UNIVERSITA' DEGLI STUDI DI PAVIA

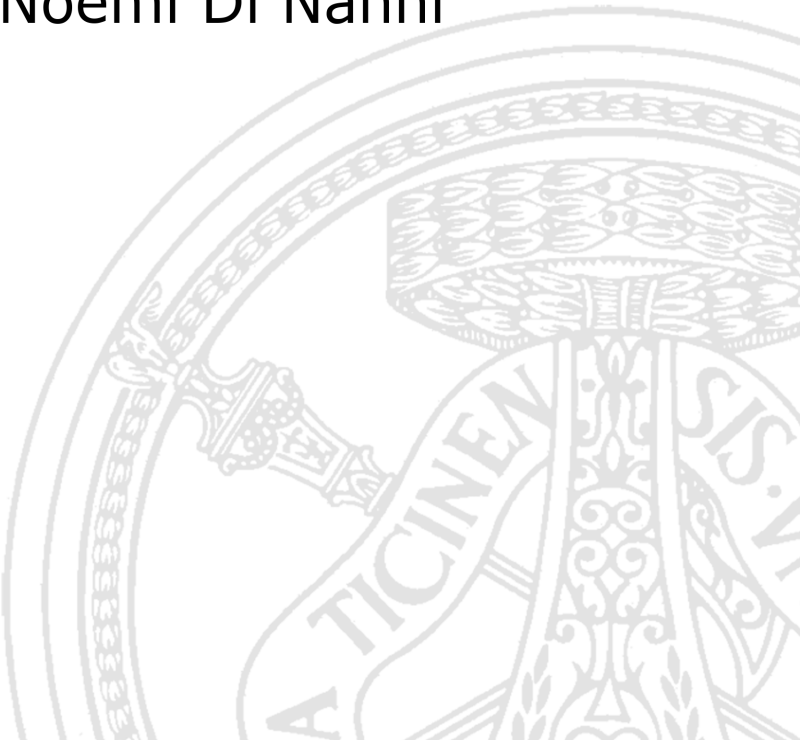
FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXII CICLO - 2019

A NETWORK DIFFUSION METHOD FOR THE INTEGRATION OF MULTI-OMICS DATA WITH APPLICATIONS IN PRECISION MEDICINE.

PhD Thesis by

Noemi Di Nanni



Advisor:
Prof. Riccardo Bellazzi

Supervisor:
Dr. Ettore Mosca

PhD Program Chair:
Prof. Riccardo Bellazzi

Acknowledgements

Four years ago, I wrote the acknowledgements of my master thesis and the last sentence was a quote from Nelson Mandela: "What counts in life is not the mere fact that we have lived; it is what difference we have made to the lives of others that will determine the significance of the life we lead." Those words have been a source of inspiration and have even encouraged me during this research activity over the years. However, this thesis has become reality thanks to the help and collaboration of different people to whom I would like to express my deepest gratitudine.

Above all, I am very grateful to my supervisor, Dr. Ettore Mosca, for his constant patience, expertise, wisdom, guidance and support throughout the time of my PhD, without which this work would not have been possible.

I would also like to thank Dr. Luciano Milanesi, for the opportunity he gave me to pursue my research activity at the Institute of Biomedical Technologies of the National Research Council. I was fortunate indeed to work in his bioinformatics group, where I found a stimulating research environment.

A big thanks goes to my tutor Prof. Riccardo Bellazzi, a talented teacher and passionate scientist, who gave me the opportunity to embark on the PhD Program in Bioengineering and Bioinformatics at the University of Pavia. His good advices and support during these years of my PhD have been invaluable both at academic and personal level.

I would also like to thank Dott.ssa Maria Grazia Daidone of Fondazione IRCCS Istituto Nazionale dei Tumori and her research group for introducing me to the interesting world of breast cancer initiating cells, for their valuable comments, advices and scientific discussion during the field work.

Financial support

This work would not have been possibile without the financial support of the following projects:
- INTEROMICS PB05 and PRIN 2015 20157ATSLF supported by Italian Ministry of Education, University and Research;
- GR-2016-02363997 supported by Italian Ministry of Health, Bando Ricerca Finalizzata Giovani Ricercatori 2016;
- LYRA 2015-0010 and FindingMS ERAPERMED2018-233 GA 779282 supported by Fondazione Regionale per la Ricerca Biomedica (Regione Lombardia);
- GEMMA 825033 supported by European Union's Horizon 2020 research and innovation programme.

List of publications

Articles in peer reviewed journals

- **N. Di Nanni**, M. Gnocchi, M. Moscatelli, L. Milanesi, E. Mosca; "Gene relevance based on multiple evidences in complex networks". *Bioinformatics*, 2019, https://doi.org/10.1093/bioinformatics/btz652.

- A. Riba, **N. Di Nanni**, N. Mittal, E. Arhné, A. Schmidt, M. Zavolan; "Analysis of protein synthesis rate and ribosome occupancy reveals determinants of eukaryotic translation speed". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2019, 116 (30), 15023-15032, https://doi.org/10.1073/pnas.1817299116.

- **N. Di Nanni**, M. Bersanelli, F. Cupaioli, L. Milanesi, A. Mezzelani, E. Mosca; "Network-based integrative analysis of genomics, epigenomics and transcriptomics in autism spectrum disorders". *International Journal of Molecular Sciences (IJMS)*, 2019, 20:13, 3363, https://doi.org/10.3390/ijms20133363.

- **N. Di Nanni**, M. Gnocchi, M. Moscatelli, L. Milanesi, E. Mosca; "isma: an R package for the integrative analysis of mutations detected by multiple pipelines". *BMC Bioinformatics*, 2019, 20:107, https://doi.org/10.1186/s12859-019-2701-0.

- **N. Di Nanni**, M.Bersanelli, L. Milanesi, E. Mosca; "Network diffusion promotes the integrative analysis of multiple omics". (under review in: *Frontiers in Genetics*).

Contributions to conference proceedings

- **N. Di Nanni**, V. Appierto, C. De Marco, V. Angeloni, M.G. Daidone, L. Milanesi, E. Mosca; "Whole-exome sequencing of breast cancer initiating cells and paired primary tumors: the impact of variant callers and filtering strategies". *14th Annual Meeting of the Bioinformatics Italian Society*, July 2016, (Best Poster Award).

- A. Riba, N. Mittal, **N. Di Nanni**, A. Schmidt and M. Zavolan; "Exploring the predictability of protein synthesis rates in the yeast Saccharomyces cerevisiare". *13th* $[BC]^2$ - *Basel Computational Biology Conference*, September 2017.

- **N. Di Nanni**, M. Gnocchi, M. Moscatelli, L. Milanesi, E. Mosca; "Network diffusion on multiple-layers: current approaches and an application to Rheumatoid Arthritis data". *Network Tools and Applications in Biology*, October 2017, DOI: 10.7287/peerj.preprints.3310v1.

- **N. Di Nanni**, V. Appierto, C. De Marco, E. Ortolan, L. Milanesi, M. Daidone, E. Mosca; "Network diffusion for the integrative analysis of multiple -omics: case studies on breast cancer". *European Human Genetics Conference*, June 2018, http://www.abstractsonline.com/pp8/#!/4652/presentation/5905.

- **N. Di Nanni**, M. Gnocchi, M. Moscatelli, L. Milanesi, E. Mosca; "Integrative analysis of multiple -omics with network diffusion". *Congresso Nazionale del Gruppo di Bioingegneria*, June 2018.

- **N. Di Nanni**, L. Milanesi, E. Mosca; "Network diffusion-based method for gene prioritization integrating multiple omics and gene networks". *15th Annual Meeting of the Bioinformatics Italian Society*, June 2018.

- E.Mosca, V.Appierto, **N. Di Nanni**, C. De Marco, L. Milanesi, MG Daidone; "Functional gene networks underlying somatic mutations in breast cancer initiating cells". *Cancer Stem Cell: Impact on Treatment*, December 2018.

- F. Genova, E. Mosca, **N. Di Nanni**, F. Cupaioli, A. Mezzelani, M. Longeri; "Identification of miRNAs and evaluation of their differential expression in Abyssinian cat amyloidosi". *International Conference of canine and feline genetics and genomics*, May 2019, https://air.unimi.it/handle/2434/648055#.XQgDmNMzbOQ.

- F. Genova, S. Nonnis, E. Maffioli, F. Grassi Scalvini, **N. Di Nanni**, F. Cupaioli, E. Mosca, A. Mezzelani, G. Sironi, LA Lyons. "Proteins and miRNA in feline renal amyloid deposits". *International Society for Animal Genetics Conference*, July, 2019.

Abstract (English)

The application of high-throughput sequencing technologies has made available data relating to different molecular entities for the same biological system (e.g. DNA, RNA, proteins, "omics-data" hereafter), allowing to obtain a more complete picture of the molecular mechanisms associated with human diseases. However, the large amount of heterogeneous data has raised the need to create bioinformatics methodologies to extrapolate information that could improve our understanding of human diseases (e.g. diagnostic biomarkers and therapeutic targets). The methods currently available do not meet all the needs of research projects, in terms of data types, data size, type of result generated, computational cost and software availability. The problem of the integration of multi-omic data is therefore nowadays an open challenge in several biomedical research projects. In this context, knowledge about the complex web of direct and indirect interactions among macromolecules at genome scale is a powerful resource to explain multiple omics measurements, highlighting the molecular mechanisms underlying diseases. The increasingly recognized importance of the use of network principles and methods for the study of human diseases has led to a new field of knowledge called "network medicine".

In this work, a new method for integration of omics data (mND) is proposed for the prioritization and classification of genes, using information coming from molecular interactions (protein-protein interaction networks) and multi-omics data. In particular, the described approach quantifies the relevance of a gene in a biological process taking into account the network proximity of the gene and its first neighbours to other altered genes in a

genome-scale gene network. Beyond the novel way of prioritizing genes, mND introduces a layer-specific gene classification to underline gene roles in each layer and suggests molecular mechanisms in relation to the datasets studied. mND has been shown to outperform other leading alternative methods in finding altered genes in network proximity in one or more layers and in recovering known cancer genes. Moreover, thanks to its versatility, the proposed method has been applied to analyze (i) multi-omics data and (ii) multiple samples of same omic type.

In the first type of application, mND was used to integrate multi-omics data of two different complex and heterogeneous diseases, such as: breast invasive carcinoma (BC) and autism spectrum disorders (ASDs). In BC, the integrative analysis of mutations and differential expression data collected from The Cancer Genome Atlas underlined a disease gene module supported by multiple biological evidences and led to enrichment in relevant pathways involved in breast cancer. In ASDs, mND integrated genomic, epigenomic and transcriptomic data obtained from several large studies on ASDs. Our study suggested a gene network significantly enriched in genes supported by one or more of the considered evidence (genomics, epigenomics, and transcriptomics) and that participate in several pathways relevant to ASDs.

In the second type of application, mND was used to integrate mutation profiles detected by means of whole-exome sequencing (WES) of subjects observed in breast cancer initiating cells. The integrative analysis allowed the identification of networks of functionally related genes that are "hot spots" of mutations in breast cancer initiating cells, molecular pathways and actionable targets. WES data analysis revealed the need to develop a software that allows the integrative analysis of mutations detected by multiple variant callers. In fact, the problem of identifying mutations from WES data of paired mutation-control samples is not simple because each variant caller encodes the same information relating to mutations in multiple ways, the number of mutations identified by each variant caller varies significantly and the overlap between the variants caller is very low. To this aim, "isma", an R package for the integrative analysis of somatic mutations detected by multiple pipelines, was also introduced. isma provides a series of functions to integrate and analyze the results of different variant callers,

to highlight the most reliable mutation sites, to quantify the consensus, to underline potential outliers and integrate evidences from publicly available mutation catalogues.

In conclusion, this research activity introduces an important advance in the class of multi-omics methods: a new way to quantifiy the relevance of genes on the basis of their complex web of interactions and multi-omics data. Importantly, mND is applicable to a broad range of data types and experimental designs. Furthermore, mND is available to the scientific community as R software package (https://www.itb.cnr.it/mnd) with an extensive documentation covering installation, usage and reproducible examples. The obtained results indicate that the proposed method could help to unravel the networks of molecular players associated with the biological process under investigation. When applied to study a human disease, the multi-omics networks found by mND are useful sources of biomarkers and actionable targets.

Abstract (Italian)

L'applicazione delle tecnologie di sequenziamento di nuova generazione ha reso disponibile grosse quantità di dati relativi a entità molecolari differenti per lo stesso sistema biologico (e.g. DNA, RNA, proteine, nel seguito "dati-omici"), allo scopo di ottenere un quadro più completo dei meccanismi molecolari associati alle malattie umane e sviluppare migliori strumenti diagnostici, prognostici e terapeutici. Tuttavia, la grande quantità di dati e la loro eterogeneità ha fatto emergere la necessità di sviluppare nuovi metodi di analisi. Infatti, i metodi attualmente disponibili non rispondono a tutte le necessità dei progetti di ricerca, in termini di tipo di dato, dimensione, tipo di risultato generato, costo computazionale e disponibilità del software. Il problema dell'integrazione di dati multi-omici è quindi oggigiorno ricorrente in progetti di ricerca in campo biomedico. In questo contesto, la conoscenza delle complesse interazioni dirette (fisiche) e indirette (funzionali) tra macromolecole su scale genomica è una risorsa importante per spiegare le evidenze biologiche. La crescente rilevanza per lo studio delle malattie umane, dell'uso di principi e metodi relativi ai network biologici ha portato al nuovo campo del sapere denominato "network medicine".

A supporto di questa necessità, in questa tesi, è stato quindi sviluppato un nuovo metodo per l'integrazione di dati omici, mND. Il metodo proposto permette di ottenere una prioritizzazione e una classificazione dei geni rilevanti nel processo biologico preso in considerazione, utilizzando le informazioni provenienti dalle interazioni molecolari (*protein-protein interaction networks*) e dai dati omici (per esempio, frequenze di mutazione e variazione a livello di espressione, nel seguito "layer"). In particolare,

l'approccio descritto quantifica l'importanza di un gene in un processo biologico tenendo conto della vicinanza nel network del gene e dei suoi primi vicini rispetto agli altri geni alterati. Oltre al nuovo modo di creare una prioritizzazione dei geni, mND introduce una classificazione dei geni specifica per layer che permette di individuare il ruolo funzionale del gene e di suggerire i meccanismi molecolari fra i geni coinvolti. E' stato dimostrato che mND ha prestazioni superiori rispetto ai metodi alternativi disponibili nell'individuare geni alterati vicini nel network in uno o più layer e nel prioritizzare geni associati al cancro. Inoltre, grazie alla sua versatilità, il metodo proposto è stato applicato per analizzare: (i) dati multi-omici e (ii) la stessa tipologia di dato omico a livello di singoli soggetti.

Nel primo tipo di applicazione, mND è stato utilizzato per integrare dati multi-omici di due differenti patologie eterogenee: tumore alla mammella (BC) e autismo (ASD). Nel BC, l'analisi integrativa delle mutazioni e dei dati di espressione differenziale raccolti da *The Cancer Genome Atlas* ha evidenziato un modulo connesso tra i geni prioritizzati da mND e ha correttamente individuato pathways coinvolti nel carcinoma mammario. Nell'ASD, mND è stato utilizzato per integrare dati di genomica, di epigenomica e di trascrittomica ottenuti da numerosi studi. La nostra analisi ha suggerito un modulo di geni significativamente arricchito di geni supportati da una o più evidenze biologiche considerate (genomica, epigenomica e trascrittomica) e che partecipano a diversi pathways rilevanti per l'ASD.

Nel secondo tipo di applicazione, mND è stato utilizzato per integrare profili di mutazione di soggetti affetti dal cancro alla mammella, di cui è stato sequenziato l'intero esoma (Whole Exome Sequecing, WES) della parte tumorale con proprietà staminali. L'analisi integrativa ha consentito l'identificazione di un network di geni arricchito di mutazioni, contenenti potenziali target per il trattamento del cancro alla mammella. Inoltre, l'analisi dei dati di WES ha fatto emergere la necessità di sviluppare un software che consentisse l'analisi integrativa delle mutazioni identificate da degli strumenti che eseguono la chiamata delle varianti (*variant caller*). In effetti, il problema di identificare le mutazioni dai dati WES non è semplice perché ogni *variant caller* codifica le stesse informazioni relative alle mutazioni in più modi, il numero di mutazioni identificate da ciascun *variant*

caller varia in modo significativo e la sovrapposizione tra i *variant caller* è molto bassa. A questo scopo è stato introdotto "isma", un pacchetto software in R per l'analisi integrativa delle mutazioni somatiche rilevate da più *variant caller*. isma fornisce una serie di funzioni per integrare e analizzare i risultati di diversi *variant caller*, di evidenziare i siti di mutazione più affidabili, di effettuare un'analisi di consenso, di evidenziare possibili outliers e di integrare siti già catalogati da altri studi.

In conclusione, questa attività di ricerca introduce un importante progresso nella classe dei metodi multi-omici: un nuovo modo per quantificare la rilevanza dei geni sulla base della loro complessa rete di interazioni e dei dati multi-omici. E' importante sottolineare che mND è applicabile a un'ampia gamma di tipi di dati e patologie. Inoltre, mND è disponibile per la comunità scientifica come pacchetto software in R (https://www.itb.cnr.it/mnd) con una vasta documentazione che copre installazione, utilizzo ed esempi riproducibili. I risultati ottenuti indicano che il metodo proposto potrebbe aiutare a districare la complessità molecolare dei networks associati al processo biologico preso in esame. Negli studi relativi alle malattie umane, i networks multi-omici individuati da mND possono fornire un utile supporto per la selezione di biomarcatori e terapie mirate.

Contents

Chapter 1

General introduction and thesis overview

1.1 Background

Current sequencing technologies allow us to collect data related to different types of molecular entities (DNA, RNA, proteins, etc.) for the same biological system. The availability of these heterogeneous datasets enables the reconstruction of a more complete picture of the molecular events associated with human diseases.
In this context, integrative approaches for the analysis of "omics" data, such as the genome, the transcriptome, the epigenome, the proteome, are valuable to support a better understanding of biological systems and the development of successful precision medicine [1–6]. The goal of precision medicine is to target the right treatments to the right patients at the right time [7] (Figure 1.1) and "-omics" approaches can facilitate the clarification of biological processes whose mechanisms are still unclear [2].
This would be especially helpful for complex and heterogeneous diseases where multiple factors are responsible for the phenotypes, such as autism

[2, 8–10] and cancer [6, 11–13].

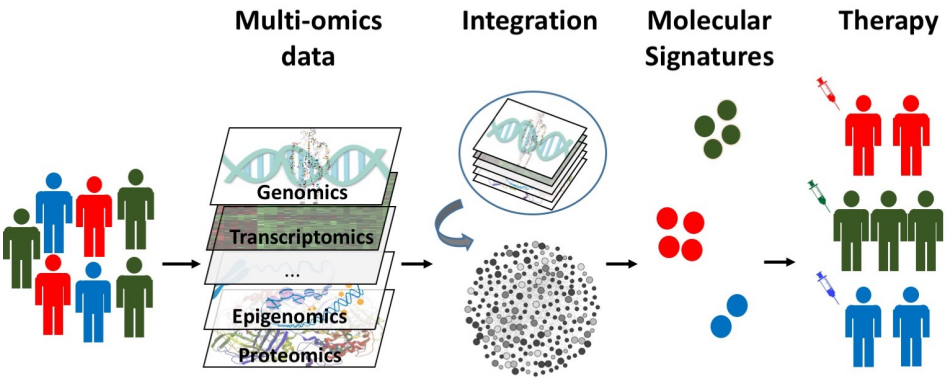


Figure 1.1: Multi-omics data integration for precision medicine. Multi-omics data are collected from patients and then integrated to develop molecular signatures (e.g. biomarkers, network/pathway signatures) that can be used to identify patients that are most likely to benefit from a treatment.

However, an ongoing challenges in the era of precision medicine and multiomics is the integration and interpretation of several "-omics" to boost understanding of biological mechanisms [5, 14–17].

There are many issues that make integrative analyses a challenge, for example: the biological interpretation and knowledge, the data-preprocessing of different existing data types and formats, the heterogeneity of omics data, the complexity of biological systems, the different type of technology and their technological limits, the missing values, the low number of biological samples and high number of biological variables.

In this scenario, biological networks, composed of direct and indirect interactions among genes, are powerful resources to explain multiple omics measurements and to highlight molecular mechanisms underlying diseases [18–20]. In particular, the principle of network diffusion, thanks to its power of quantifying network proximity considering simultaneously all the possible network paths between query genes, has been proposed to solve several problems in biological data analysis [21, 22]. Considering all possible paths among the nodes, network-diffusion caputeres the complexity of biological

networks and enable the identification of systems-level properties of molecular systems. Recently, many methods have been proposed that rely on some kind of network diffusion. In these models, scores derived from omics measurements are propagated throughout a network of molecular interactions in order to obtain a quantitative estimation of network proximity between the molecular entities involved.

However, current methods require a specific combinations of biological data and their applicability is limited by the number of omics and experimental designs. Therefore, the revolution of high throughput technologies demanded the development of more versatile methods for the integration of "-omic" data.

1.2 Thesis overview

Following the above considerations, the aim of this thesis is to develop a new approach based on network-diffusion that is able to integrate multiple biological evidences without restrinctions in terms of layer number and data types. This thesis project has been carried out at the Institute of Biomedical Technologies of the National Research Council (Segrate, MI).

The following chapters are organized as follows.

Chapter 2 introduces the fundamentals of the mathematical machinery of network diffusion, the main methods that use network diffusion processes for the integrative analysis of omics data and open issues.

Chapter 3 presents a novel network-based method for the integration of multi-omics data, called mND [23]. Firstly, algorithm is introduced; secondly, computational cost and mND R package are described; lastly, performance comparison between mND and leading alternative methods are discussed.

Chapter 4-5 present two applications of mND approach to integrate multiple biological evidences. In the first application, results emerged from several studies of genomics, epigenomics and transcriptomic on Autism Spectrum Disorders are integrated by mND method [24]. This work has

been carried out within the project "Genome, Environment, Microbiome & Metabolome in Autism: an integrated multi-omics systems biology approach to identify biomarkers for personalized treatment and primary prevention of Autism Spectrum Disorders" (GEMMA) supported by European Union's Horizon 2020 research and innovation programme. In the second application, mutation profiles observed in breast cancer initiating cells are integrated by mND method to find the networks of functionally related genes that are "hot spots" of mutations. This work has been carried out in collaboration with IRCCS Istituto Nazionale dei Tumori (Milano)[1] within the project "Integrative Mutational Analysis of patient-derived Breast Cancer Initiating Cells to disentangle tumor genetic complexity and identify actionable targets for precision medicine" (INTEROMICS BCIC-IMA) supported by Italian Ministry of Education, University and Research.

