232150678931231	0.253062948880439	0.194782608695652
"MB53"	0.236637734125171	0.251093613298338	0.203644646924829
"MB54"	0.208281573498965	0.230861723446894	0.176980198019802
"MB55"	0.201853344077357	0.224043715846995	0.166733306677329

Table S4. Classification of patients with partial datasets, Related to Figure 5. The table reports the values of the Jaccard Index (J), parametrized by the optimal θ and λ, between the 3 patients with partial datasets (MB10, MB21, MB33) and the 35 patients with complete datasets (see Methods: “Data sources of medulloblastoma genes”).

Table S5 *. Minimal set of genes, Related to Figure 5 (attached dataset). Minimal sets of altered genes associated with each one of the 38 medulloblastoma patients from (Forget et al., 2018). The labels of the original subgroups (clusters) and the ones assigned after the optimization procedure are reported.

Table S6 *. Multilayer network enrichment analysis, Related to Figure 5 (attached dataset). The table reports those associations (edges) among the minimal sets of genes that are enriched in all the patients of a cluster and unique of each cluster (WNT, SHH, G3, G4, G3-G4) for a specific layer of the multilayer network (see Methods: “Multilayer network enrichment analysis”). Association IDs are grounded in databases (see Methods: “Data sources for the construction of the multilayer network”).

* Available online at: https://doi.org/10.1016/j.isci.2021.102365.

Supplemental references

Archer, Tenley C., Tobias Ehrenberger, Filip Mundt, Maxwell P. Gold, Karsten Krug, Clarence K. Mah, Elizabeth L. Mahoney, et al. 2018. “Proteomics, Post-Translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups.” *Cancer Cell* 34 (3): 396–410.e8.

Bajpai, Akhilesh Kumar, Sravanthi Davuluri, Kriti Tiwary, Sithalechumi Narayanan, Sailaja Oguru, Kavyashree Basavaraju, Deena Dayalan, Kavitha Thirumurugan, and Kshitish K. Acharya. 2019. “How Helpful Are the Protein-Protein Interaction Databases and Which Ones*?” Cold Spring Harbor Laboratory*. https://doi.org/10.1101/566372.

Bollobás, Béla. 1998. “Ramsey Theory.” *Modern Graph Theory*. https://doi.org/10.1007/978-1-4612-0619-4_6.

Brunk, Elizabeth, Swagatika Sahoo, Daniel C. Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, et al. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3): 272–81.

Croes, Didier, Fabian Couche, Shoshana J. Wodak, and Jacques van Helden. 2006. “Inferring Meaningful Pathways in Weighted Metabolic Networks.” *Journal of Molecular Biology* 356 (1): 222–36.

Didier, Gilles, Christine Brun, and Anaïs Baudot. 2015. “Identifying Communities from Multiplex Biological Networks.” *PeerJ* 3 (December): e1525.

Didier, Gilles, Alberto Valdeolivas, and Anaïs Baudot. 2018. “Identifying Communities from Multiplex Biological Networks by Randomized Optimization of Modularity.” *F1000Research* 7 (July): 1042.

Fabregat, Antonio, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, et al. 2018. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 46 (D1): D649–55.

Forget, Antoine, Loredana Martignetti, Stéphanie Puget, Laurence Calzone, Sebastian Brabetz, Daniel Picard, Arnau Montagud, et al. 2018. “Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling.” *Cancer Cell* 34 (3): 379–95.e7.

Hamming, R. W. 1950. “Error Detecting and Error Correcting Codes.” *Bell SystemTechnical Journal*. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

Kanehisa, Minoru, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. 2019. “New Approach for Understanding Genome Variations in KEGG.” Nucleic Acids Research 47 (D1): D590–95.

Kaufman, Leonard, and Peter Rousseeuw. 1987. Clustering by Means of Medoids. https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf

Kimes, Patrick K., Yufeng Liu, David Neil Hayes, and James Stephen Marron. 2017. “Statistical Significance for Hierarchical Clustering.” Biometrics 73 (3): 811–21.

Kivela, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. 2014. “Multilayer Networks.” Journal of Complex Networks 2 (3): 203–71.

Mungall, Christopher J., Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, et al. 2017. “The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species.” Nucleic Acids Research 45 (D1): D712–22.

Oughtred, Rose, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, et al. 2019. “The BioGRID Interaction Database: 2019 Update.” Nucleic Acids Research 47 (D1): D529–41.

Signorelli, Mirko, Veronica Vinciotti, and Ernst C. Wit. 2016. “NEAT: An Efficient Network Enrichment Analysis Test.” BMC Bioinformatics 17 (1): 352.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. “Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering.” Bioinformatics 22 (12): 1540–42.

Wei, Chih-Hsuan, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. “PubTator Central: Automated Concept Annotation for Biomedical Full Text Articles.” Nucleic Acids Research 47 (W1): W587–93.

Chapter 5

Discussion

1. Addressing data scarcity using multilayer networks

The research presented in this PhD thesis pursued the main goal of exploring the potential of network biology approaches, and in particular, multilayer network analysis, to provide an efficient and interpretable integrative framework for the biomedical scenarios affected by data scarcity, specifically, rare diseases and precision oncology.

Starting to address such a challenge required of an extensive review of the current literature related to data granularity (**Chapter 2**) "**Artificial intelligence in cancer research: learning at different levels of data granularity**" (194), which focus on current applications and challenges of machine learning and AI when dealing with the different granularities of the present cancer data landscape.

During recent years, biomedical research has quickly evolved thanks to the capability of machine learning and AI-based approaches to not only deal but efficiently generate new biomedical knowledge from ever-growing resources (**Chapter 1, Section 2**). While this applies not only to oncology but to many other biomedical research fields, data-hungry algorithms only solve one part of the whole picture, with a major question arising: **How do we address knowledge discovery in scenarios characterized by data scarcity?**

This question became the central motivation of the research presented in this PhD thesis. Indeed, although the discussion introduced in Chapter 2 has a particular focus on the application to precision oncology, the presented challenges can be extrapolated to the study of other medical areas, such as rare disorders.

Data scarcity not only limits common biomedical analysis such as interpretation of disease classification and subtyping (**Chapter 4**) but largely keeps other equally important questions mostly unexplored (such as the study of the molecular determinants of disease severity, **Chapter 3**).

The methodologies introduced in this PhD thesis were developed around the clear necessity of introducing complementary biomedical knowledge to overcome the challenges of limited data availability. This way, **the first step** towards discovering novel, meaningful functional relationships in data-scarce contexts is **the integration of additional biomedical information sourced from external databases**.

A number of existing methodologies may provide efficient frameworks to achieve this task (**Introduction, Section 2**). Among them, a network biology perspective, based on multilayer networks, is chosen as it provides a robust and meaningful way for the integration of the relevant biomedical resources. Furthermore, the gene multilayer network structure (**Introduction, Section 3.4. Multilayer and complex networks**) facilitates the **second step: mapping** the limited available **patient information** on multilayer networks **and leveraging the associations displayed in this framework** for a transparent modeling of human disease**.**

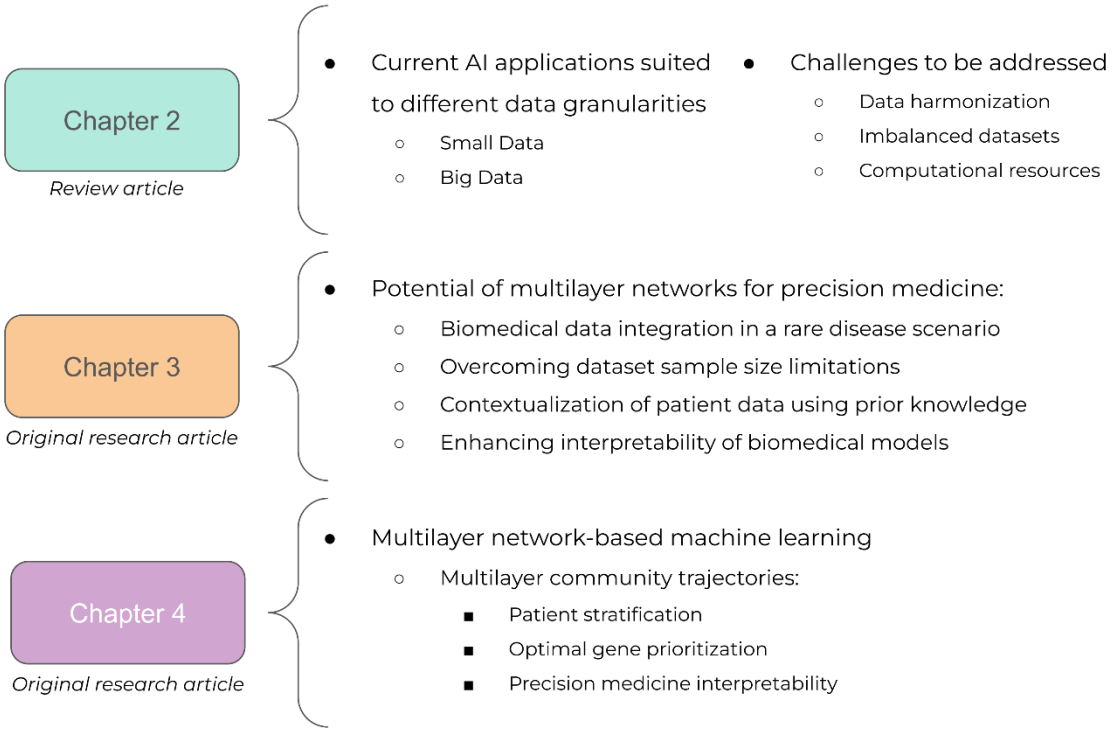


Figure 10. **Overview of the main results and objectives tackled in the research presented in chapters 2, 3 and 4.** Chapter 2 presented the main challenges of the ML applications in the biomedical fields related to data scarcity. Chapters 3 and 4 deeply delve into the potential of network biology to address such challenges.

The whole perspective to tackle the data scarcity challenge is introduced, and applied to two independent rare disease scenarios, with the multilayer network-based approaches presented in **Chapters 3** ('**Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes**') (195) and **4** (**'The multilayer community structure of medulloblastoma**') (196) (**Figure 10**). Detection of persistent associations within network community structures is introduced in Chapter 3 and is the main concept implemented in the developed algorithms (for both Chapters 3 and 4) used in different biomedical scenarios.

Globally, the novel systems biology approach presented in this PhD Thesis consists of 3 main stages (**Figure 11**):

1) **Integrative modelling of prior knowledge of interest**: The various relevant databases considered in the study are represented as individual networks, interconnected in a multilayer network.
2) **Candidate gene selection from multi-omics data:** A personalized analysis of patient omics information is performed to identify candidate genes of interest for the study. Within the context of the presented research articles, this analysis led to a selection of genes to be mapped to the multilayer network for further investigations. In Chapter 3, this selection of candidate genes is extracted from the Whole Genome Sequencing (WGS) data (i.e., specific genomic variants of each CMS patient). In Chapter 4, it corresponds to genes identified through differential analysis performed over the available patient multi-omics data.
3) **Identification of functional relationships to address specific biomedical tasks**: Persistency of network community membership (also known as multilayer community trajectories) is computed over the nodes of the multilayer network, resulting a measure of the strength of functional relationships among genes. The information provided by the multilayer community trajectories is explored in order to solve specific analytical and learning tasks, such as gene priorization (Chapter 3) and feature selection (Chapter 4).

In this view, the goal of this approach is to both integrate relevant biomedical knowledge and identify the information needed to overcome the limitations of data scarcity in analytical and modeling tasks.

The presented methodology can be viewed as a way to address this open problem, harnessing the power of graph representations to tackle emerging relevant questions: **How can the importance of the employed knowledge resources can be rationally evaluated? How should this knowledge be optimally exploited in the multilayer network beyond community detection?**.

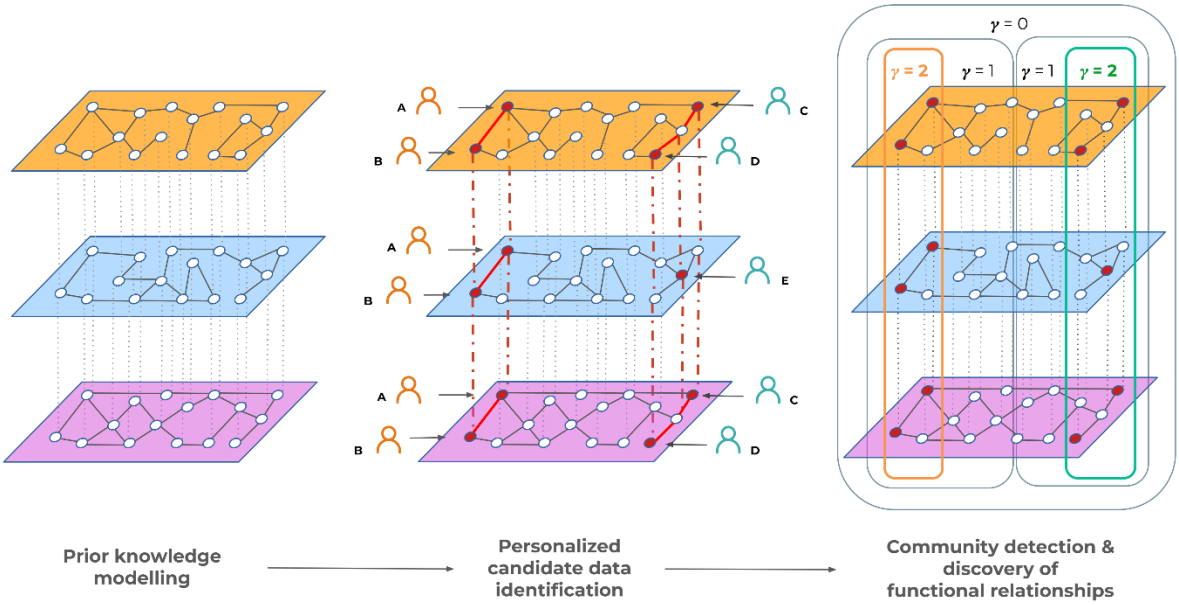


Figure 11. **Schematic representation of the methodology introduced to address data scarcity.** In the example, a disease cohort with two different subtypes is presented (depicted as orange and green patients). The first step of the methodology consists in building a multilayer network representing relevant biomedical information from external resources. Candidate information is identified for each patient's available omics information, prioritizing different target nodes for downstream analysis. Finally, the relationships existing among the target information are identified via network community analysis at various resolution (**γ**) levels.

The research article presented in Chapter 4 (**'The multilayer community structure of medulloblastoma**') solves these questions by presenting an optimization technique on top of the methodology introduced in Chapter 3, while also providing an

efficient feature selection approach using a multilayer network built on prior knowledge from databases, in order to achieve accurate patient stratification.

The concept of community trajectories, introduced in Chapter 4, within monolayer or multilayer networks, can be considered as a way of hierarchically classifying the nodes of a given graph, based on the identified community structures. The proposed optimization approach iterates over these structures to find the optimal selection of nodes that best recapitulates a ground truth (e.g., disease subtyping classification).

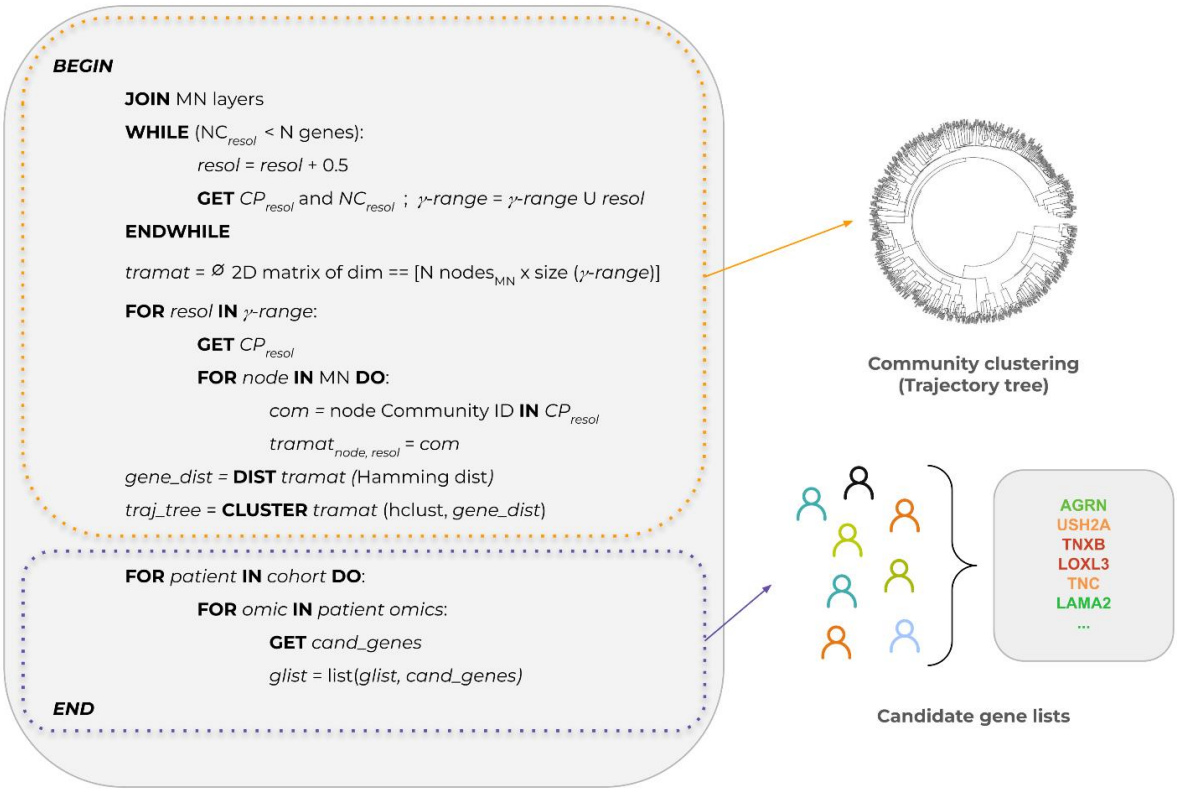


Figure 12. **Pseudocode for the first two stages of the systems biology methodology presented for the analysis of cohorts affected by data scarcity**. Highlighted in orange, is the analysis of the community structure of the multilayer network at multiple levels of modularity resolution (**γ**). Highlighted in purple, is the identification of candidate data for the patients of a given disease cohort.

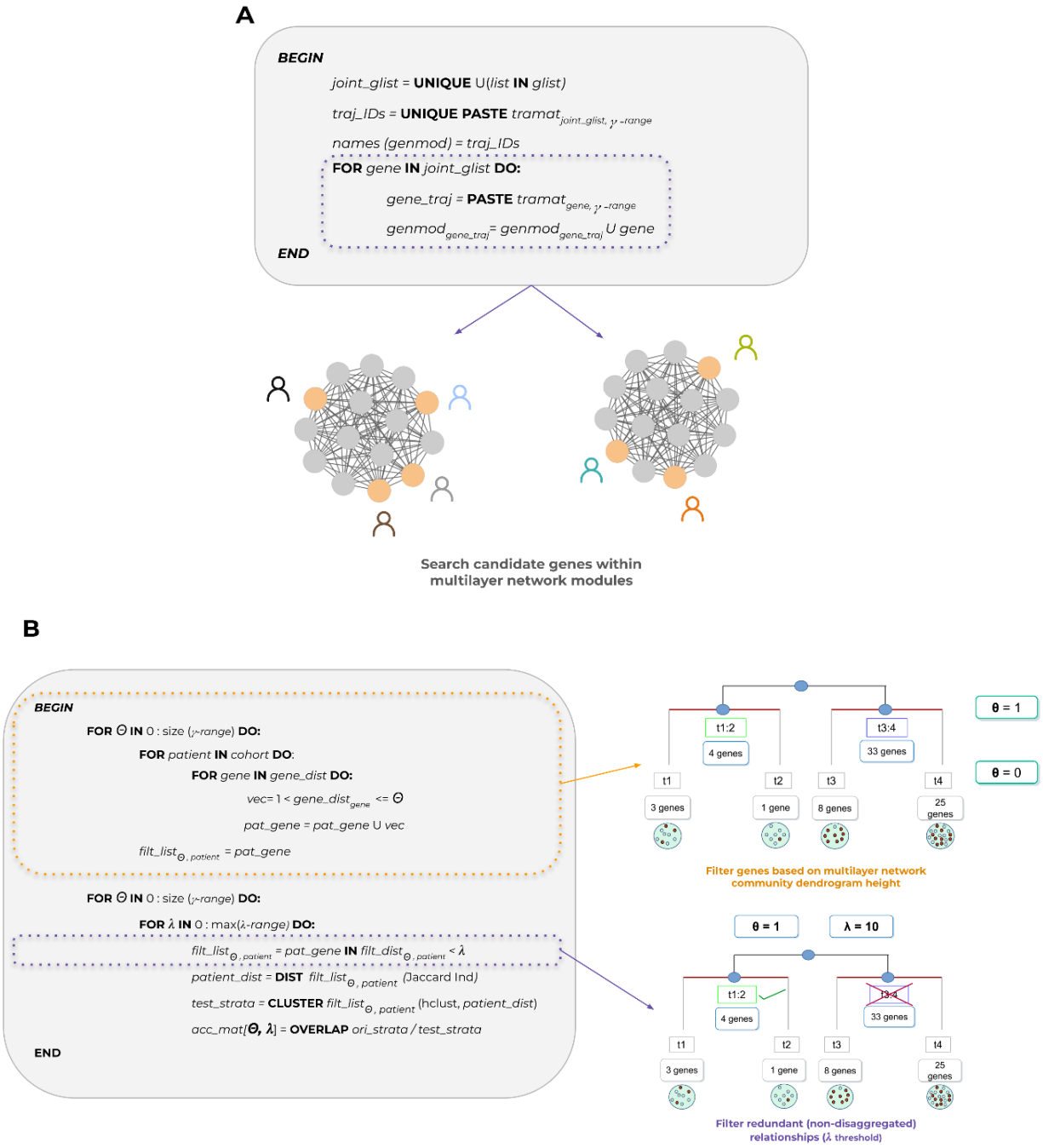


Figure 13. Pseudocode for the downstream analysis based on multilayer community trajectories. (**A**) Detection of gene modules found to share community through a whole range of modularity resolution (**γ**), the main approach undertook in Chapter 3. (**B**): Optimization procedure performed in Chapter 4 to perform optimal feature selection given a known patient sub-stratification.

In this sense, the first two stages of the presented approach (i.e., network modeling of relevant external database information and the identification of personalized candidate information) can be considered common to the analysis of any given cohort (**Figure 12**). The use of the identified topological relationships and their importance to solve specific biomedical tasks, on the other side, is specific to the particular scenario (i.e., analysis of persistent community associations for gene priorization -**Chapter 3**- and feature selection -**Chapter 4**-) (**Figure 13**).

Overall, the **main advantages** provided by the usage of the multilayer network are related to the **flexibility in modelling of biomedical associations across multiple data types** (**Table 3**).

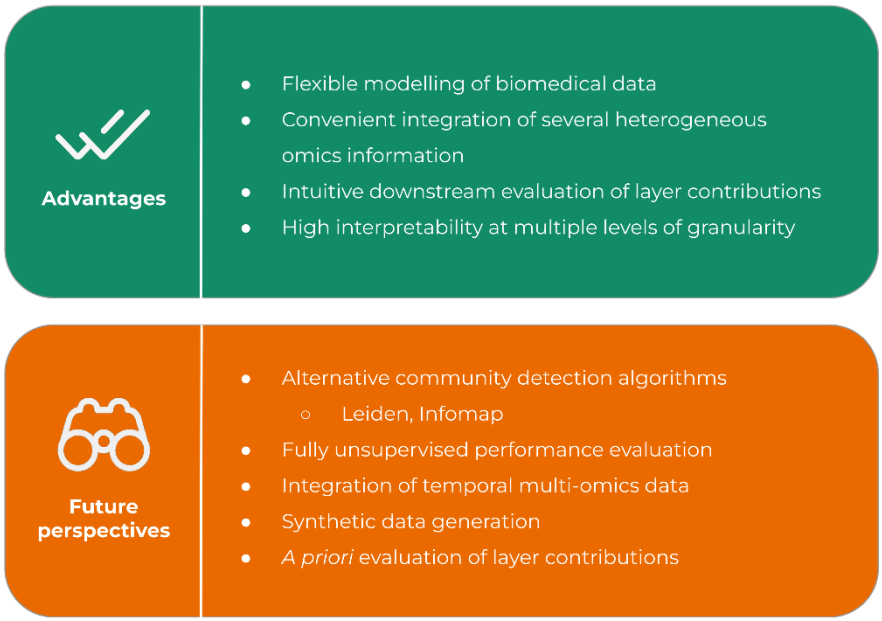


Table 3. **Advantages, limitations, and future perspectives of the presented research**. The multilayer network framework provides a highly adaptable framework for biomedical knowledge representation, enhancing and simplifying medical interpretation. Primary constraints stem from the utilization of Louvain as community detection algorithm, and the dependency on prior knowledge for downstream evaluation. As for future aspects to tackle, we can highlight the integration of longitudinal information, the production of synthetic datasets, and the careful assessment of the potential contributions of each relevant biomedical aspect for the task under study.

As we commented in section **3.2 of the introduction** (**Network-based representation of biomedical data**), graphs allow for the codification of multiple types of biological relationships existing across biomedical data, conveniently accommodating the model to the particular context under analysis.

Integrating these networks into a multilayer system provides a straightforward, natural approach to model relationships among different data levels, represented as network layers. Particularly, the gene multilayer network structure, and the detection of persistent community associations across the layers is a perfect fit for the analysis of biomedical knowledge, as gene relationships are expected to exhibit persistence across various biological facets. Incorporating gene-associated knowledge using multilayer networks simplifies the identification of genes that consistently associate layers, facilitates the clear interpretation of these relationships and enables the assessment of the contribution of each level of information.

2. Limitations of the presented research

2.1. Alternative community detection algorithms

The introduced advances from our approach present a series of limitations and future perspectives (**Table 3**) that we discuss in this section. The first issue we should comment about is the limitations associated with the usage of the Louvain algorithm as the main method for community detection. The **Leiden community detection algorithm** (197) has been recently introduced to solve a primary problem from the stochastic nature of the Louvain method: the identification of internally disconnected communities. Louvain only guarantee is to provide mutually exclusive communities, sometimes providing internally disconnected instances.

The Leiden algorithm solves this issue by adding a refinement step to the Louvain method before running Louvain algorithm's aggregation stage (**Introduction**, **Section 3.1.**). At each iteration, this refinement step checks for potential sub-strata within each detected community, thereby ensuring the existence of dense connectivity between nodes before merging. Additionally, the Leiden algorithm provides the means for a deeper exploration of the graph. This is achieved by allowing random merging of the nodes to one of the different possibilities that increase modularity, not only to the one providing the highest increase on modularity. This renders a broader analysis of the network. The new paradigm set by Leiden algorithm should then be acknowledged. While by the time research presented in Chapters 3 and for 4 started production Leiden approach was not available, its recent introduction entails an obvious need for adapting the analysis presented in those chapters to this novel update of modularity-based network community detection.

2.2. Ground truth availability and exploration of community size boundaries

The major strength of the methodology introduced in Chapter 4 is the refinement of the knowledge coming from the input personalized multi-omics data, making use of

the external biomedical information, integrated as a multilayer network. However, this approach presents two main problems. The first limitation is the evaluation procedure. Despite learning steps of the feature selection task are unsupervised, evaluation is dependent on the existence of a target feature for optimization. Although it provides an efficient approach to obtain lists of prioritized genes that best recover a particular label classification (in the particular case of Chapter 4, a previously known robust patient molecular stratification), such target label might not exist *a priori*, a typical case for rare disorders.

The second problem is related to the suitability of the lambda parameter (**λ**) for exploration of multilayer community trajectories. In the context of the specific application of multilayer network community trajectories for feature selection tasks, the utilization of this parameter provides an upper bound for prioritizing communities depending on the number of nodes of interest. While a reasonable strategy would be based on the user's choice, a heuristic to elucidate the best range of lambda values to analyze is not provided, thus potentially increasing the computational costs for the analysis.