Chapter 6 summarizes the main conclusions of this research activity, limitations and future directions.

Finally, **Appendix A** contains further results on mND's performance assessment. In **Appendix B** "isma", a new R package that integrates analysis of mutations detected by multiple pipelines, is presented [25].

[1] All patients participating in the study signed an informed consent according to the Declaration of Helsinki. The study was approved by the Ethical Review Board of Fondazione IRCCS Istituto Nazionale dei Tumori of Milan.

Chapter 2

Multi-omics data integration

The development of integrative methods is one of the main challenges in bioinformatics. The principle of network diffusion - also referred to as network propagation - thanks to its power of quantifying network proximity considering simultaneously all the possible network paths between query genes, has been proposed to solve several problems in biological data analysis. Indeed, network diffusion provides a quantitative estimation of network proximity between genes associated with one or more different data types, from simple binary vectors to real vectors. Therefore, this powerful data transformation method has also been increasingly used in integrative analyses of multiple collections of biological scores and/or one or more interaction networks. This chapter presents an overview of the state of the art of bioinformatics pipelines that use network diffusion processes for the integrative analysis of omics data, open issues and potential developments in the field. [1]

[1] The contents of this chapter are included in the manuscript entitled: *"Network diffusion promotes the integrative analysis of multiple omics"*, under review in *Frontiers in genetics*. Authors: N. Di Nanni, M. Bersanelli, L. Milanesi, E. Mosca.

2.1 Integrative analyses of multiple omics

"Omics" technologies provide data related to different types of molecular entities (e.g. DNAs, RNAs, proteins) at increasing sensitivity, down to single-cell level [26]. This offers the opportunity for integrative analyses that lead to a more comprehensive view of a biological system [5, 8]. However, integrative analyses involve several issues due to the types of biological information considered, coverage of the pool of molecular entities under investigation, data distribution types, noise and research questions that need to be addressed [16, 27, 28], just to mention a few. Therefore, the development of integrative methods is one of the main challenges in bioinformatics.

Integrative methods can be classified in three groups by objective (Figure 2.1 A): understanding of the molecular mechanisms (e.g. genes prioritization, function prediction, module detection), clustering of samples (e.g. identification of disease subtypes) or prediction of samples' outcome/phenotype (e.g. survival) [29]. These three objectives can be achieved using a single type or multiple types of omics, possibly combined with data about molecular networks (Figure 2.1 B), in a supervised or unsupervised settings.

From a methodological point of view, the arising importance of interaction networks and the type of statistical approach pave the way for a first broad classification of integrative methods. In particular, these can be divided into four broad classes depending on whether they use molecular networks and Bayesian theory: network-free non-Bayesian, network-free Bayesian, network-based non-Bayesian and network-based Bayesian [14].

Molecular networks represent a powerful framework to integrate and explain omics datasets [14, 18, 19]. Network-based methods take into account known (e.g. protein-protein interactions) and/or inferred (e.g. functional relations found by gene co-expression analysis) relations between biological variables. Significantly, network-based approaches enable the identification of system-level patterns that reflect the architecture of molecular networks. A common systems-level pattern is, for instance, the presence of gene networks that are "hot" spots of mutations in cancer and reflect the several possible combinations of mutations that are likely to lead to a common

pathological phenotype, because affect the same pathway [30, 31].

In the last decades, the mathematical machinery of network diffusion (ND) has been exploited in several network-based pipelines with different aims, like gene prioritization, gene module identification, drug target prediction and disease subtypying, thanks to its ability of amplifying association between variables (e.g. genes) that lie in network proximity. This amplification is realized by means of different methods that can be brought back to random walks, random walks with restarts or diffusion kernels. Recently, Cowen et al. [21] provided a general overview of the unifying mathematical machinery of ND, showing its broad utility in several problems of genetic research, while, previously, Wang et al. [32] described the application of ND to predict gene function and phenotype.

In this chapter, we focused on the problem of jointly analysing biological networks and multiple collections of scores ("layers") derived from omics assays, which is addressed by many pipelines relying on ND. We reviewed the integrative methods by aim, input data type, molecular network, way in which ND is exploited during the integrative analysis and application; lastly, we discussed open issues and potential developments in the field.

2.2 Molecular networks

Network-based methods require, by definition, a molecular network that enters the analysis pipeline at some point. The complex web of molecular interactions that occur within human cells is often referred to as "interactome" [30]. Such interactions can be of rather different types and are usually distinguished in two classes: *biophysical* and *functional* [33]. Biophysical interactions indicate actual molecular contact between two molecular entities, such as protein-DNA biding or protein-protein binding in a protein complex. Functional interactions indicate any kind of biologically relevant interaction (at the molecular scale), like co-expression or synthetic lethality. There is still no unique reference for the human interactome [34], but several efforts are underway. Four proteome-scale PPI interaction maps have recently been generated using different high-throughput approaches based

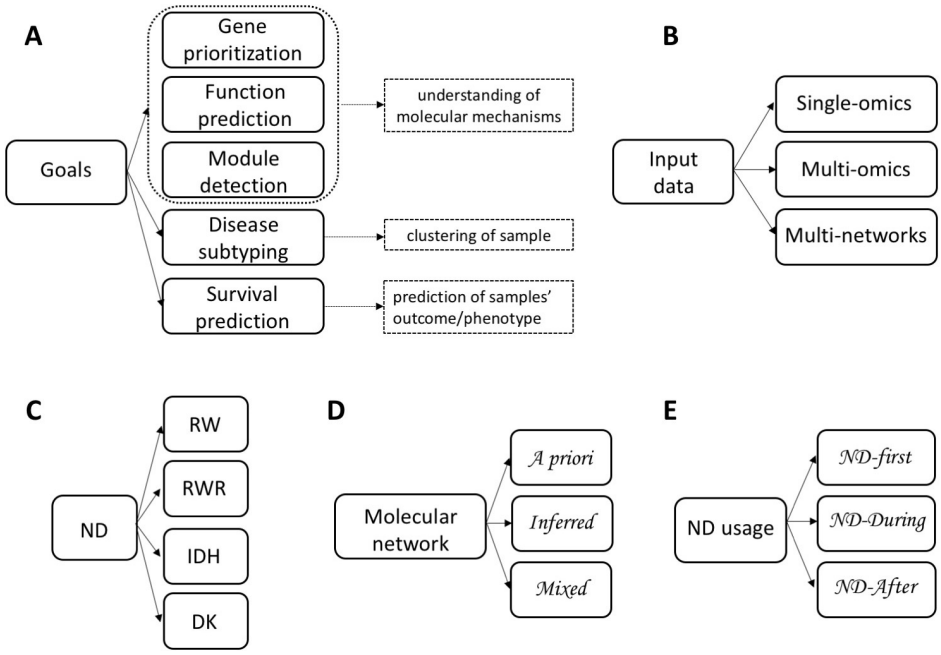


Figure 2.1: Classification of integration methods. Criteria: (**A**) Goals; (**B**) input data; (**C**) Network Diffusion (ND) model: Random Walk (RW), Random Walk with Restart (RWR); Insulated Heat Diffusion (IHD), Diffusion Kernel (DK); (**D**) Molecular network; (**E**) ND usage.

on binary interaction or complex mapping [34]. The Genotype-Tissue Expression (GTEx) project aims at the construction of a specific network for each major human tissue [35]. Projects like ENCODE97 and the Roadmap Epigenomics provide data about gene regulatory networks [36, 37]. The IMEx Consortium is an international collaboration of major public interaction data providers aimed at establishing a non-redundant set of biophysical molecular interactions [38]. In addition to *primary databases*, which collect curated experimental data from small and/or large scale studies, there are several *meta-databases*, which integrate data from several primary databases, and *prediction-databases*, which also provide predicted (biophys-

ical and/or functional) interactions obtained from the analysis of biological datasets [39]. Multiple collections of scores can be mapped on molecular networks in rather different ways, depending on data types and data analysis purposes. The resulting networks can be classified in three broad categories: multi-weighted networks, multiplex networks and networks of networks.

In a multi-weighted network, a series of weights are associated with nodes and/or links. For instance, the same biological network can be characterized by different omics weights on different layers (e.g. gene expression, methylation, somatic mutations). A multi-weighted network therefore consists of a single-layer network with multiple attributes associated with the same nodes and links, but sometimes can be referred to as a multi-layer network.

Two categories of structural multi-layer networks are multiplex networks and networks of networks. A multiplex is a collection of networks with the same set of nodes and varying intra-layer topologies and inter-layer relationships are trivially given [40]. A network of networks (sometimes also referred to as heterogeneous networks) is a collection of networks with different nodes (in principle also representing entities of different nature) with multiple types of connections (specific intra-layer links and specific inter-layer connections) [41]. The classification of multi-layer networks is indeed non-trivial; for instance, the categories described can have significant overlaps. It is possible to build hybrid networks where on a core multiplex some layer-specific nodes and links are introduced and consequently different types of inter-layer links are established; for more details about multilayer networks and their classification see the work of Kivela et al. [41].

2.3 The unifying mathematical machinery of network diffusion

Network diffusion processes can be summarized as the spreading of biological information throughout the network along network edges, initially

retained in the so-called "seed nodes". Each node will therefore gain or lose biological information according to the network proximity to the seeds and to its topological features.

From a mathematical perspective, considering a network G of n nodes, the biological information is encoded in an n-dimensional array $\boldsymbol{x}_0$ where the i-th entry accounts for the amount of biological signal initially present in node i. Therefore, $\boldsymbol{x}_0$ is defined as the initial state of the network. Then, starting from $t = 0$ up to a fixed time (finite or infinite) the state of the network $\boldsymbol{x}_T$ evolves according to the network topology until it reaches a final state $\boldsymbol{x}_T$, where, as previously mentioned, T can either a finite or an infinite time. Under the appropriate settings, when $T = \infty$, the final state of the diffusive algorithm may correspond to a steady state or steady-flow state of an associated physical model, allowing a clear interpretation of the results [42].

In general, the final state of a diffusion process consists of a graph-based transformation f_G of the initial biological information $\boldsymbol{x}_0$, which is linear in most cases so that f_G reduces to a matrix $\boldsymbol{M}_G$ and:

$$\boldsymbol{x}_T = f_G(\boldsymbol{x}_0) = \boldsymbol{M}_G \cdot \boldsymbol{x}_0 \tag{2.1}$$

The diffusion processes used by integrative methods are classified, similarly to Cowen et al. [21], on the basis of the specific transformation $\boldsymbol{M}_G$ in four categories (Table 2.1 and Figure 2.1 C):

1. Random Walk (RW): $\boldsymbol{M}_G = [\boldsymbol{A}\boldsymbol{D}^{-1}]^k$;
2. Random Walk with Restart (RWR): $\boldsymbol{M}_G = \alpha[\boldsymbol{I} - (1-\alpha)\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}]$;
3. Insulated Heat Diffusion (IHD): $\boldsymbol{M}_G = \alpha[\boldsymbol{I} - (1-\alpha)\boldsymbol{A}\boldsymbol{D}^{-1}]^{-1}$;
4. Diffusion Kernel (DK): $\boldsymbol{M}_G = e^{\alpha(\boldsymbol{D}-\boldsymbol{A})}$.

Here above, $\boldsymbol{A}$ is the adjacency matrix of the network, $\boldsymbol{D}$ is a diagonal matrix of nodes degree (number of interactions), k is the number of time-steps and $\alpha \in (0, 1)$ is a tuning parameter.

Differently from Cowen et al. [21], we choose to differentiate between RWR and IHD. In fact, the different normalization of the adjacency matrix $\boldsymbol{A}$ (symmetric for the RWR, column normalization for the IHD) implies different behaviours in the relative diffusion processes. Indeed, the RWR implies a symmetric diffusion where information flows through each link with the same intensity in each direction [43]. Conversely, IHD implies an asymmetric diffusion where information (or heat) tends to flow out from highly connected nodes much easier than from poorly connected ones [44]. Such differences in the diffusion matrix therefore imply dissimilar behaviours of information flow, mainly in relation to network hubs: at infinite time in the RWR hubs tend to naturally gather relatively more information than in the IHD, since IHD is characterized by an intrinsic hub penalization. Therefore, even if it is conceptually similar, RWR and IHD applied to complex biological networks with thousands of vertices and tens to hundreds thousands links, may present sensibly different results.

Independently from the specific kind of diffusion model, the matrix $\boldsymbol{M}_G$ is usually hard to recover analytically because it implies inverting or power-expanding a high-dimensional graph-based transition matrix: alternative numerical approaches would be needed and the direct inversion of the matrix $\boldsymbol{M}_G$ is possibly replaced with converging iterative procedures [45].

The choice of the most appropriate diffusion process depends on the goal of the analysis. For instance, if one is interested only in considering a local neighborhood of the seeds may choose RW with a finite number of steps [46], while RWR and IHD quantify network proximity to seeds considering simultaneously all the possible network paths among network nodes [44, 47].

2.4 How network diffusion is exploited by integrative methods

ND requires data about the variables ($\boldsymbol{x}_0$) and about their relations ($\boldsymbol{A}$). An important difference between integrative methods that use ND concerns the type of network in use, that is the way in which the adjacency

matrix is defined.
Three broad categories can be recognized (Table 2.1 and Figure 2.1 D):

- the topology of the network in use is defined by means of *a priori* knowledge, e.g. collected from molecular interactions databases;
- a network is *inferred* from the analysis of one or more biological datasets;
- a mixed approach that combines *a priori* and novel knowledge.

ND can be applied before, after or during the "integration step" of the analysis pipeline (Table 2.1 and Figures 2.1 E - 2.2).

In the *ND-first* approach, ND is applied to a series of collections of initial scores, each of which summarizes data of a single sample or multiple samples; the resulting collections of ND scores are subsequently integrated. An example of this approach is TieDie [48], where ND is applied to two collections of scores, one representing mutated genes while the other differentially expressed genes, on the same network; the two resulting ND score vectors are then jointly analysed and the minimum of the two ND scores of a gene is considered as the one chosen for the gene.

The *ND-after* approach consists in the application of ND after a first process of integration of different data types into a unique structure. For instance, stSVM [46] first integrates omics data and subsequently applies ND to define a global ranking of miRNA and mRNA using statistics about their differential expression integrated in a heterogeneous network.

The *ND-during* refers to the application of a type of ND in which each layer communicates information to one other. This is the case of SNF [49], in which patient similarity networks, obtained from each of their data types separately, exchange information during the ND process, leading to a unique "fused" patient network.

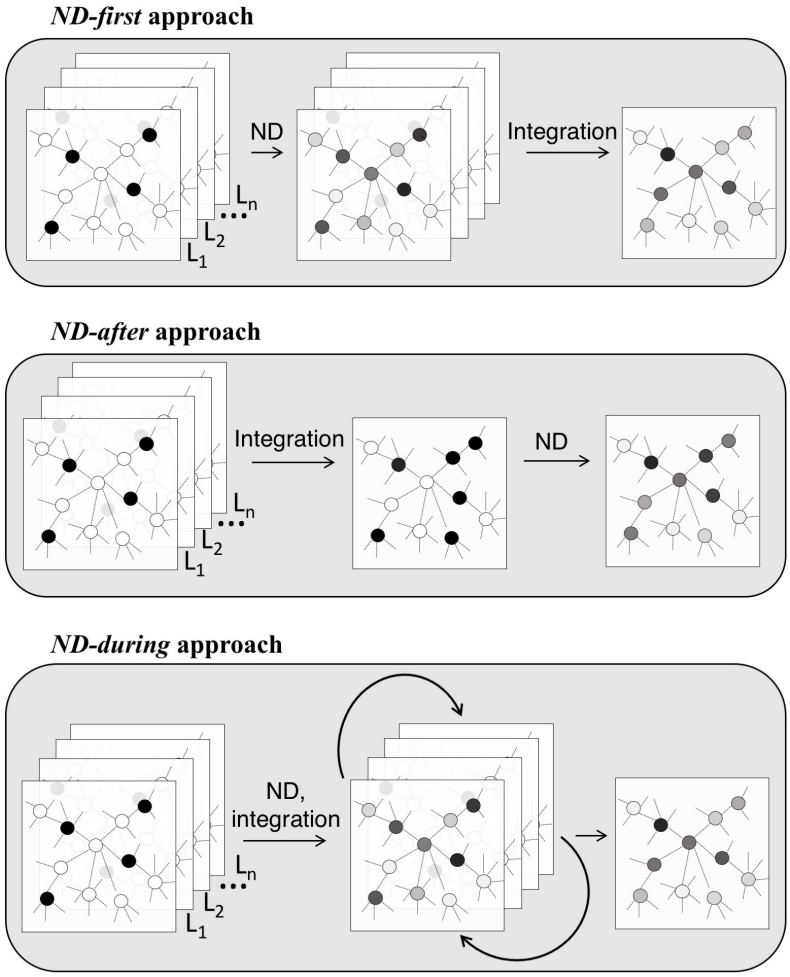


Figure 2.2: Ways in which ND enters the integrative analysis pipeline.

Method	ND	Category	Network	Goal	Input	Implementation
dmfind [42]	RWR	SO, $*$	*A priori*	MD	SM	R
EMDN [50]	RWR	MO, $*$	Inferred	MD	GM and DM	R
EPU [51]	RWR	MN, $*$	*A priori*	GP	PPI, GE, Gene Ontology, Phenotype similarity networks	NA
GeneMANIA [52]	RWR	MN, $\triangle$	*A priori*	FP	Co-expression, PPI, Genetic interaction, Co-localization, Shared protein domains	Web
Mashup [53]	RWR	MN, $*$	*A priori*	FP	PPI	Matlab
M-Module [54]	RWR	MO, $*$	Inferred	MD	GE and SM	R
NetBags [55]	DK	SO, $*$	*A priori*	DS	GE	NA
NetICS [56]	IHM	MO, $*$	*A priori*	GP	GM and AB	Matlab
NBS [47]	RWR	SO, $*$	*A priori*	DS	SM	Matlab
NBS^2 [57]	RWR	SO, $*$	Mixed	DS	SM	Phyton
RegNet [58]	RW	MO, $\triangle$	Inferred	GP	GE and CNV	R
Ruffalo [59] et al.	RWR	MO, $*$	*A priori*	GP	GE and SM	NA
Shi et al. [60]	RW	MO, $*$	Mixed	GP	GE and SM	NA
SRF [61]	RWR	MO, $*$	*A priori*	DS	GE and SM	Java
SNF [49]	DK	MO, $\bullet$	Inferred	SP, DS	GE and DM	Matlab, R
stSVM [46]	DK	MO, $\triangle$	*A priori*	PS, GP	GE and miRNA	R
TieDie [48]	IHM	MO, $*$	*A priori*	MD	GE and SM	SciPy, Matlab
WSNF [62]	DK	MO, $\bullet$	Inferred	SP, DS	GE and miRNA	R

Table 2.1: Network diffusion based methods for the integrative analyses of multiple biological layers. ND: RWR: random walk with restart, DK: diffusion kernel, IHM: insulated heat model; Category: SO: single-omics integration, MO: multi-omics integration, MN: multi-networks integrations; $*$: ND-first approach, $\triangle$: ND-after approach, $\bullet$: ND-during approach; Goal: DS: disease subtyping, GP: gene prioritization, MD: module detection, FP: function prediction, SP: survival prediction; Input: GE: gene expression, DM: DNA Methylation, SM: Somatic Mutation, CNV: copy number variations, AB: Aberration events, PPI: Protein-Protein interaction network.

2.5 Integrative methods based on network diffusion

On the basis of data types, integrative methods using ND can be distinguished in those that analyze a single type of omics, multiple omics or multiple networks (Table 2.1 and Figure 2.3).

2.5.1 Single omics

Integrative methods for the analysis of a single type of omics consider a series of molecular profiles, such as patient-wise mutation profiles.

The method called "dmfind" [42] compares ND scores obtained from a series of descriptive statistics, such as gene mutation frequencies. Subsequently, the network smoothing index (NSI) is obtained by comparison of ND scores with initial molecular profiles [42]. When applied to gene networks, NSI highlights genes in network proximity enriched by initial information according to a tuning parameter [30]. The integration is therefore realised by subtracting NSIs belonging to two patient groups (ND-first), an operation that prioritizes genes that participate in differentially enriched modules [42].

NBS (Network-Based Stratification) [47] is a method that stratifies tumor mutations finding clusters of similar patients. It applies ND to a binary somatic mutation matrix (genes-by-samples). Then, the resulting collections of ND scores are jointly analysed (ND-first) using a network-constrained non-negative matrix factorization to find patient groups. It has been applied to study 13 cancer types with exome-level mutation data [63], liver cancer [64] and in a pan-cancer genomic analysis [65].

NetBags (NETwork Based clustering Approach with Gene signatures) [55] essentially applies the strategy of NBS to a binary genes-by-samples matrix that represents the significantly expressed genes.

NBS [47] uses *a priori* knowledge of molecular interaction networks that are not cancer-specific. NBS^2 (Network-Based Supervised Stratification) [57] was proposed to address this issue. Unlike previous approaches, the weights of each molecular interaction are adjusted by a supervised strategy

so that the stratification of propagated mutation profiles after random walk is close to the pre-defined tumor subtypes.

2.5.2 Multi-omics

In multi-omics data integration each layer typically contains scores obtained from a distinct omic assay. Most methods deal with two types of layers (Figure 2.2).

2.5.2.1 Genomics and transcriptomics

Many methods tackled the problem of analysing the relation between genomic aberrations and gene expression changes.

Ruffalo et al. [59] presents a ND-based method to predict "silent" players in cancer by integration of somatic mutations and gene expression data, where a silent player is a gene neither mutated nor differentially expressed but which plays a role in cancer development and progression. Inputs are represented as two binary matrices of somatic mutation and gene expression (genes-by-samples). The authors explored several ways (e.g. the minimum, the maximum, the product, the average) of combining diffusion scores (ND-first) to obtain the features of a logistic regression model that predicts a gene's association with cancer.

Also Shi et al. [60] use patient-wise gene mutation and gene expression data to prioritize genes. The approach constructs a bipartite graph of outlying genes and mutated genes considering an influence graph (that captures *a priori* biological pathway information), mutational and expression data. A two-step diffusion is performed to calculate diffusion scores for each patient and these scores are subsequently combined (ND-first) by robust rank aggregation.

Differently from the methods described above that yield gene prioritizations, TieDIE (Tied Diffusion Through Interacting Events) [48] has been developed to identify a subnetwork that links a source gene set (S) carrying genomic alterations to a target set (T) of differentially expressed genes on the same *a priori* network. TieDIE transforms the two collections of

input scores in the corresponding ND scores and then (ND-first) the minimum of the two scores of a gene is used as the final score for that gene. TieDIE has been used to study several cancers, such as, Papillary Thyroid Carcinoma [66], Prostate Cancer [67], Leukemia [68] and in an extensive immugenomic analysis of 33 diverse cancer types [69].

Another method that seeks to identify gene modules is M-Module [54]. It infers co-expression networks from multiple data that represent disease stage transitions. Then genes are ranked in each networks *via* ND, incorporating also gene mutations as priors. In each network, ND scores are transformed in gene ranks, gene ranks into z-scores and the average z-score across all is used to obtain a final gene rank (ND-first). Gene modules are therefore identified using a graph entropy-based measure that quantifies connectivity of a module in multiple networks. Authors of M-Module proposed different variants of the algorithms: NMF-DM, in which modules of each network are discovered using a non-negative matrix factorization algorithm [70], SMMN, which uses modularity measure to discovery modules [71] and S2-jNMF a novel semisupervised joint nonnegative matrix factorization algorithm [72]. M-Module has been applied to several studies (e.g. [73–75]).

SRF [61] aims at discovering cancer subtypes by combining mutation and expression data across samples. ND is applied only to the binary matrix of gene mutations. The identification of subtypes is performed by rank matrix factorization of the ranked diffusion matrix and ranked expression matrix (ND-first).

Copy number variations (CNVs) are another type of genomics aberration that has been jointly analysed with transcriptomics. The main goal of RegNet [58] is the quantification of the impact of gene expression changes on user-defined target genes in a network inferred from gene expression and CNVs. The approach learns a regulatory network by modelling the expression level of each gene as a linear combination of the expression levels of all other potential regulator genes and the gene-specific copy number, lasso regression is used in combination with a significance test for lasso [76] to find the relevant predictors for each gene. Next, ND is applied using the learned network to quantify impacts of sample-specific gene expression changes on

other clinically relevant target genes using network-diffusion. RegNet was able to predicts novel cancer gene candidates in oligodendrogliomas [77].

2.5.2.2 Epigenomics and transcriptomics

The algorithm of M-Module is employed in EMDN framework (Epigenetic Module based on Differential Networks) [50] to characterize epigenetic modules by using differential co-methylation and co-expression networks, without incorporating genes mutations information as prior information. In this way EMDN applies ND as RW without restart, but with a symmetric normalization of the adjacency matrix.

An interesting method that aims to find disease subtypes and predict phenotypes is SNF (Similarity Network Fusion) [49]. It works without constraints for the type of input but requires that samples are matched across omics. First, networks of samples for the various types of omics are built, then, networks are fused into one network by using the non-linear method of message passing theory (KNN and graph diffusion) that iteratively updates each of the network making it more similar to other networks in each step.
Several studies in cancer have exploited SNF method to integrate GE and DM data, like: Kidney Renal Cell Carcinoma [78], medulloblastoma [79]; further, thanks to its versatility, SNF has been used to integrate other types of omics: miRNA and GE in Colorectal liver metastasis [80] and in Ovarian cancer [81]; miRNA, mRNA, lncRNA, and DNA methylation in Pancreatic Ductal Adenocarcinoma [82]; GE, miRNA and CNV in triple-negative breast cancer [83].

2.5.2.3 Transcriptomics: mRNA and miRNA

Xu et al. [62] have proposed a modification of SNF method called WSNF (Weighted Similarity Network Fusion) that takes into consideration the level of importance of genes to identify disease subtypes. WSNF constructs a miRNA-TF-mRNA regulatory network from different interaction databases, then assesses the weight of each features (miRNA, TF, mRNA),

calculated as a linear combination of two terms: ranking of features obtained using ND and expression variation across all patients in expression datasets. Weights are introduced into the formula of Euclidean distance to calculate the distance between two patients then SNF method is applied.

stSVM (smoothed t-statistic support vector machine) [46] combines *a priori* network information and omics data (miRNA and GE) to discover biomarker signature and predict disease prognosis. It smoothes gene-wise statistics from experimental data (both miRNA and gene expression) over the biological network, constructed by integration of PPI with miRNA-target gene network, using a P-step random walk kernels. A permutation test is conducted to select significant genes that will be used to train a support vector machine (SVM) classifier. It has been used in an integrative study of miRNA and GE to predict response to a monoclonal antibody in Head and Neck Squamous Cell Cancer [84].

2.5.2.4 Genomics, Epigenomics and Transcriptomics

NetICS (Network-based Integration of Multi-omics Data) [56] prioritizes cancer genes by their mediator effect, defined as the proximity of the gene to aberration events (SM, CNV, DM, a differentially expressed miRNA), differentially expressed genes and proteins in a molecular network given *a priori*. The method uses a per-sample bidirectional IHD process and initial heat vectors ($\boldsymbol{h}_1$, $\boldsymbol{h}_2$) are defined, respectively, as the number of the aberrant and differentially expressed genes of the sample.
Final scores for all genes are obtained by means of the Hadamard product of the exchanged heat matrices ($\boldsymbol{E}_1$, $\boldsymbol{E}_2$) (ND-first): $\boldsymbol{E} = \boldsymbol{E}_1 \circ \boldsymbol{E}_2$.
Lastly, diffusion scores of all samples are combined to obtain global gene ranking *via* a robust aggregation, in which a gene's rank is calculated as the sum of its per-sample ranks.

2.5.3 Integration of multiple networks

In the integration of multiple networks each layer represents a biological network. The two main applications are gene function prediction and gene

prioritization.

Mashup [53] uses ND on several protein-protein interaction networks to predict gene function and genetic interactions. It applies RWR algorithm separately on each network and then a matrix factorization based technique is used to reduce dimension of the diffusion results (ND-first). The feature learning step allows to obtain a low-dimensional feature vectors of proteins that best approximates the RWR matrix and results more robust to noise; feature vectors are used to train SVM classifiers to predict genetic interactions.

Mostafavi et al. [52] developed GeneMANIA (Multiple Association Network Integration Algorithm), a tool for predicting gene function by integration of multiple networks (e.g. co-expression, PPI, genetic interaction, co-localization, shared protein domains). Given d networks encoded as matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_d$, they are integrated into a "composite network" ($\boldsymbol{W}^{\text{comb}}$), obtained by weighted average of individual networks:

$$\boldsymbol{W}^{\text{comb}} = \sum_h \alpha_h \boldsymbol{W}_h$$

where the vector $\alpha = [\alpha_1, ..., \alpha_d]$ corresponds to network weights and is computed by solving a ridge regression problem. Then given the $\boldsymbol{W}^{\text{comb}}$ matrix, a variation of the Gaussian field label propagation algorithm (a RWR where functions of unlabeled data are predicted starting from differently labeled data and network structure) is applied to predict the gene function. GeneMANIA has been applied in several studies (e.g. [85–87]

Differently from above methods, EPU (Ensemble Positive Unlabeled learning) [51] uses a supervised learning method, that falls in the class of Positive-Unlabeled learning method, for disease gene identification by integrating multiple biological data sources (PPI, gene expression data, Gene Ontology, Phenotype-gene association data and Phenotype similarity network). ND is applied on three biological networks (Gene Expression network, PPI network, Gene ontology similarity network) to obtain weights for unlabelled genes (not associated with disease). The resulting three collections of ND scores are combined into a set of integrated scores using,

for each gene, the mean of its three ND scores (ND-first). These integrated scores are used to train three machine-learned prediction models (Weighted-KNN, Weighted-Naive Bayes, Weighted-SVM) and their results are integrated by an ensemble learning algorithm.

2.6 Perspectives and open issues

Network-diffusion based approaches have been proposed to solve several problems in biological data analysis, including integrative analyses. These methods analyse multiple collections of scores derived from different omics assays in combination with molecular networks or similarity networks, and apply ND on such networks.
The main applications include: gene function prediction; gene prioritization; identification of gene modules and molecular pathways; disease subtyping; and prediction of an outcome. In all these applications ND is a tool to transform one or more initial vectors of scores into vectors that reflect the network proximity between network nodes on which the scores are mapped. This operation provides different benefits:

- considering gene-centric datasets as a practical example, ND is a powerful way to embed the information about molecular interactions among genes into a gene-wise dataset;

- ND can be used to quantify the proximity between each pair of nodes in a global way, that is considering all possible paths among the nodes, overcoming the limits of local approaches (e.g. giving the same importance to all direct neighbours of a node) and better capturing the complexity of biological networks;

- ND highlights genes in network proximity and with high input scores. By so doing, it amplifies genetic associations according to the architecture of the molecular network, a result that offers insights in agreement with the so-called local hypothesis; that is, the hypothesis that genes that lie in network proximity within molecular networks

co-work in the development of cellular functions and are therefore co-responsible for pathological phenotypes [30];

- by a data analysis perspective, ND transforms sparse vectors into dense vectors. This operation eliminates missing values and ties, two situations that are often difficult to handle. This imputation step facilitates the joint analysis of different data types and is particularly important in the integration of multiple omics that vary in scope and coverage. For instance, mutations may affect just a few tens of genes of a tumor cell, while gene expression changes are observed for a much higher number of genes. More generally, in a multi-omic analysis of a biological process, only a subset of the genes is associated with the various types of measured alterations. In this context, ND can be used to highlight common network regions where different types of omics signals converge;

- ND is suitable to analyse patient-level molecular profiles, promoting studies within the scope of precision medicine.

ND processes, which can be brought back to four classes (paragraph 2.2), require the tuning of a parameter (k or α) that controls the diffusion process reach or the relative importance of topology and input scores. In many cases, the issue about tuning of such parameter has been solved showing that the performance of the proposed integrative method is robust to small variations of the parameter. A dependency between the optimal value and the network in use has been suggested [47].

Most methods apply ND to transform a series of input score collections to get as many collections of ND scores - in which the network topology is embedded - and, subsequently, combine the ND scores: we referred to these methods as ND-first. The combination of a series of ND scores for the same variable (e.g. a gene) is performed with simple mathematical operators, such as the mean or the minimum, or with more elaborated techniques, such as non-negative matrix factorization and support vector machines. ND scores may require a step of transformation, such as normalization, to enable the direct comparison between scores at different scale (e.g. [47]),

or ranking, to work on the relative importance rather than absolute values (e.g. [54, 60]). Other integrative methods, firstly integrate multiple data types, then use ND: we referred to these methods as ND-after. In these methods, ND is one of the last steps that lead to the final output. A third class of methods perform ND simultaneously with the integrative step (ND-during). The class of simultaneous diffusion approaches is very promising as it encodes the diffusion processes on multi-layer networks [88]. In principle, simultaneous diffusion allows to extend the classical analysis of multi-omics data on complex networks. For instance, in the case of heterogeneous networks, layer-specific nodes bring an indirect contribution to the ND scores on each other layer. Such an output is not possible neither in ND-first nor in ND-after approaches. ND-after integrative approaches build an aggregate network encoding weighted or unweighted aggregate links; such an aggregate network is therefore algebraically put together, independently from the diffusion process. The same considerations hold for ND-first approaches, but such integration issues are addressed once the ND is performed on each layer separately. Therefore, ND-after and ND-first approaches could be very informative about a specific biological analysis but they present an intrinsic lack of scalability, as the way in which properly combine and weigh networks (before or after ND) strongly depends on the biological context. Conversely, an ND-during (simultaneous) approach maintains the available biological information and avoids additional data manipulations before and after the application of the diffusive algorithm. However, simultaneous approaches may introduce computational issues as omics data size and number of layers increase.

Most of the approaches do not assess the statistical significance of ND scores. In several works it was proposed to use empirical p values [42], which also provides also the benefit of mitigating the over-estimation of hub importance. In a recent work, the calculation of empirical p values using of degree-normalized random seeds was shown to be more accurate, but computationally more demanding, than random seeds [89].

A specific combinations of omics (e.g. gene mutations and gene expression changes) and a quite specific formulation of the problem is often required. While this specificity offers advantages within the domain of the

original problem, it also poses constraints to applicability and further extension. Furthermore, efforts are still required to develop methods that combine more than two omics.

Another important issue is the reliability of interactomes. The problem of defining a reference human interactome is open in molecular biology as well as the problem of quantifying the reliability of such cell-scale reconstructions, because experimental technologies currently used to detect interactions involve a series of issues [34]; therefore a careful network selection must be made by users based on the research questions they wish to address. Moreover, some methods take into account the directions of interactions in their algorithms, but cell-scale reconstructions do not provide information about "the direction" of the interaction, which requires a deeper understanding of the mechanistic relation between the two interacting partners. Modelling this information is not trivial and comes at the cost of a relevant reduction of coverage in terms of genes that can be analysed.

In conclusion, current trends suggest that network diffusion is a tool of broad utility in omics data analysis. It is reasonable to think that it will continue to be used and further refined as new data types arise (e.g. single cell datasets) and the identification of system-level patterns will be considered more and more important in omics data analysis. However, the methods currently available do not meet all the needs of research projects: their applicability is limited by the number of omics and experimental designs.

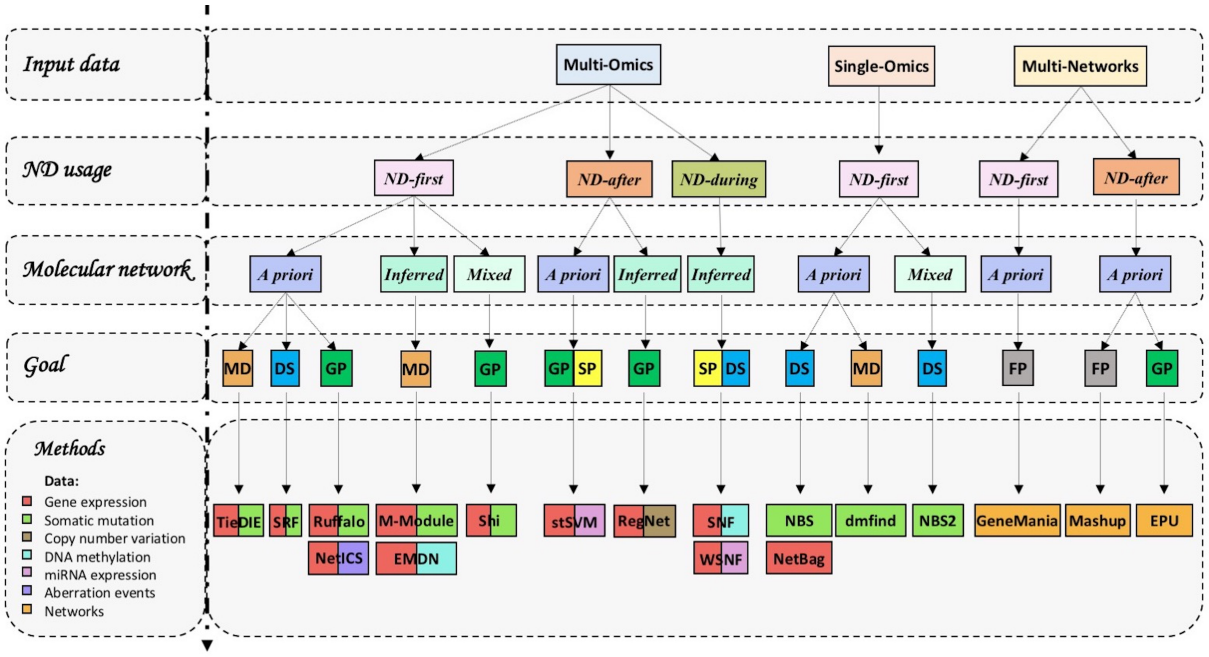


Figure 2.3: Network diffusion methods for the integrative analyses of multiple biological layers. GP: gene prioritization, MD: module detection, FP: function prediction, DS: disease subtyping; SP: survival prediction. Methods were classified according to their main use described by the respective authors.

Chapter 3

mND: gene relevance based on multiple evidences in complex networks

This chapter presents a novel use of the "mathematical machinery" of network diffusion to integrate multiple omics. A new gene score (mND) is proposed to integrate multiple biological data by quantification of genes' relevance, taking into account the network proximity of the gene and its first neighbours to other altered genes.
Since mND is applicable to a wide range of data types and experimental projects, it introduces an important advance in the class of multi-omic methods.[1]

[1]The contents of this chapter are published in: *N. Di Nanni, M. Gnocchi, M. Moscatelli, L. Milanesi, E. Mosca. (2019) "Gene relevance based on multiple evidences in complex networks". Bioinformatics, btz652.* https://doi.org/10.1093/bioinformatics/btz652.

3.1 A new network-diffusion method for the integration of multiple biological data

Multi-omics analyses, sample-level analyses, and multi-classes analyses (e.g. multiple cell clusters) demand methods to highlight the importance of altered genes considering, respectively, different types of summary information across subjects or subject-specific molecular profiles. At the same time, to explain complex patterns in these datasets (e.g. the heterogeneity of mutation profiles of tumor samples) it is important to consider the complex web of macromolecular interactions, which provides known relations among the variables (e.g. genes) under analysis. Moreover, recent studies have suggested to include the first neighbours in network-based methods for the analysis of single omics [90, 91] and multi-omics data [92].

Considering all these aspects and to overcome limitations of available methods described in *Chapter 2* (paragraph 2.6), a gene-score, named "mND", has been developed to assess gene relevance on the basis of gene position in a genome-scale network in relation to one or more types of biological evidences ("layers" hereafter) (Figure 3.1). The method allows integration of both single type of omics and multiple omics, it is based on the *ND-first* approach (*Chapter 2*, paragraph 2.4), uses the RWR algorithm (*Chapter 2*, paragraph 2.3) and the *a priori* knowledge of molecular interaction networks (*Chapter 2*, paragraph 2.4).

In particular, mND prioritizes genes considering their own importance (in proportion to original evidences) and the importance of their network location. It uses network diffusion scores that quantify the topological relevance of a gene in the context of the distribution of the considered evidences throughout the entire network and layer-specific highly "informative" first neighbours. Therefore, genes are ranked considering their relevance within each layer (e.g. number of mutations, p-values from differential expression analysis), their network proximity to other relevant genes as well as the layer-specific-relevance of their neighbours; the statistical significance of the gene scores defined by mND (mND score) is assessed by dataset permutations.

Furthermore, in addition to producing a global gene ranking, mND introduces a new method to help unravel the role of a gene in each layer by classifying it as a member of a module of high scoring genes, linker of high scoring genes or, lastly, high scoring but isolated gene.

Unlike current methods, mND can be used in integrative analysis of different types of omics (e.g. mutation, CNV and expression changes) or multiple samples of same omic type (e.g. patient-level mutational analysis), without particular constraints on the number of layers and layer type. It works on a general gene-by-sample input matrix, where each column is a vector of scores representing different data types (e.g. genomics, transcriptomics) or the same type (e.g. fold changes or p-values from single cell clusters).

The R package "mND" is available at URL: https://www.itb.cnr.it/mnd.

3.2 Algorithm development

The calculation of mND score requires an undirected interaction network G and a matrix of initial scores $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L]$, in which $\boldsymbol{x}_i \in \mathbb{R}^n$ with $i = 1, 2, ..., L$ is the score vector over all vertices of G. The computation of mND consists of five steps (Figure 3.1).

3.2.1 Network diffusion

Input scores $\boldsymbol{X}$ are smoothed by ND (RWR), obtaining the corresponding network-constrained scores $\boldsymbol{X}^*$, using the following iterative procedure, where the subscript $q \in [0, \infty)$ indicates the current iteration and $\boldsymbol{X}_0 = \boldsymbol{X}$:

$$\boldsymbol{X}_{q+1} = \alpha \boldsymbol{W} \boldsymbol{X}_q + (1 - \alpha) \boldsymbol{X}_0, \boldsymbol{X}^{\text{ss}} = \lim_{q \to \infty} \boldsymbol{X}_q \tag{3.1}$$

where $\alpha \in (0, 1)$ is a scalar that weights the relative importance of topology and input scores, and $\boldsymbol{W}$ is the symmetric normalized form of the adjacency matrix $\boldsymbol{A}$:

$$w_{ij} = \frac{a_{ij}}{\sqrt{d_i} \cdot \sqrt{d_j}} \tag{3.2}$$

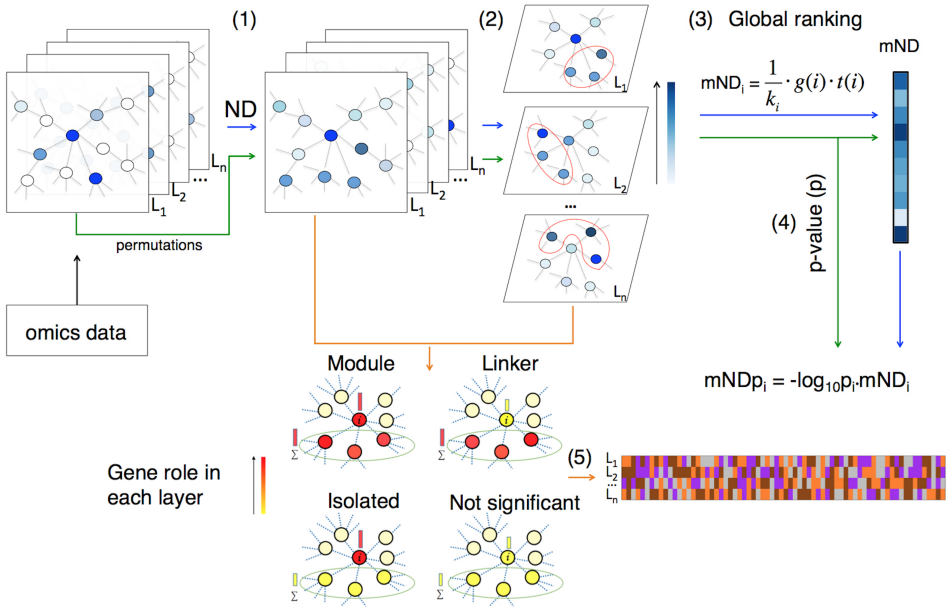


Figure 3.1: Flowchart of the analysis pipeline with mND. (1) Network-diffusion is applied to the original dataset, composed of multiple layers $L_1, L_2, ..., L_n$ (e.g. different types of omics or multiple samples of same omic type); (2) Identification of the top k neighbours for each gene in each layer; (3) Calculation of mND score; (4) Empirical p-value assessment; (5) Classification of genes across layers.

where $a_{ij} \in \boldsymbol{A}$ are the elements of the adjacency matrix and (d_i, d_j) are the degrees of the corresponding genes.
The final matrix $\boldsymbol{X}^{\text{ss}}$ is the matrix $\boldsymbol{X}_{q+1}$ that satisfies the termination criterion $\max(|\boldsymbol{X}_{q+1} - \boldsymbol{X}_q|) < 10^{-6}$.

To enable direct multiplication of values belonging to different layers, $\boldsymbol{X}^{\text{ss}}$ is column-wise normalized by the maximum of each column, obtaining the matrix $\boldsymbol{X}^*$.