Overcoming both limitations is critical for the potential application of this approach to any patient sample cohort, both in supervised and unsupervised scenarios. Focusing on solving these issues, a new version of this feature selection pipeline is currently under development, within the context of an ongoing collaborative effort with the Computational Biology Group of the Department of Biosystems Science and Engineering (D-BSSE) from ETH Zürich (Basel, Switzerland). Initial results of the approaches aiming to address both limitations mentioned earlier are presented and discussed in **Annex II**.

3. Future perspectives

3.1. Overlapping communities

An interesting perspective to address in multilayer network community detection is the analysis of overlapping modular structures. Classically, community detection algorithms such as Louvain or Leiden, center on the analysis of **disjoint communities** (i.e., each element belongs to a unique module) (198). Although we presented the considerable utility of such concept, many real networks present nodes that belong to multiple communities, which can be described to be **overlapping**.

Although the participation of biomedical interactors in several biological processes is well-known, only recently overlapping community detection has become a trending subject in network science (199), with some algorithms already adapting this functionality to the multiplex network level. Indeed, many biological interactors have roles in multiple biological processes, and therefore overlapping communities may be able to keep information that mutually-exclusive modules may lose (200). An interesting example is provided with the **Infomap** algorithm, a community detection procedure based on the minimization of an alternative quality function, the **map equation** (201), which describes the description length of flows obtained via random walks. Overall, Infomap provides a framework for the identification of multilayer community trajectories) by varying the time of the Markov-chain process. Selection of lower Markov time values favor the identification of community structures with higher number of lower size modules, while longer Markov times will favor lower number of modules with bigger sizes, as some of the nodes will not be encoded in the descriptors (202). Application of overlapping multilayer community detection to the medical field is yet in initial stages but could represent a way to overcome potential biases and biological information loss coming from the identification of disjoint communities. This way, the introduction of overlapping community identification to the research presented in chapters 3 and 4 should be a primary objective for future studies.

3.2. Integration of temporal multi-omics data

Modelling of multiple simultaneous aspects is an additional feature of multilayer network frameworks (190). One of the most prominent aspects for application is the analysis of temporal-varying networks. Many biological processes (e.g., gene expression, protein interaction, gene regulation and disease spreading) present changing dynamics over time, yet graph-based modelling of biological processes tends to focus on static frameworks (209), losing the information related to the temporal dimension.

Modelling of time-dependent biological processes has been subject of huge interest during the last decade: patient networks for example, allow for detection of individuals sharing molecular and phenotypic features (138). As features characterizing patient clusters may change over-time, patient clusters may present dynamical cluster identity. Tracking those variations has enable the inception of the concept of **disease trajectories**, which identify patients that follow similar evolution in their analyzed features (203), allowing for a finer recovery and understanding of longitudinal biomedical data (204). Extension of the methodologies presented in this PhD dissertation to include longitudinal information should be a main objective for future studies: identification of disease trajectories is key for obtaining finer knowledge on the processes underlying disease progression in complex disorders.

3.3. Synthetic data generation

Another potential application of the presented methodologies is on the generation of synthetic rare disease data (205). The usage of synthetic data has recently attracted the attention of researchers as it holds great potential to enhance the training of deep learning technologies in cases of highly unbalanced data (206). The approach presented for gene priorization (**Chapter 4**) may be helpful for the efficient selection of features for synthetic data generation, contributing to mitigate the problems caused by a high dimensionality of the search space to the small amount of available data.

3.4. Evaluation of layer contributions

While the presented methods have their major strength in discovering novel relationships between limited patient data, making use of external information, there is not an optimal *a prori* procedure for selecting such external data. As a result, the solution often relies on the usage of general resources and, when having a proper justification, specific meaningful resources for the disorder.

For example, in the article presented in Chapter 3, we modeled our multilayer network based on information from metabolomics, interactome and biological pathways. This is because CMS affects proteins mediating the signaling processes leading to normal neuromuscular junction development, thus giving a clear rationale for the focus on those three layers. Such explanations are also accompanied by an evaluation of the individual involvement of each layer to the detected multilayer communities.

Indeed, using additional target layers is always in the scope of integrative biomedical studies, and may provide promising new insights when such studies are of a more exploratory nature (for example, Buphamalai et al. (27) provided a remarkable analysis to find rare disease-specific patterns across a multiplex network including 46 unique database layers).

However, there is a general need for the network biology field for addressing the absence of a proper way for predicting the importance of each data layer, which may help provide new ways for careful selection of relevant external information, and most importantly, understand the way in which this information selection can affect the resulting topological structure of the network. This may prevent undesired effects coming from redundant biomedical information, potentially avoiding biased results.

4. Closing remarks: Implications for precision medicine

The research highlighted in this PhD thesis demonstrated the potential of network-based data integration to overcome the limitations associated with data-scarce scenarios. Providing solutions for the analysis of such biomedical scenarios, is crucial not only because of the impact it has for the patients suffering those conditions, but for the application of precision medicine to similar cases and other diseases presenting molecular relationships.

The knowledge stored in biomedical databases emerges as the pivotal target for addressing data-scarce scenarios. These knowledge resources serve as the starting point for uncovering previously undetected relationships among the available patient information, effectively assisting in the reconstruction of the 'missing information' that often impedes research in these specific cases (207).

In this sense, modelling and integration of external biomedical information with the patient-specific data emerges as the key solution for the data scarcity challenge in precision medicine. Although we have chosen multilayer networks for this integration due to their advantages in interpretability, there are several approaches that can efficiently accomplish the same task (as discussed in **Section 2** of the **Introduction**). However, it is important to note that all these equivalent frameworks must always be constructed upon the foundation provided by existing biomedical knowledge.

The presented research has multiple potential implications of multilayer network modelling for the successful application of precision medicine. These studies can be thought as examples of the benefits coming from the usage of external biomedical information for obtaining a more predictive, preventive, and personalized standard for clinical research, helping to establish the P4 medicine paradigm pursued by precision medicine.

Let us consider the applied research presented in Chapters 3 and 4 as illustrative examples. In terms of **predictive** power, the presented methodologies allowed us new target genes for predicting severe manifestations of CMS and the identification of clinical subtypes of medulloblastoma. These prioritized genes can thereby be used for the prognostic analysis and **prevention** of severe clinical evolution of these rare diseases. Moreover, the identified molecular interactors, additionally offer new insights into understanding the underlying reasons for the observed beneficial effects of drugs with no prior biomedical explanation. This scenario is of particular interest since discovery of **personalized treatments** for cohorts with limited availability of patients often relies on 'trial and error' testing, based on clinical expertise. In this sense, the methodologies introduced with this PhD Thesis hold huge potential for **drug repurposing**, which is crucial in a context where production of new, personalized treatments for cohorts presenting limited patient numbers is clearly hindered by its low cost-effectiveness.

An additional aspect to consider in rare disease research is the importance of the analysis of diseases presenting overlapping features. For example, the identified candidate genes in Chapter 3 were known to be involved in multiple disorders with features overlapping CMS manifestations (i.e., other rare myopathies, Ehler-Danlos syndrome, myasthenia gravis). This fact highlights the medical impact of the study of information exchange between rare diseases and conditions with higher prevalence. In this sense, the usage of the multilayer network system can help identify shared mechanisms between multiple diseases, providing new insights at multiple levels, such as the identification of new therapeutic targets and the analysis of the molecular biology underlying disease comorbidities.

Overall, the ongoing progress of multilayer network-based methodologies is pioneering a whole new set of approaches for the analysis of biomedical information in multiple precision medicine contexts, enhancing our understanding of the molecular biology underlying disease complexity.

Chapter 6

Conclusions

1. This PhD thesis shows the importance of the integration of the limited available data from small size cohorts with external biomedical information coming from relevant large-scale knowledge resources, to solve the challenge posed by data scarcity. This essential principle is the building foundation for the development of novel computational approaches focused on solving biomedical contexts constrained by patient-specific data availability.

2. Multilayer networks are a highly valuable methodology for the integration of patient-specific data with complementary information extracted from relevant biomedical resources, allowing for a more comprehensive understanding of the specific biomedical processes underlying the complex manifestations of a particular disease.

3. In the case of the study centered on the analysis of severity in Congenital Myasthenic Syndromes, our results show that topological analysis of multilayer networks (particularly, detection of persistent community associations across multiple levels of resolution) can be used for the discovery of novel genetic disease modifiers affecting relevant functional processes in the disease, enabling for a finer interpretation of the molecular biology leading to the differences in phenotypic severity in these rare diseases.

4. In this sense, we identified patient-specific gene variants on damaged neuromuscular junction (NMJ) interactors affecting patients with severe phenotypic of the disease. These genes are functionally connected to known causative processes. Furthermore, our study revealed compound heterozygous variants affecting post-synaptic acetylcholine receptor presentation, providing insights into potential therapeutic targets. Additionally, we uncovered a previously unknown functional role of the gene USH2A at the NMJ level, supported by experimental evidence using a zebrafish model.

5. In the case of the study on medulloblastoma, the introduction of a novel concept in complex graph theory -the multilayer community trajectory- shows the potential of

multilayer networks in feature selection tasks. The multilayer network, and more particularly, this new topological feature, offer a new way of achieving high performances in the recovery of known disease subtypes, minimizing the feature space employed for patient clustering.

6. The research studies presented in this PhD thesis also revealed new challenges to be addressed by the systems biology field. This includes the need of alternative community detection heuristics applicable to multilayers, new methods for time-dependent dynamical processes, and in general new approaches combining the power of multilayer networks with the state-of-the-art developments in AI, namely generative AI. In summary, the significant developments introduced can help to pioneer approaches for the application of precision medicine in multiple new biomedical research areas.

References

1. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018 May;19(5):299–310.

2. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012 Mar 16;148(6):1293–307.

3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

4. Mirnezami R, Nicholson J, Darzi A. https://doi.org/10.1056/NEJMp1114866. Massachusetts Medical Society; 2012 [cited 2022 Sep 5]. Preparing for Precision Medicine. Available from: https://www.nejm.org/doi/pdf/10.1056/NEJMp1114866

5. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Pers Med. 2013;10(6):565–76.

6. Hood L, Heath JR, Phelps ME, Lin B. Systems Biology and New Technologies Enable Predictive and Preventative Medicine. Science. 2004 Oct 22;306(5696):640–3.

7. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. 2020 Jun 1;121:103761.

8. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021 Jan 1;19:949–60.

9. Eddy S, Mariani LH, Kretzler M. Integrated multi-omics approaches to improve classification of chronic kidney disease. Nat Rev Nephrol. 2020 Nov;16(11):657–68.

10. Li CX, Wheelock CE, Sköld CM, Wheelock ÅM. Integration of multi-omics datasets enables molecular classification of COPD. Eur Respir J [Internet]. 2018 May 1 [cited 2022 Sep 7];51(5). Available from: https://erj.ersjournals.com/content/51/5/1701930

11. Ranek JS, Stanley N, Purvis JE. Integrating temporal single-cell gene expression modalities for trajectory inference and disease prediction. Genome Biol. 2022 Sep 5;23(1):186.

12. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.

13. Kaufmann M, Evans H, Schaupp AL, Engler JB, Kaur G, Willing A, et al. Identifying CNS-colonizing T cells as potential therapeutic targets to prevent progression of multiple sclerosis. Med. 2021 Mar 12;2(3):296-312.e8.

14. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022 Feb;22(2):114–26.

15. Zhu Z, Zhang S, Wang P, Chen X, Bi J, Cheng L, et al. A comprehensive review of the analysis and integration of omics data for SARS-CoV-2 and COVID-19. Brief Bioinform. 2022 Jan 1;23(1):bbab446.

16. Griggs RC, Batshaw M, Dunkle M, Gopal-Srivastava R, Kaye E, Krischer J, et al. Clinical research for rare disease: Opportunities, challenges, and solutions. Mol Genet Metab. 2009 Jan 1;96(1):20–6.

17. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet. 2013 Oct;14(10):681–91.

18. Villalón-García I, Álvarez-Córdoba M, Suárez-Rivero JM, Povea-Cabello S, Talaverón-Rey M, Suárez-Carrillo A, et al. Precision Medicine in Rare Diseases. Diseases. 2020 Dec;8(4):42.

19. Might M, Crouse AB. Why rare disease needs precision medicine—and precision medicine needs rare disease. Cell Rep Med [Internet]. 2022 Feb 15 [cited 2023 Jul 11];3(2). Available from: https://www.cell.com/cell-reports-medicine/abstract/S2666-3791(22)00030-1

20. Delavan B, Roberts R, Huang R, Bao W, Tong W, Liu Z. Computational drug repositioning for rare diseases in the era of precision medicine. Drug Discov Today. 2018 Feb 1;23(2):382–94.

21. Luque J, Mendes I, Gómez B, Morte B, López de Heredia M, Herreras E, et al. CIBERER: Spanish national network for research on rare diseases: A highly productive collaborative initiative. Clin Genet. 2022;101(5–6):481–93.

22. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. J Gen Intern Med. 2014 Aug;29(Suppl 3):780–7.

23. Plant D, Barton A. Machine learning in precision medicine: lessons to learn. Nat Rev Rheumatol. 2021 Jan;17(1):5–6.

24. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022 Jan 1;23(1):bbab454.

25. Liu B, Wei Y, Zhang Y, Yang Q. Deep Neural Networks for High Dimension, Low Sample Size Data. 2017;2287–93.

26. Mitani AA, Haneuse S. Small Data Challenges of Studying Rare Diseases. JAMA Netw Open. 2020 Mar 23;3(3):e201965.

27. Buphamalai P, Kokotovic T, Nagy V, Menche J. Network analysis reveals rare disease signatures across multiple levels of biological organization. Nat Commun. 2021 Nov 9;12(1):6306.

28. Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data. 2019 Oct 31;6(1):251.

29. Li Y, Chen L. Big Biological Data: Challenges and Opportunities. Genomics Proteomics Bioinformatics. 2014 Oct;12(5):187.

30. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan 1;107(1):1–8.

31. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct;431(7011):931–45.

32. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020 Jan 8;48(D1):D835–44.

33. McKusick VA. Mendelian Inheritance in Man and Its Online Version, OMIM. Am J Hum Genet. 2007 Apr 1;80(4):588–604.

34. Tomczak K, Czerwińska P, Wiznerowicz M. Review
The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Onkol. 2015;2015(1):68–77.

35. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017 Aug 18;9(1):75.

36. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207–10.

37. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013 Jan 1;41(D1):D991–5.

38. Papatheodorou I, Moreno P, Manning J, Fuentes AMP, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020 Jan 8;48(D1):D77–83.

39. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.

40. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016 Jan 4;44(D1):D184–9.

41. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011 Jan 1;39(suppl_1):D152–7.

42. Altelaar AFM, Munoz J, Heck AJR. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet. 2013 Jan;14(1):35–48.

43. Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff Å, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods. 2014 Jan;11(1):59–62.

44. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. PROTEOMICS. 2015;15(5–6):930–50.

45. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D480–9.

46. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

47. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D529–41.

48. Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D581–9.

49. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

50. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

51. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan 7;50(D1):D439–44.

52. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med. 2021 Oct;27(10):1666–9.

53. Harrigan GG, Goodacre R. Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis: Its Role in Biomarker Discovery and Gene Function Analysis. Springer Science & Business Media; 2003. 354 p.

54. Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol. 2018 Mar;36(3):272–81.

55. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, et al. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res. 2019 Jan 8;47(D1):D614–24.

56. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature. 2019 Jul;571(7766):489–99.

57. Heard E, Martienssen RA. Transgenerational Epigenetic Inheritance: myths and mechanisms. Cell. 2014 Mar 27;157(1):95–109.

58. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020 Jan 8;48(D1):D882–9.

59. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007 Oct;449(7164):804–10.

60. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The Integrative Human Microbiome Project. Nature. 2019 May;569(7758):641–8.

61. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012 Apr;13(4):260–70.

62. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017 Feb 13;2(5):1–7.

63. Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun. 2017 Oct 10;8(1):845.

64. Sanos SL, Bui VL, Mortha A, Oberle K, Heners C, Johner C, et al. RORγt and commensal microflora are required for the differentiation of mucosal interleukin 22–producing NKp46+ cells. Nat Immunol. 2009 Jan;10(1):83–91.

65. Sawa S, Lochner M, Satoh-Takayama N, Dulauroy S, Bérard M, Kleinschek M, et al. RORγt+ innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota. Nat Immunol. 2011 Apr;12(4):320–6.

66. Thaiss CA, Zmora N, Levy M, Elinav E. The microbiome and innate immunity. Nature. 2016 Jul;535(7610):65–74.

67. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nat Rev Microbiol. 2018 Mar;16(3):143–55.

68. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med. 2019 Jun;25(6):1012–21.

69. Wang P, Zhang S, He G, Du M, Qi C, Liu R, et al. microbioTA: an atlas of the microbiome in multiple disease tissues of Homo sapiens and Mus musculus. Nucleic Acids Res. 2023 Jan 6;51(D1):D1345–52.

70. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203–11.

71. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat Med. 2020 Jan;26(1):52–8.

72. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol. 2019 May 1;20(5):728–40.

73. Fedorov A, Longabaugh WJR, Pot D, Clunie DA, Pieper S, Aerts HJWL, et al. NCI Imaging Data Commons. Cancer Res. 2021 Aug 15;81(16):4188–93.

74. Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data. 2023 Jan 19;10(1):41.

75. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021 Jan 1;19:4538–58.

76. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019 Jan;18(1):41–58.

77. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D668–72.

78. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan 1;40(D1):D1100–7.

79. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023 Jan 6;51(D1):D1373–80.

80. Li F, Yin J, Lu M, Mou M, Li Z, Zeng Z, et al. DrugMAP: molecular atlas and pharma-information of all drugs. Nucleic Acids Res. 2023 Jan 6;51(D1):D1288–99.

81. Lilienfeld DE, Lilienfeld AM, Stolley PD. Foundations of Epidemiology. Oxford University Press; 1994. 390 p.

82. Carrell DS, Halgrim S, Tran DT, Buist DSM, Chubak J, Chapman WW, et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. Am J Epidemiol. 2014 Mar 15;179(6):749–58.

83. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207–17.

84. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005 Jan 1;33(suppl_1):D428–32.

85. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D687–92.

86. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D457–62.

87. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan 8;49(D1):D613–21.

88. Ates HC, Yetisen AK, Güder F, Dincer C. Wearable devices for the detection of COVID-19. Nat Electron. 2021 Jan;4(1):13–4.

89. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Pers Med. 2018 Sep;15(5):429–48.

90. Son D, Lee J, Qiao S, Ghaffari R, Kim J, Lee JE, et al. Multifunctional wearable devices for diagnosis and therapy of movement disorders. Nat Nanotechnol. 2014 May;9(5):397–404.

91. Iqbal SMA, Mahgoub I, Du E, Leavitt MA, Asghar W. Advances in healthcare wearable devices. Npj Flex Electron. 2021 Apr 12;5(1):1–14.

92. Melillo P, Izzo R, Orrico A, Scala P, Attanasio M, Mirra M, et al. Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. PLOS ONE. 2015 Mar 20;10(3):e0118504.

93. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data. 2020 May 25;7(1):154.

94. López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the Integration of Omics and Non-Omics Data. Genes. 2019 Mar;10(3):238.

95. Pazos F, Chagoyen M, Seoane P, Ranea JAG. CoMent: Relationships Between Biomedical Concepts Inferred From the Scientific Literature. J Mol Biol. 2022 Jun 15;434(11):167568.

96. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009 Jan 20;1(1):2.

97. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021 Jan 1;19:3735–46.

98. Oltvai ZN, Barabási AL. Life's Complexity Pyramid. Science. 2002 Oct 25;298(5594):763–4.

99. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. Npj Digit Med. 2020 Mar 26;3(1):1–5.

100. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021 Jan 5;12(1):124.

101. Pierre-Jean M, Deleuze JF, Le Floch E, Mauger F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief Bioinform. 2020 Dec 1;21(6):2011–30.

102. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet [Internet]. 2017 [cited 2022 Oct 19];8. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2017.00084

103. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, et al. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018 Oct 1;34(10):790–805.

104. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Interpretable machine learning: definitions, methods, and applications. Proc Natl Acad Sci. 2019 Oct 29;116(44):22071–80.

105. Lee D, Seung HS. Algorithms for Non-negative Matrix Factorization. In: Advances in Neural Information Processing Systems [Internet]. MIT Press; 2000 [cited 2022 Nov 10]. Available from: https://papers.nips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html

106. Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform. 2021 Mar 1;22(2):1604–19.

107. Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer. 2009 Aug;42(8):30–7.

108. Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics. 2018 Jun 19;19(1):233.

109. Kriebel AR, Welch JD. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun. 2022 Feb 9;13(1):780.

110. Hamamoto R, Takasawa K, Machino H, Kobayashi K, Takahashi S, Bolatkan A, et al. Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. Brief Bioinform. 2022 Jul 1;23(4):bbac246.

111. Lu X, Zhang K, Van Sant C, Coon J, Semizarov D. An algorithm for classifying tumors based on genomic aberrations and selecting representative tumor models. BMC Med Genomics. 2010 Jun 22;3(1):23.

112. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med Genomics. 2012 Dec 31;5(1):66.

113. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, et al. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Syst. 2019 May 22;8(5):380-394.e4.

114. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014 May 29;15(1):162.

115. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016 Jul 1;17(4):628–41.

116. Zhou G, Ewald J, Xia J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Res. 2021 Jul 2;49(W1):W476–82.

117. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016 Oct 10;2(1):1–13.

118. Revilla L, Mayorgas A, Corraliza AM, Masamunt MC, Metwaly A, Haller D, et al. Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis. PLOS ONE. 2021 Feb 8;16(2):e0246367.

119. Bayjanov JR, Doornbos C, Ozisik O, Shin W, Queralt-Rosinach N, Wijnbergen D, et al. Integrative analysis of CAKUT multi-omics data [Internet]. bioRxiv; 2023 [cited 2023 Jul 13]. p. 2023.06.29.547015. Available from: https://www.biorxiv.org/content/10.1101/2023.06.29.547015v1

120. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009 Nov 15;25(22):2906–12.

121. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013 Mar 12;110(11):4245–50.

122. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017 Jun 15;169(7):1327-1341.e23.

123. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative Subtype Discovery in Glioblastoma Using iCluster. PLOS ONE. 2012 Apr 23;7(4):e35236.

124. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun;14(6):e8124.

125. Lu J, Cannizzaro E, Meier-Abt F, Scheinost S, Bruch PM, Giles HAR, et al. Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia. Nat Cancer. 2021 Aug;2(8):853–64.

126. Forny P, Bonilla X, Lamparter D, Shao W, Plessl T, Frei C, et al. Integrated multi-omics reveals anaplerotic rewiring in methylmalonyl-CoA mutase deficiency. Nat Metab. 2023 Jan;5(1):80–95.

127. Janocha K, Czarnecki WM. On Loss Functions for Deep Neural Networks in Classification [Internet]. arXiv; 2017 [cited 2023 Jun 27]. Available from: http://arxiv.org/abs/1702.05659

128. Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. p. 765–9.

129. Lee J, Liu C, Kim J, Chen Z, Sun Y, Rogers JR, et al. Deep learning for rare disease: A scoping review. J Biomed Inform. 2022 Nov 1;135:104227.

130. Hira MT, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Sci Rep. 2021 Mar 18;11(1):6265.

131. Mersha TB, Afanador Y, Johansson E, Proper SP, Bernstein JA, Rothenberg ME, et al. Resolving Clinical Phenotypes into Endotypes in Allergy: Molecular and Omics Approaches. Clin Rev Allergy Immunol. 2021 Apr;60(2):200–19.

132. Gao M, Liu S, Qi Y, Guo X, Shang X. GAE-LGA: integration of multi-omics data with graph autoencoders to identify lncRNA–PCG associations. Brief Bioinform. 2022 Oct 27;bbac452.

133. Bodein A, Scott-Boyer MP, Perin O, Lê Cao KA, Droit A. Interpretation of network-based integration from multi-omics longitudinal data. Nucleic Acids Res. 2022 Mar 21;50(5):e27.

134. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018 Nov 27;19(6):1370–81.

135. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017 May;545(7655):505–9.

136. Milenković T, Lai J, Pržulj N. GraphCrunch: A tool for large network analyses. BMC Bioinformatics. 2008 Jan 30;9(1):70.

137. Staudt CL, Sazonovs A, Meyerhenke H. NetworKit: A tool suite for large-scale complex network analysis. Netw Sci. 2016 Dec;4(4):508–30.

138. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014 Mar;11(3):333–7.

139. Wang C, Lue W, Kaalia R, Kumar P, Rajapakse JC. Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. Sci Rep. 2022 Sep 14;12(1):15425.

140. Narayana JK, Aogáin MM, Ali NABM, Tsaneva-Atanasova K, Chotirmall SH. Similarity network fusion for the integration of multi-omics and microbiomes in respiratory disease. Eur Respir J [Internet]. 2021 Aug 1 [cited 2022 Nov 15];58(2). Available from: https://erj-ersjournals-com.sire.ub.edu/content/58/2/2101016

141. Mac Aogáin M, Narayana JK, Tiew PY, Ali NABM, Yong VFL, Jaggi TK, et al. Integrative microbiomics in bronchiectasis exacerbations. Nat Med. 2021 Apr;27(4):688–99.

142. Eng SWM, Olazagasti JM, Goldenberg A, Crowson CS, Oddis CV, Niewold TB, et al. A Clinically and Biologically Based Subclassification of the Idiopathic Inflammatory Myopathies Using Machine Learning. ACR Open Rheumatol. 2020 Feb 10;2(3):158–66.

143. Guo Y, Zheng J, Shang X, Li Z. A Similarity Regression Fusion Model for Integrating Multi-Omics Data to Identify Cancer Subtypes. Genes. 2018 Jul;9(7):314.

144. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [Internet]. arXiv; 2013 [cited 2023 Jan 2]. Available from: http://arxiv.org/abs/1301.3781

145. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2014 [cited 2023 Jan 2]. p. 701–10. (KDD '14). Available from: https://doi.org/10.1145/2623330.2623732

146. Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. Bioinformatics. 2017 Jul 15;33(14):i190–8.

147. Pio-Lopez L, Valdeolivas A, Tichit L, Remy É, Baudot A. MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach. Sci Rep. 2021 Apr 22;11(1):8794.

148. Dickison ME, Magnani M, Rossi L. Multilayer Social Networks. Cambridge University Press; 2016. 215 p.

149. De Domenico M, Granell C, Porter MA, Arenas A. The physics of spreading processes in multilayer networks. Nat Phys. 2016 Oct;12(10):901–6.

150. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56–68.

151. Barabási AL. Network Science by Albert-László Barabási [Internet]. 2016 [cited 2022 Dec 12]. Available from: http://networksciencebook.com/

152. Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977;40(1):35–41.

153. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959 Dec;1(1):269–71.

154. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. 2019 Feb 1;35(3):497–505.

155. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015 Feb 20;347(6224):1257601.

156. Ranea JAG, Perkins J, Chagoyen M, Díaz-Santiago E, Pazos F. Network-Based Methods for Approaching Human Pathologies from a Phenotypic Point of View. Genes. 2022 Jun 17;13(6):1081.

157. Chagoyen M, Ranea JAG, Pazos F. Applications of molecular networks in biomedicine. Biol Methods Protoc. 2019 Jan 1;4(1):bpz012.

158. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008 Oct;2008(10):P10008.

159. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004 Feb 26;69(2):026113.

160. Yang Z, Algesheimer R, Tessone CJ. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. Sci Rep. 2016 Aug 1;6:30750.

161. Arenas A, Fernández A, Gómez S. Analysis of the structure of complex networks at different resolution levels. New J Phys. 2008 May;10(5):053039.