3.2.2 Neighbours selection

For each gene i, the top $k_i = min(k, d_i)$ first neighbours with the highest diffusion scores in each layer l are selected as representatives of the network proximity of the neighbourhood of i to the original scores in layer l, and their network diffusion scores are summed:

$$\mathrm{T}(i,l) = \max\left\{\sum_{j\in C} a_{ij}x^*_{jl} | C \in S\right\} \tag{3.3}$$

where $x^*_{jl} \in \boldsymbol{X}^*$ with $j = 1, 2, ..., N$ is the network-constrained value of j-th gene in l-th layer ($j \neq i$), S is the set of all k_i-subsets of $1, 2, ..., N$ and $0 < \mathrm{T} \leq k_i$.

3.2.3 Integration

At this point, the mND score for gene i is calculated as the product between the sum of its network constrained scores (term $\mathrm{g}(i)$) and the sum of the contributions of its top k first neighbours (term $\mathrm{t}(i)$):

$$\mathrm{mND}_i = \frac{1}{k_i}\mathrm{g}(i)\mathrm{t}(i) = \frac{1}{k_i}\left(\sum_{l=1}^{L} x^*_{il}\right)\left(\sum_{l=1}^{L}\mathrm{T}(i,l)\right) \tag{3.4}$$

where L is the total number of layers and $0 < \mathrm{mND}_i \leq L^2$.

3.2.4 Significance assessment

The matrix $\boldsymbol{X}$ is permuted Π times by swapping its rows and the corresponding values of $\mathrm{mND}^\dagger_i$ are used to calculate empirical p-values, defined as the fraction of times that $\mathrm{mND}^\dagger_i$ is equal or greater than mND_i:

$$p_i = \frac{1 + \#\{\mathrm{mND}^\dagger_i \geq \mathrm{mND}_i\}}{\Pi + 1} \tag{3.5}$$

The product of p_i and mND_i provides a gene score weighted by its estimated statistical significance, as previously described [93]:

$$\mathrm{mNDp}_i = -\log_{10}(p_i) \cdot \mathrm{mND}_i \tag{3.6}$$

3.2.5 Classification

The distribution of initial and diffused scores is used to provide a layer-specific gene classification that may suggest functional roles and offer mechanistic insights in relation to the datasets studied.
A gene i is classified by evaluating the membership of the gene in two gene sets H_l and N_l which define, respectively, the high scoring genes according to original data ($\boldsymbol{X}$) and neighbour information (T).
The gene set H_l is composed of the high scoring genes in layer l of $\boldsymbol{X}$, defined using a layer specific criterion (e.g. the differentially expressed genes at $p < 0.05$).
The gene set N_l is composed of the genes with the highest:

$$\mathrm{tp}_{i,l} = -\log_{10}(p^t_{il}) \cdot \mathrm{T}(i,l) \tag{3.7}$$

where p^t_{il} is the empirical p-value calculated comparing T to $\mathrm{T}^\dagger$, the latter obtained with permuted $\boldsymbol{X}$. The use of empirical p-value to scale T overcomes the issue of ties due to genes with equal values of T
The cardinality of N_l can be defined in different ways, considering:

- an *ad hoc* number of top values (e.g. in proportion to $|H_l|$);
- a threshold value for p^t_{il};
- a combination of the two previous criteria.

The gene i is ISOLATED if it is in H_l but its neighbourhood is not in N_l (Figure 3.2). If both the gene and its neighbourhood are in, respectively, H_l and N_l the gene is part of a high scoring module and therefore is termed MODULE (Figure 3.2). If the gene is not in H_l but its neighbourhood is in N_l, then it is named as LINKER (Figure 3.2).

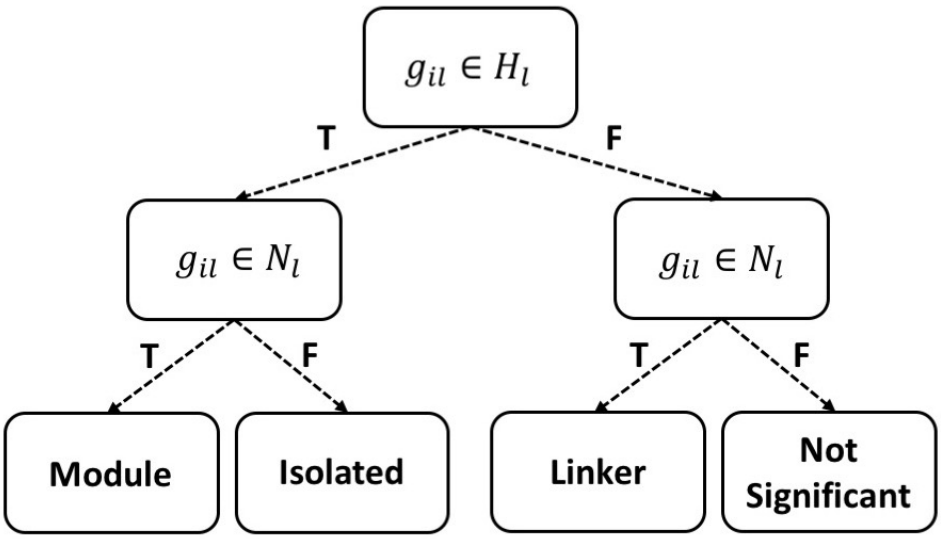


Figure 3.2: Flowchart of gene classification across layers g_{il} represents the i-th gene in l-th layer; T: TRUE, F: FALSE.

3.2.6 Optimization of k value

Lastly, the value of k can be optimized selecting a value that yields connected networks enriched in initial scores.
To this aim, the Ω function at the basis of network resampling method [42] is adapted, it calculates a network score considering top ranking genes and shows to which extent such network score is expected if links among genes are shuffled (keeping the same degree distribution). Such function (designated here as Ω_0) is applied to the original scores $\boldsymbol{X}(R_{kn}, l)$ in layer l associated with the top n genes R_{kn} ranked by mND using a particular k value:

$$\Omega_0(\boldsymbol{X}(R_{kn}, l), \boldsymbol{A}(R_{kn}, l)) = \boldsymbol{X}(R_{kn}, l)^T \boldsymbol{A}(R_{kn}) \boldsymbol{X}(R_{kn}, l) = \omega_{knl} \quad (3.8)$$

where $\boldsymbol{A}(R_{kn})$ is the adjacency matrix relative to R_{kn}.
The resulting value ω_{knl} increases as the initial scores of the top ranking connected genes increases.

We define the global trend of ω_{knl} over all layers among the top $i \in 1, 2, ..., n$ ranking genes at varying k, summing the ω_{knl} values of each layer, normalized by the maximum value observed in such layer using different k:

$$\omega'_{ki} = \sum_{l=1}^{L} \frac{\omega_{knl}}{\max_{ki}(\omega_{knl})} \tag{3.9}$$

The non-decreasing trend of ω'_{ki} varies in the interval $[0, L]$ and highlights the effect of k on the connectivity of top n ranking genes found by mND and the presence of high initial scores in such gene networks.

3.3 Computational cost

The computational cost of mND depends on interactome size (number of nodes and links), number of layers and number of permutations used in significant assessment. In particular, ND is the rate-limiting step, which is repeated several times during significance assessment. For example, the computation of ND using STRING (11 796 genes and 309 850 links) on 2 layers of initial scores required approximately 30s on a server with dual Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz, 64GB DDR4 2133 MHz memory and disk storage on Lustre Filesystem; the whole analysis, involving 1000 permutations, took about 1 hour and 50 minutes on 4 cores. See Appendix Table A.1 for additional details and further examples.

3.4 mND R package

To make the proposed method available to the scientific community, mND has been implemented as an R software package with extensive documentation covering installation, use and reproducible examples. The R package is free and open source, available online under the GNU License at URL: https://www.itb.cnr.it/mnd. The use of R should facilitate integration with other existing bioinformatics tools to further processed the

results generated by mND (e.g. assessment of the presence of a significant gene module, functional characterization of gene networks).

To facilitate the usage, the package contains a tutorial with detailed examples:

```
vignette('mND')
```

3.4.1 Input of mND R package

The following data are required to run the integrative analysis with mND (Figure 3.3, red boxes):

- $\boldsymbol{A}$: adjacency matrix of undirected interaction network G;
- $\boldsymbol{X}_0$: score matrix;

Moreover, the analysis requires four parameters:

- α: smothing factor (see paragraph 3.2.1, $\alpha = 0.7$ by default);
- k: number of top neighbour to consider ($k = 3$ by default);
- r: number of permutations of the input matrix;
- cores: number of cores to run analysis in parallel (cores = 1 by default).

To run the optional gene classification, the following inputs are required:

- H_l: high scoring genes names in each layer of $\boldsymbol{X}_0$;
- top_{N_l}: number of genes with the highest neighbourhoods to define the gene set N_l (see paragraph 3.2.5);
- α_{N_l}: significance level on the empirical p-value to define the gene set N_l (see paragraph 3.2.5).

Examples data of adjacency matrix and score matrix are given in the package that you can load as follow:

```
data(A,X0)
```

The first contains molecular interactions retrieved from STRING [94]; the second reports gene mutations ($\boldsymbol{x}_1$) and gene expression variation ($\boldsymbol{x}_2$) in breast cancer collected from TCGA.

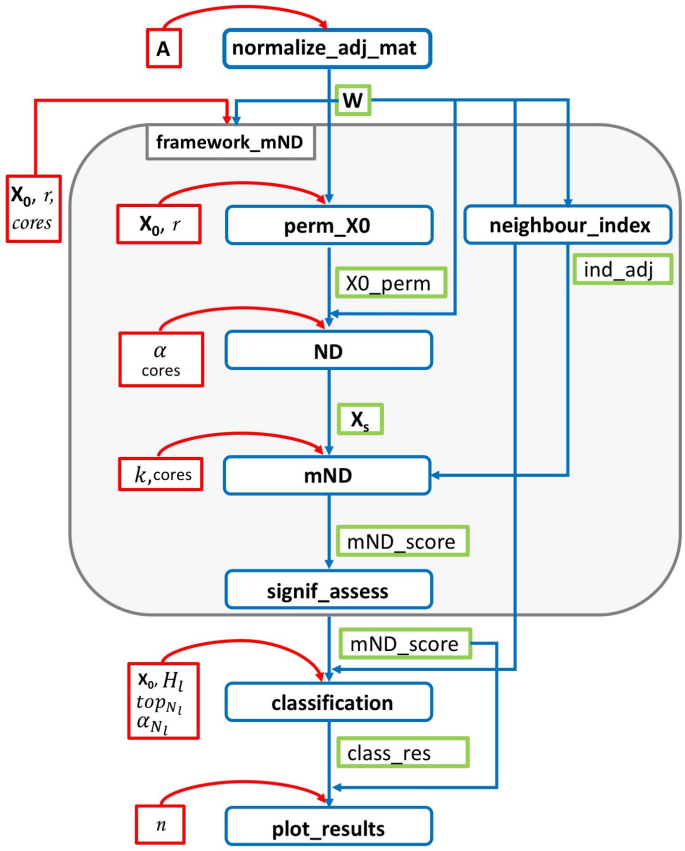


Figure 3.3: mND R package Blue boxes: function name; Red boxes: function input; Green boxes: function output; Grey box: analyses that can be easily carried out through the wrapper function "framework_mND ".

3.4.2 Functions

The mND package provides ten functions (Table 3.1, Figure 3.3 blue and grey boxes). To calculate the mNDp_i score (Equation 3.6), the first seven function of Table 3.1 should be applied to input data in this order:

```
W <- normalize_adj_mat(A)
X0_perm <- perm_X0(X0, r, W)
Xs <- ND(X0_perm, alpha, W, cores)
ind_adj <- neighbour_index(W)
mND_score <- mND(Xs, ind_adj, k, cores)
mND_score <- signif_assess(mND_score)
```

Most of the analyses can be easily carried out through the wrapper function "framework_mND" (Figure 3.3, grey box) that calculates permutations of $\boldsymbol{X_0}$, applies network-diffusion on data, computes the mND score and the relative empirical p-value.

```
W <- normalize_adj_mat(A)
mND_score <- framework_mND(X0, W, k, r, cores)
```

Outputs of function can be used to classify genes in each layer with the "classification" function:

```
class_res <- classification(mND_score, X0, Hl, topNl
   , alphaNl)
```

Furthermore, results could be visualized and saved with the "plot_results" function that gives in output the following plots:

- genes ranked by mND score and the corresponding p-value;
- gene networks composed of the top n ranking genes;
- gene classification for the top 100 ranking genes across layers.

Lastly, the function "optimize_k" provides the opportunity to optimize the k value. It calculates the mND score for different k values ("k_val"), evaluates which value of k yields connected networks enriched in initial

scores and generates a plot that shown the trend of ω'_{ki} values (see paragraph 3.2.6) among the top ranked genes by mND ("top") at varying values of k:

```
k_val <- seq(1,5,1)
k_results <- optimize_k(Xs, X0, k_val, ind_adj, W,
    top)
```

Function	**Description**
normalize_adj_mat	Perform symmetric normalization of $\boldsymbol{A}$.
perm_X0	Perform r permutations of $\boldsymbol{X}_0$.
ND	Perform network diffusion.
neighbour_index	Return indices of neighbours for each gene.
mND	Calculate mND score
signif_assess	Perform significance assessment.
classification	Perform gene classification.
plot_results	Return plots with results.
framework_mND	Wrapper function: Calculate permutations of $\boldsymbol{X}_0$, applies network-diffusion on data, computes the mND score and the relative empirical p-value.
optimize_k	Perform k optimization.

Table 3.1: List of functions of the R package mND $\boldsymbol{A}$: adjacency matrix; r: number of permutations; $\boldsymbol{X}_0$: input matrix of scores

3.5 Performance assessment

Performance of mND with respect to existing methods was evaluated considering the general problems of locating high scoring genes in network proximity across multiple layers (paragraph 3.5.2) and recovering known cancer genes in four cancer types, using two types of omics and a single type of omics at patient-level (paragraph 3.5.4).

The former is involved in several applications in which multi-omics datasets are explained relying on the architecture of intracellular circuits, underlying "hot" gene modules (e.g. disease modules) supported by multiple layers of information, while the latter has addressed a problem considered by recent network-based methods for the analysis of multi-omics datasets. Sensitivity of the mND results to the value of α and k was also evaluated (paragraph 3.5.3).

Lastly, in the paragraph 3.5.5, it is shown that the application of mND to rank genes based on mutations and expression changes in breast cancer points to relevant pathways underlying the disease, providing a more complete picture than each individual omics on its own.

3.5.1 Data Source

3.5.1.1 Molecular interactions

Three sources of interactions were considered, abbreviated as STRING (11 796 genes; 309 850 interactions) [94], GH (13 244; 138 045) [95] and WU (6 016; 128 150) [96]. Native identifiers were mapped to Entrez Gene [97] identifiers using the R package "org.Hs.eg.db" [98].

3.5.1.2 Analysis of somatic mutations and gene expression variations

Somatic mutations (SM) and gene expression (GE) data from matched tumour-normal samples (blood for SM and solid tissue for GE) were collected from The Cancer Genome Atlas (TCGA) [99] (Figure 3.4) for breast invasive carcinoma (BC), lung squamous cell carcinoma (LUSC), prostate

adenocarcinoma (PRAD), and thyroid carcinoma (THCA), using the R packages TCGAbiolinks [100] and isma [25] (see Appendix B) and considering the human genome version 38 (hg38).

Mutation Annotation Format files were obtained from 4 pipelines: Muse [101], Mutect2 [102], SomaticSniper [103], Varscan2 [104]. Only mutation sites detected by at least two variant callers were considered. Gene mutation frequencies were calculated as the fraction of subjects in which a gene was associated with at least one mutation. Gene expression data were obtained using the TCGA workflow "HTSeq-Counts". The R package limma [27] was used to normalize and quantify differential expression in matched tumor-normal samples, yielding log-fold changes, the corresponding p-values and FDRs (Bonferroni-Hochberg method).

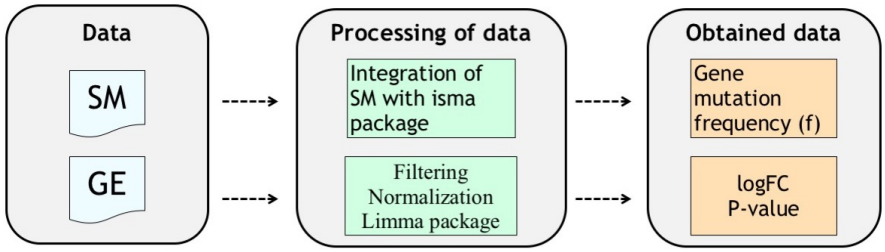


Figure 3.4: Analysis of TCGA data. Somatic Mutations: SM and gene expression (GE) data from matched tumour-normal samples were collected from TCGA and processed as reported.

The four cancers datasets were considered in two tasks: the analysis of two types of omics, mutations and expression changes, and the analysis of mutation profiles of multiple patients. In the first task, $\boldsymbol{x}_1$ was defined as gene mutation frequencies while $\boldsymbol{x}_2$ as $-\log_{10}(\text{FDR})$. In the second task each layer $\boldsymbol{x}_i$ was represented by mutation profiles of subjects, defined as the number of mutation sites in each gene. In all analysis, empirical p-values were calculated on a total of 1 000 permutations (the input matrix and 999 random permutations of it).

In the joint analysis of mutations and expression changes in BC (paragraph 3.5.5), the two sets of high scoring genes (H_1, H_2) were defined

considering, respectively, all genes with at least one mutation (1 238 genes) and the top 1 200 differentially expressed genes (FDR $< 10^{-7}$).

3.5.2 Finding significant genes that lie in network proximity

To assess the ability of mND in finding high scoring genes in network proximity across multiple layers, two types of real signal (gene mutation frequencies and log fold changes) were assigned to gene modules of different size and modularity, corresponding to real pathways (Figures 3.5,3.6-3.7 A).

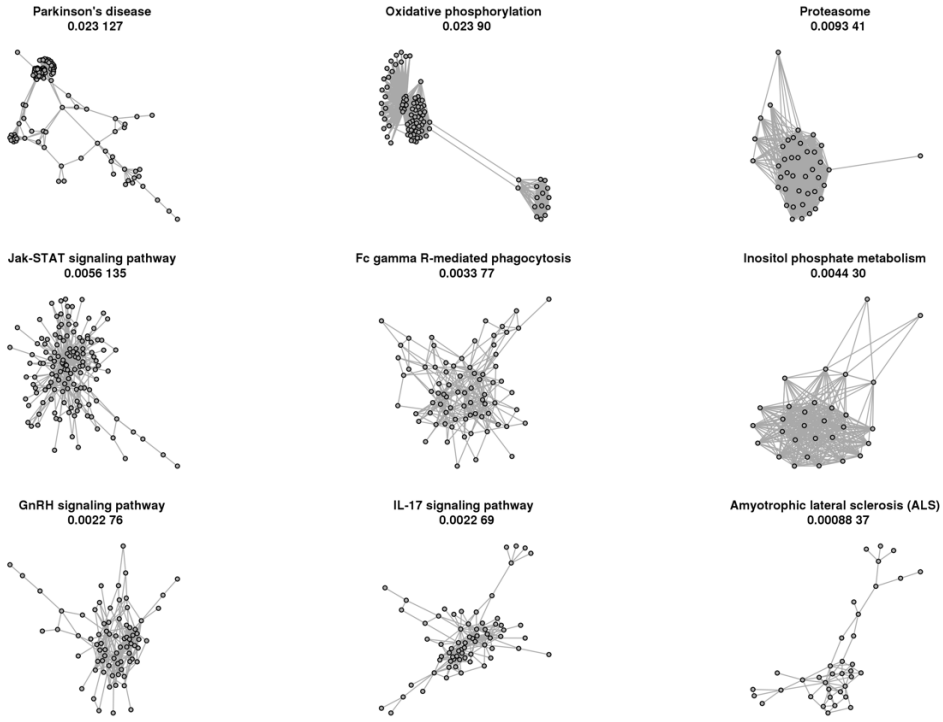


Figure 3.5: Gene modules. Largest connected component of biological pathways from KEGG database [105] in GH interactome. The two quantities below pathway name are modularity (as defined in Clauset et al. [106] and implemented in R function "modularity") and size (number of genes).

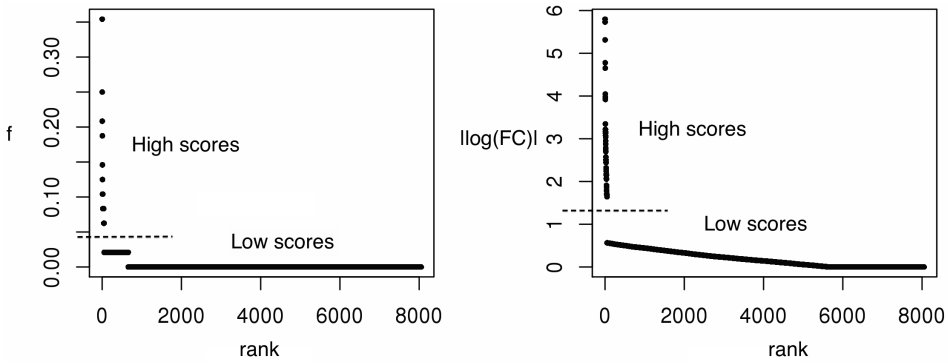


Figure 3.6: Initial gene scores assigned to gene modules. Two types of high and low scores that were randomly assigned to gene modules, derived from gene mutation frequency across subjects (left) and absolute fold changes between matched tumor-normal samples (right) from TCGA breast cancer data.

Each gene module was defined as the largest connected component obtained considering the genes associated with a biological pathway (from KEGG database [105]) and all interactions among them in GH interactome (Figure 3.5). The highest and lowest values of gene mutation frequencies ($\boldsymbol{x}_1$) and fold changes ($\boldsymbol{x}_2$) calculated from BC data (see above) were used to define, respectively, high scoring genes and low scoring genes (Figure 3.6). High scoring values were randomly assigned to genes of each module independently for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, thus to obtain a specific percentage (e.g. 10%) of high scoring genes within the module in each layer. Unused high scoring values were assigned to genes outside the module and, lastly, low scoring values were assigned to the remaining genes within and outside the module.

In each of the resulting configurations, the recall values obtained by mND were compared to those obtained by other methods:

- the product of ND scores ("NDPROD") between the two layers (as in Ruffalo et al. [59], *Chapter 2* paragraph 2.5.2.1): $x^*_{i,1} \cdot x^*_{i,2}$, where $x^*_{i,l}$ is the network-constrained value of i-th gene in l-th layer;

- the minimum of ND scores ("NDMIN") for each gene i between the two

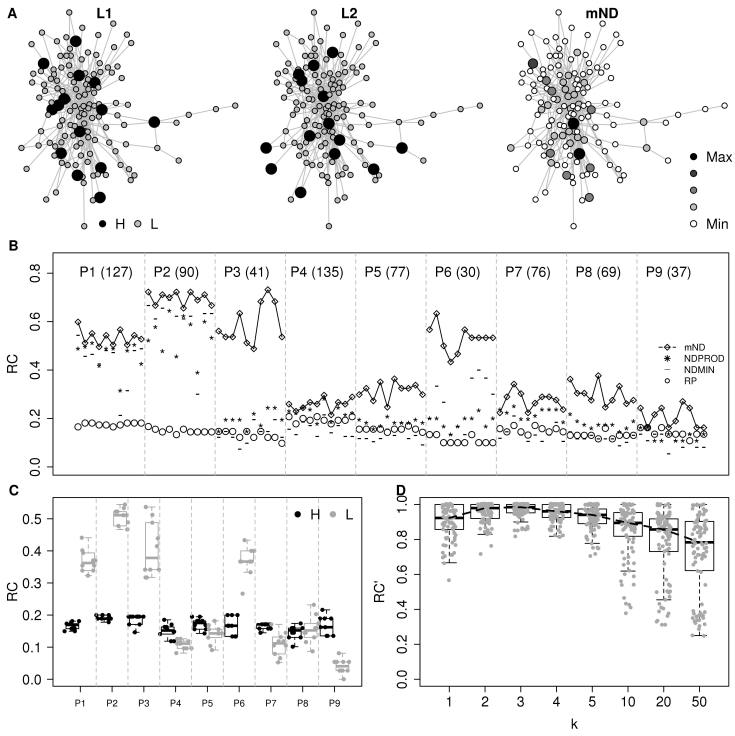


Figure 3.7: Performance in ranking high scoring genes in network proximity. (**A**) Example of a gene module with its high scoring genes (H, black) in each of the two layers and the resulting mND score; only genes belonging to the module and links occurring among such genes are reported. (**B**) Recall values for 10 signal permutations for each of the 9 modules (P1, P2, ..., P9), using mND score and other methods; the number between parentheses after module id is module size. (**C**) Recall values, shown separately for high scoring genes and other genes in each module. (**D**) Recall values normalized by the highest recall found for each input configuration at varying number of neighbors (k). (**A-D**) These results were obtained using interactome GH.

layers (as in TieDIE [48], *Chapter 2* paragraph 2.5.2.1): $min(x^*_{i,1}, x^*_{i,2})$, where $x^*_{i,l}$ is the network-constrained value of i-th gene in l-th layer;

- the rank product ("RP") of initial scores: $\sqrt{r(x_{i,1}) \cdot r(x_{i,2})}$, where $r(x_{i,l})$ is the rank of i-th gene in l-th layer.

The recall was defined as the fraction of module genes ranked (by the assessed method) among the top M genes, where M is the module size.

The rank product (RP) was successful in identifying genes with high scoring values in at least one of the two layers (Figure 3.7 B), but typically missed other module genes with lower values. NDPROD, a multi-omic approach described in [59] and corresponding to using only the term g(i) in Equation 3.4, led to better performance than RP in more than half of the cases, and equal or even low performance in others, indicating the failure to identify high scoring genes in favour of genes in network proximity to the module, but outside of it (Figure 3.7 B). Similarly, NDMIN, the multi-layer combination strategy underlying TieDIE method [48], yielded recall values that are higher or lower than RP depending on gene module and signal distribution. Instead, mND determined the highest recall in almost all cases. This result underlines the importance of using gene neighbourhoods, i.e. the term t(i) in Equation 3.4 (Figure 3.7 B). Importantly, the performance of mND is the result of spotting both high scoring genes (almost all) plus other module genes with low score, but relevant topological position (Figure 3.7 C).

Overall, a small number of neighbours (parameter k in Equation 3.4) was sufficient to guarantee the highest performances (Figure 3.7 D), which were observed around $k = 3$. A similar trend was observed when finding significant genes lying in network proximity over 3 layers (Figure 3.8).

To assess whether the results obtained in ranking high scoring genes lying in network proximity (Figure 3.7) were limited to the interactome in use (GH), the same analyses were repeated using a different interactome (STRING). The same patterns were observed in terms of mND performance, types of genes found and the relation between performance and k parameter (Appendix Figure A.1).

This is a significant result, considering the relevant differences between the two interactomes, and further underlines the wide applicability of mND in analysing multiple biological evidences spread in complex networks.

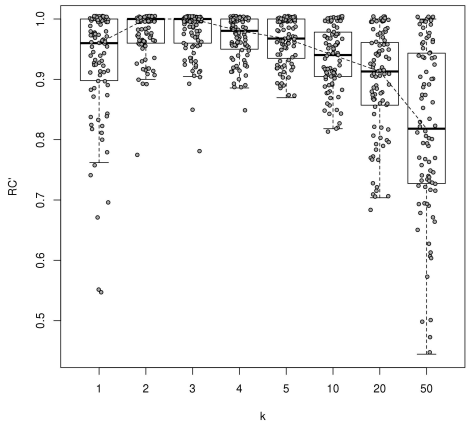


Figure 3.8: Recall values in the analysis of 3 layers. Recall values normalized by the highest recall found for each input configuration at varying number of neighbors (k). This analysis was carried out like described in Chapter 3.5.2, but using 3 layers of mutation frequencies, like in a hypothetical analysis of three cancer subtypes or a hypothetical pan-cancer analysis.

3.5.3 Sensitivity of mND results to parameters α and k

mND depended on two parameters: α, that weights the contribution of the two addends in Equation 3.1, and k, the maximum number of neighbours that are considered in the calculation of mND score (Equation 3.3-3.4)

Parameter α was set to 0.7, a value that represents a good trade-off between diffusion rate and computational cost, and determined consistent results in previous studies [43,47,107,108]. However, the sensitivity of mND to α was estimated and it was found that varying α by 10% resulted in highly correlated mND scores and only a few different genes (6-8%) among the top 100 (Table 3.2 and Figure 3.9).

In the paragraph 3.5.2, mND obtained the highest performance in finding significant genes in network proximity over two or more layers considering just a few top ranking neighbours ($2 \leq k \leq 5$). Overall, $k = 3$ (i.e. at most 3 neighbours) determined the highest performances in such

α	0.63	0.7	0.77
$\langle\%\rangle$	6.0	0	7.5

Table 3.2: Sensitivity of mND to α Average percentage ($\langle\%\rangle$) of genes that change within the top 100 ranked by mND in 90 runs varying α by $\pm 10\%$.

problem. This observation can be explained considering 3 neighbours a reasonable trade-off to include multiple neighbours without penalizing high degree genes. However, the sensitivity of mND to the value of k was evaluated and it was found that varying k of one unit had only minor effects on mND scores, which are highly correlated and indeed differ of only a few ($\sim$ 4-6) genes among the top 100 (Table 3.3 and Figure 3.10).

k	2	3	4
$\langle\%\rangle$	5.8	0	4

Table 3.3: Sensitivity of mND to k Average percentage ($\langle\%\rangle$) of genes that change within the top 100 ranked by mND in 90 runs varying k of 1 unit.

An opportunity to further optimize the value of k relies in selecting a value that yields connected networks enriched in initial scores (paragraph 3.2.6).

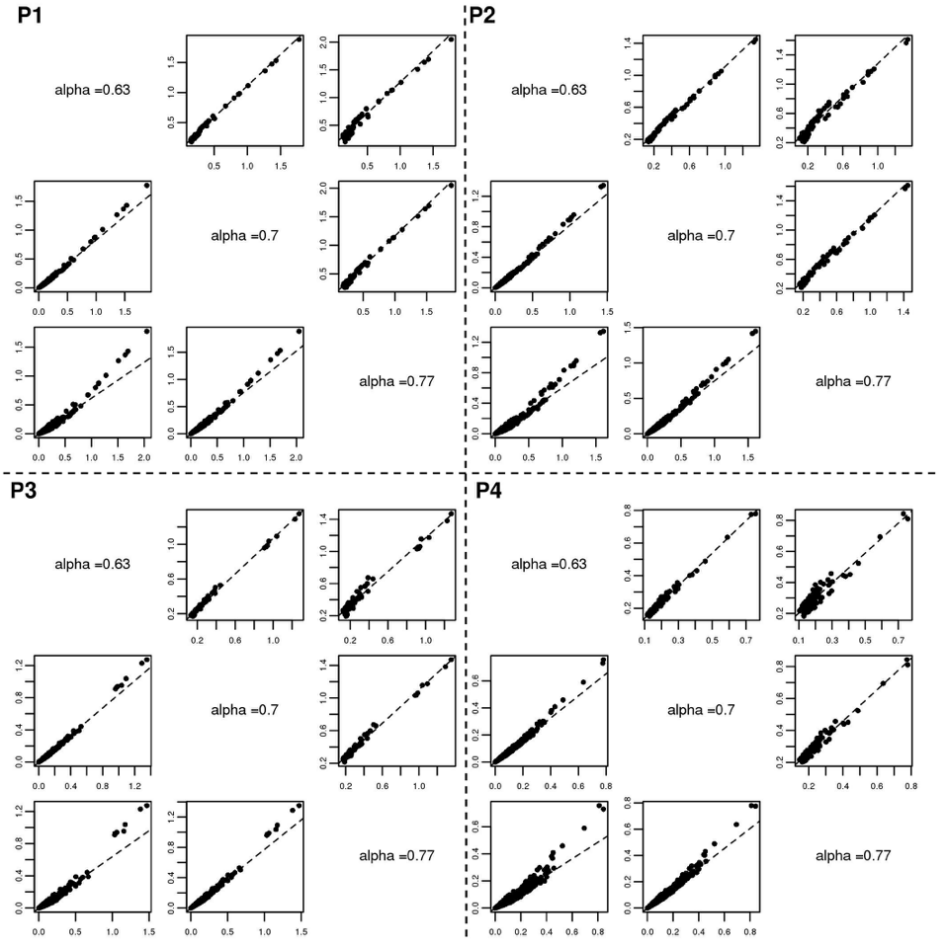


Figure 3.9: Sensitivity of mND to α parameter. Correlation of mND scores at varying α, reported for all genes (below diagonal) and top 100 genes only (above diagonal) in four examples (P1-P4) of the analysis described in the paragraph 3.5.2.

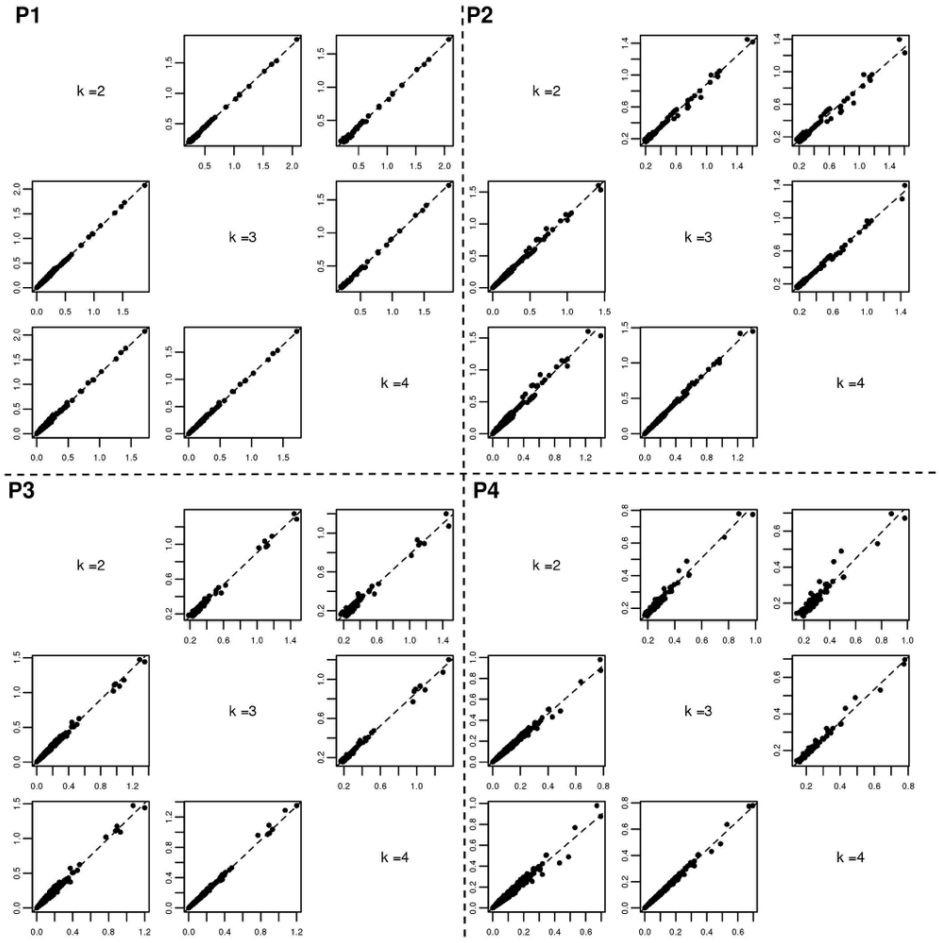


Figure 3.10: Sensitivity of mND to the value of k. Correlation of mND scores at varying k, reported for all genes (below diagonal) and top 100 genes only (above diagonal) in four examples (P1-P4) of the analysis described in the paragraph 3.5.2

3.5.4 Recovering known cancer genes

The performance of mND was also evaluated in the problem of recovering known cancer genes at low false positive rates (FPRs), by calculation of the partial area under the ROC curve (pAUC).
The pAUC of mND was compared with those obtained by other network-based methods, like NDPROD, NDMIN and NetICS [56] (*Chapter 2* paragraph 2.5.2.4), and considering four cancer types (BC, LUSC, PRAD, THCA).

The pAUC accounts for the number of true positives that score higher than the n-th highest scoring negative, measured for all value from 1 to n:

$$\mathrm{pAUC_n} = \frac{1}{n\mathrm{TP}} \sum_{i=1}^{n} \mathrm{TP_i} \tag{3.10}$$

where TP is the total number of known cancer genes and $\mathrm{TP_i}$ is the number of true positives that score higher than the i-th highest scoring negative [109].

Genes mutations associated with cancer were collected from COSMIC [110] and previous studies [111,112]. Differentially expressed genes were derived from Bioexpress [113], considering log2-fold change between matched primary tumor-normal samples greater than or equal to 1 and FDR < 0.05.

Considering mutations and expression changes as input, mND reported higher pAUC than other methods in all four cancer types considered (Figure 3.11). Furthermore, performance was studied using mutational profiles only as input and, also in this case, mND reported better performance than other methods (Appendix Figure A.2) and gene classification underlined the presence of several linkers with a relevant role in BC (Appendix Figure A.3). For instance, the deletion of *HIC-1*, never found mutated in the dataset under analysis but spotted as linker in 15 subjects, has been demonstrated to promote BC [114, 115]; *FYN* has been proposed as a prognostic marker in ER+BC [116] and promotes mesenchymal phenotypes of basal types BC cells [117].

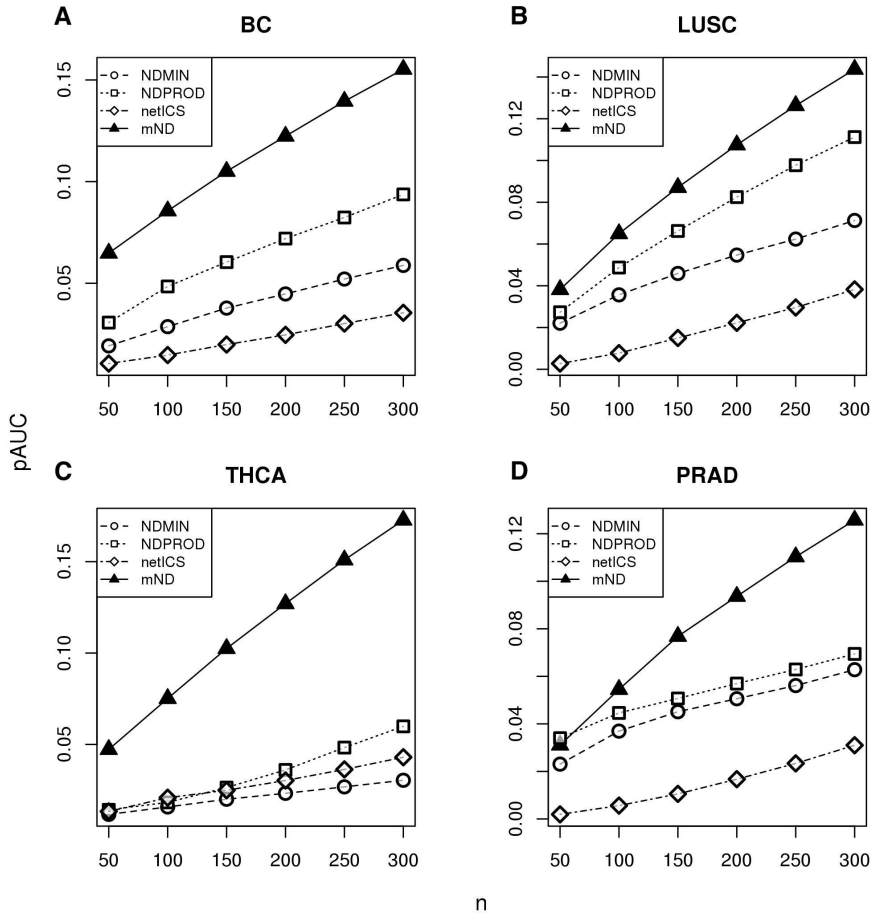


Figure 3.11: Performance in recovering known cancer genes. Partial AUC (pAUC) at varying number of top false positive ranking genes (n) in the analysis of mutations and expression changes in four cancer types. (**A-D**) These results were generated using interactome WU.

3.5.5 Gene networks enriched in mutations and expression changes in breast cancer

As a proof of principle, mND was applied to find functionally related genes on the basis of gene mutation frequency (layer 1, L_1) and gene expression variation (layer 2, L_2) in BC.

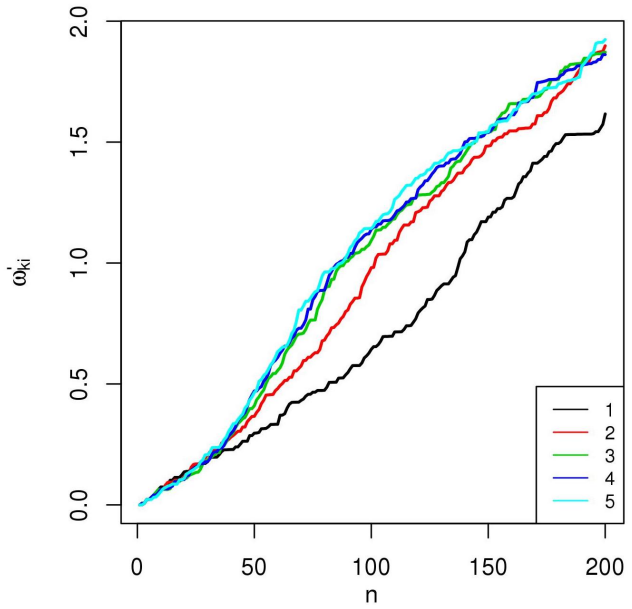


Figure 3.12: Effect of varying k on the enrichment of the top gene networks in high initial scores. Trend of ω'_{ki} values (see *Chapter 3*, paragraph 3.2.6) among the top 200 genes found by mND at varying values of k in the analysis of mutations and gene expression changes in BC.

$k = 3$ has been observed as a reasonable choice to obtain connected gene networks enriched in genes with the highest mutation frequencies and expression variations (see paragraph 3.2.6 and Figure 3.12).

Genes highly ranked by mND (Figure 3.13 A) include those that were relevant according to initial scores in both layers (Figure 3.13 B, green rhombuses, e.g. *CCNB1, TOP2A*), as well as those that were high scoring

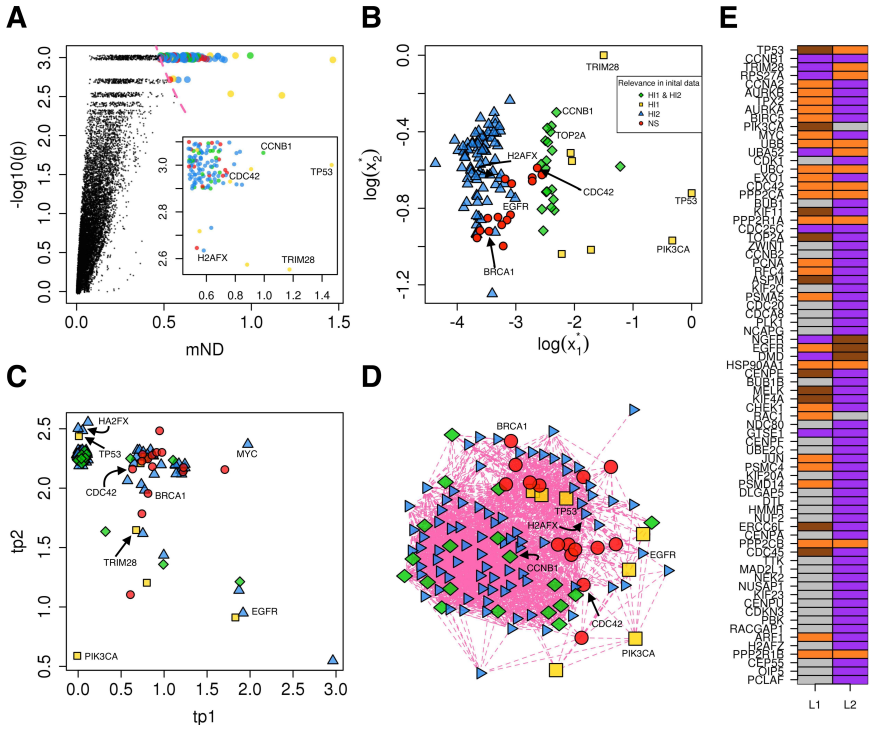


Figure 3.13: Analysis of mutations and expression changes in BC. (**A**) mND score and empirical p-value; the red dashed line indicates the top 123 genes (subplot); colors and shapes have the same meaning of panel B. (**B**) Gene diffusion scores of the top 123 genes ranked by mND. (**C**) tp values (Equation 3.7) for the two layers. (**D**) Gene network composed of the top 123 genes ranked by mND; colors and shapes have the same meaning as in panel B. (**E**) Classification of genes across layers (only the top 75 ranked genes are shown for clarity); brown: isolated; orange: linker; purple: module; grey: not significant. (**A-D**) Layer 1 (L_1): mutations; Layer 2 (L_2): expression variations. H_1, H_2: sets of genes with high initial scores in respectively L_1 and L_2. Green rhombuses: genes belonging to H_1 and H_2; blue triangles: genes belonging only to H_1; yellow rectangles: genes belonging only to H_2; red shapes: genes neither in H_1 nor inH_2. These results were generated using interactome STRING.

in one of them (Figure 3.13 B, e.g. *EGFR* and *PIK3CA*) and linker genes (Figure 3.13 B, red circles), which have low initial values, but lie in relevant network proximity to significantly altered genes. Interestingly, top scoring linker genes include genes already known to be involved in BC, such as *CDC42* and *BRCA1* (Figure 3.13 B-C).