162. Greene LH, Higman VA. Uncovering Network Systems Within Protein Structures. J Mol Biol. 2003 Dec 5;334(4):781–91.

163. Zhang Q, Ma C, Gearing M, Wang PG, Chin LS, Li L. Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. Acta Neuropathol Commun. 2018 Mar 1;6(1):19.

164. Azeloglu EU, Iyengar R. Signaling Networks: Information Flow, Computation, and Decision Making. Cold Spring Harb Perspect Biol. 2015 Apr;7(4):a005934.

165. Altieri DC. Survivin, cancer networks and pathway-directed drug discovery. Nat Rev Cancer. 2008 Jan;8(1):61–70.

166. Amara A, Frainay C, Jourdan F, Naake T, Neumann S, Novoa-del-Toro EM, et al. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. Front Mol Biosci [Internet]. 2022 [cited 2023 Jun 20];9. Available from: https://www.frontiersin.org/articles/10.3389/fmolb.2022.841373

167. Schmid R, Petras D, Nothias LF, Wang M, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun. 2021 Jun 22;12(1):3832.

168. Jourdan F, Breitling R, Barrett MP, Gilbert D. MetaNetter: inference and visualization of high-resolution metabolomic networks. Bioinformatics. 2008 Jan 1;24(1):143–5.

169. Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. Nat Commun. 2019 Mar 13;10(1):1197.

170. Kruse K, Sewitz S, Babu MM. A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. Nucleic Acids Res. 2013 Jan 1;41(2):701–10.

171. Pancaldi V, Carrillo-de-Santa-Pau E, Javierre BM, Juan D, Fraser P, Spivakov M, et al. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol. 2016 Jul 8;17(1):152.

172. Choy MK, Javierre BM, Williams SG, Baross SL, Liu Y, Wingett SW, et al. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. Nat Commun. 2018 Jun 28;9(1):2526.

173. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A. 2007 May 22;104(21):8685–90.

174. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms–disease network. Nat Commun. 2014 Jun 26;5(1):4212.

175. Urda-García B, Sánchez-Valle J, Lepore R, Valencia A. Patient stratification reveals the molecular basis of disease comorbidities [Internet]. medRxiv; 2021 [cited 2023 Jun 27]. p. 2021.07.22.21260979. Available from: https://www.medrxiv.org/content/10.1101/2021.07.22.21260979v1

176. Faner R, Cruz T, López-Giraldo A, Agustí A. Network medicine, multimorbidity and the lung in the elderly. Eur Respir J. 2014 Sep;44(3):775–88.

177. Sánchez-Valle J, Tejero H, Fernández JM, Juan D, Urda-García B, Capella-Gutiérrez S, et al. Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. Nat Commun. 2020 Jun 5;11(1):2854.

178. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet [Internet]. 2019 [cited 2023 Jun 18];10. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.00317

179. Faner R, Cruz T, Casserras T, López-Giraldo A, Noell G, Coca I, et al. Network Analysis of Lung Transcriptomics Reveals a Distinct B-Cell Signature in Emphysema. Am J Respir Crit Care Med. 2016 Jun 1;193(11):1242–53.

180. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020 Feb;17(2):147–54.

181. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLOS ONE. 2010 Sep 28;5(9):e12776.

182. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017 Nov;14(11):1083–6.

183. Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. Npj Syst Biol Appl. 2019 Apr 23;5(1):1–12.

184. Didier G, Brun C, Baudot A. Identifying communities from multiplex biological networks. PeerJ. 2015 Dec 22;3:e1525.

185. Amoroso N, La Rocca M, Bellantuono L, Diacono D, Fanizzi A, Lella E, et al. Deep Learning and Multiplex Networks for Accurate Modeling of Brain Age. Front Aging Neurosci [Internet]. 2019 [cited 2022 Nov 28];11. Available from: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00115

186. Hajiseyedjavadi S, Lin YR, Pelechrinis K. Learning embeddings for multiplex networks using triplet loss. Appl Netw Sci. 2019 Dec;4(1):1–16.

187. Wang XW, Chen Y, Liu YY. Link Prediction through Deep Generative Model [Internet]. bioRxiv; 2020 [cited 2022 Nov 28]. p. 247577. Available from: https://www.biorxiv.org/content/10.1101/247577v4

188. Cantini L, Medico E, Fortunato S, Caselle M. Detection of gene communities in multi-networks reveals cancer drivers. Sci Rep. 2015 Dec 7;5(1):17386.

189. Trimbour R, Deutschmann IM, Cantini L. Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS [Internet]. bioRxiv; 2023 [cited 2023 Sep 13]. p. 2023.06.09.543828. Available from: https://www.biorxiv.org/content/10.1101/2023.06.09.543828v1

190. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. J Complex Netw. 2014 Sep 1;2(3):203–71.

191. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature. 2010 Aug;466(7307):761–4.

192. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science. 2010 May 14;328(5980):876–8.

193. Magnani M, Hanteer O, Interdonato R, Rossi L, Tagarelli A. Community Detection in Multiplex Networks. ACM Comput Surv. 2021 May 8;54(3):48:1-48:35.

194. Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. Mol Oncol. 2021;15(4):817–29.

195. Núñez-Carpintero I, O'Connor E, Rigau M, Bosio M, Azuma Y, Topf A, et al. Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes [Internet]. bioRxiv; 2023 [cited 2023 Jan 21]. p. 2023.01.19.524736. Available from: https://www.biorxiv.org/content/10.1101/2023.01.19.524736v1

196. Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. iScience [Internet]. 2021 Apr 23 [cited 2021 Nov 4];24(4). Available from: https://doi.org/10.1016/j.isci.

197. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):5233.

198. Fortunato S. Community detection in graphs. Phys Rep. 2010 Feb 1;486(3):75–174.

199. Xie J, Kelley S, Szymanski BK. Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv. 2013 Aug 30;45(4):43:1-43:35.

200. De Domenico M, Lancichinetti A, Arenas A, Rosvall M. Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems. Phys Rev X [Internet]. 2014 Aug 13 [cited 2021 Feb 17];5(1). Available from: http://arxiv.org/abs/1408.2925

201. Edler D, Bohlin L, Rosvall M. Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap. Algorithms. 2017 Dec;10(4):112.

202. Kheirkhahzadeh M, Lancichinetti A, Rosvall M. Efficient community detection of network flows for varying Markov times and bipartite networks. Phys Rev E. 2016 Mar 9;93(3):032309.

203. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014 Jun 24;5(1):4022.

204. Lambert J, Leutenegger AL, Jannot AS, Baudot A. Tracking patient clusters over time enables to extract all the information available in the medico-administrative databases [Internet]. medRxiv; 2022 [cited 2023 Jan 24]. p. 2022.08.05.22278468. Available from: https://www.medrxiv.org/content/10.1101/2022.08.05.22278468v1

205. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021 Jun;5(6):493–7.

206. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. Npj Digit Med. 2020 Nov 9;3(1):1–13.

207. Fernandez-Novo S, Pazos F, Chagoyen M. Rare disease relations through common genes and protein interactions. Mol Cell Probes. 2016 Jun 1;30(3):178–81.

Annex I

Statement from the co-directors on the contributions of the PhD candidate

The impact factor of the journals, as well as their ranking (as of February the 2nd, 2023) including the presented research (or under peer-review process, which is the case for Chapter 3 and Annex III) presented in this PhD Thesis can be found, for each corresponding chapter, under the section named as '**Publication Record**'. The information is extracted from Clarivate's Journal Citation Reports ™.

Additionally, the particular work performed by the PhD candidate in each publication is stated under the section '**Contribution of the PhD candidate**' for each of the article chapters.

The PhD co-directors, Dr. Davide Cirillo and Prof. Alfonso Valencia, acknowledge this information, as well as the fact that none of the presented articles have been previously used by any co-authors for the obtention of a PhD degree.

Signatures of the co-directors

Dr. Davide Cirillo Prof. Alfonso Valencia

Annex II

Addressing ground truth availability and enhancing the exploration of community size boundaries

This annexed chapter presents and discusses the initial results towards solving the limitations related to the pipeline introduced in **Chapter 4** and commented in the **Section 2.2.** of the **Discussion** of this PhD thesis.

This ongoing collaborative effort, started during an internship period awarded with an EMBO short-term fellowship, aims to address the main limitations of the mentioned approach. This collaboration involves the Computational Biology Group of the Department of Biosystems Science and Engineering (D-BSSE) from ETH Zürich (Basel, Switzerland) as hosting institution. During the early stages of the internship, the primary focus of the work was to leverage the pipeline presented on Chapter 2 for the analysis of a hepatocellular carcinoma (HCC) cohort provided by the host group (1,2).

To address the first of the previously mentioned limitations (i.e., **the dependency on the availability of a ground truth patient stratification**), the pipeline is applied in a fully unsupervised manner, utilizing an unsupervised clustering of the patient cohort instead of relying on the already known sample classification. The evaluation process involves two stages: analyzing the accuracy of the unsupervised clustering recovery, and secondly, conducting the comparison of the enriched functional activities shared between both clustering structures. The original clustering of the target cohort, described in (2) and based on Bayesian mixture modeling, is compared with the optimized unsupervised clustering derived from our pipeline, which takes advantage of the multilayer network structure analysis.

With this goal in mind, the optimization pipeline was applied to an unsupervised hierarchical clustering of the samples derived from patient multi-omics data. Available omics data for the cohort include proteomics, RNA-seq, phosphoproteomics and whole exome sequencing, considering single nucleotide variants (SNVs) and copy number variation (CNVs). This way, the methodology performance is also tested using mixed data types.

Using Monte-Carlo based bootstrapping resampling (3), the unsupervised clustering identified four significant clusters. Subsequently, we applied the optimization pipeline from Chapter 4 to this clustering structure, identifying optimal values of 3 and 33, respectively, for parameters **θ** and **λ**. The results yielded high accuracy (approximately 92%) on the recovery of the unsupervised clustering, with a gene dimensionality reduction of approximately 65% (**Figure 1A**). Strikingly, network enrichment analysis (4) revealed that the prioritized genes associated to each cluster exhibited enrichments in functional characteristics related to those previously described for the clusters detected in (2) (**Figure 1B-C**). However, although the pipeline demonstrated its capability to recover meaningful disease knowledge for the identified HCC clusters, it also highlighted the challenge of predicting the optimal range of λ values for exploration.

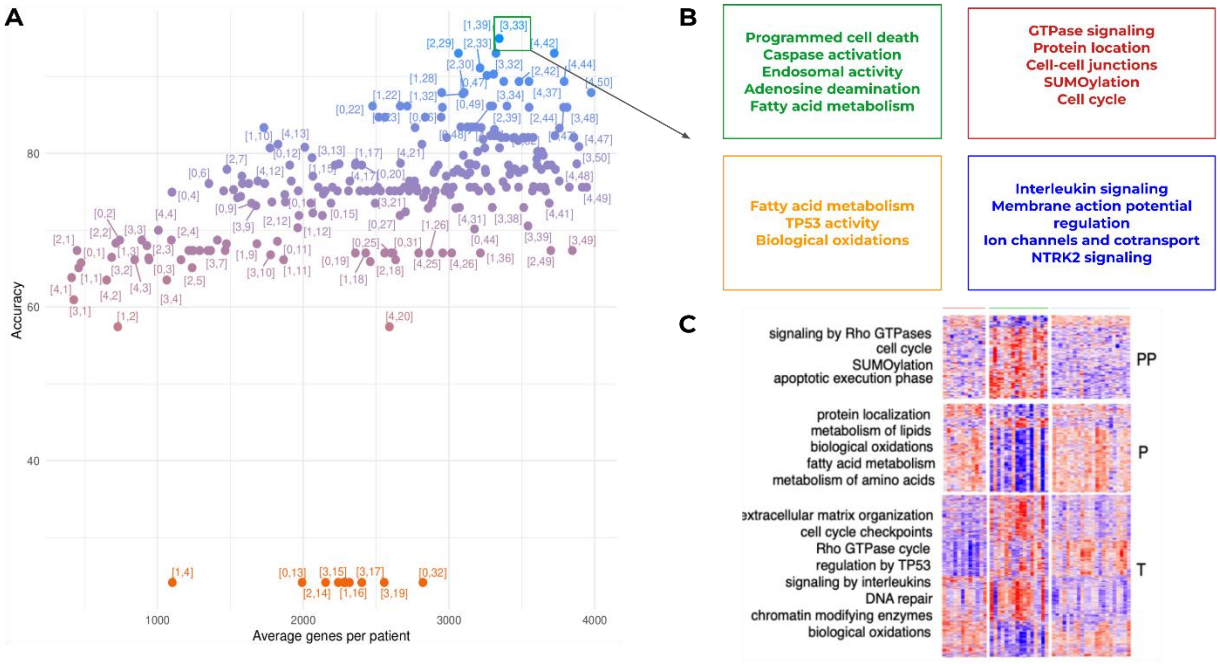


Figure 1. (A) Parameter optimization for the HCC sample cohort. Values next to each point highlight the corresponding **[θ,λ]** combination. Y axis represents the accuracy for the recovery of the unsupervised hierarchical clustering structure, X axis indicates the average number of altered genes per patient sample used for the unsupervised hierarchical clustering. (**B**) Enriched functional pathways for each patient sample cluster obtained with the optimal values for dimensionality reduction. (**C**) Enriched pathways for the clusters observed in (2).

To enhance the methodology and explore the space of the identified multilayer community trajectories further, our current work is focused on developing a new version of the pipeline where we evaluate the importance of the found trajectories at each value of the θ (which controls the changes in modularity resolution, with lower values representing increasingly more intimate association at the multilayer network level) for each sample under optimization. The aim is to prioritize genes found within the most valuable multilayer community trajectories for the given cohort.

As a way of replacing the exploration introduced with the λ parameter, the new iteration of the algorithm is based on distance assessment using the Random Walk with Restart (RWR) metric. Our approach is based on the removal of altered patient's genes that are found in the multilayer community trajectories existing at a particular level of resolution. We generate patient-specific multilayer networks where the altered genes of the sample under optimization are absent, and evaluate how the connectivity of the genes of the same trajectories (that are not altered in the patient) are affected. This assessment is computed as the mean of the RWR probability ratio scores of the trajectory genes (**Ω**), between the original multilayer network and the patient-specific multilayer network (**Figure 2**).

With the implementation of this novel approach, we are now able to score the entire space of multilayer community trajectories existing at a specific θ value. This allows to filter genes based on the distribution of the mean **Ω** RWR score for the corresponding multilayer community trajectories, thereby effectively exploring tentative distribution thresholds (**Figure 3**).

Strikingly, the new pipeline highlighted 2 and 70% as the optimal combination of parameters for **θ** and the **Ω** RWR score threshold, respectively. With these optimal values, accuracy for the recovery of the unsupervised hierarchical clustering increased to approximately 94%, and most importantly, the gene dimensionality reduction also increased to about 75% (**Figure 3**).

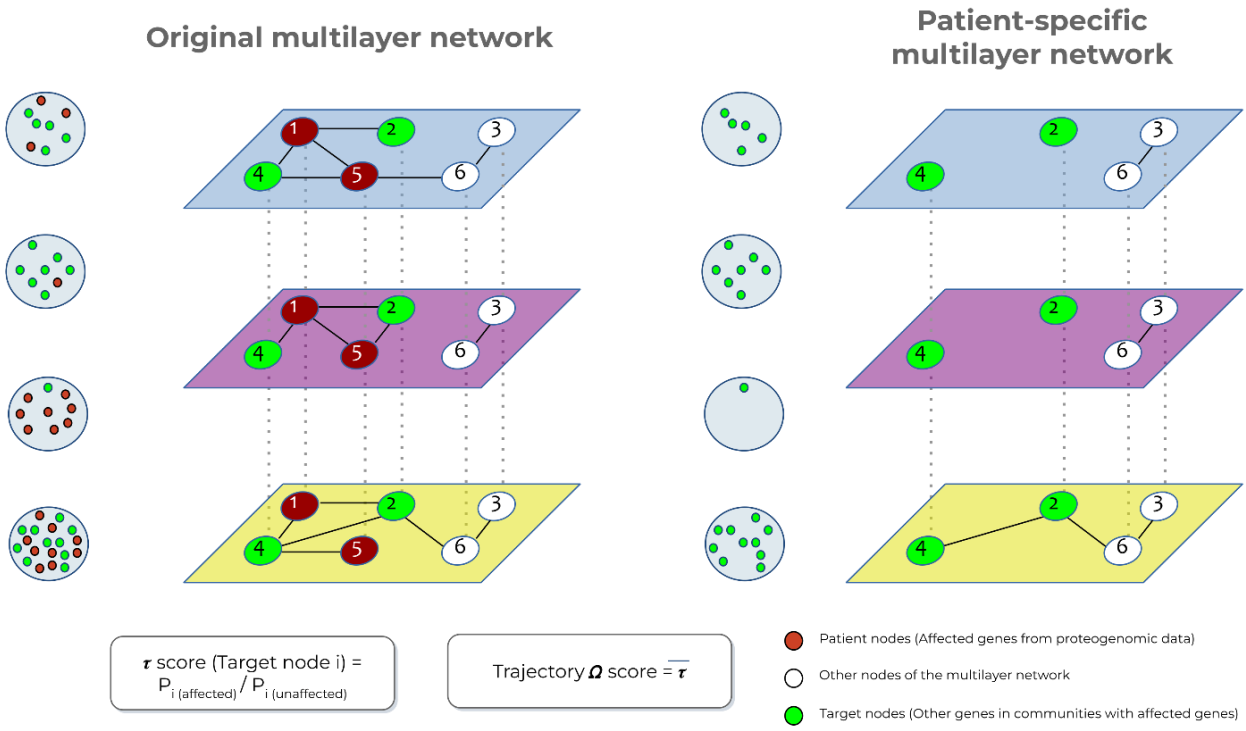


Figure 2. Schematic representation of the patient-wise RWR multilayer network analysis. The RWR score for a given node represents the probability of reaching the node when performing random walks through the multilayer network from a given set of starting nodes, which are called **seeds**. In our analysis, the target nodes under evaluation are the genes found in the same multilayer community trajectories as the genes that are found to be altered in any of the patient's proteogenomic data (**nodes depicted in green**). Seed nodes for the RWR include the union of the altered genes from the patient under optimization (**nodes depicted in red**) and the other genes from the multilayer network that happen to be outside of multilayer community trajectories with altered genes (**nodes depicted in white**). For each target node, we obtain its RWR probability score. In parallel, we build a second multilayer network where the altered genes from the patient (**nodes in red**) are removed, and recompute the RWR probability score for the target nodes. We compute the ratio of the RWR score (**τ**) for each target node between both multilayer networks, and the overall RWR score for each trajectory (**Ω**) as the mean of the **τ** score of each target node of the trajectory.

Summing up, development of the new version of our previous feature selection gene pipeline is independent of previously existent patient sub-stratifications for the evaluation of the optimization procedure, effectively making the whole process an unsupervised technique.

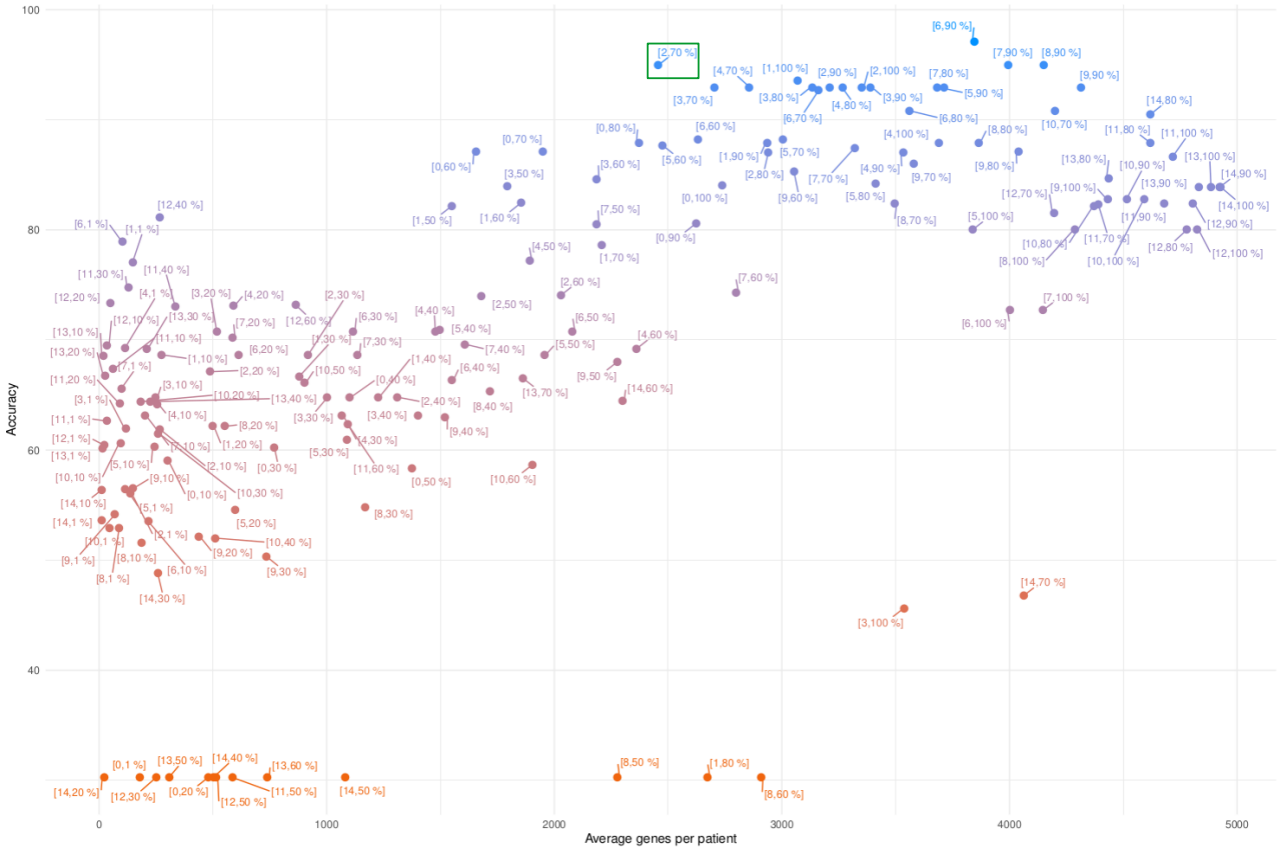


Figure 3. Parameter optimization results for the patient-wise RWR multilayer network analysis. Values next to each point highlight the corresponding [θ,**Ω** threshold] combination. The best performing combination is highlighted in green.

Moreover, we defined a new way for evaluating the importance of the multilayer community trajectories for a given cohort in this procedure, exploring the whole space of functional implication of the genes from each multilayer community trajectory in a personalized manner. Both facts are important milestones for this procedure because they enhance the potential general usage for any patient sample cohort with available personalized omics data, both in supervised and unsupervised scenarios.

As for future work to be undertaken, the main priority is on reducing the computational costs of the new pipeline. Ideally, this would allow for general parallelization of the approach in HPC environments with minimal resource usage, considerably easing and accelerating the generation and RWR evaluation of the patient-wise multilayer networks, which is the main current drawback of the procedure.

References

1. Ng CKY, Dazert E, Boldanova T, Coto-Llerena M, Nuciforo S, Ercan C, et al. Proteogenomic characterization of hepatocellular carcinoma. bioRxiv; 2021p. 2021.03.05.434147. Available from: https://www.biorxiv.org/content/10.1101/2021.03.05.434147v1

2. Suter P, Dazert E, Kuipers J, Ng CKY, Boldanova T, Hall MN, et al. Multi-omics subtyping of hepatocellular carcinoma patients using a Bayesian network mixture model. PLOS Comput Biol. 2022 Sep 6;18(9):e1009767.

3. Kimes PK, Liu Y, Hayes DN, Marron JS. Statistical Significance for Hierarchical Clustering. Biometrics. 2017 Sep;73(3):811–21.

4. Signorelli M, Vinciotti V, Wit EC. NEAT: an efficient network enrichment analysis test. BMC Bioinformatics. 2016 Sep 5;17(1):352.

Annex III

The PENGUIN approach to reconstruct protein interactions at enhancer-promoter regions and its application to prostate cancer

Publication Record

This annexed chapter presents the preprint of the original research article *'The PENGUIN approach to reconstruct protein interactions at enhancer-promoter regions and its application to prostate cancer'* currently under revision at the Nature Communications journal (2021 Journal Impact factor: 17.694; Q1 in the Multidisciplinary sciences field. Rank: 6/74) (2022 Journal Impact factor: 16.6; Q1 in the Multidisciplinary sciences field. Rank: 6/73).

Co-authors & affiliations

Alexandros Armaos[1,*], François Serra[2,3*], Iker Núñez-Carpintero[2], Ji-Heui Seo[4], Sylvan C. Baca[4], Stefano Gustincich[1], Alfonso Valencia[2,5], Matthew L. Freedman[4], Davide Cirillo[#,2], Claudia Giambartolomei[#,1], Gian Gaetano Tartaglia[#,1,5,6]

1. Istituto Italiano di Tecnologia, Via Enrico Melen 83, Building B, 7th floor, 16152 Genoa, Italy

2. Barcelona Supercomputing Center, Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain

3. Josep Carreras Leukaemia Research Institute, Badalona, Barcelona, Spain

4. Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA 02215, USA.

5. ICREA - Institució Catalana de Recerca I Estudis Avançats, Pg. Lluís Companys 23, 08010 Barcelona, Spain

6. Sapienza University Rome, Biology and Biotechnologies Department C. Darwin, P.le Aldo Moro 5, 00185

* These authors contributed equally

Corresponding author

Current reference

Armaos A, Serra F, Núñez-Carpintero I, Seo JH, Baca SC, Gustincich S, et al. The PENGUIN approach to reconstruct protein interactions at enhancer-promoter regions and its application to prostate cancer. *bioRxiv*; 2023. p. 2022.10.20.512998.

Available from: https://www.biorxiv.org/content/10.1101/2022.10.20.512998v2

Contribution of the PhD Candidate

The PhD Candidate, as second main author of the publication, suggested the edge-based network clustering approach undertaken (performed by Dr François Serra). Additionally, the PhD Candidate designed and performed, in consultation with the main co-authors (Dr. Alexandros Armaos and Dr. François Serra) and his co-supervisors, the protein specificity significance analysis described in section **2.4.** (**Involvement of E-P protein interactomes in tumor-related functional processes**) (**Methods: Enriched intermediate nodes within each cluster**) as well as the functional gene set enrichment analysis described in the same section (**Methods: Functional gene set enrichment analysis**).

Article abstract

Here we introduce Promoter-ENhancer-GUided Interaction Networks (PENGUIN), a method to uncover protein-protein interaction (PPI) networks at enhancer-promoter contacts. By integrating H3K27ac-HiChIP data and tissue-specific PPI information, PENGUIN enables cluster enhancers-promoter PPI networks (EPINs) and pinpoint actionable factors.

Validating PENGUIN in cancer (LNCaP) and benign (LHSAR) prostate cell lines, we observed distinct CTCF-enriched clusters, which identifies diverse chromatin conformations. In LNCaP, we found an EPIN cluster enriched with oncogenes and prostate cancer-associated SNPs. We uncovered a total of 208 SNPs in LNCaP EPINs and used CRISPR/Cas9 knockout and RNAi screens to confirm their relevance.

PENGUIN's application in prostate cancer demonstrates its potential for studying human diseases. The approach allows exploration in different cell types and combinations of GWAS data, offering promising avenues for future investigations. In conclusion, PENGUIN provides valuable insights into the interplay between enhancer-promoter interactions and PPI networks, facilitating the identification of relevant genes and potential intervention targets.

1. Introduction

Enhancer-promoter (E-P) interactions play a crucial role in orchestrating gene expression and ensuring the proper regulation of cellular processes. DNA-binding proteins (DBPs), including transcription factors (TFs), act as key players in this regulatory network by binding to enhancers and bridging additional protein interactions between enhancers and promoters. In this work we define Enhancers-Promoter protein-protein Interaction Network (EPIN) as the local interactome connecting a single promoter with all its interacting enhancers. EPIN interactions are facilitated by various types of intermediate proteins, such as co-activators (e.g., mediators), chromatin structural proteins (e.g., cohesin), and noncoding RNA-binding proteins.

While protein-protein interactions (PPIs) have been extensively studied (1,2), the integration of chromatin architecture information, specifically through chromosome conformation capture (3Clike) techniques, with PPI analysis is still in its early stages. Joint investigations of chromatin loops and PPIs are crucial for prioritizing functional interactions (3). However, it is important to note that many of these studies often lack the necessary biological context at various levels.

As of today, the characterization of context specific intermediate PPIs involved in disease pathways and their association with DBPs remains largely unanswered (4). Previous studies have highlighted the significance of disrupted E-P loops in several human disorders (5–7). In cancer, enhancers are frequently subject to sequence and structural variations, leading to the dysregulation of TFs and chromatin modifiers, which contribute to oncogenesis (8). Consequently, targeting these enhancer-driven mechanisms holds great promise for therapeutic interventions in cases such as Prostate Cancer (PrCa) (9). In this context, advanced techniques such as HiC and its derivative HiChIP (10), in combination with ChIP-seq, could enable the identification and characterization of specific chromatin interactions between enhancers and promoters. In particular, H3K27ac-HiChIP has emerged as a powerful tool designed

to detect and amplify E-P interactions and has been successfully employed to uncover susceptibility genes associated with cancer, including PrCa (11).

To characterize protein interactions that take place at the E-P contacts, we developed the Promoter-ENhancer-GUided Interaction Networks (PENGUIN) approach. For each promoter annotated in the genome and covered by at least one HiChIP interaction, PENGUIN builds an EPIN by integrating several sources of information: (1) high-resolution chromatin interaction maps enriched for a marker of active E-P activity (H3K27ac-HiChiP); (2) tissue-specific physical nuclear PPIs; (3) high-quality curated binding motifs of protein-DNA interactions; (4) tissue specific gene expression, used as a filter of protein data.

To prove the usefulness of our PENGUIN approach, we applied it to uncover EPINs in a PrCA cell line, androgen-sensitive human prostate adenocarcinoma cells (LNCaP), and validate our findings in comparison to a benign prostate epithelial cell line (LHSAR). PrCa is the 2nd most common cancer in men (12). Its distinct hormone-dependent nature is characterized by high expression and frequent genetic amplification of AR. AR is a regulator of homeostasis and proteases transcription, such as KLK3 encoding PSA (Prostate-Specific Antigen).

AR gene is also a principal therapeutically targeted oncogene in PrCa (13). Increased genetic instability resulting in chromosomal rearrangements and high frequency of mutations are deemed indicative of PrCa aggressiveness (14) for which there is need of ad hoc treatments (15). Recurrent mutations in *FOXA1*, involved in prostate organogenesis and regulator of AR transcription, have been observed in several populations (16,17). Hundreds of PrCa-associated single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (GWAS), including genomic regions within tumor suppressor genes and oncogenes, such as *MYC* (18). However, the functional relationship between most of these SNPs and PrCa pathophysiology is unknown. This missing part of the picture, together with the growing evidence of abnormal transcriptional programs driven by genetic instability,

led us to investigate the role of chromatin architecture in PrCa. In particular, we focused on the nuclear proteins potentially involved in transcriptional regulation through the interaction of promoters and non-coding regulatory elements, enhancers.

By clustering together promoters with similar EPIN structures, PENGUIN identified 273 promoters whose genes are enriched in PrCa fine-mapped SNPs, known PrCa oncogenes, and ChIP-Seq-validated binding sites of transcriptional repressor CTCF. The proteins that populate such EPINs constitute putative PrCa-related factors, some of which have not been previously described to be associated with PrCa SNPs or oncogenes. Moreover, the EPINs detected by PENGUIN enable the characterization of distinct molecular cascades enriched in PrCa SNPs at E-P contacts. These represent new potential molecular targets in PrCa that cannot be identified through conventional analytical procedures, such as E-P contacts and GWAS overlap. To explore our results we made a dedicated server available at https://penguin.life.bsc.es/.

Our methodology, focusing at the specific EPIN resolution level, reveals a new relation between 3D genome conformation and disease phenotype. This new relation allows PENGUIN to propose new directions in the molecular characterization of chromatin interactions as well as in the definition of potential targets for molecular screening towards disease treatment.

2. Results

2.1. The PENGUIN framework

PENGUIN builds EPINs by leveraging multiple sources of information. Specifically, it integrates diverse datasets:

(1) High-resolution chromatin interaction maps that capture active promoter-enhancer interactions, highlighting the dynamic nature of gene regulation.

(2) Tissue specific physical nuclear protein-protein interactions (PPIs), enabling the exploration of the intricate molecular associations within the nucleus.

(3) Curated binding motifs of protein-DNA interactions, providing insights into the specific interactions between proteins and DNA.

(4) Gene expression levels, identifying active elements with the interaction networks (**Figure 1**).

With this comprehensive approach, PENGUIN reconstructs EPINs by clustering enhancers that interact with the same promoter based on PPIs. Each EPIN consists of three distinct types of nodes: promoter-bound nodes, encompassing proteins with DNA binding motifs present in the promoter region; enhancer-bound nodes, comprising proteins with DNA binding motifs in the enhancer sequences; and intermediate nodes, representing proteins that interact with either the promoter-bound or enhancer-bound nodes but lack direct DNA binding motifs on the promoter or enhancers.

By integrating these diverse nodes, PENGUIN provides a holistic view of the intricate molecular landscape within EPINs. This approach enables the exploration of the interplay between DNA-binding proteins, enhancers, and intermediate proteins, shedding light on the regulatory mechanisms that shape gene expression and ultimately influence cellular functions.

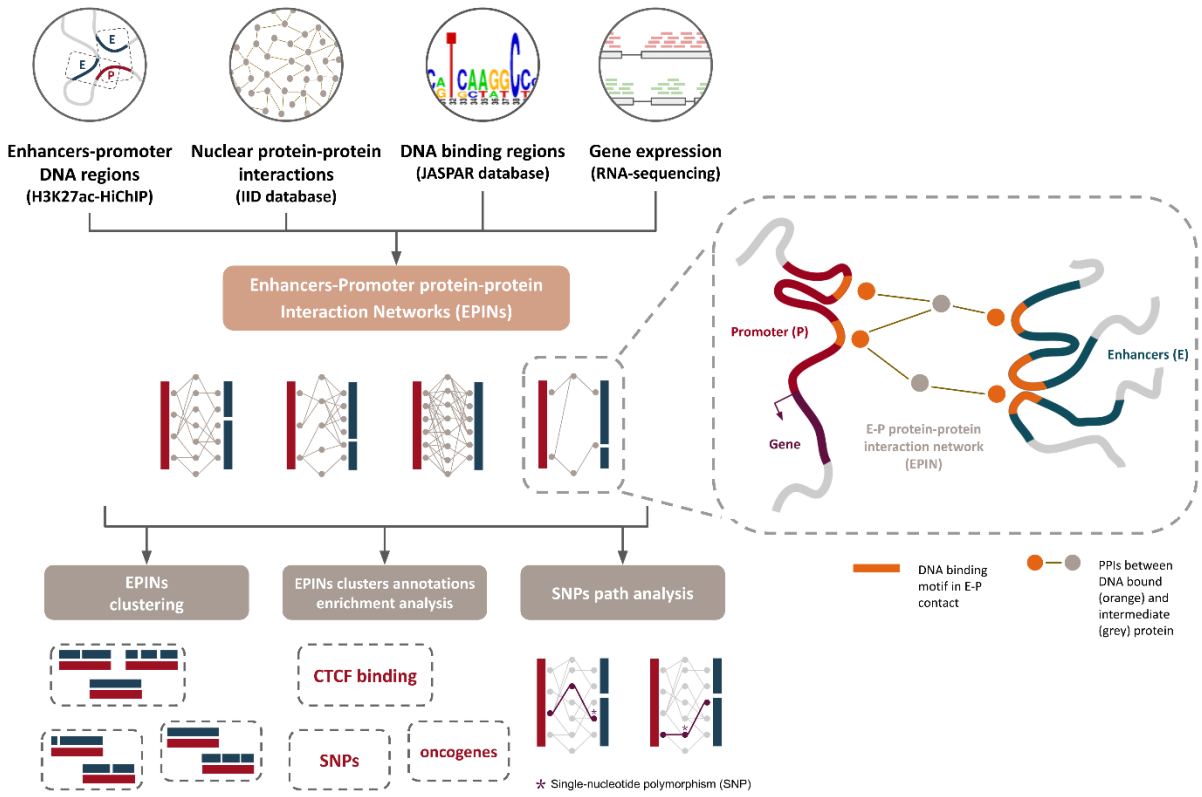


Figure 1. General overview of the PENGUIN workflow and downstream analyses. PENGUIN input consists of HiChIP data (in this work, H3K27ac in LNCaP or LHSAR cell lines), tissue-specific nuclear protein-protein interactions, PPIs (in this work, cancer and normal prostate PPIs from IID database), curated DNA-binding motifs (in this work, motifs from JASPAR database), and gene expression profiles (in this work, RNA-sequencing data in LNCaP or LHSAR cell line). PENGUIN output consists of Enhancer-Promoter protein-protein Interaction Networks (EPINs). Downstream analyses are designed to address specific questions related to prostate cancer (PrCa), namely the identification of clusters of promoters based on EPIN similarity, their enrichment in distinct annotations (CTCF binding from ChIP-seq peaks, PrCa associated SNPs, and PrCa oncogenes), and finally the formulation of mechanistic hypothesis based on SNPs path analysis. In the inset, we report a schematic representation of an enhancer-promoter protein-protein interaction network (EPIN) reconstructed with PENGUIN for a given E-P contact detected by H3K27ac-HiChIP. Promoter and enhancer DNA binding motifs found in HiChIP regions after enhancer prioritization and the corresponding bound proteins are indicated in orange; their physical interactions with other factors of the EPIN (in gray) are represented as gray lines.

2.2. PENGUIN identifies PrCa clusters of protein interaction based on chromatin contacts

We leveraged 24,547 E-P contacts (30,416 after refinement and prioritization, **Methods; Figure S1**) identified using H3K27ac-HiChIP data in LNCaP, 810 binding motifs from 639 DNA-binding proteins, and 31,944 prostate-specific, experimentally validated, physical and nuclear PPIs (filtering out proteins from unexpressed genes, **Methods; Figure S2**) to construct 4,314 EPINs using the PENGUIN clustering approach outlined in Figure 2 (**Methods**). Each EPIN is centered around one promoter that we found to be contacted by a median of 4 enhancers, with a maximum of 93 enhancers for the promoter of the gene *CRNDE* (**Table S1**). Altogether, the 4,314 EPINs contain a total of 8,215 interactions (edges) among a total of 885 proteins (nodes) that are expressed in LNCaP (**Methods**). A mean of 36% proteins found in these EPINs are encoded by differentially expressed genes in LNCaP versus LHSAR (**Methods and Table S1**). Overall, 751 out of the 885 proteins represent intermediate nodes, with 127 of them acting both as intermediate and as DNA-bound nodes in different EPINs (**Table S2**). 261 unique DNA-binding proteins have predicted binding sites in at least one of the anchors of enhancers and promoters. A mean of 32.8 (s.d. 11.5) distinct DBPs were identified per promoter anchor with *SP1, EGR1, SP2* being the most represented; and a mean of 24.8 (s.d. 7.69) were predicted per enhancer anchor with SP1, IRF1 and TFAP2A being the most represented. A mean of 1.43 (normalized) promoters (0.88 s.d.) are shared among enhancers, with a maximum of 15 promoters for the same enhancer. To identify communalities and differences among the 4,314 EPINs in LNCaP, we performed an unsupervised, hierarchical clustering based on edge composition (Ward's linkage method, **Methods**). Using this approach, we identified 8 clusters of promoters with specific networks (**Table S1,Table S3, Figure 2 and Figure S3 and Figure S4**). The decision to divide the hierarchical tree into 8 clusters was based on the analysis of cluster characteristics, achieved by varying the number of clusters (**Figure 2C**).

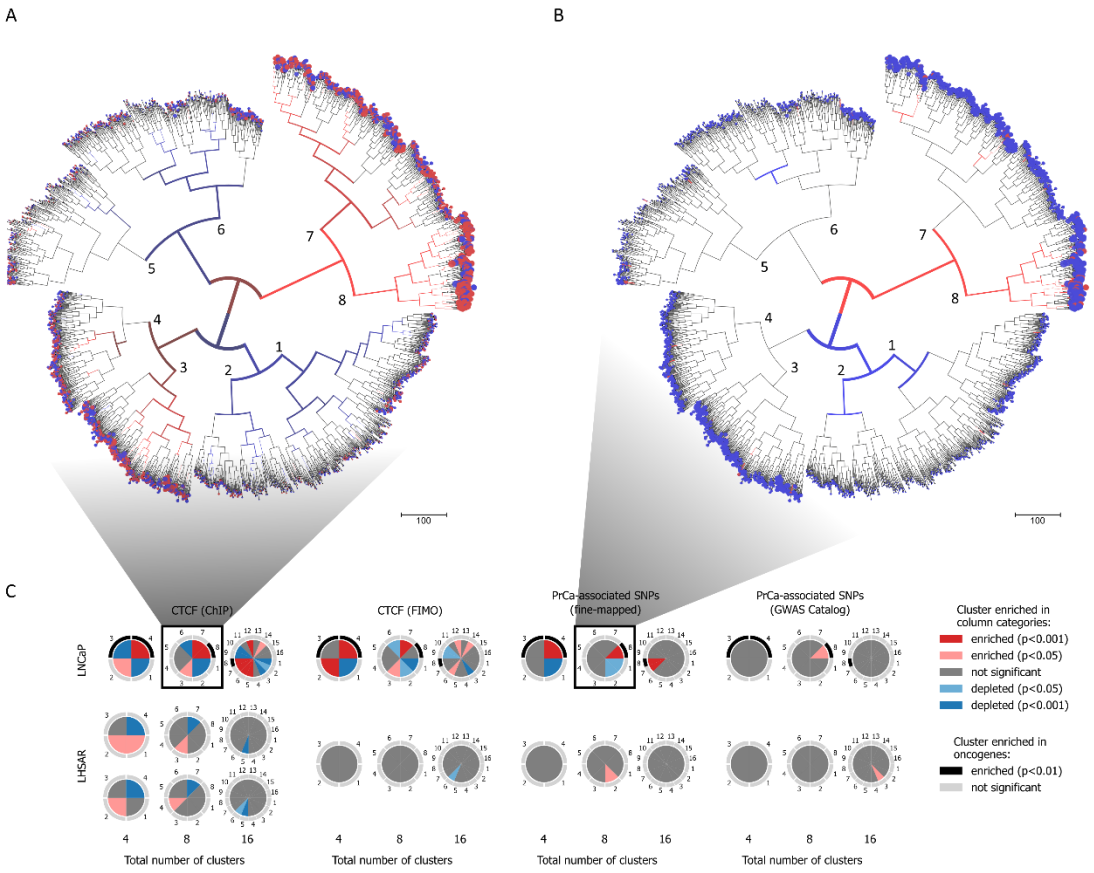


Figure 2. Clustering of the promoters originating the PENGUIN reconstructed EPINs. Clustering is based on edge composition of the EPINs. Leaf radius is proportional to network size. Color code (two-sided Fisher's exact test): red, enriched; blue, depleted; The figure is generated using ETE3 68. (A) Enrichment of PrCa SNPs in enhancers. We identified one PrCa SNP enriched cluster (GWAS+; cluster 8), and multiple PrCa SNP depleted (GWAS-; clusters 1, 2) and neutral (GWAS=; clusters 3, 4, 5, 6, 7) clusters. (B) Enrichment of CTCF ChIP-seq binding sites. We identified multiple CTCF enriched (CTCF+; clusters 3, 7, 8), depleted (CTCF-; clusters 1, 2, 6) and neutral (CTCF=; clusters 4, 5) clusters. (C) Clustering analysis on LNCaP (Top) and LHSAR (bottom) reconstructed EPINs. Pie-charts represent clustering results for a distinct total number of clusters used to partition the hierarchical clustering tree (4, 8, 16). Numbered pie-slices represent the different clusters, and their color gradients encode the significance of enrichment (shades of red), depletion (shades of blue) or neutral (gray) of the overlap with distinct annotations (ChIP-Seq CTCF peaks, predicted CTCF binding sites by FIMO, PrCa-associated SNPs from fine-mapping and GWAS). Clusters significantly enriched with previously known oncogenes are annotated with black arcs. All enrichments have been estimated using two-sided Fisher's exact test.

2.3. Characterization of PrCa clusters identified by PENGUIN

We characterized the 8 clusters using PrCa specific annotations. We used the previously described 95% credible set of SNPs (henceforth referred to as PrCa SNPs) across 137 PrCa associated regions fine-mapped from the largest publicly available GWAS summary statistics (N=79,148 cases and 61,106 controls (19)). By comparing each cluster with all other clusters, we found a significant enrichment of PrCa SNPs in one specific cluster (cluster 8 or GWAS+ cluster; two-sided Fisher's exact test, **Methods**). Interestingly this enrichment is exclusively due to SNPs in enhancers (**Table 1**). Our results show that E-P interactions containing PrCa SNPs are clustered together (red branches in **Figure 2A**) indicating that they have similar characteristics in the way their PPI networks are wired. We found that most pairwise interactions (67.5%, or 5,550 out of 8,215 edges) are found in all clusters but establishing different topologies.

PrCa SNPs overlaps	Odds Ratio (OR)	p-value
Only enhancers	11.329	1.80e-12
Only promoters	1.139	0.6
Either enhancers or promoter	8.551	2.68e-11
Both enhancers and promoter	0	1

Table 1. Enrichment of PrCa SNPs in cluster 8 (GWAS+) when considering SNPs overlapping enhancers, promoters, either or both.

We identified the protein interactions that are enriched in each cluster and estimated the significance of overrepresentation of each edge in a cluster compared to all others (**Methods**). GWAS+ cluster (cluster 8 in **Figure 2; Figure S5**) exhibits the lowest number of promoters and distinctive network characteristics (**Table S3A, Figure S3**). Nonetheless, per promoter, it displays the largest number of edges (p-value < 1e-16) and intermediate nodes (p-value < 1e-16), in line with its greater number of enhancers per promoter (p-value < 1e-16), see **Figure S4**.

We then assessed whether PENGUIN clustering was influenced by super-enhancer-like regions sharing target promoters in given clusters. Although the distribution of enhancers per hotspots is similar among our 8 clusters (**Figure S4G**), the GWAS+ cluster has fewer single enhancers (enhancer at more than 15 kb from any other enhancer). The average number of promoters targeted by each hotspot for all our 3,752 defined enhancer hotspots was 1.83 promoters targeted per hotspot. When measured considering only the promoters in given EPIN clusters, the values were: 1.29 for cluster 1, 1.28 for cluster 2, 1.25 for cluster 3, 1.24 for cluster 4, 1.22 for cluster 5, 1.21 for cluster 6, 1.34 for cluster 7 and 1.27 for cluster 8. In this case, values were very similar between EPIN clusters.

Moreover, the EPINs of the GWAS+ cluster have the lowest values of node-level centrality measures, namely betweenness and degree (**Figure S3**). The degree of a node measures the amount of connections it has, while the betweenness centrality measures the amounts of shortest paths that pass through it. Low values of betweenness and degree indicate a lower amount of connections among different nodes of the network. Betweenness and degree are significantly different across clusters (Kruskall-Wallis test p-value < 1e-16), but not with respect to the ensemble of all EPINs, which indicates that, despite the high number of shared pairwise interactions (67.5% of edges), the wiring of the cluster-specific EPINs are distinctive.

Since CTCF is a major actor in the formation and maintenance of transcriptionally productive E-P interactions (20,21), we tested the clusters identified by PENGUIN for enrichment in CTCF binding. For this analysis we used CTCF ChIP-seq peaks, from the same cell line (LNCaP), from the ENCODE project instead of predictions based on DNA-binding motifs (Methods).

We found that the enriched interactions with CTCF peaks, that we call CTCF+, cluster together (red branches in **Figure 2C, Figure S5**), suggesting that the presence of CTCF in chromatin interactions results in the formation of characteristic PPI networks between the promoter and its enhancers.

CTCF+ clusters overlap the GWAS+ cluster (**Figure 2C, Figure S5**), suggesting that CTCF mediated interactions could be more functionally relevant to PrCa. In particular, GWAS+ cluster (representing 6% of the total number of promoters considered) is the only one presenting the unique and significant enrichment in CTCF binding, PrCa SNPs, and oncogenes coincidentally (**Table 2, Table S3 , Figures S5**).

Cluster	Number of genes	CTCF	OR CTCF	P-value CTCF	PrCa SNPs	OR PrCa SNPs	P-value PrCa SNPs	Number of oncogene promoters	OR oncogenes	P-value oncogenes
1	825	-	0.617	1.91e-9	-	0.28	2.46e-2	8	1.17	0.67
2	399	-	0.613	9.65e-6	-	0.00	2.00e-2	5	1.54	0.38
3	544	+	1.348	1.35e-3	=	0.80	8.27e-1	2	0.39	0.31
4	491	=	1.084	4.09e-1	=	0.51	3.60e-1	4	0.94	1.00
5	465	=	0.841	9.12e-2	=	0.75	8.14e-1	1	0.23	0.17
6	641	-	0.664	4.24e-6	=	0.38	1.03e-1	1	0.16	0.03
7	676	+	1.655	2.12e-9	=	1.42	3.18e-1	5	0.84	1.00
8	273	+	3.287	3.64e-20	+	11.33	1.80e-12	11	6.48	1.04e-5

Table 2. Enrichment of PrCa SNPs, CTCF ChIP-seq binding sites ("CTCF" in the header), and other PrCa annotations (oncogene promoters and PrCa SNPs from GWAS Catalog) across the eight clusters identified by PENGUIN. Cluster 8 is enriched in CTCF binding, PrCa SNPs, and oncogenes. Symbols code: +, enriched; -, depleted; =, neutral. OR: Fisher's exact test Odds Ratio.

This cluster is enriched in the Hippo signaling pathway (KEGG:04390) (Bonferroni-corrected p-value=1.56e-3), WNT Signaling Pathway (KEGG:04310) (Bonferroni-corrected p-value=9.57e-3) and Pathways in cancer (KEGG:05200) with genes such as BCL2L1, MYC, FOS (Bonferroni-corrected p-value = 0.047) (**Methods, Table S5**). Interestingly GWAS+ cluster, or any other cluster, did not significantly stand out in terms of overall expression level (**Figure S2**) or, notably, in terms of fraction of differentially expressed genes (**Figure S2**).

To explore the potential connection between our clustering approach and the presence of trans-eQTLs, we used the trans-eQTLs reported from the largest eQTL study available (large-scale meta-analysis in up to 31,684 blood samples from 37 eQTLGen Consortium cohorts in Whole Blood, (22)) and defined a region an 'eQTL hotspot' when associated to more than 3 genes (**Methods**). We observed an enrichment of eQTL hotspots across all clusters (**Figure 5SE**, empirical p-value < 0.0001), but not specifically for cluster GWAS+ (**Figure 5SF**).

In conclusion, PENGUIN enabled the identification of a cluster of E-P contacts whose EPINs are uniquely enriched in PrCa SNPs, ChIP-seq CTCF peaks, and oncogenes (a.k.a. GWAS+ cluster or cluster 8, **Figure 2 and Table 2**). It should be emphasized that our findings demonstrate consistent results also when employing PrCa-associated SNPs from the GWAS catalog, in which case we also identified cluster 8 as significantly enriched (**Methods, Table S6**).

2.4. Baseline comparisons and assessment of PENGUIN specificity

Among the 273 promoters belonging to the identified GWAS+ cluster (cluster 8 in **Figure 2A**), 11 belong to known oncogenes, *FOXA1, ZFHX3, CDKN1B, KDM6A, BRCA2, CDH1, CCND1, NKX3-1, BAG4, MYC, GATA2* (**Methods**).

We compared enrichment of PrCa functional annotations in the reconstructed networks with and without inclusion of intermediate proteins. Including intermediate proteins allows increasing the number of retrieved PrCa-related oncogenes in GWAS+ cluster from 6 to 11 and increasing significance of enrichment indicating improved specificity (**Table S4**). We then compared our results with the simple overlap of the genomic regions of E-P contacts and known oncogene promoters [see **Table S1**, which also reports on the overlaps of E-P contacts with CTCF peaks (in both enhancers and promoters, see **Methods**), and PrCa SNPs (in enhancers)].

In this scenario, only 30 promoters (12 overlapping the GWAS+ cluster) would be identified that overlap both PrCa SNPs and CTCF peaks. Of these, just 3 are promoters of known oncogenes (and only one, ZMYM3, is not in the GWAS+ cluster).

To explore the cell and disease-specificity of our results we applied PENGUIN on LHSAR, a benign prostate epithelial cell line. We performed H3K27Ac HiChIP experimental data and applied the PPI clustering procedure to explore functional relationships within the clusters. We then proceeded to apply PENGUIN to identify clusters of EPINs based on their edges (**Methods**). As the selection of an exact number of clusters in a given tree could be considered an important variable in our analysis, we examined various cluster numbers (4, 8, 16). We investigated the presence of cluster enrichment in GWAS and CTCF (**Table S3B**). Our analysis did not reveal any cluster enrichment in GWAS and CTCF within the benign prostate control LHSAR. Moreover, we did not observe a significant increase in the number of identified oncogenes in LHSAR (**Figure 2B**). These results lead us to conclude that PENGUIN, along with the integration of intermediate PPI networks, significantly enhances the identification of candidate PrCa-related SNPs affecting key elements in chromatin architecture.

Despite the high similarity in PPIs between LHSAR and LNCaP cells (Jaccard index of 0.85), their clustering based on H3K27Ac HiChIP data revealed distinct EPINs (**Figure 2B**). This finding highlights the sensitivity of our method in capturing subtle differences within EPINs. To further validate this, we conducted additional statistical analyses on PPIs across different cancer cell types.

By examining the overlap between PPI networks, we discovered significant variations that were highly specific to each cell type (**Figure S6**). This observation not only reinforces the reliability of the differences found in LHSAR and LNCaP cells but also suggests that our results can be expected in other cellular contexts provided the required H3K27ac-HiChIP information, which is currently unavailable in most cases.

To further investigate the significance of intermediate PPI networks, we conducted clustering analysis exclusively based on HiChIP interactions. Specifically, we utilized the list of enhancer IDs, denoted by their genomic coordinates, within each EPIN (**Figure S7**).

Our findings unequivocally demonstrate that the exclusion of intermediate PPI networks substantially diminishes the number of identified oncogenes. This outcome strongly suggests that the information conveyed by the PPI network plays a crucial role in the classification of EPINs and their correlation with phenotypic traits.

2.5. Involvement of E-P protein interactomes in tumor-related functional processes

We analyzed the functional enrichment of the set of 885 proteins composing the universe of nodes used in the EPINs of LNCaP. 43 out of these 885 proteins are encoded by one of the 122 known PrCa oncogenes (32 intermediates, 7 DBPs among which *MGA, ETV4, ETV1, GATA2, ETV3, ERF, NKX3-1*, and 4 of both types among which *TP53, MYC, FOXA1, AR*; see **Methods and Table S2**). In total, 11 out of 885 have been targeted by PrCa-specific drugs (source: DrugBank; protein targets: *ESR2, ESRRA, AR, PARP1, NFKB2, NFKB1, NCOA2, NCOA1, AKT1, TOP2A, TOP2B*; drugs: Estramustine, Genistein, Flutamide, Nilutamide, Bicalutamide, Enzalutamide, Olaparib, Custirsen, Amonafide); and 190 out of 885 are targets of non-prostate drugs indicating the possibility of re-purposing.

Considering the genes encoding for 477 out of 751 intermediate proteins with annotations for KEGG pathways retrieved using g:Profiler (23), 41 were annotated in the prostate cancer pathway (KEGG:05215) (adjusted p-value = 3.62E-24), which annotates a total of 97 genes (**Methods and Table S7**). We next studied specific protein enrichments in the nodes of the EPINs of each identified cluster (**Table S8**). Although intermediates are ubiquitous and generally shared among all clusters, we could identify 22 significantly specific proteins enriched in the GWAS+ cluster (**Methods**).