To assess whether genes highly ranked by mND are in significant network proximity, we used network resampling [42]: this computational approach calculates a network score considering top ranking genes and shows to which extent such network score is expected if links among genes are shuffled (keeping the same degree distribution). This procedure confirmed that genes highly ranked by mND are in significant network proximity (Appendix Figure A.4): in particular, a dense module of 123 genes was identified (Figure 3.13 D).

Beyond gene global ranking, mND classified genes in each layer as members of a module, linkers, or isolated genes, on the basis of the amount of signal found in the genes themselves and their neighbours. Complementing the global ranking with layer-by-layer information on gene positions, such classification helps clarifying genes role in the context of the alterations detected and suggests possible underlying molecular mechanisms (Figure 3.13 E). For instance, *TP53* is classified as "isolated" and clearly emerges as a gene with primary role in BC, not only because of its mutation, but also because its functional partners are differentially expressed (it is classified as linker in gene expression layer). *CDC42* is classified as linker in both layers: it neither carries a relevant amount of mutations nor is among the top differentially expressed genes, but its interacting partners are highly enriched in both mutations and differential expression. Interestingly, *CDC42* is an important molecule in luminal BC, with prognostic significance [118]. Among genes highlighted as modules, we found *PIK3CA* (a highly mutated gene in BC [119]), highly ranked on the basis of mutations; other genes play a role according to one type of alteration only, like *CDCA8* [120], which emerged as being involved specifically in terms of differential expression, being a member of a differential expression module.

Lastly, the genes prioritized by mND were characterized in terms of biological pathways. Pathways were downloaded from the KEGG database

[105]. A total of 331 human pathways with at least 5 genes were considered. The number of genes prioritized in each pathway by mND, by gene expression (x_2), ND scores of gene mutation frequencies (x_1) and gene expression (x_2^*), were quantified for different numbers of top ranking genes ($n = 50, 100, 150, 250, 300$).

For each pathway and value of n, the difference $\mathrm{DP}(n)$ between the number of genes (D) found by mND and the best of the other approaches was quantified as:

$$\mathrm{DP}(n) = D_{\mathrm{mND}}(n) - \max(D_{x_2}(n), D_{x_1^*}(n), D_{x_2^*}(n)) \tag{3.11}$$

Interestingly, mND found relatively more genes than each omics considered independently ($\langle \mathrm{DP} \rangle > 0$), in pathways like: KEGG 'Breast Cancer' and signal transduction ways known to have a relevant role in BC (Figure 3.14), 'Cell Cycle' [121], 'Hippo signalling pathways' [122], 'FoxO Signalling pathways' [123], 'p53 Signalling pathways' [124], 'PI3K-Akt signalling' [125] and 'Proteasome' [126].

Therefore, the joint analysis of the two omics led to enrichment in relevant pathways, compared to single omics on its own, a result that underlies the added value of combining multiple evidences with mND.

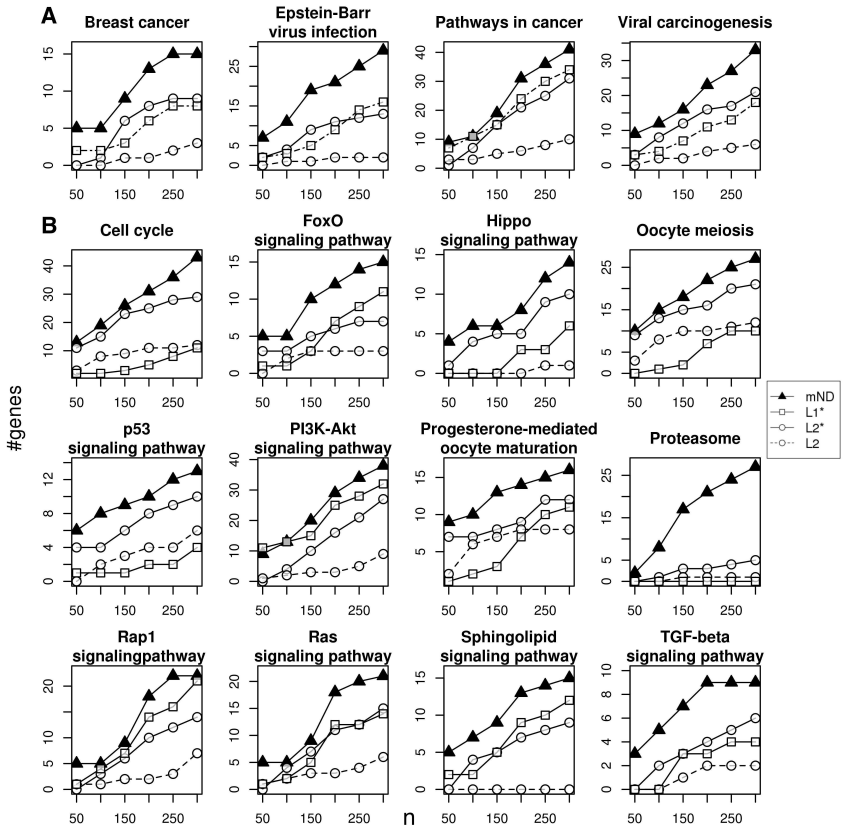


Figure 3.14: Pathways enriched in mutated genes and/or differentially expressed genes in BC. Number of genes found by mND and single omics analyses (L_1^*, L_2^* and L_2) in each pathway at varying number of top ranking genes considered (horizontal axis, n); L_1: mutations; L_2: gene expression variations; the asterisk distinguishes between gene ranking by original data and the corresponding network diffusion scores. (**A**) Disease pathways; (**B**) Other pathways. (**A-B**) Pathways from KEGG database.

Chapter 4

Integrative analysis of genomics, epigenomics and transcriptomics in autism spectrum disorders.

Current studies suggest that ASDs may be caused by many genetic factors and integrative analysis of multiple omics could provide a more comprehensive view of the disease. This chapter presents an integrative network-based analysis, based on mND algorithm, of genes reported as associated with ASDs by studies that involved genomics, epigenomics and transcriptomics.[1]

[1]The contents of this chapter are published in: *N. Di Nanni, M. Bersanelli, F. Cupaioli, L. Milanesi, A. Mezzelani, E. Mosca. (2019) "Network-based integrative analysis of genomics, epigenomics and transcriptomics in autism spectrum disorders". International Journal of Molecular Sciences (IJMS), 20:13, 3363.* https://doi.org/10.3390/ijms20133363.

4.1 The complex molecular basis of ASDs

Autism spectrum disorders (ASDs) are among the most common neurodevelopmental disorders. ASDs are characterized by impaired social interactions, repetitive behavior and restricted interests, and are often in comorbidities with other conditions such as epilepsy, mental retardation, inflammation and gastrointestinal disorders. Despite the high heritability of ASDs is well established, the exact underlying causes are unknown in at least 70% of the cases [127]. Large genome-wide association studies (GWAS), CNV testing and genome sequencing yielded many non-overlapping genes, a fact that underlines the complex genetic heterogeneity of ASDs [127] and reflects the architecture of intracellular networks, in which several possible combinations of genetic variations are likely to lead to a common pathological phenotype [30, 31].

The identification of key molecular pathways that link many ASDs-causing genes is of prominent importance to develop therapeutic interventions [127]. In this context, network-based and pathway-based analyses provide functional explanations to non-overlapping genes and narrow the targets for therapeutic intervention [128]. The rich functional pathway information emerging from such analyses might unearth common targets that are amenable to therapy [127].

The application of ND to genes associated with ASDs from genetic data had led to the identification of gene networks and pathways particularly enriched in disease genes [108]. Interestingly, several genes predicted as relevant in such study are now included in the SFARI Gene database [129], which provides curated information on all known human genes associated with ASDs.

In addition to genetics, several reports have suggested a role for epigenetic mechanisms in ASDs etiology [130, 131]. Recent studies have also demonstrated the utility of integrating gene expression with mutation data for the prioritization of genes disrupted by potentially pathogenic mutations [132, 133]. More generally, the integrative analysis of multiple omics has emerged as an approach that can be crucial to unravel the mechanism of this complex disease [5, 8].

While the analysis of epigenomics and transcriptomics from brain-derived samples can provide important insights into potential mechanisms of disease etiology, there are relevant limitations with these types of studies (e.g. quality of autopsy-derived tissue, sample size, influence of life experience and cause of death) [131]. These barriers have been overcome by analysing blood samples and recent blood-based works have shown the usefulness of this alternative approach to gather insights into ASDs [131, 134–136].

Considering all these aspects, an integrative network-based analysis was performed using results emerged from several studies on ASDs, based on genomics, epigenomics and transcriptomics (Figure 4.1). Firstly, following the hypothesis of the omnigenic model [31], genetic data were analyzed to introduce a graduated scale of gene relevance in relation to core genes for ASDs. Subsequently, omics data were integrated using mND algorithm (*Chapter 3*, paragraph 3.2) and a gene network significantly enriched in genes supported by one or more of the considered evidence (genomics, epigenomics and transcriptomics) was proposed. The gene network involved genes that participate in several pathways relevant to ASDs, which were distinguished by type (or types) of alteration from which they are affected. Collectively, the network-based meta-analysis provided a prioritization of the large number of genes proposed to be associated with ASDs, based on genes' relevance within the intracellular circuits, the strength of the supporting evidences of association with ASDs and the number of different molecular alterations affecting genes.

4.2 Data sources

The following paragraphs describe the source of data used in the analysis: molecular interactions and genes associated with ASDs on the basis of genomics, epigenomics and transcriptomics.

For each omics data, the genes were divided into two groups: those supported by strong evidence ("-MAJOR" group) and the others ("-MINOR").

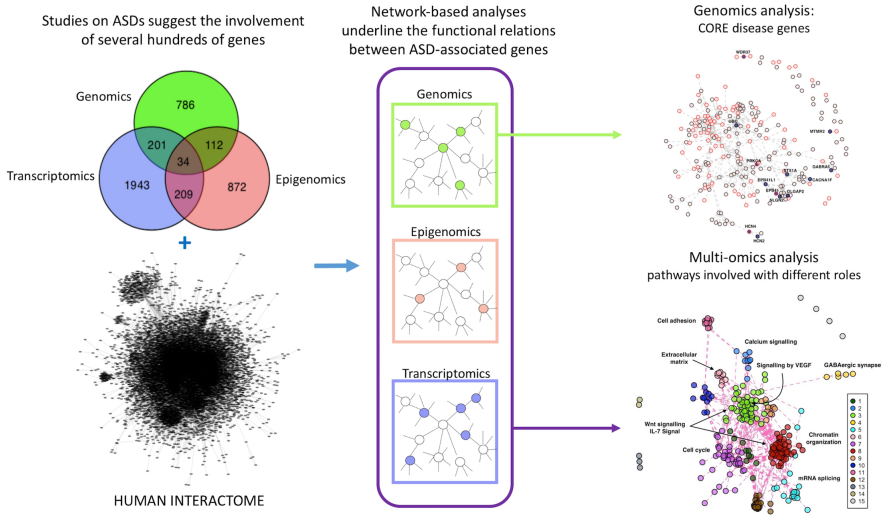


Figure 4.1: Overview of the integrative network-based analysis.

4.2.1 Molecular interactions

Molecular interactions were collected from STRING database [94], for a total of 12 739 genes and 355 171 links with high confidence (score $\geq$ 700). In case multiple proteins mapped to the same gene identifier, only the pair of gene identifiers with the highest STRING confidence score was considered.

4.2.2 Genomics data

Genes associated with ASDs on the basis of genomics evidences were collected from SFARI Gene database [129], two recent large studies [137, 138] and a series of previous studies summarized in [108], for a total of 1 133 genes (Table 4.1).

The SFARI Gene scoring system classifies genes on the basis of the strength of the supporting evidences as: "syndromic" (S), "high confidence" (1), "strong candidate" (2), "suggestive evidence" (3), "minimal evidence" (4), "hypothesized but untested" (5) and "Evidence does not support a role"

(6).

Genes classified as S, 1, 2, 3, 1S, 2S, 3S and 4S were assigned to the GENOMICS-MAJOR evidence group (334 genes).

Genes belonging to the GENOMICS-MINOR (799 genes) group were collected from: Mosca et al. [108], in which genes associated with SNPs, mutations and CNV emerging from several large studies are reported; the meta-analysis study of GWAS of over 16 000 individuals with ASDs [137]; and the whole-exome sequencing study of rare coding variation in 3 871 autism cases and 9 937 ancestry-matched or parental controls [138]. Native gene identifiers were converted to Entrez Gene [97] identifiers and only genes occurring in STRING network were considered in network-based analyses.

4.2.3 Epigenomics

Genes associated with ASDs at epigenomics level were collected from a previous study [131], in which the authors performed a case-control meta-analysis of blood DNA methylation among two large case-control studies of autism (796 ASDs cases and 868 controls) using METAL software [139] on the probes that were present in both studies.
All genes found by their meta-analysis with $p < 10^{-3}$ were assigned to EPIGENOMICS-MAJOR group while the genes with $10^{-3} \leq p < 5 \cdot 10^{-3}$ were assigned to EPIGENOMICS-MINOR (Table 4.1). Native gene identifiers were converted to Entrez Gene [97] identifiers and only genes occurring in STRING network were considered in network-based analyses.

4.2.4 Transcriptomics

Genes associated with ASDs at transcriptomics level were collected from the four studies [134, 140–142] reported in [143], in which the original authors generated blood-based gene expression profiles from microarray experiments with sample sizes greater than 40 and provided list of differentially expressed genes.
Following the approach by Saeliw et al. [143], genes reported as differentially expressed in at least two studies were assigned to the TRANSCRIPTOMICS-

Type of evidence	Description	Subjects	Number of genes			
			Initial		Selected	
			**	*	**	*
G	SFARI [129]	-	404	1087	334	799
G	Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules [108]	-				
G	Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder [137]	15 954				
G	Synaptic, transcriptional and chromatin genes disrupted in autism [138]	13 808				
E	Case-control meta-analysis of blood DNA methylation and autism spectrum disorder [131]	1 654	416	1444	272	955
T	Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders [140]	116	330	3045	256	2131
T	Blood gene expression signatures distinguish autism spectrum disorders from controls [134]	285				
T	Disrupted functional networks in autism underlie early brain mal-development and provide accurate classification [141]	147				
T	Gene expression in blood of children with autism spectrum disorder [142]	47				

Table 4.1: Datasets considered in this study. Selected: number of genes for which at least a high confidence interaction with any other gene is catalogued in STRING database. G: genomics; E: epigenomics; T: transcriptomics. **MAJOR evidence; *MINOR evidence

MAJOR group while the other differentially expressed genes were assigned to the TRANSCRIPTOMICS-MINOR group (Table 4.1). Native gene identifiers were converted to Entrez Gene [97] identifiers and only genes occurring in STRING network were considered in network-based analyses.

4.3 Genomics analysis

Recently, the "omnigenic model" was proposed to explain the inheritance of complex diseases [31]. In this model, the genes whose genetic damage will tend to have the strongest effects on disease risk are considered CORE genes, while those genes that have a minor impact on disease risk are des-

ignated as PERIPHERAL. The number of PERIPHERAL genes may be large, as consequence of the multiple ways in which these genes may interact with CORE throughout cell regulatory networks.
Importantly, such classification may be on a graduated scale rather than simply binary [31].

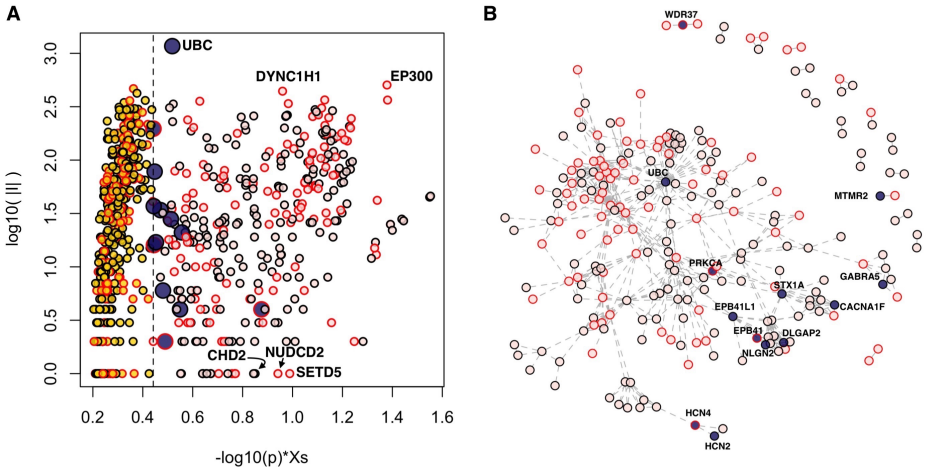


Figure 4.2: Genes in network proximity to the CORE genes of ASDs. (**A**) Diffusion score (Xs) normalized by its empirical p-value (horizontal axis) and number of interactions (|I|, vertical axis); only genes with are shown. (**B**) Connected components of "CORE+13" network. (**A-B**) Blue points: 13 genes of "CORE+13"; pink points: CORE genes; yellow points: significant genes outside "CORE+13" genes; red border of points: genes supported at transcriptomic and/or epigenetic levels

In this context, ND provides an opportunity to define quantitatively the degree of peripherality of all genes in relation to the CORE genes, exploring all possible network paths among genes in intracellular networks.
ND was applied on the human interactome defined by the high confidence functional and biophysical interactions catalogued in STRING and considering as CORE genes, those classified in SFARI as "syndromic", "high confidence", "strong candidate", "suggestive evidence" and "syndromic minimal evidence", for a total of 334 genes; as PERIPHERAL genes, the other

799 genes proposed to have a role in ASDs.

Several genes were found with a significant network proximity to CORE genes (Figure 4.2 A). By a topological point of view, among these genes both hubs (genes that establish many interactions, such as *UBC*, *DYNC1H1*, *EP300*) and genes with a lower number of connections (e.g. *CHD2*, *NUDCD2*, *SETD5*) were found, nevertheless important for the information flow within the network (Figure 4.2 A).

Interestingly, 13 genes obtained scores comparable to those of CORE genes (Figure 4.2). These results indicate that these 13 genes closely interact with the CORE genes and, in almost all cases, the number of interactions that these genes establish with the CORE genes is significant (Table 4.2). From now on, the set of CORE genes and the 13 genes closely related to the CORE genes will be called as "CORE+13". In the CORE+13 gene network, the 13 genes act as linkers between groups of CORE genes not directly connected with each other; for instance, *WDR37* (WD repeat domain 37) links *PACS1* (phosphofurin acidic cluster sorting protein 1) and *PACS2* (phosphofurin acidic cluster sorting protein 2). The resulting largest connected component involves 204 genes, while the remaining 143 genes are mostly isolated or form very small modules of 2 or 3 genes.

It has been checked whether any of these 13 genes, currently not included in the highest categories of SFARI, is nevertheless classified in other categories corresponding to a lower degree of evidence or has been reported in other network-based analyses of ASDs data [144–147]. Six genes were found belong to the categories designated as "minimal evidence" or "hypothesized but untested", and eight genes were proposed as part of gene networks associated with ASDs (Table 4.2), providing further evidences in favor of these genes.

The association with ASDs for 12 of the 13 genes is supported at genomic level (Table 4.2). In addition, *HCN4* (hyperpolarization activated cyclic nucleotide gated potassium channel 4) was found with epigenetic modifications in the study of Andrews et al. [131], while *PRKCA* (protein kinase C alpha) was found both epigenetically modified [131] and differentially expressed [141]. *WDR37* does not have supporting evidences at genomic level, but was found differentially expressed [140].

Symbol	Description	|I|	|Ic|	|CORE|	p	G	E	T	SFARI score	Other modules
HCN4	Hyperpolarization activated cyclic nucleotide gated potassium channel 4	4	2	334	$3.91 \cdot 10^{-3}$	*	*	0	-	-
DLGAP2	DLG associated protein 2	21	8	334	$3.10 \cdot 10^{-3}$	*	0	0	4	[144–146]
HCN2	Hyperpolarization activated cyclic nucleotide gated potassium and sodium channel 2	4	1	334	$1.01 \cdot 10^{-1}$	*	0	0	-	-
UBC	Ubiquitin C	1168	43	334	$1.41 \cdot 10^{-2}$	*	0	0	-	[147]
NLGN2	Neuroligin 2	28	8	334	$4.04 \cdot 10^{-7}$	*	0	0	4	[144]
WDR37	WD repeat domain 37	2	2	334	$6.85 \cdot 10^{-4}$	0	0	*	-	-
MTMR2	Myotubularin related protein 2	6	1	334	$1.47 \cdot 10^{-1}$	*	0	0	-	-
EPB41L1	Erythrocyte membrane protein band 4.1 like 1	34	9	334	$1.55 \cdot 10^{-7}$	*	0	0	-	[145]
GABRA5	Gamma-aminobutyric acid type A receptor alpha5 subunit	17	4	334	$8.43 \cdot 10^{-4}$	*	0	0	5	[145]
STX1A	Syntaxin 1A	78	10	334	$3.47 \cdot 10^{-5}$	*	0	0	4	[145, 147]
EPB41	Erythrocyte membrane protein band 4.1	16	6	334	$4.14 \cdot 10^{-5}$	*	0	**	-	[147]
CACNA1F	Calcium voltage-gated channel subunit alpha1 F	37	6	334	$3.63 \cdot 10^{-4}$	*	0	0	4	[145]
PRKCA	Protein kinase C alpha	197	11	334	$1.48 \cdot 10^{-2}$	*	*	*	4	-

Table 4.2: The 13 genes that closely interact with the CORE genes of ASDs. |I|: number of interactors; |Ic|: number of interactors that are CORE genes; p: hypergeometric probability of observing |Ic| in a hypergeometric experiment; G: genomics; E: epigenomics; T: transcriptomics; **MAJOR; *MINOR; 0: no evidence; SFARI score: "minimal evidence" (4), "hypothesized but untested" (5); Other modules: reference of gene-networks studies of ASDs in which the gene is mentioned. The total number of genes considered is equal to the interactome size: 12 739 genes.