Functional enrichment analysis of these 22 proteins revealed significant relationships with tumorigenic processes (**Table S9**). KEGG Prostate cancer pathway (KEGG:05215) appears highly enriched (adjusted p-value = 1.27e-2) together with other pathways related to tumors such as Colorectal cancer (KEGG:05210, adjusted

p-value = 3.20e-5), Pancreatic cancer (KEGG:05212, adjusted p-value = 9.54e-4) and Breast cancer (KEGG:05224, adjusted p-value = 7.06e-4). KEGG pathway KEGG:04919 (Thyroid hormone signaling pathway) is an additional highly enriched pathway (adjusted p-value = 2.57e-4). Thyroid hormones have been previously described as modulators of prostate cancer risk (24–27). Pathway KEGG:05200 (called Pathways in cancer) appears as the fourth most enriched KEGG concept (adjusted p-value= 3.63e-4). Other classical tumorigenic pathways, such as WNT signaling pathway (KEGG:04310, adjusted p-value = 1.27e-2) and TGF-beta signaling pathway (KEGG:04350, adjusted p-value = 8.21e-4) appear to be enriched. In this regard, recent studies analyzed the involvement of WNT signaling in the proliferation of prostate cancer cells (28,29), as well as the involvement and TGF-beta signaling (30,31).

Furthermore, we examined the functional enrichment of significantly central proteins across all other clusters. This analysis was conducted to facilitate functional comparisons across different clusters (**Methods** 'Functional gene set enrichment analysis'). This analysis revealed no enrichments for clusters 1, 2, 4, 5, and 6 (cluster 5 does not have significantly central proteins). This observation can be attributed to the higher number of central proteins in these clusters (365 in cluster 1, 283 in cluster 2, and 318 in cluster 6) compared to the other clusters (3 in cluster 3, 7 in cluster 7, and 22 in cluster 8). Despite having a similar number of significantly central proteins to cluster 8 (30 proteins), cluster 4 does not show any enrichment.

Moreover, of the clusters presenting enrichments (i.e., clusters 3 and 7), only cluster 7 presents enrichments related to those observed in cluster 8 (for example, KEGG prostate cancer pathway is enriched, adjusted p-value = 2.041e-2; **Figure S8**). As commented, cluster 7 presents only 7 significantly central intermediate proteins (*CREBBP, CTNNB1, GSK3B, KAT5, MAPK1, PIN1, SMAD2*), out of which, 6 overlap with those significantly central in cluster 8 (only PIN1 is absent).

2.6. SNPs path analysis in the E-P protein interactomes

Next, we sought to perform an analysis of the SNPs found along the paths within each EPIN (**Methods**). In this analysis, a path in a network is a sequence of edges joining a sequence of nodes connecting the promoter and the enhancers of an EPIN (**Figure 3A**). We distinguish between two possible scenarios based on the location of the SNPs within the paths:

> **(1)** PrCa SNPs fall in the DNA binding motifs found in enhancers, indicating a possible dysregulation of TFs binding and activity (**Figure 3B**).
>
> **(2)** PrCa SNPs in the genomic regions of the genes that encode for the intermediate nodes of the EPINs, indicating a possible alteration of the PPIs (**Figure 3C**).

The first analysis aims to identify the location of enhancers that could be targeted by genetic perturbation techniques such as CRISPR/Cas9. The second analysis aims to identify the proteins that are potentially affected by mutations so as to enhance our understanding of prostate cancer biology. Overall, we characterized all PrCa SNPs falling within any path that connects enhancers to a promoter (rs4962419 was found in both scenarios analyzed). In the following, we discuss the two scenarios and report on the *MYC, CASC11* and *GATA2* promoters as illustrative examples.

2.7. Network paths with PrCa SNPs in enhancer binding motifs

We sought to detect SNPs located in the DNA binding motifs found in the enhancers of the EPINs. Based on previous evidence (32,33), our hypothesis is that SNPs in enhancers could disrupt the binding of proteins such as TFs having an impact on their interactome. In **Table S10** we list the 36 PrCa SNPs falling within 60 DBP motifs in enhancer regions linking 34 different promoters whose EPINs include 5,184 edges. Among these, we identified 17 PrCa SNPs falling within 16 EPINs (1,894 edges) belonging to the GWAS+ cluster that had at least one PrCa SNP in their enhancers.

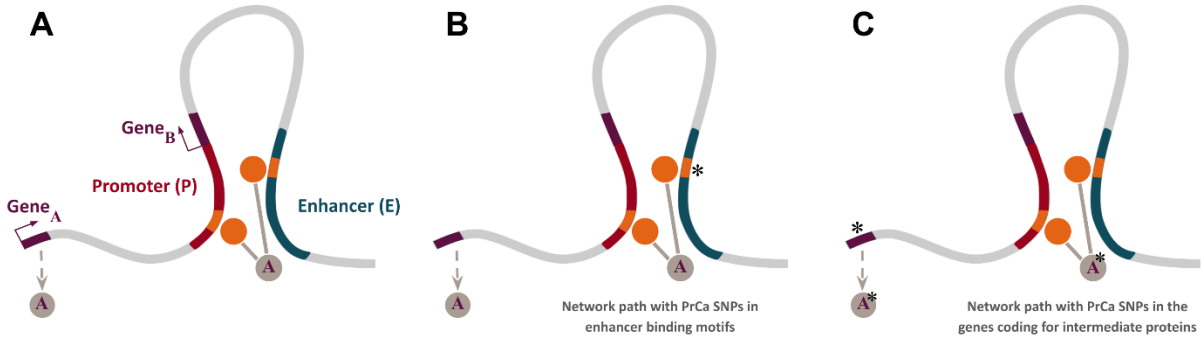


Figure 3. Schematic representation of different types of network paths found in the EPINs reconstructed by PENGUIN. In general, a network path is defined by an intermediate protein (gray circle), encoded by a gene (dark red line; Genei), that interacts with DBPs (orange circles) with binding motifs (orange lines) on the enhancer (green line) and the promoter (red line) of another gene (dark red line; Genej) (**A**). If a PrCa SNP (asterisk) falls in the enhancer binding motif, the interaction between the DBP and the enhancer may be disrupted and possibly its interactions (**B**). If a PrCa SNP (asterisk) falls in the gene that encodes for the intermediate protein, the gene product could be affected and possibly its interactions (**C**). Colors are consistent with **Figure 1**.

Several of these EPINs have promoters of differentially expressed genes (such as *DLL1, STOM* and *SEC11C* in the GEPIA tumor/normal dataset; *ID2, RPS27, SEC11C, CASZ1, CRTC2, C5* and *STOM* in the LNCaP/LHSAR dataset; **see Methods, Differential Gene Expression**).

To establish the biological significance of the identified SNPs, we leveraged data from previous pooled genome-wide CRISPR/Cas9 knockout and RNAi screens conducted in prostate cancer LNCaP cells, available in the DepMap database (https://depmap.org/, DepMap ID: ACH000977). These screens provide essentiality scores, which quantify the relevance of specific gene networks to the proliferation of LNCaP cells. In our analysis, we retrieved essentiality scores for genes in prostate tissue from DepMap and compared three distinct gene sets:

(**1**) The genes (EPIN promoters) prioritized in **Table S10**;

(**2**) All genes (EPIN promoters) included in our analysis and;

(**3**) All genes available in the DepMap database.

Remarkably, we observed significant differences in the essentiality scores (Z-scores) among these sets, with lower Z-scores indicating a higher degree of gene essentiality (**Figure 4A**). This analysis aligns with the RNAi findings, demonstrating a significant decrease in essential scores for genes containing the SNPs listed in **Table S10** (**Figure 4B**). Furthermore, the GSEA analysis unveiled a noteworthy enrichment (p-value = 0.0017) for these EPIN promoters that harbor intermediate nodes with SNPs at their genomic location (as indicated in the supplementary **Table S10**) (**Figure 4C**). Among the top essential genes, the CRISPR/Cas9 and RNAi screens prioritize the following ones : *GATA2-AS1, CASZ1, MYC, KRT8, GTPBP4-AS1, MFN2, CTBP2*, and *ID2*.

Finally, at the level of intermediate proteins, we also found some encoded by genes reported to be differentially expressed. We observed that the mean proportion of intermediates that are differentially expressed is on average 40% (**Figure S4**). We tested whether promoters belonging to the GWAS+ cluster were significantly enriched for intermediate protein encoding for differentially expressed genes (**Methods**). Among the 16 EPINs belonging to the GWAS+ cluster that had at least one PrCa SNP in their enhancers, 11 contain expression data to study potential direct effects of the SNPs. In this subset we found 4 EPINs differentially expressed in promoters (3 also differentially expressed in intermediates: *CASZ1, ID2, SEC11C*), and 4 EPINs only differentially expressed in intermediates: *MIIP, MRPL14, MYC, TMEM63B* (**Table S1**). The differential expression of intermediates makes it easier to identify interesting and potentially novel cases. For instance, *MYC* is not differentially expressed but it has differentially expressed intermediates.

2.8. Network paths with PrCa SNPs in the genes coding for EPIN nodes

In this analysis, we identify EPINs with PrCa SNPs falling within genes that encode either for intermediate or anchor bound nodes (**Table S11**), indicating a potential alteration of PPIs involved in E-P contacts.

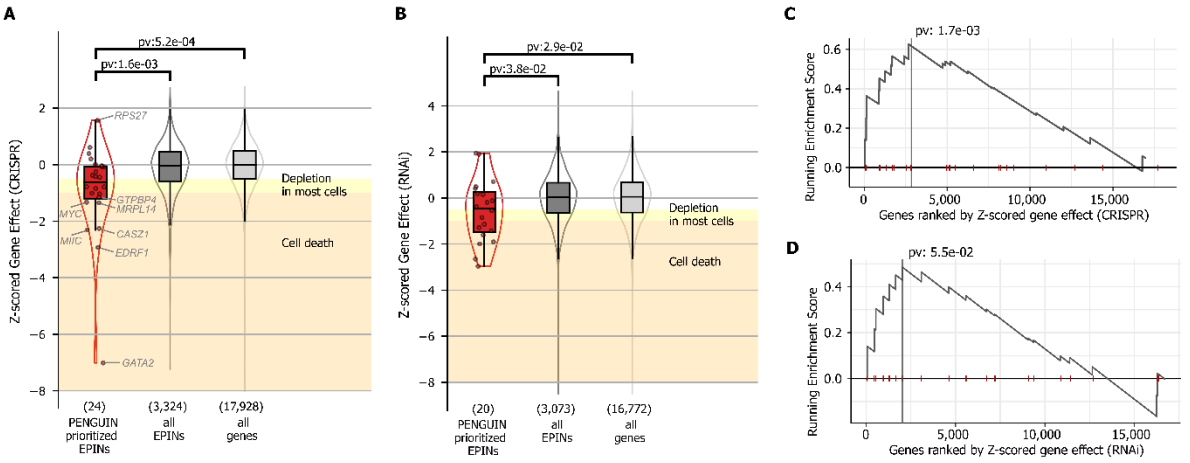


Figure 4. Validation of SNPs prioritized by PENGUIN. CRISPR/Cas9 knockout and RNAi screens provide Z- scores to quantify the relevance of a specific gene network to proliferation of LNCaP cells. (**A**) CRISPR/Cas9 knockout analysis indicates that intermediate SNPs prioritized by PENGUIN occur in genes essential for LNCaP (significance calculated with Mann-Whitney test). Genes with the strongest effect are displayed. (**B**) RNAi analysis shows milder but significantly consistent results with CRISPR/Cas9 knockout. (**C**) Gene Set Enrichment Analysis (GSEA) indicates that SNPs prioritized by PENGUIN occur in the most essential genes identified by CRISPR/Cas9 knockouts. (**D**) GSEA indicates that SNPs prioritized by PENGUIN occur in the most essential ones based on the RNAi screen. for C and D, the statistical significance of the enrichment of a gene set within the ranked gene list is reported.

We found that the GWAS+ cluster has the highest proportion of PrCa SNPs in these nodes compared to all other clusters (mean = 53.2, SE = 18.0, p-value <= 0.01, **Table S12**). The EPINs of *STK40* and *GATA2* promoters in GWAS+ cluster display the highest fraction of EPIN nodes with PrCa SNPs in their corresponding genes encoding them (**Table S1**). We use the SNP paths to link 172 PrCa SNPs falling within the gene bodies of 26 genes of which 7 are known oncogenes (*MAP2K1, CHD3, AR, SETDB1, ATM, CDKN1B, USP28*).

We identify edges that are most enriched in our GWAS+ cluster which could be pointing to essential links between the gene encoding for the node and containing a PrCa predisposing SNP at a particular EPIN. For example, we identify the link between *MDM4* containing SNP rs35946963 (PrCa p-value 1e-24) and *TP53* (34) and

between *KDM2A* containing SNP rs12790261 (PrCa p-value 1e-7) and *BCL6* (35) and ARNT continuing SNP rs139885151 (PrCa p-value 3e-13) and *HIF1A* (36).

We integrated information from pQTL associations between the 172 PrCa SNPs and protein levels (**Methods**). Two intermediate proteins (*CREB3L4, MAP2K1*) were associated with PrCa SNPs falling within the gene encoding for them (p-value of association with proteins were 7.75e-86 for *CREB3L4* and 2.40e-5 for *MAP2K1*). We identified 3 out of 26 promoter EPINs (*TRIM26, MEIS1, POU2F2*) with suggestive evidence (p-value < 1e-5) of association between the PrCa SNP with the PENGUIN-linked promoter EPIN, pointing to the cancer promoting mechanistic action of these variants: gene with SNPs in *POU2F2* linked to the EPIN promoter of gene *PHGDH* (SNP with lowest p-value rs113631324 = 3.80e-8); gene with SNPs in *TRIM26* and EPIN promoter of gene *RRM2* (SNP with lowest p-value rs2517606 = 2.69e-7); gene with SNPs in *MEIS1* and EPIN promoter of gene *STOM* (SNP with lowest p-value rs116172829 = 8.19e-6).

We note that, unlike SNPs in enhancers, whose effect can be directly assessed by CRISPR/Cas9 or RNAi assays, the impact of SNPs on intermediate nodes is more complicated to estimate due to their shared involvement in multiple gene networks.

In fact, it is worth mentioning that among the 885 proteins identified by PENGUIN, 751 serve as intermediate nodes (section **PENGUIN identifies PrCa clusters of protein interactions based on chromatin contacts**). This overlapping functionality further complicates the prediction of SNP effects on these intermediate nodes.

2.9. Examples: SNPs path analysis of *MYC, CASC11* and *GATA2* promoters

From HiChIP data, the *MYC* promoter (chr8:128747814-128748813) is in contact with 73 enhancer regions among which one holds the SNP rs10090154 (p-value of association with PrCa = 1.4e-188). This SNP is located in the binding motif of the transcription factor *FOXA1* on the *MYC* EPIN enhancer. The integration of PrCa SNPs information highlights paths in the EPIN of *MYC* that are particularly compelling in the context of the disease (red line in **Figure 5; Figure S9**).

The promoter region of MYC binds 8 proteins *TFAP2C, KLF5, RBPJ, SP1, ZBTB14, ATF6, ZBTB7A, PRDM1* and contains 17 protein interactors (dots in **Figure 4**) that might be affected by the possible disruption of its binding motif, namely, *HMGA1, RCC1, TFAP4, NFIC, PBX1, HOXB9, NFIX, NACC1, RARA, PIAS1, RPA2, H2AFY, RECQL, SATB2, CREB1, AR*. The gene encoding for *FOXA1* is differentially expressed, along with other interactors (**Table S10; Methods**). Interestingly, 24 PrCa SNPs fall within the genomic region of *AR* (marked by an asterisk next to the gene name), all with p-values of association with PrCa below 1e-11 (**Table S11**).

AR is targeted by several drugs used in the treatment of prostatic neoplasms, such as apalutamide, bicalutamide, diethylstilbestrol, enzalutamide, flutamide, and nilutamide (triangle in the **Figure 4A, source: DrugBank**). Notably, mutations in FOXA1 enhancers were previously shown to alter TF bindings in primary prostate tumors (33). And, also in line with our observations, FOXA1 enhancer region has been previously reported to be coupled to MYC (37) and has been shown to have a strong binding of AR (38).

We report two additional examples, the EPINs for the promoters of *CASC11* (**Figure S10A**) and *GATA2* (**Figure S10B**). The EPIN of *CASC11* promoter is also affected by variant rs10090154, the same well-known variant associated with risk of developing prostate carcinoma that we introduced with *MYC* EPIN (39,40) (**Table S10**). Interestingly, ***CASC11*** is known to enhance prostate cancer aggressiveness and is regulated by *C-MYC* (41), while being close to the MYC gene on chromosome 8. The promoter binds 6 proteins: *TFAP2C, SP3, SP1, PKNOX1, NR2C2* and *KLF5*.

Potentially affected protein interactors of the EPIN include: *HMGA1, PIAS1, AR, RARA*, and *PBX1*. *GATA2* is an interesting case given its essentiality score from DepMap (Z-score=-7.01). Its EPIN presents up to 11 intermediates affected by PrCa related SNPs, namely *TCF4, CTBP2, AR, ARNT, TCF7L2, CDKN2A, NEDD9, ANKRD17, MEIS1, MDM4* and *CHD3*.

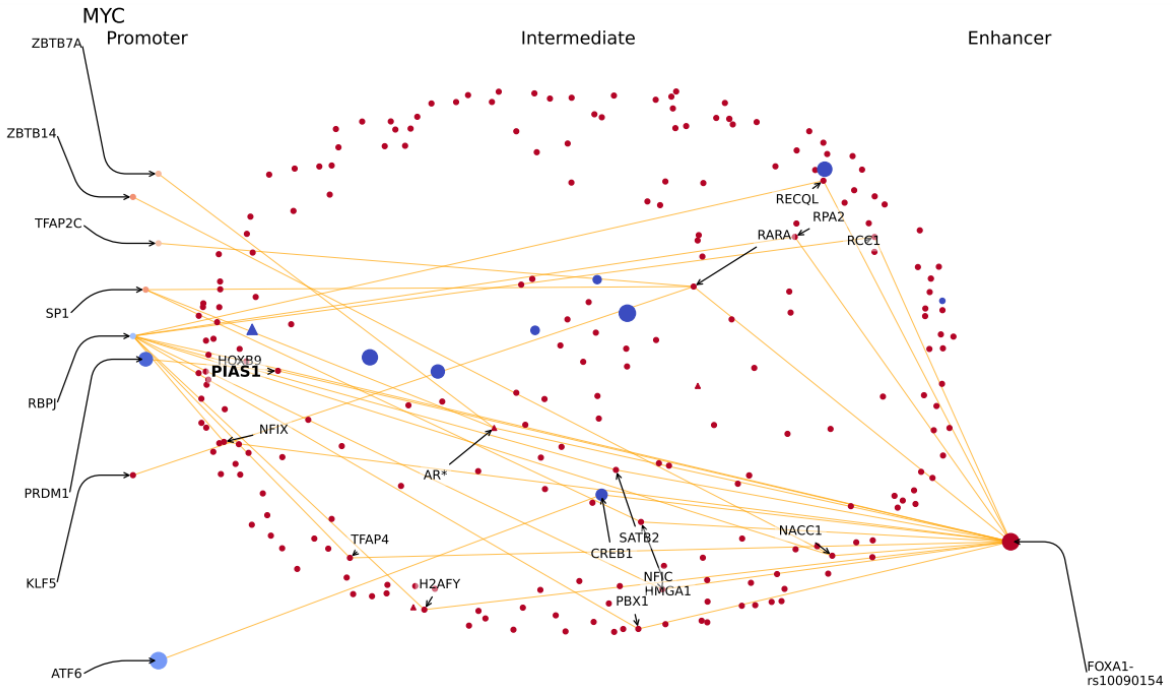


Figure 5. Reconstructed protein interactions between MYC promoter and its enhancers. DBPs with binding motifs on the promoter region are aligned on the left, while those with binding motifs on the enhancers are aligned on the right. In the middle, proteins that connect DBPs through a shortest path. Each dot represents a protein. Color, size and shape codes are explained in the Tutorial section of the PENGUIN web service at https://penguin.life.bsc.es/. In this figure, only the edges of network paths with PrCa SNPs in enhancer binding motif are represented (orange lines). Such PrCa SNPs are indicated beside the name of the enhancer-bound DBP (e.g., *FOXA1*-rs10090154); PrCa SNPs in intermediate proteins are indicated with an asterisk (e.g., *AR*); the proteins found to be enriched in the GWAS+ cluster are highlighted in bold (e.g., *PIAS1*); druggable proteins from DrugBank are indicated as triangles.

The role of *GATA2* as mediator of AR signaling in AR-dependent prostate cancer, as well as its role as a potential target for treatment development (42) has been previously described, as silencing of the gene is known to affect other relevant genes such *as C-MYC* and *AURKA* (Chiang et al., 2014). Proteins bound to the promoter region include: *ZBTB7A, ZBTB33, TCF3, SF1, NR2C2, KLF3, EGR1, E2F1* and *CREB1*, but most importantly, the EPIN presents AR bound to the enhancer region, which, as we pointed out with *MYC* EPIN, is the target of several PrCa treatments.

3. Discussion

Here we introduced the PENGUIN approach that operates on the premise that the EPIN network structure connecting a promoter and its enhancers can serve as a distinctive signature associated with specific functional profiles and diseases. Our assumption is grounded in earlier research that has demonstrated the correlation between 3D loop topology and chromatin state or gene expression (43). We propose PENGUIN as a molecular approach to study variations in structural characteristics of chromatin loops, establishing a direct link to disease-related phenomena. By integrating the PPI network information, the method offers valuable insights into the underlying mechanisms driving these distinctive features and their relevance to disease progression.

Previous approaches have already incorporated PPI networks with GWAS hits to enhance their analysis. For instance, Ratnakumar et al. (44) identified proteins that exhibited an enrichment of PPIs with GWAS hits. In a recent study, Dey et al. (45) demonstrated the benefits of employing strategies that capture both distal and proximal gene regulation in prioritizing disease-related genes. In addition, alternative methods have amalgamated information from 3D chromatin interactions and GWAS SNPs to establish connections between intergenic SNPs and gene regulation in cancer contexts (3,46,47).

These approaches have contributed to unraveling the relationship between genetic variations, chromatin organization, and disease. In contrast, our method takes a unique approach by being completely agnostic to the presence of SNPs. It combines information from PPI networks and enhancer-promoter interactions derived from H3K27ac-HiChIP data within a unified framework. This integrative approach allows us to leverage both the protein interaction landscape and the regulatory interactions between enhancers and promoters, leading to a comprehensive understanding of the molecular mechanisms underlying disease.

By utilizing PPI networks, we were able to reveal a distinct set of genes associated with PrCa that would have remained undiscovered using other methods. Notably, the intermediate nodes within this PPI network possess intrinsic properties that can be leveraged for the classification and characterization of E-P chromatin loops. Thus, our study demonstrates the capability of PENGUIN to group genes based on their involvement in PrCa, even in the absence of any prior information.

This breakthrough opens up an uncharted avenue towards comprehending and identifying unsuspected biological markers in disease. In particular, the genes identified within the cluster exhibiting the highest enrichment in SNPs associated with PrCa (the GWAS+ cluster) can be considered promising candidate oncogenes or potential partners of oncogenes. It is conceivable that these genes may share "onco-enhancers," which are enhancers contributing to tumorigenic activity.

For instance, PENGUIN can be used to identify trans-acting factors (e.g., interaction cascades of TFs and chromatin regulators) that could be targeted by drugs, or cis-acting factors (e.g., DBPs with binding motifs in regulatory elements) whose DNA binding affinity could be modified through knock-outs via CRISPR for therapeutic intervention. Moreover, unlike traditional TF enrichment analysis which detects general enrichments of particular proteins, PENGUIN can help identify the specific protein cascade potentially disrupted at enhancer loci for the disease under study.

Overall, our findings highlight the potential of PENGUIN in unveiling previously unknown gene networks and provide valuable insights into the identification and characterization of biomarkers in various diseases, including PrCa.

To validate our findings, we have used cell-line specific datasets, androgen-sensitive human prostate adenocarcinoma cells (LNCaP) or a normal prostate epithelial cell-line (LHSAR). Each of the sources of information could be directly or indirectly related to the specific cell-lines used in this study:

(**1**) H3K27ac-HiChiP in LNCaP and in LHSAR;

(**2**) Prostate-specific PPIs;

(**3**) DNA binding motifs extracted from publicly-available datasets but filtered by our cell-type specific interacting 1 kb promoter-enhancer regions and;

(**4**) gene expression on cell-line for filtering PPI networks.

The comparison of the results in cancer cell-line (LNCaP) to the results in a benign cell line (LHSAR) support our PrCa cell-specific findings. In LHSAR we found a significant association between the obtained clusters and the presence of CTCF, pointing towards the correct classification of EPINs into biologically relevant categories. However, this same clustering in the benign LHSAR cell-line did not reveal any association to PrCA, neither at the level of PrCa-SNPs, nor at the level of specific oncogenes. Future analyses could explore the use of clustering E-P loops with PENGUIN using other methods and sources for each of these layers. For example, we have used as input enhancer-promoter loops cell-specific H3K27Ac HiChIP experiments (strict calling of loops and prioritization), to maximize our true positives in the input data. The input for the PENGUIN clustering approach can also be constituted by enhancer-promoter links measured from other experimental methods aside from HiChIP or even using computational methods. We leave this for subsequent analyses.

In this work, we use a targeted approach and use the information on association of SNPs from fine-mapping as an annotation to our clusters. Specifically, we identify potential SNP paths from defined PrCa associated regions. SNP paths link genes in a network through a path that either starts from TF binding sites in enhancers or passes through proteins from the intermediate EPIN network that would have SNP in their gene bodies. This approach adds a new dimension in the contextualization of GWAS-associated SNPs using the EPIN looping realm.

It is important to mention our primary objective was to shed light on specific links that could be disrupted by PrCa-predisposing variants, such as CTCF bindings that connect promoters to their enhancers, or intermediate structural proteins that play a role in the E-P network. Further investigation is required to gain a comprehensive understanding of the biology and mechanisms underlying these crucial links. For this purpose, and to facilitate the exploration of SNP pathways associated with prostate cancer, we developed a user-friendly web interface accessible at https://penguin.life.bsc.es/.

This platform serves as a tool for convenient investigation into the pathways influenced by SNPs in the context of prostate cancer. It is also intriguing to observe that, while PENGUIN successfully identifies clusters of EPINs significantly associated with PrCA, the gene expression analysis did not reveal any significant trends. At first glance, this observation may appear contradictory to our definitions of EPIN clusters and the core concept of EPIN itself. However, considering the evidence presented by our analysis, we believe that PENGUIN enables the detection of cancer associations with heightened sensitivity compared to traditional differential expression analyses. The ability of PENGUIN to capture intricate associations between EPINs and cancer surpasses the limitations of relying solely on gene expression changes, offering a more comprehensive understanding of the underlying molecular mechanisms involved in cancer development and progression.

Our analysis comes with some caveats to keep in mind. Firstly, we relied on data from the HiChIP technique for capturing enhancer-promoter (E-P) interactions, protein-DNA interactions from FIMO, and tissue-specific protein-protein interactions from the integrated interactions database (IID). The comprehensiveness of these datasets is inherently limited by the scope and constraints of the underlying databases and methodologies employed. Furthermore, our approach focuses on networks involving proteins with known edges, resulting in a consideration of only those proteins. Additionally, for the purpose of visualization, we have condensed the number of

reported proteins and have presented only one intermediate protein (expanded one edge away).