Collectively, it has been observed that a significant number of "CORE+13" genes emerged as associated with ASDs at epigenomics level ($p = 2.63 \cdot 10^{-4}$, hypergeometic test; Table 4.3) and at transcriptomics level ($p = 1.22 \cdot 10^{-3}$; hypergeometic test; Table 4.3). The observation that different types of alterations refer to the same genes further stresses the role of these genes in

A	**B**	$\|\mathbf{A}\|$	$\|\mathbf{B}\|$	$\|\mathbf{U}\|$	$\|\mathbf{A} \cap \mathbf{B}\|$	$\langle\|\mathbf{A} \cap \mathbf{B}\|\rangle$	$P(\mathbf{x} \geq \|\mathbf{A} \cap \mathbf{B}\|)$
CORE+13	CORE+13(E)	347	1227	12739	54	3.27	$2.63 \cdot 10^{-4}$
CORE+13	CORE+13(T)	347	2387	12739	88	6.37	$1.22 \cdot 10^{-3}$
G	E	1133	1227	12739	146	109	$1.09 \cdot 10^{-4}$
G	T	1133	2387	12739	235	212	$3.95 \cdot 10^{-2}$
E	T	1227	2387	12739	243	230	$1.66 \cdot 10^{-1}$
G**	E**	334	272	12739	15	7.13	$5.47 \cdot 10^{-3}$
G**	T**	334	256	12739	15	6.71	$3.12 \cdot 10^{-3}$
E**	T**	272	256	12739	5	5.47	$6.42 \cdot 10^{-1}$

Table 4.3: Overlaps among the lists of genes associated with ASDs. G: genomic; E: epigenomics; T: transcriptomics; **MAJOR; *MINOR; CORE+13(E) and CORE+13(T) indicate genes belonging to the "CORE+13" set and which are supported by E and T respectively.

ASDs. These results are in line with those of previous studies that suggest a potential role of genetic factors in contributing to DNA methylation differences in ASDs [131]. Moreover, blood-derived epigenetic changes observed in genes whose sequence variations are associated with ASDs are more likely to have a common function across tissues, compared to those not related to genetic changes [136].

As for the 13 predicted genes (Table 4.2) that closely interact with the CORE genes of ASDs, they mainly belong to different neuronal pathways and are especially involved in synaptic function and plasticity that, if impaired, could actively contribute to the pathogenesis of ASDs and/or to their comorbidities. Genes encoding for ion channel were found among these genes, and the role of various ion channel gene defects (channelopathies) is known in the pathogenesis of ASDs. For instance, *HCN2* and *HCN4* belong to hyperpolarization-activated cyclic nucleotide-gated (HCN) channels family, encoding for non-selective voltage-gated cation channels, and they are strongly expressed in the brain. These channels establish the slow native pacemaker currents contributing to membrane resting potentials, input resistance, dendritic integration, synaptic transmission and neuronal excitability. Interestingly, it seems that *SHANK3*, strongly linked to ASDs, works in organization of HCN-channels [148] and that its expression negatively influences those of *HCN2* [149], so variations in *SHANK3* gene are reflected in pacemaker current abnormalities. In addition, variants in

HCN1, another member of the HCN family, were detected in patients with epileptic encephalopathy and clinical features of Dravet syndrome, intellectual disability, and autistic features [150].

Some of the predicted genes, such as *EPB41* and *EPB41L1*, take part in cytoskeleton and synaptic structures. *EPB41* is the founding member of the large family of proteins that associate with membrane proteins and cytoskeleton and in neurons is involved in protein-protein interactions at synaptic level. It interacts with NRXN1 and NRXN2, as well as NLGN1, -2, -3 and -4X. These proteins act at presynaptic and post synaptic level and causative variations in *NRXN1* and *-2* [151, 152] and *NLGN2* (also in CORE+13 genes), *-3* and *-4X* [153, 154] have already been described in ASDs. Furthermore, EPB41L1 (highly expressed in the brain) and the ionotropic glutamate receptor GRIA1, were listed in the 13 predicted and in CORE genes, respectively, interact thus contributing to glutamate neurotransmission. An alteration of glutamate neurotransmission was found in ASDs. Interestingly, EPB41L1 is associated with mental retardation, deafness autosomal dominant 11 and dutosomal dominant Non-syndromic intellectual disability.

Then again, DLGAP2 is a member of the postsynaptic density proteins (as SHANK3), probably involved in molecular organization of synapses and signalling in neuronal cells, with implications in synaptogenesis and plasticity. In particular, DLGAP2 could be an adapter protein linking the ion channel to the sub-synaptic cytoskeleton. Animal models demonstrated that *DLGAP2* has key role in social behaviors and synaptic functions [155]. Case studies also report rare *DLGAP2* duplications in ASDs [156–158]. Then again, *DLGAP2* gene has an important paralog, *DLGAP1*, already associated with ASDs. DLGAP1 proteins interact with other ASDs-associated proteins such as DLG1, DLG4, SHANK1, SHANK2 and SHANK3 [144]. Moreover, the analysis of rare copy number variants in ASDs found numerous de novo and inherited events in many novel ASDs genes including DLGAP2 [146].

Among the 13 predicted genes, syntaxin-1A (*STX1A*) is also involved in synaptic signaling. This gene encodes for part of complex of proteins mediating fusion of synaptic vesicles with the presynaptic plasma membrane.

A dysregulation of *STX1A* expression [159–161] has been reported in high functioning autism and Asperger Syndrome. A significant association between three *STX1A* SNPs and Asperger syndrome was recently described. These SNPs could alter transcription factor binding sites both directly and through other variants in linkage disequilibrium [162].

The list of predicted genes includes *GABRA5*. It transcribes for the subunit 5 of GABA receptor alpha whose reduced expression and reduced protein level have been described in autism [163], and SNPs of this gene are biomarkers of symptoms and developmental deficit in Han Chinese with autism [164]. The inclusion of this gene in the CORE list strengthens the evidences of imbalance between excitatory and inhibitory neurotransmission in ASDs and abnormalities in glutamate and GABA signaling as possible causative pathological mechanisms of ASDs.

Few of these predicted genes encode for proteins involved non-neuronal specific signalling pathways, which are also important for ASDs: *PRKCA*, *WDR37* and *UBC*. PRKCA regulates many signalling pathways such as cell proliferation, apoptosis, differentiation, tumorigenesis, angiogenesis, platelet function and inflammation. A meta-analysis performed on de novo mutation data of 10,927 individuals with neurodevelopmental disorders found an excess of missense variants in *PRKCA* gene [165]. The *WDR37* gene encodes a member of a protein family that is involved in many cellular processes such as cell cycle progression, signal transduction, apoptosis, and gene regulation. WDR37 is a nuclear protein ubiquitous expressed and particularly abundant in cerebellum and whole brain. There are no direct evidences for ASDs development and WDR37 - however recently, it has been demonstrated that WDR47 shares functional characteristics with PAFAH1B1, which causes lissencephaly. PAFAH1B1 also constitutes a key protein-network interaction node with high-risk ASDs genes expressed in the synapse that can impact synaptogenesis and social behaviour [166].

The analysis confirms the importance of X-linked gene in the aetiopathogenesis of ASDs. Mutations of *CACNA1F* (located at Xp11.23), mainly cause X-linked eye disorders. Since the role of various ion channel gene defects (channelopathies) in the pathogenesis of ASDs is becoming evident, the deep resequencing of these functional genomic regions has been

performed. These studies revealed potentially causative rare variants contributing to ASDs in *CACNA1F*. Then again, *CACNA1D*, an important paralog of *CACNA1F*, displayed de novo missense variants in ASDs probands from the Simons Simplex Collection [167]. Moreover, being the gene X-linked, could contribute to the sex bias of ASDs.

4.4 Multi-omics Analysis

We assessed the significance of the overalps among the lists of genes associated with ASDs by genomics, epigenomics, and trascriptomics evidences. We observed significant overalps between the list of genes from genomics and those supported by epigenomics or transcriptomics (Table 4.3). The intersection among the three gene lists consists of 40 genes, 34 of which are included in the considered interactome (shortly "SHARED") (Figure 4.3).

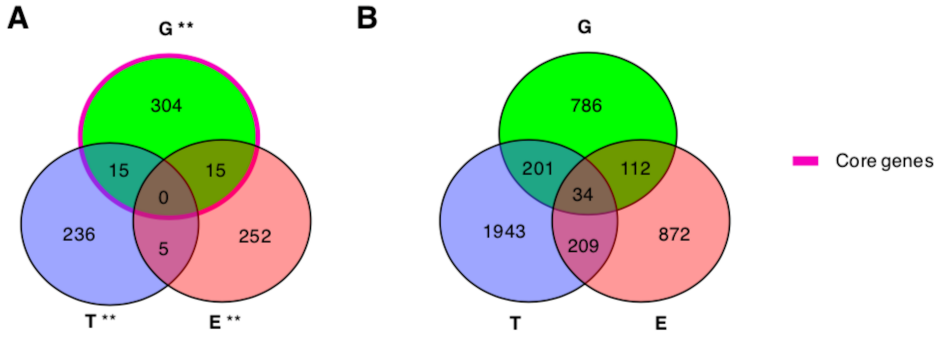


Figure 4.3: Overlaps among genes associated with ASDs by genomics, epigenomics and transcriptomics. G: genomics; E: epigenomics; T: transcriptomics. **MAJOR

Out of the SHARED genes, 26 do not interact directly with any other SHARED gene, while 8 genes form three connected components composed of: *TRAPPC6B*, *DYNC1H1*, *TRAPPC9* and *CSNK1D*; *GNAS* and *PRKCA*; *EP400* and *TRRAP* (Figure 4.4).

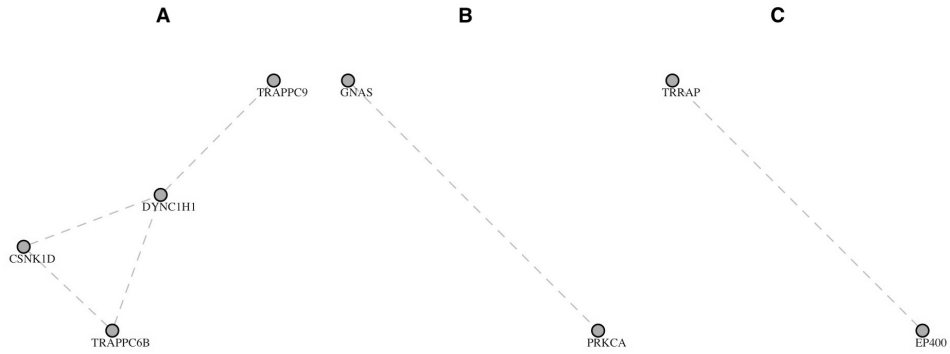


Figure 4.4: Interactions among the SHARED genes. Only 8 of the 34 SHARED genes interact directly with at least another SHARED gene.

In order to find modules of functionally related genes supported by one or more types of evidences ("layers" from now on), mND algorithm (paragraph 3.2) was used to obtain the final integrative score that summarizes the relevance of each gene in relation to its location in the interactome and its network proximity to other genes associated with ASDs in one or more layers (genomics, epigenomics and transcriptomics).

Therefore, the genes-by-layer input matrix $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$ was defined, where each element x_{ij} was set to: 1 if the gene i was member of a "-MAJOR" group in layer j (paragraphes 4.2.2 - 4.2.4); 0.5 if the gene i was member of a "-MINOR" group in layer j (paragraphes 4.2.2-4.2.4); and 0 if the gene was i was not associated with ASDs in layer j. ND was applied to $\boldsymbol{X}_0$ using the genome-wide interactome represented by the symmetric normalized adjacency matrix $\boldsymbol{W}$ (Equation 3.1) and α parameter was set to 0.7, after that the mND score was calculated with the parameter k set equal to 3 (see paragraph 3.5.2); therefore, for each gene i, the top 3 direct neighbours of i with the highest diffusion scores in each layer were considered. Statistical significance of gene scores was assessed by empirical p values, calculated using 1000 permutations of the input matrix $\boldsymbol{X}_0$.

At the top of the resulting genome-wide ranking, genes with significant scores (Figure 4.5 A) were found.

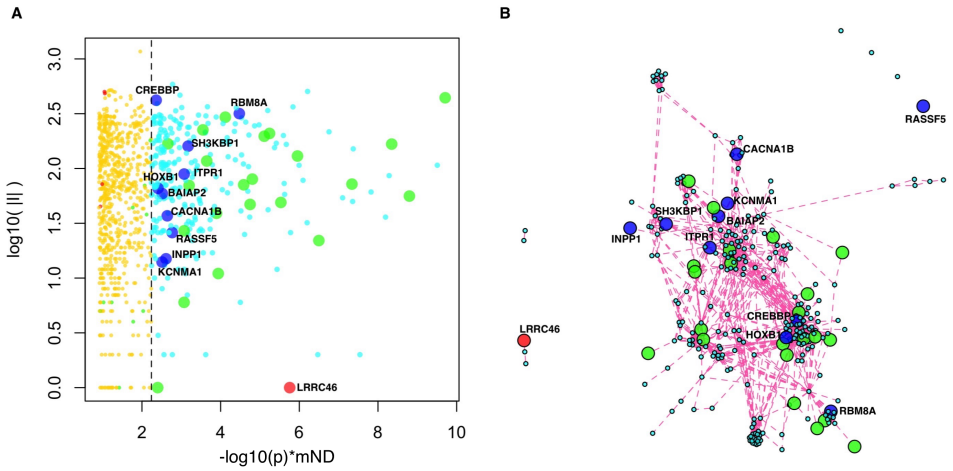


Figure 4.5: Integrative multi-omics analysis. (**A**) Global network diffusion scores (horizontal axis) and number of interactions (vertical axis) of the top ranking genes; the vertical dashed line separates the top 275 genes belonging to the INT-MODULE (higher scores, on the right) from the other genes (lower scores, on the left). (**B**) Network of the top 275 genes (INT-MODULE). Green circles: SHARED genes; blue circles: genes included in SFARI categories 4 and 5; red circle: *LRRC46*.

To assess whether these highly ranked genes formed significantly connected gene modules, network resampling [42] was applied. We found a multi-omics integrative gene module (INT-MODULE) involving a total of 275 genes (Figure 4.6) strongly supported by genomics, epigenomics and transcriptomics.

The largest connected component (266 genes) of INT-MODULE connects 22 SHARED genes, which do not establish direct interactions with each other if considered in isolation (Figure 4.5 B).

The INT-MODULE was compared with gene networks proposed by other studies on ASDs different in terms of input data and analysis approach [144–147]. 157 genes occurred in at least one of such networks and a total of the 144 INT-MODULE genes belong to the highest scoring SFARI categories. Furthermore, ten other genes of the network are currently clas-

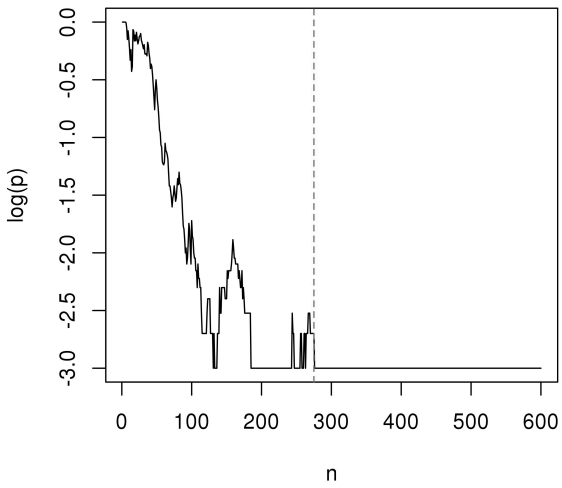


Figure 4.6: Network resampling on the genes top ranked by the multi-omics analysis. Logarithm of p-value (y-axis) calculated for each rank of the gene list (x-axis) ordered by decreasing values of the global diffusion score.

sified in SFARI as "minimal evidence" and "hypothesized but untested" (Table 4.4, Figure 4.7) and are also supported by epigenomics and/or transcriptomics; and 7 of these 10 genes were reported by other network-based analyses (Table 4.4). The INT-MODULE also includes *LRRC46* (leucine rich repeat containing 46), the only gene of the module that does not occur in any of the input gene lists (Figure 4.5, 4.7).

To functionally characterize the INT-MODULE, we partitioned its largest connected component (266 genes) in topological clusters and assessed both the enrichment of each cluster in terms of molecular pathways and the types of evidences associated with each cluster.

Topological community identification was performed using methods based on different rationales such as modularity/energy function optimization, edge removal, label propagation, leading eigenvector and random walks. Modularity was quantified using Newman definition [169], with functions implemented in igraph R package [170]. Several community detection strategies were explored and the highest modularity was found with a partition

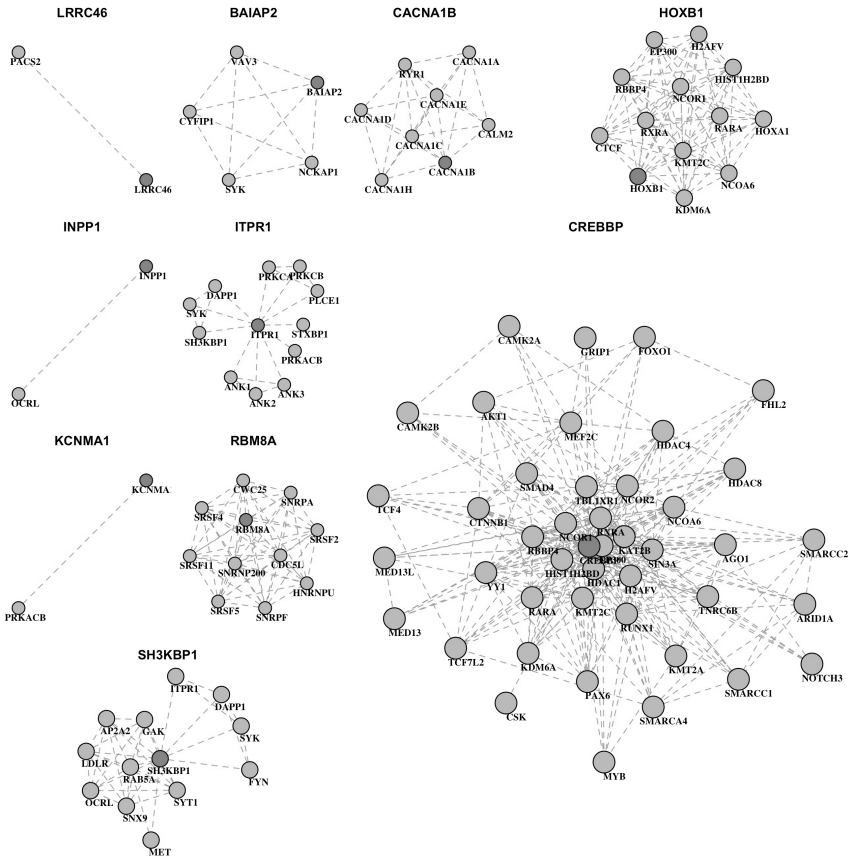


Figure 4.7: Interactors of some genes belonging to the INT-MODULE. Only the interactors that are member of the INT-MODULE are shown, while other interactors are not reported. See main text.

of 12 clusters (Figure 4.8, Figure 4.9).

Pathway analysis was carried out using gene-pathway associations from Biosystems [171] and MSigDB Canonical Pathways [172]. Each pathway was assessed for over-representation of genes from each cluster using the hypergeometric test (R functions "phyper" and "dhyper"). Nominal p values were corrected for multiple testing using Bonferroni-Hochberg method

Symbol	Description	#Im	G	E	T	SFARI score	Other modules
BAIAP2	BAI1-associated protein 2	4	*	*	0	5	[145, 168]
CACNA1B	calcium voltage-gated channel subunit alpha1B	7	0	**	0	4	[145, 168]
CREBBP	CREB binding protein	43	0	0	**	5	[144, 145, 147]
HOXB1	homeobox B1	12	0	*	*	5	[144]
I*NPP1*	inositol polyphosphate-1 -phosphatase	1	0	**	0	4	[144, 145]
ITPR1	inositol 1,4,5-trisphosphate receptor type 1	11	*	*	0	4	[145, 168]
KCNMA1	potassium large conductance calcium-activated channel, subfamily M, alpha member 1	1	0	**	0	4	[147]
RASSF5	Ras association domain family member 5	0	0	**	**	4	-
RBM8A	RNA binding motif protein 8A	10	0	**	*	5	-
SH3KBP1	SH3-domain kinase binding protein 1	12	*	0	**	5	-

Table 4.4: INT-MODULE genes SFARI. G: genomics; E: epigenomics; T: transcriptomics. **MAJOR; *MINOR. Im: number of interactors within the INT-MODULE; SFARI score: "minimal evidence" (4), "hypothesized but untested" (5). Other modules: reference of gene-networks studies that also associated the gene to ASDs

(R function "p.adjust"), obtaining q values.
The enrichment of each cluster in terms of a type A of evidence (e.g. genomics) was quantified as the ratio between the fraction of genes supported by A in the cluster and the fraction of genes supported by A in the INT-MODULE.

Therefore, the largest connected component of the network was partitioned in 12 subgroups or topological sub-modules (Figure 4.9).

The two largest clusters are composed of 61 (cluster #8) and 53 (cluster #3) genes, and are characterised by a similar proportion of supporting evidences (Figure 4.9 B). These two central clusters contain genes that are part of the same pathways, such as Wnt signalling pathway (#8: $q = 1.83 \cdot 10^{-2}$;

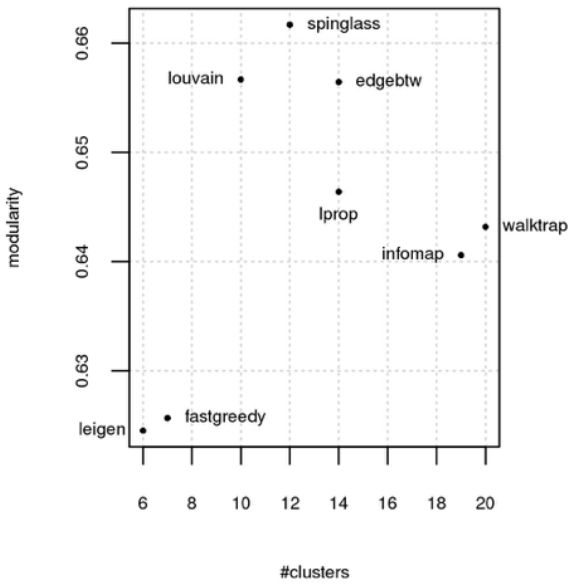


Figure 4.8: Modularity of different partitions of the INT-MODULE. Modularity and number of clusters obtained partitioning the INT-MODULE by means of 8 community detection algorithms.