Moreover, it is worth mentioning that our study focuses on E-P interactions within a stable environment (LNCaP cells), representing a snapshot in time. While this field is still undergoing active research and further exploration, existing literature suggests that E-P interactions can exhibit minimal and quantitatively small changes in these conditions. Thus, while interpreting our findings, it is essential to consider the limitations of the utilized databases and methodologies, the specific protein selection, the condensed visualization approach, and the stable cellular context in which the E-P interactions were examined.

In conclusion, the PENGUIN approach employed in this study to investigate PrCa in LNCaP cells has the potential to be applied to the study of other human diseases, given the availability of similar data. This approach can be extended to explore different scenarios, such as different cell types or combinations of GWAS data, offering a promising avenue for future investigations.

For instance, utilizing E-P dataset from another prostate cancer cell line would allow the identification of target genes regulated by enhancers from diverse cell types. These target genes can be prioritized using a genome-wide collection of disease-specific risk SNPs. The networks generated by PENGUIN provide a molecular understanding of the associations involved in cancer-related chromatin dynamics, making them well-suited for training advanced machine learning models like graph neural networks (GNNs). We propose potential intermediates in PrCa that engage in E-P networks within cancer cells and present opportunities for therapeutic intervention. High-throughput functional studies could validate the impact of genetic perturbations on thousands of enhancers simultaneously. As shown in our analysis, leveraging CRISPR-Cas9 technology would enable precise editing of specific genomic regions, facilitating targeted investigations and further elucidating the functional consequences of these genetic perturbations.

Acknowledgements

The authors are grateful to José María Fernández González (Barcelona Supercomputing Center) for the crucial guidance with the PENGUIN web server. They also thank Biola Javierre's lab at the Josep Carreras Leukaemia Research Institute for the support, the 'RNA initiative' at IIT and all the members of Tartaglia's lab at CRG, Sapienza and IIT.

Funding

The research leading to these results has been supported by the European Research Council [RIBOMYLOME_309545 and ASTRA_855923], the H2020 projects [IASIS_727658 and INFORE_825080], and the project ONCOLOGICS (ERA Net Grant 779282, ERAPERMED2020- 036; and Departament de Salut-Generalitat de Catalunya support, SLD040/20/000001).

CG has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754490 – MINDED project.

I.N.C. was supported by a grant for pre-doctoral contracts for the training of doctors (Project ID: SEV-2015-0493-18-2) (Grant ID: PRE2018-083662) from the Spanish Ministry for Science, Innovation and Universities.

Methods

Conformation capture and E-P interactions

We used Hi-C followed by chromatin immunoprecipitation (HiChIP) targeting H3K27Ac in LNCaP cells (androgen-sensitive prostatic carcinoma cell line) across 5 biological replicates including 1 billion reads as previously described (11) (**Table 3**). As a comparison, we also performed H3K27Ac HiChIP on LHSAR (Prostate epithelial cells overexpressing androgen receptor), across three replicates including 309 million reads. HiChIP, an efficient proteinmediated chromatin-conformation assay, was performed following the procedure described (10). The alignment, processing and loop calling from raw fastq files (paired-end data) was performed as previously described (11). Briefly, HiC-Pro (48) was used to map the HiCHiP trimmed reads and extract unique interactions; FitHiChIP (49) was used to identify significant interactions with a predefined set of peaks from H3K27ac ChIP-seq in LNCaP to refine accurate anchor ranges.

	LNCaP	LHSAR
HiChIP H3K27ac	5 replicates[11]	3 replicates
RNA-seq	2 replicates [50]	2 replicates [50]
ChIP-seq H3K27ac	1 replicate [11]	2 replicates
ChIP-seq CTCF	2 replicates [51]	2 replicates*[51]

Table 3. Genomic datasets used in the work. Data with no references was generated for this study. (*) Not from LHSAR but from human epithelial cells of the prostate.

We used q-value < 0.01 and a 5 kb resolution and considered only interactions between 5 kb and 3 Mb as previously described (11). In this analysis, we restricted to a stringent global background estimation to reduce as much as possible the number of false-positive interactions. The corresponding FitHiChIP specifications used were “IntType=3” (the peak-to-all) for the foreground, meaning at least one anchor to be in the H3K27 peak, and “UseP2PBackgrnd=1” (the peak-to-peak (stringent)) for the

global background estimation of expected counts and contact probabilities for each genomic distance for learning the background and spline fitting. We identified 49,565 significant interactions (FitHiChIP, FDR < 0.01) for LNCaP, and 12,053 for LHSAR. We categorized interactions by overlapping anchors with transcription start sites (TSS) and enhancers identified by H3K27ac ChIP-seq as previously described (11). Briefly, we first extended anchors by 5 kb on either side; we defined promoter regions around the TSS (+/- 500 bases) using RefSeq hg19 (see Data Availability); we defined enhancer regions using regions from H3K27ac ChipSeq in the same cell.

Specifically, these were 49,638 and 53,561 enhancer regions, respectively from H3K27ac LNCaP in regular media (union of narrow and broad peaks) and from H3K27ac LHSAR. We note that the enhancer anchors at this stage of the analysis are of length 15 kb, due to 5 kb resolution of the HiChIP data analysis and additional 5 kb padding added to anchors on either side. We labeled the promoters and enhancer regions that overlap either right or left anchors, and considered E-P if only one anchor overlaps a promoter and the other an enhancer region. For LNCaP, out of the 49,565 significant interactions, we considered 18,151 E-P interactions. For LHSAR, out of the 12,052 significant interactions, we considered 5,435 E-P interactions. It is important to emphasize that our study relies solely on enhancers defined by our own HiChIP experiments, rather than relying on annotated enhancers or external definitions from ENCODE. We further prioritized E-P interactions to 1 kb regions and discarded from enhancers the 1 kb bins with fewer HiChIP interactions with the promoter (see E-P HiChIP prioritization section). We obtain 30,416 and 4,497 E-P interactions of 1 kb each for LNCaP and LHSAR respectively. The 15 kb original E-P interactions dataset contained a mean of 1.6 (1.3 s.d.) promoter anchors per enhancer anchor (after prioritization of enhancer anchor to 1 kb region, mean of 1.4 (0.9 s.d.) promoters per enhancer). There were 11,127 (17,683 prioritized 1 kb regions) enhancer anchors in total; 7,341 (12,385 prioritized 1 kb regions) enhancer anchors are contacted by one promoter anchor with a maximum of 21 promoter anchors (15 using prioritized enhancer regions) sharing the same enhancer.

E-P HiChIP prioritization

In order to reduce experimental artifacts in the context of our EPINs, we developed a specific prioritization method. This prioritization starts by normalizing the data assuming, as most used capture-C normalizations (ICE (52), Vanilla, or KR (53)) that all biases (e.g., GC content, number of restriction sites, mappability, or in the case of HiChIP, immunoprecipitation bias) can be corrected together. For this normalization step, we assume that there is a specific bias per any 1 kb genomic loci (**see Figures S1A and S1B**). This bias causes the difference between a theoretical expected number of interactions (Exy between loci *X* and *Y*) and the observed number of interactions (Oxy between loci *X* and *Y*). In this representation we can define a system of 9 equations involving three 1 kb loci in the promoter (exactly from TSS -1 kb to TSS +2 kb) and fifteen 1 kb loci on the enhancer side. This system of equations is then solved using Sequential Quadratic Programming (SQP) (54). The procedure is repeated in an overlapping window manner along the 15 kb of the enhancer, always against the target 1 kb of the promoter and its two 1 kb neighboring loci. Before the normalization step, we observed a different interaction pattern for interactions below 10 kb (**Figure S1C**) due, in part, to the contiguity of restriction-enzyme fragments or chromatin persistence length. As these interactions may also be a source of bias in the construction of a PPI network, we removed them from our study. We applied the normalization to the remaining interactions and observed a better correlation between genomic distance and interaction count (**Figures S1D**).

In order to compare with standard normalization procedure we applied the ICE normalization52 to our dataset (using TADbit (55) 1 kb resolution; filtering bins with less than 100 di-tags - 75% of the genome lost even using a threshold 10 times below the recommended (53)). Because of the sparsity of the genomic matrix the normalization did not fully converge (ICE was not able to completely balance the average di-tag counts per bin (52)). Next we applied the following normalization to our loops dataset, with few modification in order to rescue as much signal as possible: 1- in the promoter site, as our definition of promoter is exact (TSS to TSS +1 kb), we

corrected using the average of the two bins spanning over this 1 kb region 2- on the enhancer site, as most of the 1 kb loci were excluded by the normalization filter we also averaged the ICE bias over the whole region. Even with these modifications, only half the original data was recovered. However, the correlation between genomic distance and number of interactions was significantly improved with respect to raw data. Overall, the correlation value observed with ICE was similar to the one measured for our normalization (**Figure S1E**). We believe however that, for this dataset and for our methodology, our normalization procedure represents an improvement as it considers the exact promoter regions (not partially overlapping 1 kb bins) and minimizes the loss of promoter-enhancer data.

The normalized profile of interactions was finally used to prioritize the most interacting 1 kb loci on the 15 kb enhancer (**Figure S1F**). The selected 1 kb regions are referred to as prioritized enhancer regions.

DNA binding motifs

DNA binding motifs were retrieved from JASPAR (Fornes et al. 2019), an open-access database of curated, non-redundant binding profiles of DBPs (a.k.a. motifs) stored as position frequency matrices (PFMs). To detect the binding motifs, we used FIMO from the MEME-suite software (Grant et al. 2011), with p-value <= 1e-4 and q-value <= 5e-2 cutoffs. JASPAR **contains 810 DNA binding motifs of 640 proteins** that overlap the E-P contacts identified with HiChIP.

Gene expression data

We assayed RNA sequencing (RNA-seq) in the cell line LNCaP and LHSAR for two replicates using the VIPER pipeline as previously described (11), and fragments per kilobase of transcript per million mapped reads (FPKM) values were calculated for 20,114 RefSEQ genes. Genes with expression levels above the threshold of 0.003 in both replicates were considered in the entire analysis (**Figure S2**).

Depending on the dataset, this expression lower-bound may be modified in different use cases, for instance based on specific insights or based on a differential analysis between conditions. In this work, we used FPKM instead of more direct measures as we set our threshold very low and did not want to enrich our dataset with very long, virtually unexpressed, transcripts.

Protein-protein interaction network

We obtained protein-protein interactions (PPIs) from the Integrated Interactions Database (IID) (56). To better contextualize the interactome information, we combined the annotations of the PPIs from IID database with the LNCaP gene expression data. As for the IID annotations, we applied the following selection criteria. First we selected interactions annotated as “experimental” in the “evidence type” field and identified by at least two independent biological assays reported in the “methods” field. Then, we filtered only for interactions in the prostate or in prostate cancer cells and between nuclear proteins. Finally, we retain proteins whose gene expression levels were FPKM > 0.003 in both replicates (this cut-off removes ~30% of the genes). In total, 14,221 proteins from a pool of 20,111 human protein coding genes meet the gene expression criteria. The combination of the above filtering criteria (gene expression, using only nuclear, prostate cancer or prostate and experimentally by 2 methods) resulted in an unweighted network of **31,944 prostate-specific nuclear PPIs among 4,295 proteins** (56).

Similarly, for the comparison with the LHSAR cell line we reconstructed the PPI interaction networks with PPIs from the same database (IID) having the following annotation criteria: “experimental” in the “evidence type” field and identified by at least two independent biological assays reported in the “methods” field. Then, we filtered only for interactions in the prostate cells and between nuclear proteins. Finally, we retain PPIs between proteins whose LHSAR gene expression levels were FPKM > 0.003 in both replicates. In total 29,316 PPIs representing 4,363 proteins were used

for the EPIN reconstruction in the LHSAR cell line. The Jaccard Index between the two resulting PPIs between LNCaP and LHSAR is 0.852.

The PENGUIN pipeline

We set up graph-based approach, called Promoter-ENhancer-GUided Interaction Networks (PENGUIN), to reconstruct individual networks of protein interactions that might occur between one promoter (P) and its contacting enhancers (E), that we call E-P protein-protein Interaction Networks (EPINs). To reconstruct the EPINs, PENGUIN integrates information about chromatin contacts, protein-DNA binding, and protein-protein interactions (PPIs).

For the case under study in this work (prostate cancer, PrCa), chromatin contacts information comes from H3K27Ac HiChIP of LNCaP cells (4,314 promoters and 5,789 enhancer regions; see Methods, “Conformation capture and E-P interactions"), protein-DNA binding information (53,54) comes from the JASPAR database (810 DNA binding motifs of 640 proteins; see **Methods**, “DNA binding motifs”), and PPIs information comes from the IID database (31,944 prostate-specific nuclear PPIs among 4,295 proteins; see Methods, “Protein-protein interaction network”) further filtered using LNCaP RNA-seq data (see **Methods**, “Gene expression data”).

The reconstruction of EPINs follows these steps: for each E-P contact, (1) DNA binding motifs are detected in the corresponding sequences of promoter and enhancer regions; (2) a subnetwork of PPIs is selected containing all promoter-bound proteins, all enhancer-bound proteins, and all their intermediate interactors, with a maximum of 1 intermediate node between enhancer and promoter bound DNA binding proteins; (3) intermediate interactors are discarded if they only connect promoter-bound proteins or enhancer-bound proteins.

Using the provided PrCa information, PENGUIN reconstructed 4,314 EPINs consisting of a total of 9,141 PPIs among 885 proteins of which 751 are intermediate proteins linking promoter-bound and enhancer-bound proteins.

Node centrality measures

In several analyses we employed two measures of node centrality, namely betweenness and degree. **Betweenness** is a measure of centrality in a graph based on shortest paths. For every pair of nodes in a connected graph, there exists at least one shortest path between the vertices such that either of the number of edges that the path passes through is minimized.

The **degree** of a node in a network is the number of connections it has to other nodes; the degree distribution is the probability distribution of these degrees over the whole network.

Clustering EPIN

We defined EPIN clusters by taking into account their edge content. Each edge consists of an individual pairwise PPI as defined previously. We collected the full universe of edges using all existent edges between all promoter EPINs (the union graph). Then we computed the distance between EPINs by counting the number edges shared over the total number of edges in our predefined universe of edges.

Finally, we performed clustering using this distance matrix from all possible combinations of EPIN pairs. The clustering was performed using Ward's linkage method. Each leaf in the obtained cluster represents a promoter EPIN.

Identifying enriched functional annotations in EPIN clusters

We performed two-sided Fisher's exact tests on every single branch of the dendrogram representing the obtained hierarchical clustering. We evaluated the enrichment of any feature (CTCF binding sites by ChIP-seq, PrCa SNPs from curated GWAS, PrCa oncogenes) in the leaves under a branch of interest compared to those in the rest of the tree. For the enrichment in CTCF binding, we used CTCF peaks from an external dataset but in the same cell line (see CTCF ChiP-Seq peaks). We considered an EPIN to be CTCF-positive (CTCF+), if a CTCF peak was found in a 10 kb region around its promoter and around 10 kb of at least one of its enhancer regions.

For the GWAS feature, we require the presence/overlap of a PrCa-associated SNP (see Genome-wide association data) in at least one of the enhancers of an EPIN. Two-sided Fisher's exact tests were used to calculate the odds ratio (OR) and enrichment p-values for presence of PrCa annotations within the identified clusters.

Druggability information

We extracted information for target druggability from DrugBank (57). The use of each drug was obtained from the Therapeutic Target Database (58). We annotated each protein node that is a target of drugs that are assigned as Approved or under Clinical Trials (Phase 1, 2, 3) or Investigable for Prostate Cancer, as PrCa druggable.

CTCF ChiP-Seq peaks

CTCF ChIP-seq peaks for LNCaP cell line were retrieved from ENCODE51 project (https://www.encodeproject.org/) for the same Genome assembly, hg19 (GEO references: *GSM2827202* and *GSM2827203*). Overlaps of the CTCF binding sites with enhancer and promoter anchors allowed a 10 kb gap between them. Since CTCF ChiP-seq peaks for LHSAR cell line were not available in ENCODE, we retrieved from ChIP Atlas (https://chip-atlas.org/) two distinct sets (GEO references: *GSM2825573* and *GSM2825574*) of CTCF peaks (of same Genome assembly hg19) for prostate epithelial cells at a q-value of 1e-10 (**Table 3**). We used these two sets independently and in concatenation when comparing the clustering results between LNCaP and LHSAR. These narrow peaks were mapped on the enhancer regions using the python package PyRanges (see "E-P contacts" section). For both cases, LNCaP and LHSAR, the narrow peaks were considered as the CTCF binding sites.

PrCa SNPs

To explore enrichment of SNPs associated to PrCa across the identified clusters, and to identify the SNP paths, we used the previously reported 95% credible set (11) from fine-mapping 137 previously-associated PrCa regions using a Bayesian statistical method PAINTOR (59) employing the largest PrCa genome-wide association studies

(GWAS) (N = 79,148 cases and 61,106 controls) (60). This set was composed of 5,412 distinct SNPs (rsid). We will refer to these as PrCa SNPs. Note that this set also includes SNPs that do not reach genome-wide-filters of p-value significance. We illustrate the location of the associated PrCa regions and number of PrCa SNPs in **Figure S11**.

We did not find a significant correlation between the number of PrCa SNPs in the regions and the number of PrCa SNPs we prioritized in this work (Pearson r=0.2, p-value=0.06 and Pearson r=0.1, p-value=0.3 for Tables S10 and S11, respectively). We mapped the SNP location to prioritized enhancer regions anchor locations with a window of 10 kb. 518 out of 5,412 overlap our prioritized enhancer regions; 18 of them overlap our promoter regions. In total 218 prioritized enhancers and 14 promoters overlap a PrCa SNP.

SNP paths (PrCa SNPs in enhancer binding motifs)

A path in a network is a sequence of edges joining a sequence of nodes. We detected PrCa SNPs located in the DNA binding motifs in the enhancers, and identified the corresponding SNP paths (linked edges and nodes) for each EPIN promoter. For SNP paths analyses and the web-browser, we used all PrCa SNPs in the 95% credible set. There were 36 PrCa SNPs falling in enhancer binding motifs across clusters 3, 4, 5, 6, 7, 8. To report the most interesting cases in the Tables and Results, we used the subset of those passing genome-wide significance of p-value for PrCa association < 5e-8. There were 15 PrCa SNPs falling in enhancer binding motifs across clusters 3, 5, 6, 7, 8.

SNP paths (PrCa SNPs in intermediate proteins)

We detected PrCa SNPs falling within genes that encode for intermediate nodes, and identified the corresponding SNP paths (linked edges and nodes)for each EPIN promoter. For SNP paths analyses and the web-browser, we used all PrCa SNPs in the 95% credible set.

PrCa GWAS enrichment using GWAS Catalog and comparison with other diseases

This analysis had two aims: 1) explore whether we could replicate our finding and identify the GWAS enriched cluster using a different source for the GWAS; 2) to compare the GWAS signal for different diseases. We estimated enrichment of SNPs overlapping the enhancers in each of the identified clusters by exploring the NHGRI GWAS Catalog associations (61). First, we retrieved GWAS data and filtered the traits according to their “umlsSemanticTypeName” as defined in DisGeNet database (62) to one of the following: "Mental or Behavioral Dysfunction", "Neoplastic Process", "Disease or Syndrome", "Congenital Abnormality; Disease or Syndrome", "Disease or Syndrome; Congenital Abnormality", "Disease or Syndrome; Anatomical Abnormality".

We considered only traits with at least 10 genome-wide-significant SNPs (unadjusted p-value < 5e-8). We mapped the SNP location to prioritized enhancer anchor locations with a window of 10kb. 104 diseases had SNPs overlaps and 17 of them have more than 10 SNP overlapping (**Table S5**). For each cluster, we tested enrichment of disease-associated SNPs using Fisher tests and considered significant p-value < 0.01 and OR > 1.

Trans-eQTL hotspots

We retrieved trans-eQTLs reported in the largest meta-analysis with up to 31,684 blood samples from 37 eQTLGen Consortium cohorts in whole blood in (22). We grouped enhancers by collapsing when they were separated by less than 20 kb, thereby creating ‘enhancer clusters’. To qualify as a trans-eQTL hotspot, the enhancer clusters had to contain a SNP associated with at least 3 different genes. We quantified the normalized mutual information (NMI) between the hotspot-related enhancer clusters and our 8 EPIN clusters.

In order to infer deviation from expected by chance and estimate an empirical p-value, we randomized 10 thousand times the association between each enhancer and its

corresponding EPIN cluster and computed the NMI between each randomized EPIN clustering and the observed hotspot-related enhancer clustering. Additionally, we checked if a given cluster was significantly enriched in trans-eQTL hotspots. For this purpose we applied a Fisher test to our pool of enhancers comparing the two contingencies, inside/outside a given cluster, and inside/outside a trans-eQTL hotspot.

Oncogenes Gene list

We used a previously identified list of 122 Genes ("PrCa_GeneList_Used.csv") known to be somatically mutated in PrCa oncogenesis (37 out of 4,314 promoters considered). As previously described (11), the 122 oncogenes are a set of prostate cancer–genes curated from three large-scale PrCa studies that show evidence of somatically acquired mutations, at both localized and advanced prostate cancer, known and recurrently altered in localized prostate cancer and metastatic prostate cancer.

Super-enhancer-like regions

We defined enhancer hotspots as groups of enhancers separated by less than 15 kb, and identified 3,752 enhancer hotspots using *bedtools cluster*.

Enriched edges within each cluster

Two-sided Fisher's exact tests were used to compute odds ratios and p-values of the edges and nodes in the eight different clusters. Specifically, each edge or node was tested for presence/absence in a cluster compared to all others. Therefore, one edge or node can be enriched in one or more than one cluster, it cannot be enriched in all clusters.

Enriched intermediate nodes within each cluster

We computed protein importance for each cluster in terms of two network centrality measures: betweenness and degree. For each protein we obtain both betweenness

and degree specificity ratios in order to equitably quantify internal protein centrality differences between the clusters. For each of the found clusters we independently estimated the specificity of the observed protein centrality measures ("Betweenness" and "Degree"). For a given protein (Pi) in a particular cluster (Cj), we define the specificity as the ratio between the mean centrality value of Pi inside the fraction of networks belonging to Cj ; divided by the mean centrality value of Pi for the fraction of networks outside of the cluster Cj.

Specificity ratio (P*i*, C*j*) = (mean (P*i* centrality in C*j* networks) + 1) / (mean (Pi centrality in non-C*j* networks) + 1)

We assessed protein specificity ratio significance for each cluster upon random network cluster generation. Aiming to assess the significance of the different specificity ratios for the proteins within each cluster, we developed a significance analysis test based on random cluster subsamplings. In order to compute the significance of a given protein specificity ratio (P*i*) within a particular cluster of analysis (C*j*), we performed 1000 random network samplings to produce random network clusters containing the same number of networks as the real cluster being analyzed (i.e., if the real cluster contains 100 networks, the random clusters generated will contain 100 random networks out of the 4,314 clustered networks). Within each of those 1000 random clusters, we compute the corresponding protein specificity ratios, with the p-value representing the probability of finding the protein specificity ratio to be higher or equal to the real value computed for the particular cluster of interest (C*j*).

We also performed Fisher tests to assess enrichment for the presence of the node in the cluster (Fisher test p-value < 0.01). EP300 was excluded from the enrichment test as the presence of that node was not significantly enriched (Fisher test p-value <0.01). 22 proteins (*SMAD2, KAT5, NCOR2, MAPK8, SMAD4, CREBBP, CTNNB1, PGR, HDAC3, HDAC2, GSK3B, UBA52, UBE2I, JUND, PIAS1, XRCC5, CDK6, XRCC6, MAPK1, FOS, HIF1A and MAPK3*) were found to be significantly specific for both betweenness and degree ratios (p-value < 0.01 for both centrality measures and over-

represented in this cluster using Fisher tests) and used as input for the functional gene set enrichment analysis presented **as Table S9**).

We provide the full results of the centrality significance analysis for each cluster in github:

https://github.com/bsc-life/penguin_software/tree/main/Protein_Significance_analysis

Functional gene set enrichment analysis

Functional enrichment analysis was performed using the g:GOST module from g:Profiler, a web tool to perform simultaneous gene set enrichment analysis across multiple biomedical databases (23). We query the web service using the R implementation available from gprofiler2 package. g:GOST performs cumulative hypergeometric tests of an input gene set against preprocessed database-specific gene sets. The code for this analysis is available as a Jupyter Notebook that can be accessed in github:

https://github.com/bsc-life/penguin_software/tree/main/gProfiler_GSEA/Supplementary_Tables_5_7_9_and_Significantly_Central_Protein_Enrichment_Analysis.ipynb

We set alternative backgrounds for the gene set enrichment analysis, depending on the analysis. For the analysis presented as Table S5, where we run the web service to test functional enrichment of the genes associated to the promoter networks from cluster 8, the background is set to the 4,314 genes associated with the clustered EPINs. For the analysis presented as **Table S7**, where we test for general functional enrichment of all different proteins forming the EPINs, we run the web service considering only annotated genes for the statistical domain scope. Finally, for the analysis presented as **Table S9**, where we test the functional enrichment of the significantly central (p-value < 0.01 for both degree and betweenness centrality) proteins of networks from GWAS+ cluster, the background is formed by the very limited set of 751 unique intermediate proteins forming the EPINs. We additionally provide, within the very same Jupyter Notebook, comparative dot plots presenting the

functional enrichment analysis of significantly central proteins of each cluster under **Table S9** setting.

Reported adjusted p-values correspond to Benjamini-Hochberg correction for multiple testing, with adjusted p-values ≤ 0.05 considered to be significant. Gene set enrichment analysis results are provided for KEGG pathways, Reactome, Gene Ontology, Wikipathways, TRANSFAC, miRTarBase, Human Protein Atlas, CORUM and Human Phenotype Ontology. For the enrichment analysis of significantly specific proteins of the GWAS+ cluster, we provided as input the 22 previously described proteins. For the enrichment analysis of the GWAS+ cluster, we provided as input all genes associated with the EPIN promoters in cluster GWAS+.

Differential Gene Expression

We integrated data from EPIN promoters with differential gene expression (DE) from two sources. DE analysis on prostate cancer tumor versus normal was downloaded from GEPIA: http://gepia2.cancer-pku.cn/#degenes, which use the TCGA and GTEx projects databases to compare gene expression between tumor and normal tissues under Limma, both under and over expressed. We used the default thresholds of log2FC of 1 and qvalue cut-off of 0.01. These data covered 84 out of 885 genes encoding for intermediates in PENGUIN and 413 out of 4,314 promoter EPINs. DE analysis of RNA-Seq on LHSAR (an immortalized prostate epithelial line overexpressing androgen receptor) versus LNCaP was performed as previously described. Briefly, RNA-seq data were processed using the VIPER pipeline (63). Reads were aligned to the hg19 human genome built with STAR. FPKM values were calculated with Cufflinks for 20,114 RefSEQ genes included in the VIPER repository. Differential expression analysis was performed with the DESeq2 R package (64). 15,650 genes with DE data covered 884 of the 885 genes encoding for intermediates in PENGUIN and 3,286 genes out of 4,314 promoter EPINs.

We annotated whether the EPIN promoters themselves and the genes encoding the intermediate proteins in our data were DE using either of the two databases. We

considered as DE those genes passing |log2 fold change| > 1 and adjusted p-value <= 0.01. For the LNCAP/LHSAR dataset, we could compute a Fisher test of enrichment of differentially expressed genes encoding for intermediate proteins within each EPIN promoter versus within the SNP paths (we could not compute this for the GEPIA since we did not have the full dataset of covered genes). The genes that were not passing these filters were considered non-DE and the genes not covered by the two datasets were excluded from the enrichment analysis described next. For each EPIN we calculated the fraction of DE intermediates within the SNP paths, and we estimated the enrichment of those compared to the fraction of DE intermediates in the full EPIN network.

To find the enrichment of DE genes in SNP paths (PrCa SNPs in intermediate proteins) compared to those in the entire EPIN, we computed as enrichment the ratio of Fraction1 / Fraction2, where:

> Fraction1 = (number of DE intermediates within SNP paths) / (number of covered intermediates within SNP paths)
>
> Fraction2 = (number of DE intermediates the EPIN) / (number of covered intermediates in the EPIN).

We identify as enriched EPIN genes those passing enrichment ratio (“enrichment_DE_deseq_SNP.bs.TF.path”) > 1.

pQTL look-up

We downloaded summary statistics with genome-wide association between SNPs and 4907 proteins reported in the deCODE study (Ferkingstad et al. 2021) and annotated with pQTL association the SNPs we identified falling in either in enhancer binding sites or in node genomic locations. The deCODE pQTL summary statistics data contained information on 4,907 proteins and 186 (201 PrCa SNPs out of the 213 PrCa SNPs we looked up were in the data and 186 also matched by alleles). 808 out of the 4,314 genes promoters ("Gene_network") and 278 out of the 885 gene

intermediates (in total 997 out of 4,918 genes promoters and coding for intermediates in our networks) have information on associations with their respective coded proteins covered by the pQTL deCODE data.

Gene dependency and gene effect metrics

Gene Effect and Gene Dependency metrics were downloaded from the DepMap portal (https://depmap.org/portal/). We used both the RNAi (66) and CRISPR (67) datasets.

Data availability

RefSeq hg19 from UCSC Genome Browser is available at the following URL: **http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=694977049_xUU5i1QkIJ50dj5miBt9wkAYuxN3&clade=mammal&org=&db=hg19&hgta_group=genes&hgta_track=knownGene&hgta_table=knownGene&hgta_regionType=genome&position=&hgta_outputType=selectedFields&hgta_outFile Name=knownGene.gtf**

All EPINs and related statistics can be downloaded through the PENGUIN web service at **https://penguin.life.bsc.es/**.

All the raw listed in Table 3, as well as the corresponding processed and metadata for LHSAR and LNCaP related to H3K27ac (HiChIP) and RNAseq have been deposited in GEO. CTCF ChIP-Seq data used in this work comes from ENCODE51 with references GSM2827202, GSM2827203 for LNCaP and GSM2825573, GSM2825574 for the human epithelial cells or prostate that we use to infer CTCF-bindings in LHSAR GSM2825573, GSM2825574.

Code availability

Source code of the related to the PENGUIN protocol is available at github: **https://github.com/bsc-life/penguin_software.**

Source code of the related to the PENGUIN web service is available at github: **https://github.com/bsc-life/penguin_analytics**

R (v.4.2.0) and Python were extensively used to analyze data and create plots. biomart / ensembl from biomaRt package Ensembl hg19 data for overlaps of SNPs with intermediates.

None of the authors have competing financial or non-financial interests.

References

1. Zhang, K., Li, N., Ainsworth, R. I. & Wang, W. Systematic identification of protein combinations mediating chromatin looping. *Nat. Commun*. 7, 12249 (2016).

2. Wang, R. et al. Hierarchical cooperation of transcription factors from integration analysis of DNA sequences, ChIP-Seq and ChIA-PET data. *BMC Genomics* 20, 296 (2019).

3. Liu, N. et al. Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics Chromatin* 14, 41 (2021).

4. Deng, W. & Blobel, G. A. Manipulating nuclear architecture. *Curr. Opin. Genet. Dev*. 25, 1– 7 (2014).

5. Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb. Perspect. Biol*. 7, a019356 (2015).

6. Norton, H. K. & Phillips-Cremins, J. E. Crossed wires: 3D genome misfolding in human disease. *J. Cell Biol*. 216, 3441–3452 (2017).

7. Krumm, A. & Duan, Z. Understanding the 3D genome: Emerging impacts on human disease. *Semin. Cell Dev. Biol*. 90, 62–77 (2019).

8. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* 16, 483–493 (2016).

9. Chen, X., Ma, Q., Shang, Z. & Niu, Y. Super-enhancer in prostate cancer: transcriptional disorders and therapeutic targets. *NPJ Precis Oncol* 4, 31 (2020).

10. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* 13, 919–922 (2016).

11. Giambartolomei, C. et al. H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *Am. J. Hum. Genet*. 108, 2284–2300 (2021).

12. Rebello, R. J. et al. Prostate cancer. *Nat Rev Dis Primers* 7, 9 (2021).

13. Tan, M. H. E., Li, J., Xu, H. E., Melcher, K. & Yong, E.-L. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacol. Sin*. 36, 3–23 (2015).

14. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163, 1011–1025 (2015).

15. de Bono, J. et al. Olaparib for Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* 382, 2091–2102 (2020).

16. Adams, E. J. et al. FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature* 571, 408–412 (2019).

17. Parolia, A. et al. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* 571, 413–418 (2019).

18. Ahmadiyeh, N. et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U. S. A.* 107, 9742–9746 (2010).

19. Schumacher, F. R. et al. Author Correction: Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 51, 363 (2019).

20. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A.* 111, 996–1001 (2014).

21. Pugacheva, E. M. et al. CTCF mediates chromatin looping via N-terminal domaindependent cohesin retention. *Proc. Natl. Acad. Sci. U. S. A.* 117, 2020–2031 (2020).

22. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* 53, 1300–1310 (2021).

23. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 47, W191–W198 (2019).

24. Mondul, A. M. et al. Circulating thyroxine, thyroid-stimulating hormone, and hypothyroid status and the risk of prostate cancer. *PLoS One* 7, e47730 (2012).

25. Hsieh, M.-L. & Juang, H.-H. Cell growth effects of triiodothyronine and expression of thyroid hormone receptor in prostate carcinoma cells. *J. Androl.* 26, 422–428 (2005).

26. Lehrer, S., Diamond, E. J., Stone, N. N. & Stock, R. G. Serum thyroid-stimulating hormone is elevated in men with Gleason 8 prostate cancer. *BJU Int.* 96, 328–329 (2005).

27. Hellevik, A. I. et al. Thyroid function and cancer risk: a prospective population study. Cancer Epidemiol. Biomarkers Prev. 18, 570–574 (2009).

28. Ma, F. et al. Autocrine canonical Wnt signaling primes noncanonical signaling through ROR1 in metastatic castration-resistant prostate cancer. *Cancer Res.* (2022) doi:10.1158/0008-5472.CAN-21-1807.

29. Wei, X. et al. Paracrine Wnt signaling is necessary for prostate epithelial proliferation. *Prostate* 82, 517–530 (2022).

30. Natani, S. et al. Activation of TGF-β - SMAD2 signaling by IL-6 drives neuroendocrine differentiation of prostate cancer through p38MAPK. *Cell. Signal.* 91, 110240 (2022).

31. Xi, X. et al. High expression of small nucleolar RNA host gene 3 predicts poor prognosis and promotes bone metastasis in prostate cancer by activating transforming growth factorbeta signaling. *Bioengineered* 13, 1895–1907 (2022).

32. Speedy, H. E. et al. Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nat. Commun.* 10, 3615 (2019).

33. Zhou, S. et al. Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. Nat. Commun. 11, 441 (2020).

34. Mejía-Hernández, J. O. et al. Targeting MDM4 as a Novel Therapeutic Approach in Prostate Cancer Independent of p53 Status. *Cancers* 14, (2022).

35. Liu, L., Liu, J. & Lin, Q. Histone demethylase KDM2A: Biological functions and clinical values (Review). *Exp. Ther. Med.* 22, 723 (2021).

36. Mandl, M. & Depping, R. ARNT is a potential direct HIF-1 target gene in human Hep3B hepatocellular carcinoma cells. *Cancer Cell Int.* 17, 77 (2017).

37. Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res.* 73, 4180–4184 (2013).

38. Jia, L. et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* 5, e1000597 (2009).

39. Conti, D. V. et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet*. 53, 65–75 (2021).

40. Cheng, I. et al. 8q24 and prostate cancer: association with advanced disease and meta-analysis. *Eur. J. Hum. Genet.* 16, 496–505 (2008).

41. Capik, O. et al. CASC11 promotes aggressiveness of prostate cancer cells through miR-145/IGF1R axis. *Prostate Cancer Prostatic Dis.* 24, 891–902 (2021).

42. Rodriguez-Bravo, V. et al. The role of GATA2 in lethal prostate cancer aggressiveness. *Nat. Rev. Urol.* 14, 38–48 (2017).

43. Galan, S., Serra, F. & Marti-Renom, M. A. Identification of chromatin loops from Hi-C interaction matrices by CTCF-CTCF topology classification. *NAR Genom Bioinform* 4, lqac021 (2022).

44. Ratnakumar, A., Weinhold, N., Mar, J. C. & Riaz, N. Protein-Protein interactions uncover candidate 'core genes' within omnigenic disease networks. *PLoS Genet*. 16, e1008903 (2020).

45. Dey, K. K. et al. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genom* 2, (2022).

46. Javierre, B. M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* 167, 1369–1384.e19 (2016).

47. López de Maturana, E. et al. A multilayered post-GWAS assessment on genetic susceptibility to pancreatic cancer. Genome Med. 13, 15 (2021).

48. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16, 259 (2015).

49. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun*. 10, 4221 (2019).

50. Baca, S. C. et al. Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. *Nat. Commun*. 12, 1979 (2021).

51. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).

52. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* 9, 999–1003 (2012).

53. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680 (2014).

54. Virtanen, P. et al. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 352 (2020).

55. Serra, F. et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* 13, e1005665 (2017).

56. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* 44, D536–41 (2016).

57. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082 (2018).

58. Zhou, Y. et al. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res*. 50, D1398–D1407 (2022).

59. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical finemapping studies. *PLoS Genet.* 10, e1004722 (2014).

60. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet*. 50, 928–936 (2018).

61. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005-D1012 (2019).

62. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855 (2020).

63. Cornwell, M. et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* 19, 135 (2018).

64. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).

65. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* 53, 1712–1721 (2021).

66. Tsherniak, A. et al. Defining a Cancer Dependency Map. Cell 170, 564–576.e16 (2017).

67. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* 49, 1779–1784 (2017).

68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33, 1635–1638 (2016)

Supplementary Figures

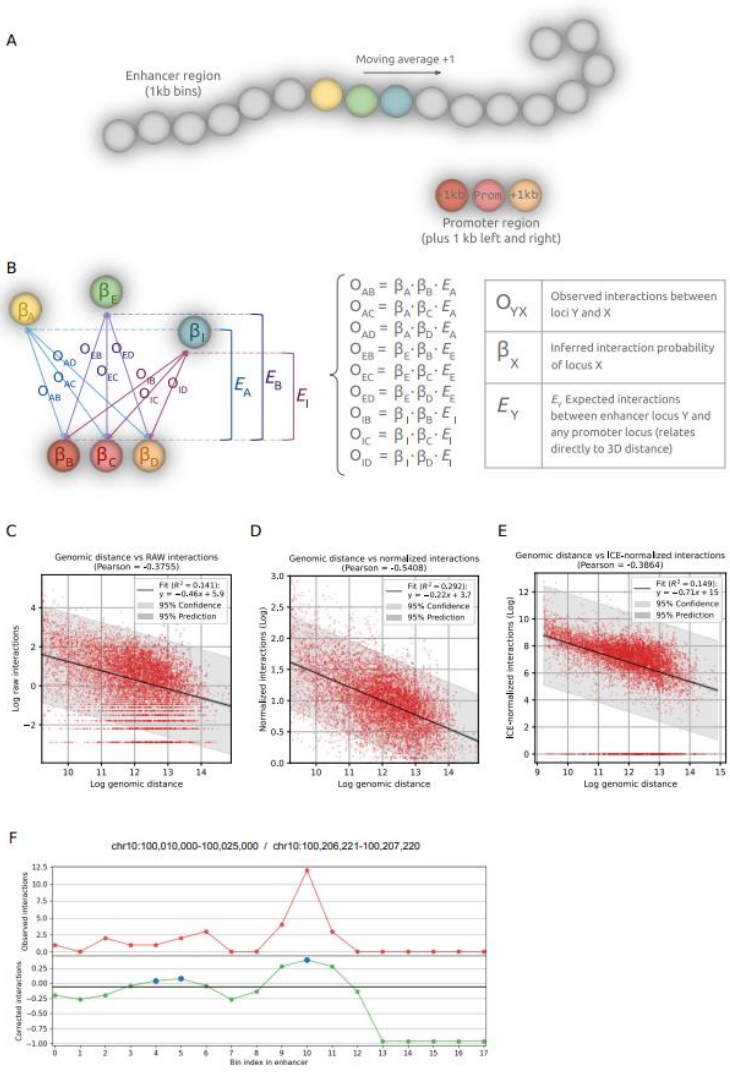


Figure S1: HiChIP promoter-enhancer loop prioritization. (**A**) Schematic representation of a promoter enhancer loop, where the promoter is represented by a 1kb bead surrounded by two neighbor beads, and the enhancer by fifteen 1kb beads. (**B**) Representation of the parameters taken into account to compute the expected number of interactions between a 1 kb loci from the enhancer and the 1kb loci from the promoter. (**C**) Correlation between the genomic distance between enhancer and promoter and the number of interactions. (**D**) Same as (**C**) but with data normalized by the strategy explained in B, (**E**) Same as (**C**) but with data normalized by ICE (**F**) Example of profile of raw interactions along an enhancer (top), and normalized interactions (bottom). Only beads highlighted in blue in the bottom plot would be used (prioritized) in the PENGUIN analysis.

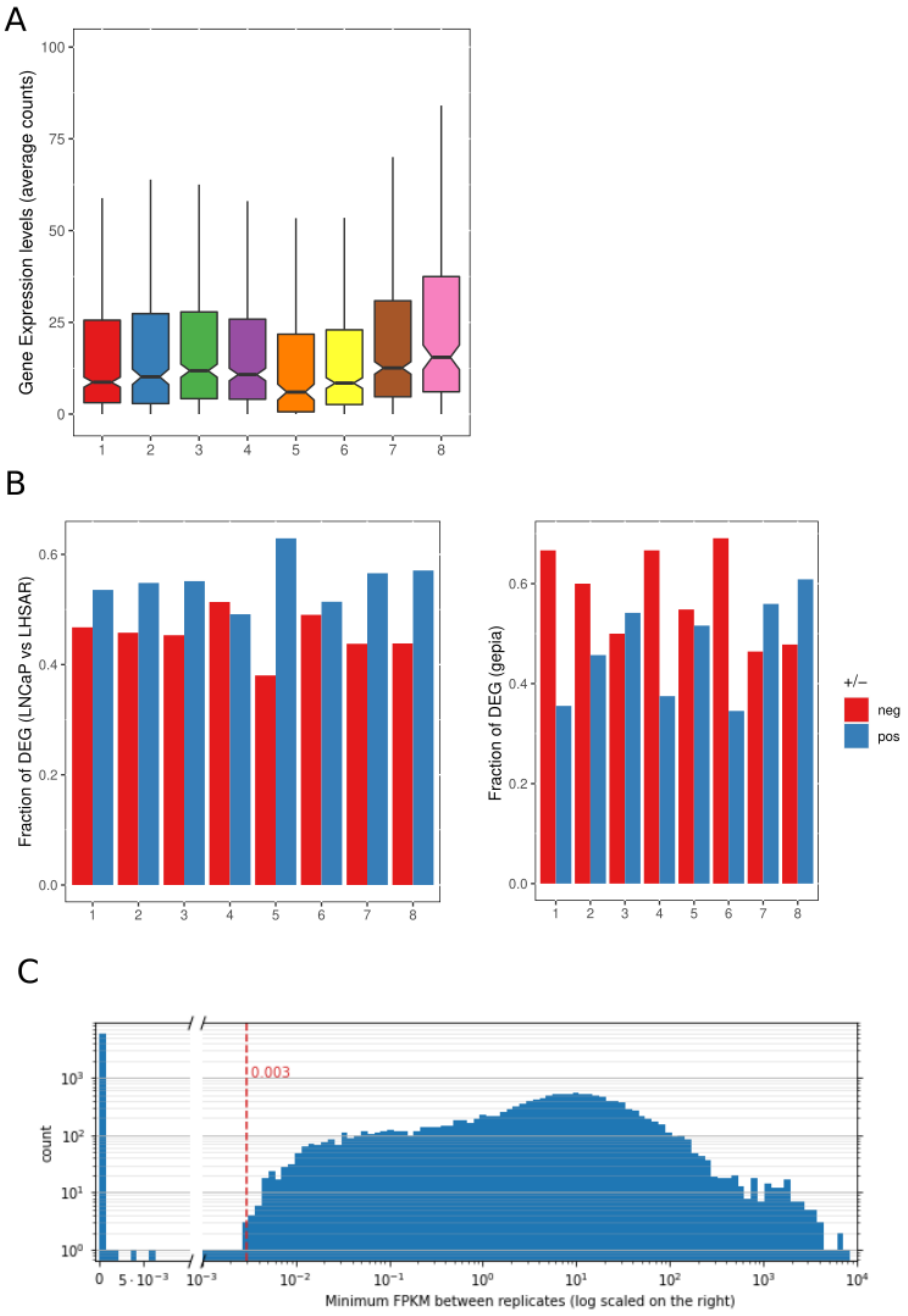


Figure S2: General statistics on gene expression analysis. (**A**) Gene expression per cluster and per EPIN's representative gene. (**B**) Fraction of differentially expressed promoter EPINs (DEG) in each cluster. Positive DEG in blue, negative DEG in red. (**C**) Distribution of expression, showing, per gene, the minimum value between the two RNA-seq replicates (methods). The red dotted line represents the threshold to consider a gene to be expressed.

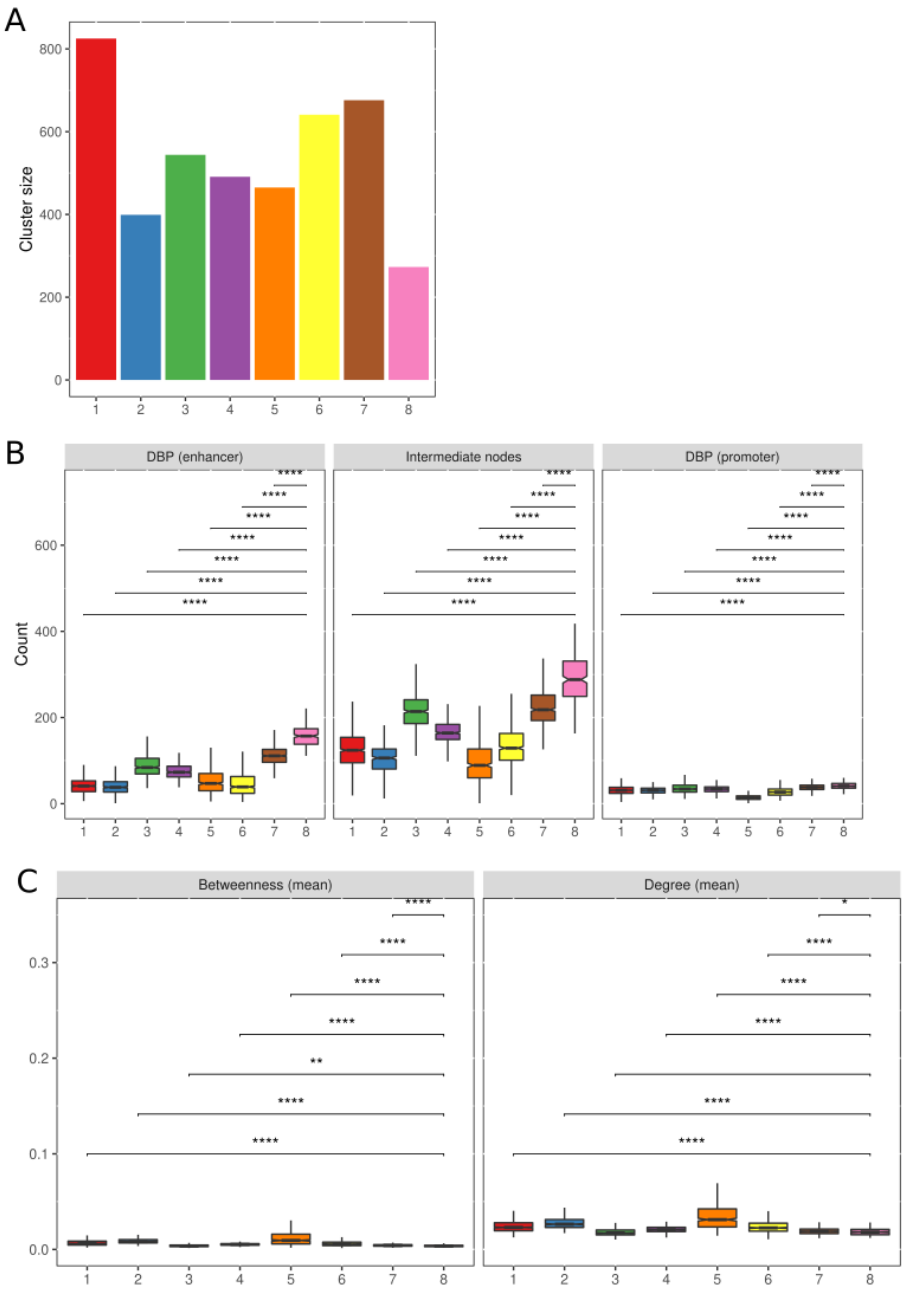


Figure S3: Descriptive statistics on EPIN clusters. (**A**) number of EPINs per cluster, each EPIN is composed of one promoter and at least one enhancer. (**B**) Total number of DNA binding proteins (DBPs) potentially bound to enhancers (left), intermediate nodes from the PPI network between the promoter and its enhancers (center), and DBPs potentially bound to promoter (right), per EPIN cluster. (**C**) Centrality measures on intermediate nodes of EPINs per cluster. Namely betweenness (left) and degree (right); stars on the top indicate the degree of significance after a t-test test (*: p-value<0.0001).

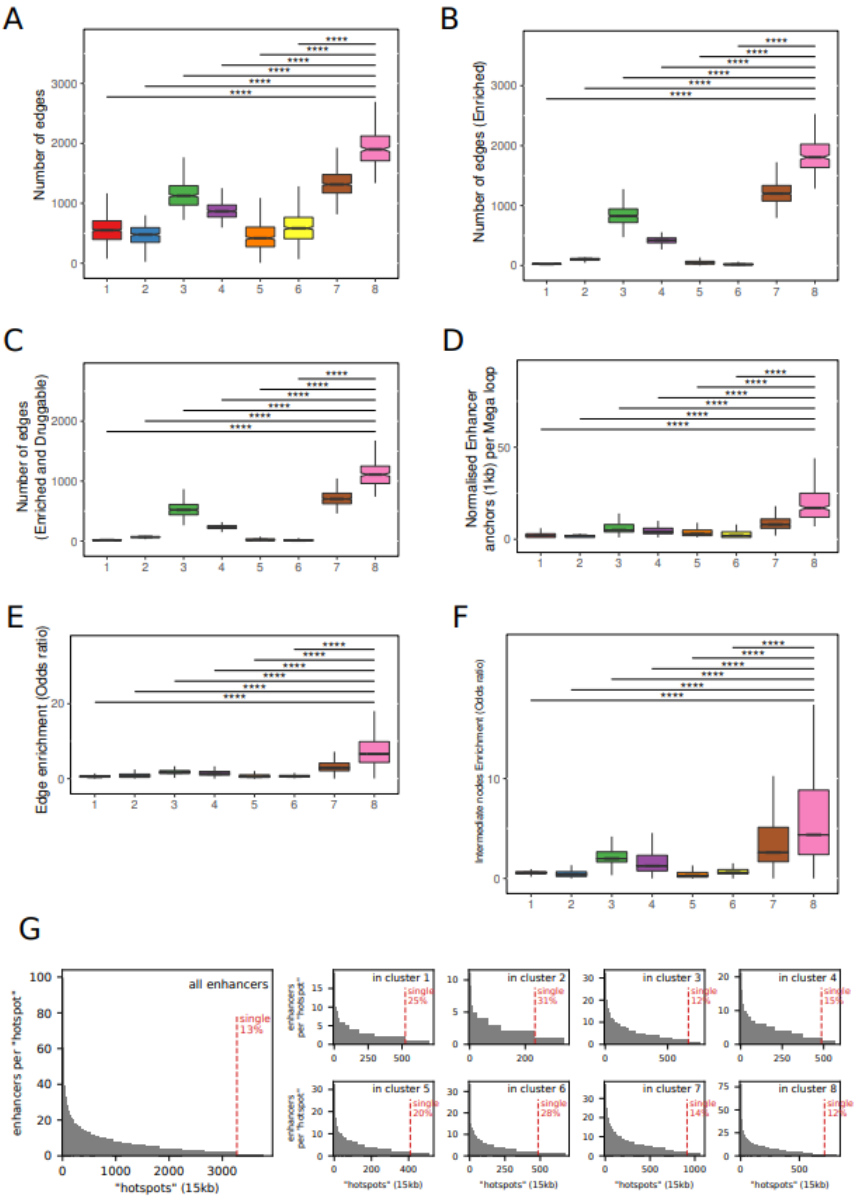
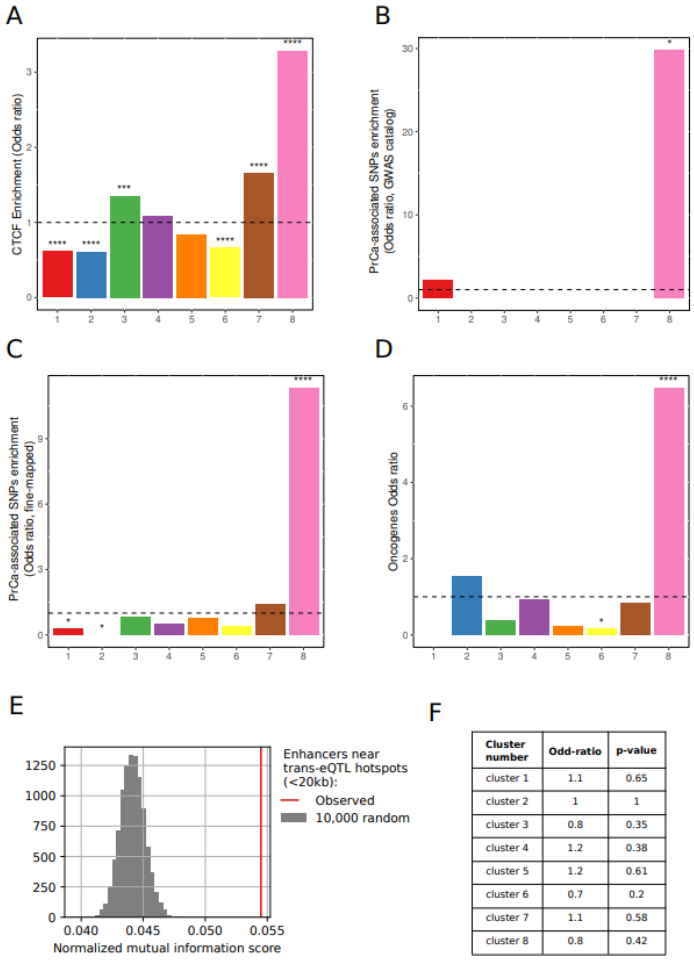


Figure S4: Statistics on EPIN edges. (**A**) Number of edges per EPINs grouped by clusters (boxplots). (**B**) Same with enriched edges (Fisher enrichment test in one cluster with respect to the others, see methods). (**C**) Same as B filtered by edges containing a druggable (see methods) target protein. (**D**) . (**E**). (**F**). Significance levels depicted represent the same as in Figure S3. (**G**) Number of prioritized enhancers per enhancer hotspots. Hotspots are defined as groups of enhancers separated by less than 15kb. Dotted red line shows the proportion of enhancers that are isolated. The different panels show enhancers in the whole genome (left), and in each of our 8 defined clusters (smaller panes on the right).



Cluster number	Odd-ratio	p-value
cluster 1	1.1	0.65
cluster 2	1	1
cluster 3	0.8	0.35
cluster 4	1.2	0.38
cluster 5	1.2	0.61
cluster 6	0.7	0.2
cluster 7	1.1	0.58
cluster 8	0.8	0.42

Figure S5: Enrichment of EPIN clusters in biologically relevant features. (**A**) CTCF. (**B**) GWAS SNPs. (**C**) GWAS SNPs paintor. (**D**) Oncogene Odd ratio. In all panels, stars represent significance of a fisher test against all clusters (*: p-value <0.05; **: p-value<0.01; ***: pvalue<0.001; ***: p-value<0.0001). (**E**) Relationship between trans-eQTL hotspots and the 8 clusters using the concept of normalized mutual information. We focused on enhancers derived from our EPINs, which were associated with trans-eQTL hotspots located within a proximity of less than 20kb. A relatively weak correlation coefficient of 0.0546 if found between the 8 clusters and the hotspots defined by their proximity to trans-eQTL hotspots. Randoms were generated by shuffling the association between enhancers and EPIN clusters. (**F**) We investigated whether a specific cluster exhibits a significant enrichment of trans-eQTL hotspots. For this employed a Fisher test, comparing two contingencies within our list of enhancers: those within or outside a given cluster, and those within or outside a trans-eQTL hotspot.

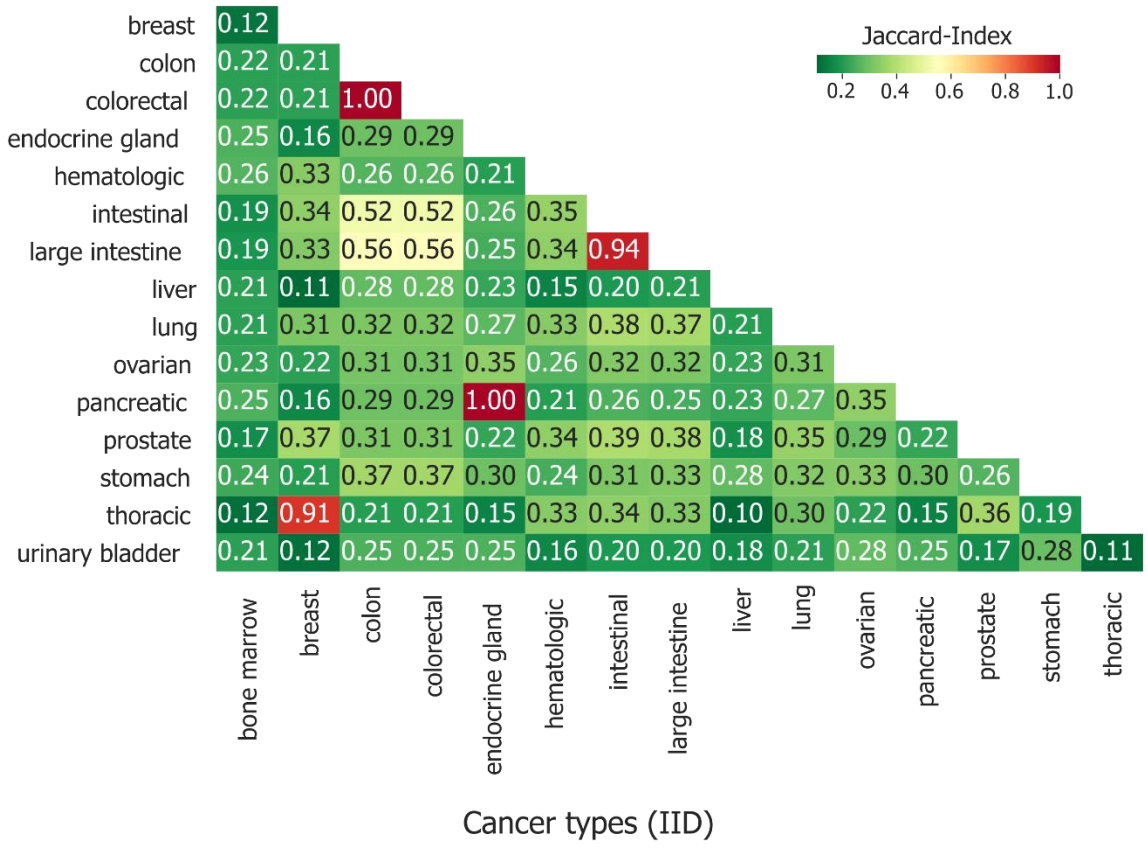


Figure S6. PPI networks comparison. Statistical analyses on PPIs across cancer cell types available at http://iid.ophid.utoronto.ca/. Using the Jaccard index we studied the overlap between PPI networks observing significant variations that were highly specific to each cell type. The results show that the PPIs used in PENGUIN vary significantly depending on the cell types of interest.

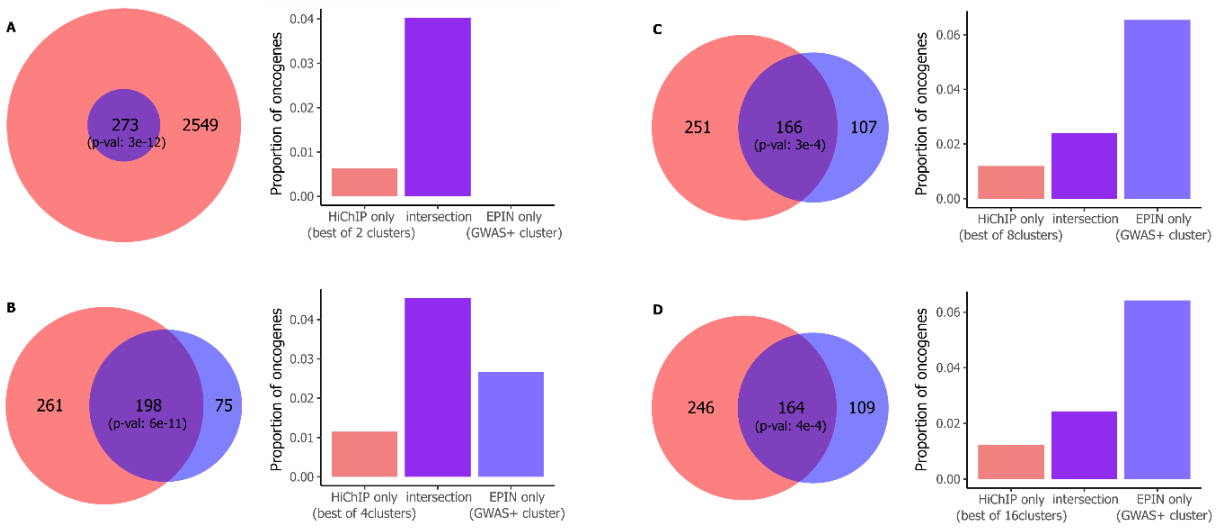


Figure S7. Comparison between clustering based on full EPINs (blue) and using only HiChIP data (no intermediate PPI network) (red). For each clustering strategy, only the cluster most enriched in PrCa SNPs and CTCF peaks is used in the comparison. The comparison is conducted in terms of the proportion of known PrCa oncogenes in the two sets, considering various cluster numbers within the red set (2, 4, 8, and 16 clusters), and only one cluster set (8 clusters for the blue set). Each panel (**A-D**) illustrates a Venn diagram showing the intersection (purple) between the red set and the blue set, and the corresponding fraction of oncogenes as a bar plot. The fraction of oncogenes that are unique to the red set ("HiChIP only") is consistently lower than the fraction of oncogenes that are unique to the blue set ("EPIN only"). Moreover, when compared with 8 and 16 clusters of the red set, the fraction of oncogenes of the "EPIN only" subset is higher than the intersection, indicating a relative gain in oncogenes retrieval when PENGUIN is employed. The significance of the intersection was estimated with a hypergeometric test considering the union of the two sets as the background.

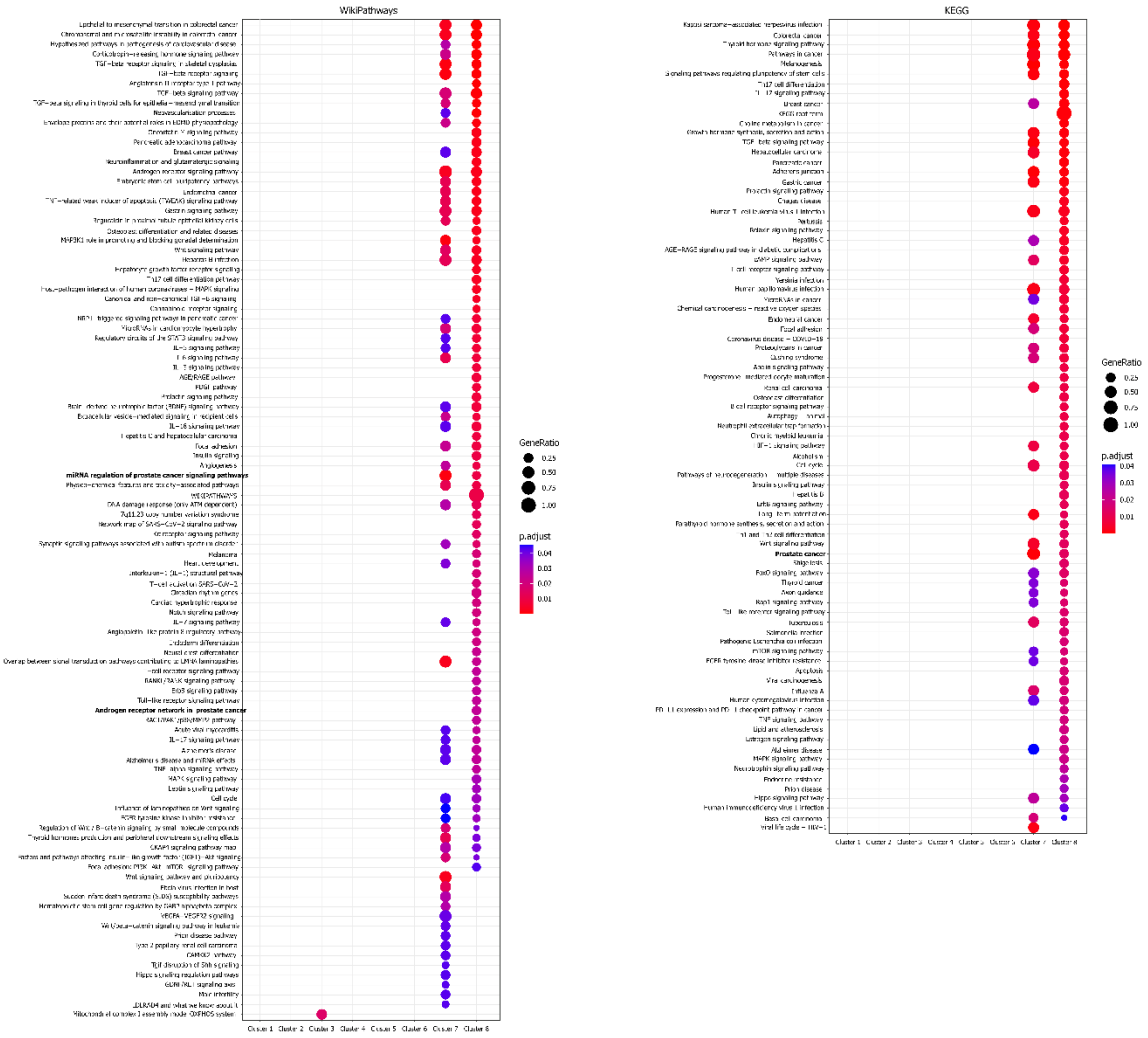


Figure S8. Functional enrichment analysis using g:Profiler to compare the central proteins across different clusters. Two databases of pathways were interrogated, WikiPathways (left 4), and KEGG (right 5). Overall clusters 1, 2, 3, 4, and 6 did not show any enrichments, possibly due to their higher number of central proteins compared to clusters 5, 7, and 8. Among the clusters with enrichments, only cluster 7 showed similarities to cluster 8, such as enrichment in prostate cancer (adjusted p-value = 2.0e-2). Cluster 8 also shows a significant prostate specific WikiPathway Androgen receptor network in prostate cancer.

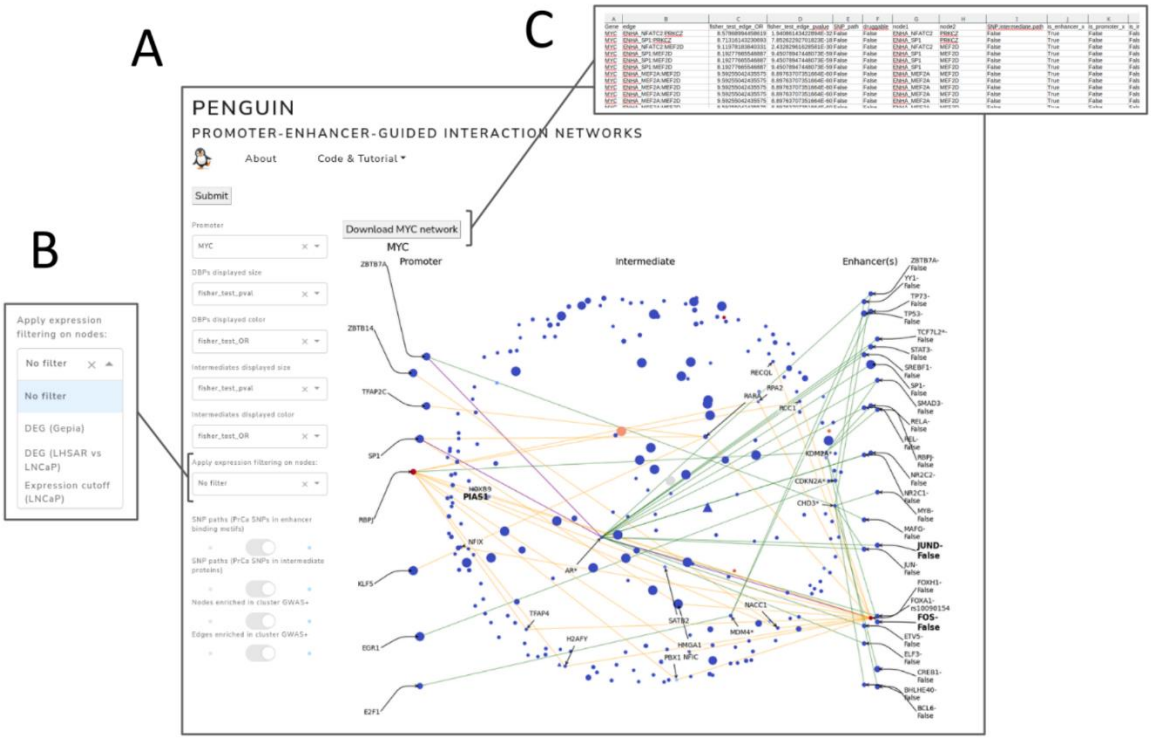


Figure S9. PENGUIN web server. (**A**) Screenshot of the main page of the PENGUIN web server where the EPIN of the MYC promoter is visualized with SNP paths highlighted (orange, PrCa SNPs in enhancer binding motifs; green, PrCa SNPs in intermediate proteins; purple, both). (**B**) Example of displaying node filtering options based on gene expression (deregulated genes, DEGs, identified using Gepia resource, LHSAR versus LNCaP). (**C**) Option to download network and associated statistics for each of the over 4 thousands PrCa EPINs available.

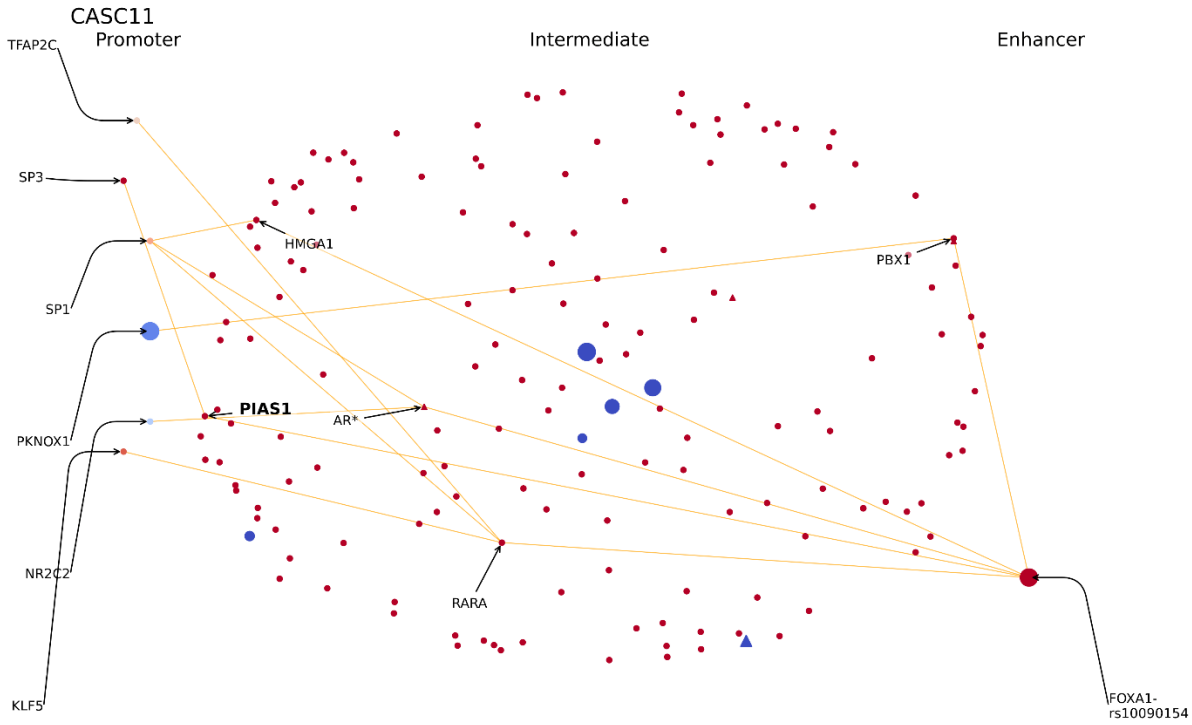


Figure S10A. The CASC11 example. The EPIN of CASC11 promoter is affected by variant rs10090154, the same well-known variant associated with risk of developing prostate carcinoma that we introduced with MYC EPIN. The promoter binds 6 proteins: TFAP2C, SP3, SP1, PKNOX1, NR2C2 and KLF5. Potentially affected protein interactors of the EPIN include: HMGA1, PIAS1, AR, RARA, and PBX1. The yellow lines represent the set of edges that bridge promoter-bound DBPs and intermediate proteins with enhancer-bound DBPs with PrCa SNPs falling in their binding motif.

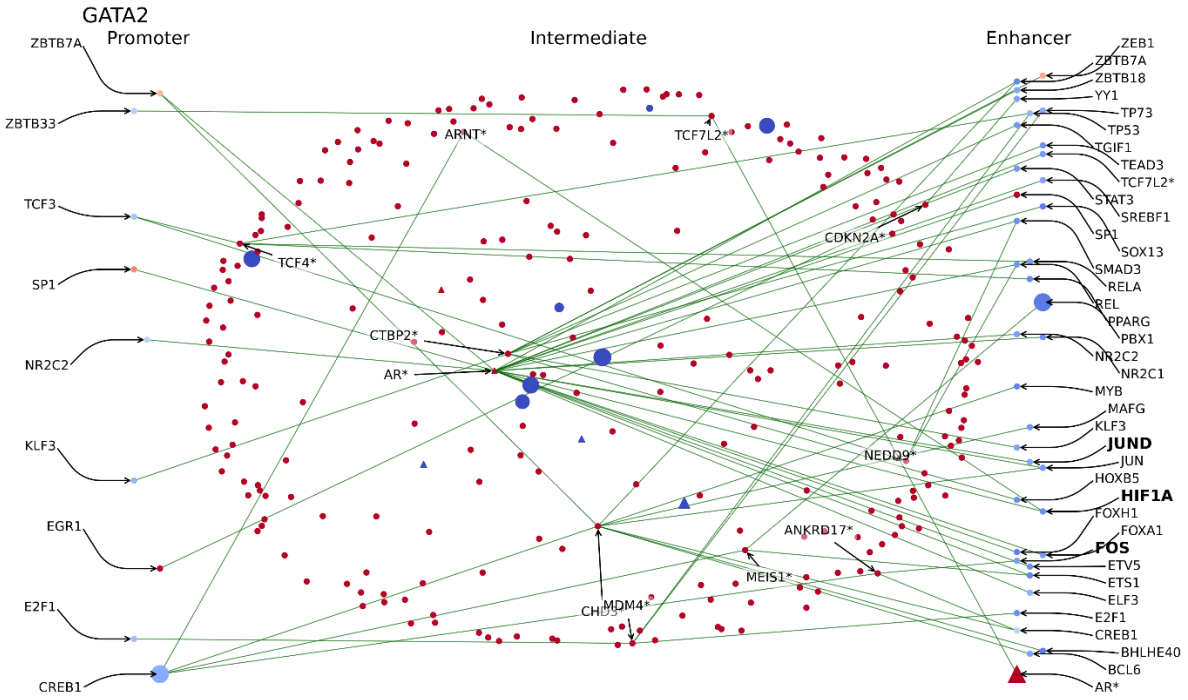


Figure S10B. The GATA2 example. GATA2 EPIN presents up to 11 intermediates affected by PrCa related SNPs, namely TCF4, CTBP2, AR, ARNT, TCF7L2, CDKN2A, NEDD9, ANKRD17, MEIS1, MDM4 and CHD3.Proteins bound to the promoter region include: ZBTB7A, ZBTB33, TCF3, SF1, NR2C2, KLF3, EGR1, E2F1 and CREB1, but most importantly, the EPIN presents AR bound to the enhancer region, which, as we pointed out with MYC, is the target of several PrCa drugs. The green lines represent the set of edges that bridge promoter-bound DBPs with enhancer-bound DBPs through intermediate proteins with PrCa SNPs falling in the genomic region of the corresponding coding gene.

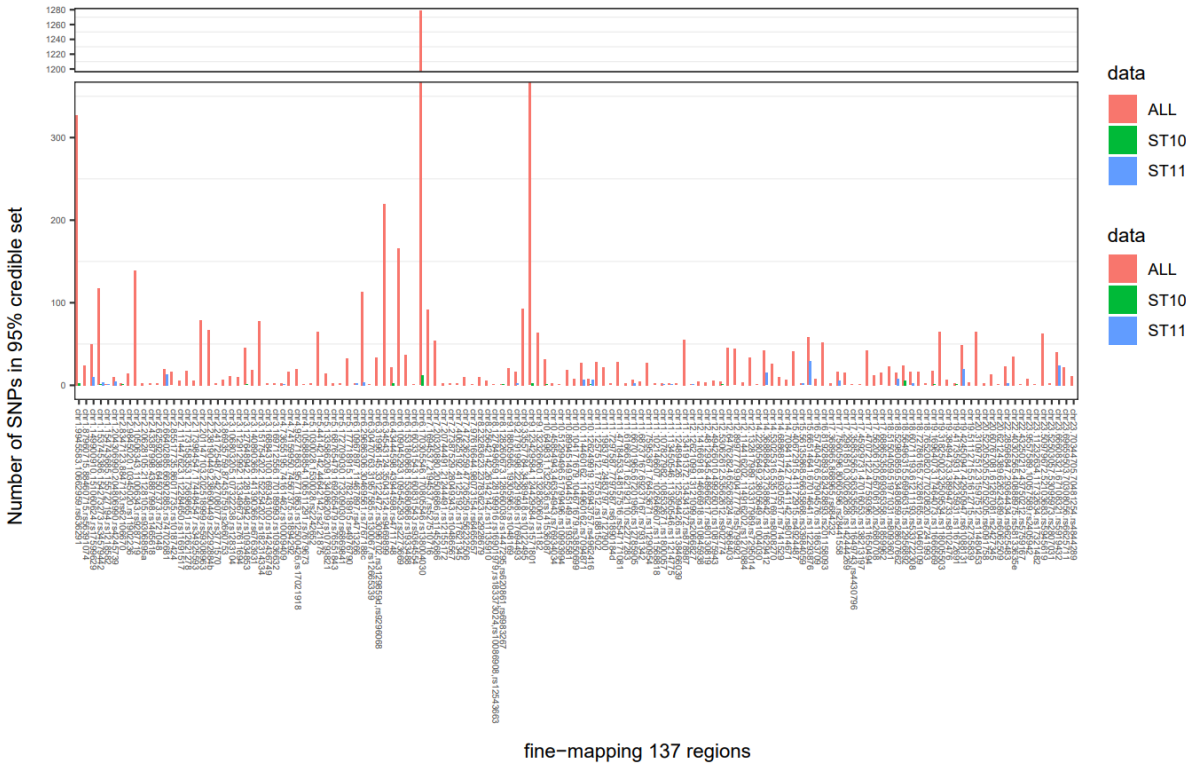
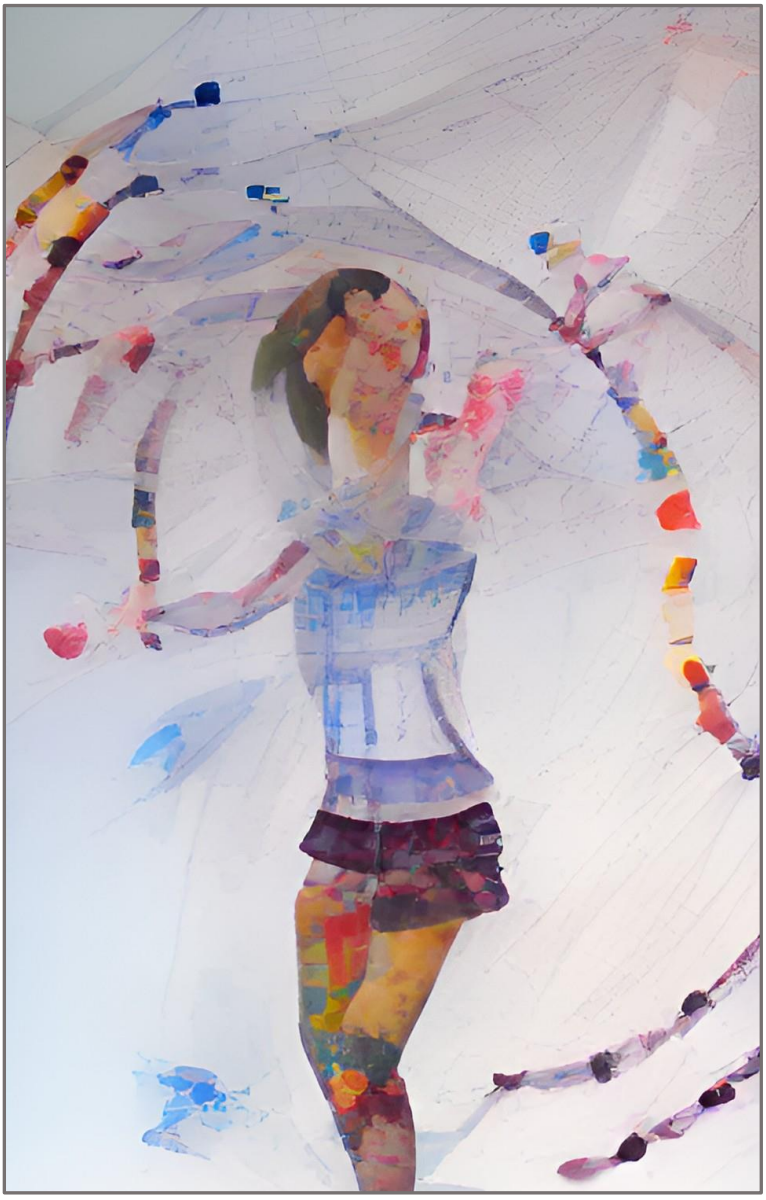


Figure S11. Fine mapped regions. x-axis illustrates 137 regions previously associated with PrCa; y-axis the number of PrCa SNPs (95% credible set) in each region, across ALL (5,412 PrCa SNPs) in red. Color-coded parallel bars in green and blue illustrate the location of the PrCa SNPs identified in ST10 and ST11 and characterized by PENGUIN. No significant correlation (Pearson r=0.2, p-value=0.06 and Pearson r=0.1, p-value=0.3, for ST10 and ST11, respectively) was identified between the number of PrCa SNPs in the regions and the number of PrCa SNPs we prioritized in this work.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universitat de Barcelona