#3: $q = 1.79 \cdot 10^{-5}$) and IL-7 Signal Transduction (#8: $q = 4.54 \cdot 10^{-2}$; #3: $q = 6.02 \cdot 10^{-4}$), but are also marked by specific pathways. In particular, among the pathways specifically enriched in cluster #8 and #3, Chromatin organization ($q = 1.27 \cdot 10^{-27}$) and Signaling by VEGF ($q = 5.41 \cdot 10^{-23}$) were found respectively. Cluster #7 (41 genes) is the most enriched in differentially expressed genes and significantly associated with pathways involved in cell cycle processes. Cluster #5 is mainly enriched in genes associated with epigenetic and transcriptional changes, and marked by mRNA splicing ($q = 1.92 \cdot 10^{-12}$). Cluster #6 is particularly enriched in genes with epigenetic changes and associated with Extracellular matrix organization ($q = 6.10 \cdot 10^{-5}$). Cluster #2 (9 genes) is supported at genomics and epigenomics levels and enriched in genes of Calcium signalling path-

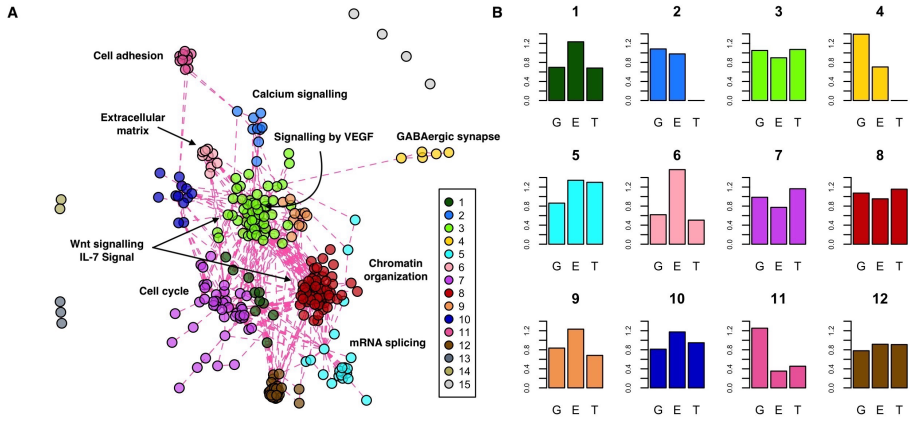


Figure 4.9: Functional characterization of the INT-MODULE. (**A**) Topological clusters; #1-12: Clusters of the largest connected component; #13,14: Two clusters of three and two genes, respectively; #15: The remaining four genes. (**B**) Enrichment (vertical axis) of each cluster in terms of genes supported by genomics (G), epigenomics (E), and transcriptomics (T): A value of 1 indicates the same proportion within the cluster and in the whole INT-MODULE.

way ($q = 4.88 \cdot 10^{-12}$). Lastly, clusters #11 and #4 are composed of genes associated with ASDs mainly at genetic level, which, respectively, control GABAergic synapse (#4: $q = 3.90 \cdot 10^{-6}$) and encode for Cell adhesion molecules (#11: $q = 1.58 \cdot 10^{-5}$) active in the neuronal system.

This analysis suggests a different role of the sub-modules by function and by association with one or more types of alterations. For example, cluster #3, equally supported by all three types of evidence, includes genes that belong to inflammatory mediator regulation of transient receptor potential (TRP) channels. Inflammation and immune system dysfunctions are in comorbidity with ASDs, and TRP canonical channel 6 (TRPC6) is emerging as a functional element for the control of calcium currents in immune-committed cells and target tissues, influencing leukocytes tasks. Interestingly, the TRPC6 is also involved in neuronal development and variants in *TRPC6* gene (within CORE gene) were found in patients with ASDs. Moreover, MeCP2, a transcriptional regulator whose mutations cause Rett

syndrome, was found abundant in TRPC6 promoter region resulting a transcriptional regulator of this gene [173] TRPC6, in turn, activate neuronal pathways including BDNF, CAMKIV, Akt and CREB, also involved in ASDs [174].

It is possible to conclude that the integrative analysis of the large number of genes reported by the studies on ASDs allowed the prioritization of a series of genes interconnected by functional relations and associated with one or more types of molecular alteration, which might unearth common targets that are amenable to therapy [127].

Chapter 5

Integrative analysis of somatic mutations in breast cancer initiating cells.

Cancer is a heterogeneous and complex disease, characterized by genetic and phenotypic differences within each individual tumor (intratumor heterogeneity) and among patients (intertumor heterogeneity). Recently, network-based analyses have been successfully applied in cancer research, allowing to highlight conserved patterns (gene networks and pathways) among the heterogeneous molecular alterations that play an essential role in tumorigenesis and tumor progression. This chapter presents an integrative network-based analysis, based on mND algorithm, of mutation profiles observed in breast cancer initiating cells derived from the primary tumors of 11 subjects, which led to the identification of gene networks enriched in mutations, containing potentially valuable targets for breast cancer treatment.[1]

[1]Some contents of this chapter are part of a manuscript in preparation by the following authors (in alphabetical order): *V. Angeloni, V. Appierto, M.G. Daidone, C. De*

5.1 Network-based analysis to explain the genetic heterogeneity of breast cancer initiating cells.

Understanding the complexity of cancer is an ongoing challenge due to the genetic heterogeneity both between and within tumours that could have a negative impact on the response to anticancer therapies and the clinical outcomes [175–177]. Moreover, there is emerging evidence that an additional source of heterogeneity comes from a small population of tumor cells displaying self-renewal and tumor initiation power, also known as tumor-initiating cells, that played a key role in the tumorigenic, growth and spread of the tumor [176–183]. Efforts are required to identify cancer-relevant genes and related pathways from somatic mutation profiles that are exceptionally sparse [184]. Indeed, analysis of data generated from The Cancer Genome Atlas [99] and the International Cancer Genome Consortium [185] have shown that only a few well-studied driver genes are frequently mutated among subjects ($f > 10\%$), in contrast to a "long-tail" of many genes mutated that are found mutated at low frequency ($f < 5\%$) [44, 186].

In this context, the INTEROMICS BCIC-IMA project planned an integrative mutational analysis of patient-derived breast cancer initiating cells (BCICs) to disentangle tumor genetic complexity and identify actionable targets for precision medicine. In particular, primary tumors (PTs), PTs-derived BCICs and control samples (blood) from 11 subjects underwent whole-exome sequencing (WES) (Figure 5.1 A).

Considering all these aspects, mND algorithm was applied to identify the networks and pathways enriched in mutations in BCIC. Firstly, WES data were processed with a specific in-house bioinformatics pipeline based on multiple mutation callers (Figure 5.1 B). Subsequently, all mutation sites obtained by five variant callers were integrated and analyzed by "isma", an R package developed by us [23] within this research activity. Lastly, mutation profiles observed in BCICs were integrated using mND algorithm

Marco, N. Di Nanni, L. Milanesi, E. Mosca; "Integrative mutational analysis of patient-derived breast cancer initiating cells to disentangle tumor genetic complexity and identify actionable targets for precision medicine."

as described in *Chapter 3* (Figure 5.1 C). The integrative analysis identified gene networks significantly enriched in BCICs genes and significant pathways (Figure 5.1 D) that could be targeted by drugs (Figure 5.1 E).

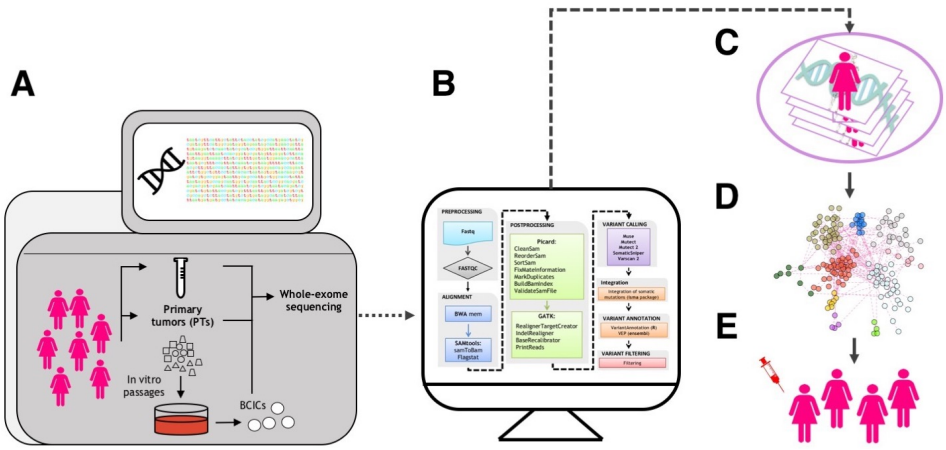


Figure 5.1: Overview of integrative analysis of somatic mutations in breast cancer initiating cells. (**A**) BCICs were isolated from PTs as non-adherent mammospheres and propagated at early in vitro passages, WES was performed; (**B**) WES data were processed with a specific in-house bioinformatics pipeline; (**C**) mutation profiles were integrated by mND; (**D**) Functional characterization of gene networks enriched in mutations; (**E**) Identification of actionable targets.

5.2 Mutation profiles of BCICs

5.2.1 Whole exome sequencing

Primary BC tumors were collected from BC consented patients that underwent surgical resection at Fondazione IRCCS Istituto Nazionale dei Tumori [2]. In particular, fresh frozen primary breast tumors character-

[2] All patients participating in the study signed an informed consent according to the Declaration of Helsinki. The study was approved by the Ethical Review Board of Fon-

ized with respect to the intrinsic molecular subtypes with corresponding aliquots of non-tumoral specimens (e.g., buffy coats) were collected from 11 patients (Figure 5.1 A). BCICs were isolated from primary tumors as non-adherent mammospheres and propagated at early in vitro passages (from p3 to p10). WES was performed by the DNA sequencing services of Fasteris (Swiss) and Genomix4Life s.r.l (Salerno, Italy), using the Illumina technology and Agilent SureSelect XT kit on paired BCICs, PTs and normal samples (blood-derived buffy coats).

5.2.2 Sequencing data analysis and variant detection

WES data were processed with a specific in-house bioinformatics pipeline based on multiple mutation callers (Figure 5.2), which consisted of mainly six phases: pre-processing of raw data, alignment of paired-end reads, post-processing analysis, variant calling, variant annotation and filtering.

In particular, sequencing files were quality controlled with FastQC, then paired-end reads were aligned to hg19 human genome [187] using BWA-mem tool [188]. Alignment files require postprocessing steps (e.g. sorting of the reads according their genomic location, marking of removing duplicate reads, indexing of the bam file, local realignment of reads around candidate indels [189]) that were processed by two tools developed at the Broad Institute: Picard [190] and GATK [191].

Since it has been shown that calling mutations using different pipelines results in a low consensus [192–196], somatic mutation call from WES data were detected using five bioinformatics pipelines for matched tumor-normal samples: GATK Mutect version 1 and 2 [102], Varscan2 [104], SomaticSniper [103] and Muse [101]. The considered variant callers are based on different models: Mutect version 1, Muse and SomaticSniper detect only single nucleotide variations, while Mutect version 2 and Varscan2 call both single nucleotide variations (SNPs) and insertions/deletions (INDELs). Therefore, all mutation site obtained by the variant callers were integrated by "isma", an R package developed by us [25] (see Appendix B), that allows to

dazione IRCCS Istituto Nazionale dei Tumori of Milan.

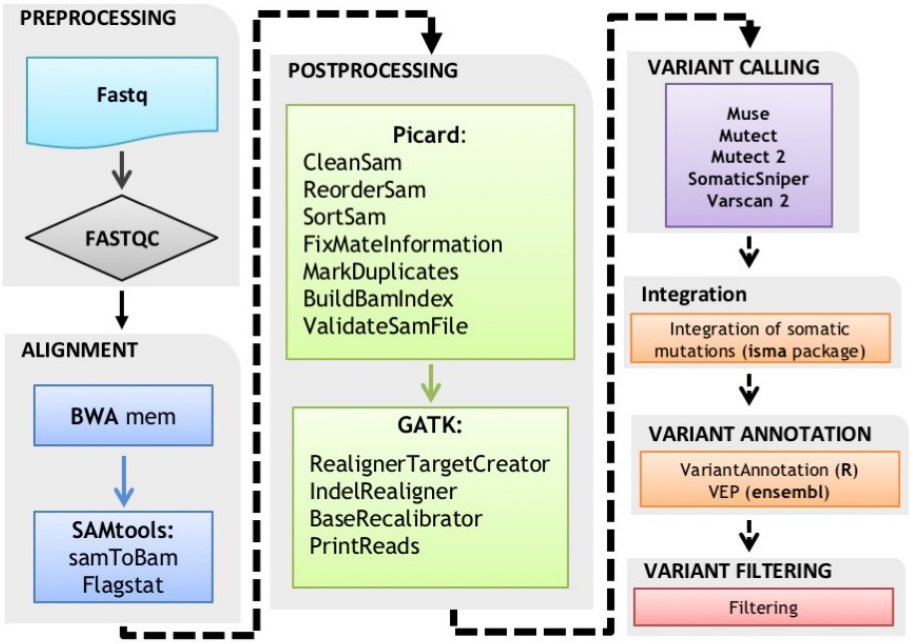


Figure 5.2: Bioinformatics pipeline The pipeline involving mainly six steps: pre-processing of raw data, alignment, post-processing, variant discovery, annotation and filtering of variants.

integrate and analyze the information obtained by multiple variant callers. After integration, mutation sites were annotated by the Variant Ensembl Predictor tool [197], which besides the gene annotation provides the consequence of variants on the protein sequence.

Lastly, we applied filtering strategies to reduce the number of mutation sites likely to be false positives. In particular, the following criteria was applied:

- minimum number of reads supporting variant call in PTs/BCICs samples: 10;

- minimum number of reads supporting variant call in normal samples: 10;

- minimum number of reads supporting alternative allele in PTs/BCICs samples: 3;

- variant allele frequency in normal samples less than 0.01;

- the 12 most severe consequences as estimated by Ensembl [198] (Figure 5.3, red boxes), that is those with "HIGH" and "MODERATE" impact (with the exception of "regulatory_region_ablation" , classified as "MODERATE" but ranked 30-th [198] out of 35 in decreasing order of severity).

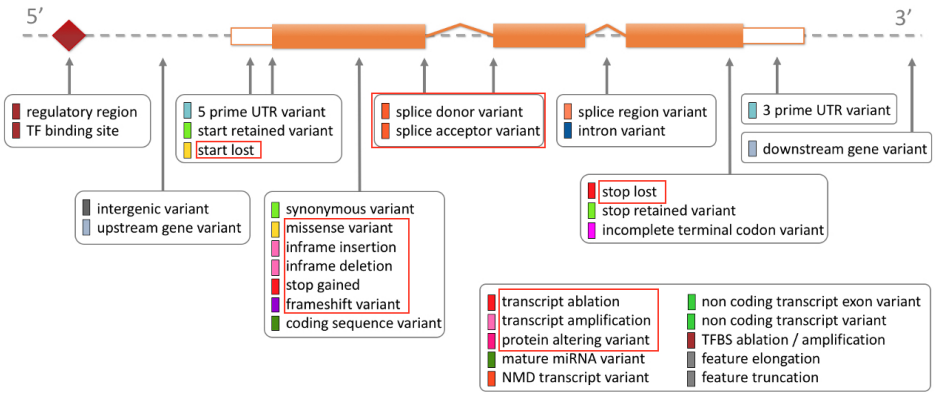


Figure 5.3: Ensembl Variation - Variant consequences The diagram shows the location of each display term relative to the transcript structure [198]. The red boxes indicate the selected consequences type calls as filtering criteria.

5.3 Integration of somatic mutations detected by multiple variant callers with "isma"

Since recent studies have recommended to analyze the same Next Generation Sequencing data using multiple callers, the lists of mutations encoded in Variant Call Format (VCF) [199] generated by five variant callers (Mutect version 1 and 2 [102], Varscan2 [104], Muse [101] and SomaticSniper [103]) were integrated by the isma R package [25].

To integrate VCF files, the configuration file (*Appendix B*, paragraph B.2) was generated with the following fields:

- the file name and path of files;
- mutation caller identifier (muse, mutect1, mutect2, somsnip, varscan2);
- subject identifier from which both normal and tumor samples derive from (e.g. 1_PT, 1_BCIC);
- a variable that defines groups of samples (BCIC, PT);
- tumor sample name reported in the VCF file (e.g. "TUMOR", ID of tumor sample),
- normal sample name reported in the VCF file (e.g. "NORMAL", ID of normal sample).

The function "pre_process" read the input VCF files reported in the configuration file and generated a single list of mutations sites, in which variants were merged from different callers, experimental evidences from different VCF files were harmonized, unique site identifiers were generated and some new fields were added (e.g. the total number of reads supporting a site in each sample, variant allele frequency, INDEL size and mutation type). The isma package was also used to quantify the consensus among variant callers, investigate the possible presence of outliers and common patterns, integrate information of sites and genes already catalogued by TCGA studies. In particular, the function site_analysis was used to obtain

the statistics at sites level, like the overlap among subjects and callers, the detection of outlier subjects and tools, the number of sites detected by each tool and the occurrence of mutation sites across callers and subject, calculating a consensus measure among callers in each subject.

The overall consensus plot (Figure 5.4 A) reported that mutect2 found the highest number of sites, the 70% of which was not reported by other callers; furthermore, the site co-occurrence matrices (Figure 5.4 B) underlined how mutect2 detected up to 4 times more mutation sites than other tools, while somatic sniper shared the majority of its calls with other tools.

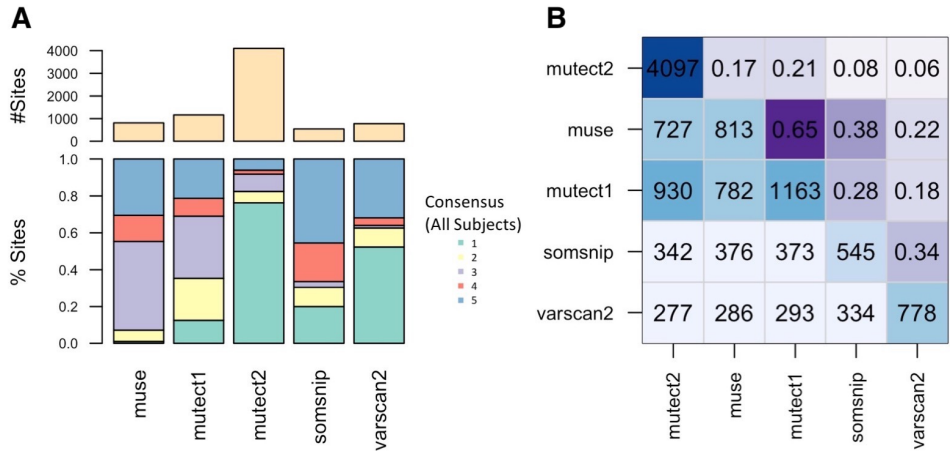


Figure 5.4: Consensus among somatic mutation callers. (**A**) Overall consensus among pipelines; (**B**) Site co-occurrence plot: site co-occurrence among mutation callers (below diagonal) and corresponding similarity between callers (Jaccard index, above diagonal).

The detailed consensus plot per subject showed the total amount of sites, consensus on sites for each subject and subject recognized as outlier according to site number (defined by the inter-quartile range, Tukey's fences), imbalance in site number across tools, imbalance in consensus among tool and tool consensus score (Equation B.1). The plot indicated the absence of hypermutated subjects (Figure 5.5 A) and that the amount of mutation sites varied from patient to patient. Several subjects displayed an imbalance of

calls among the pipelines (Figure 5.5 B) and a few tools were recognized as outlier in four subjects (9_BCIC, 10_BCIC, 11_BCIC, 8_PT). Furthermore, there were subjects with a relevant (e.g. 5_BCIC, 5_PT) or poor (e.g. 7_BCIC, 8_BCIC) consensus among tools (Figure 5.5 C), summarized by CS score (5.5 D).

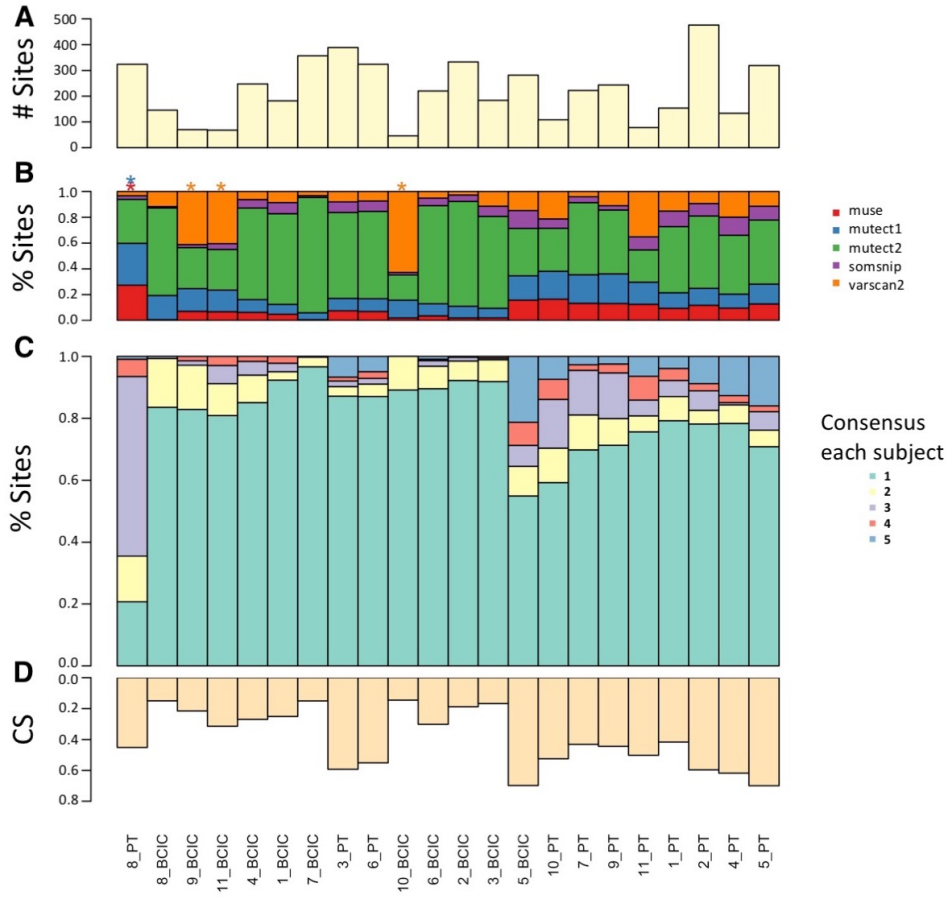


Figure 5.5: Detailed consensus plot. (**A**) Number of mutation sites; (**B**) Fraction of sites called by different pipelines; (**C**) Tool Consensus across subjects: (**D**) Consensus score (CS). (**A-D**) Asterisks indicate outliers.

Lastly, the function gene_analysis was used to perform analysis at gene level, providing information like gene mutation frequency across subjects and its dispersion estimated considering multiple callers. Gene co-occurence among subjects underlined similarity between mutation profiles (Figure 5.6 A), for example 6_PT shares more mutated genes with 2_PT subject than 7_PT; the variability of such co-occurrences due to the use of different callers is quantified as the coefficient of variation (above diagonal). The analysis of gene mutation frequency (f, defined as the fraction of subjects with at least one mutations) highlighted genes with a more or less variable frequency (f) and dispersion across callers (Figure 5.6 B), for example the genes in the red box shown a particularly higher variation of f across tools (higher than 0.20).

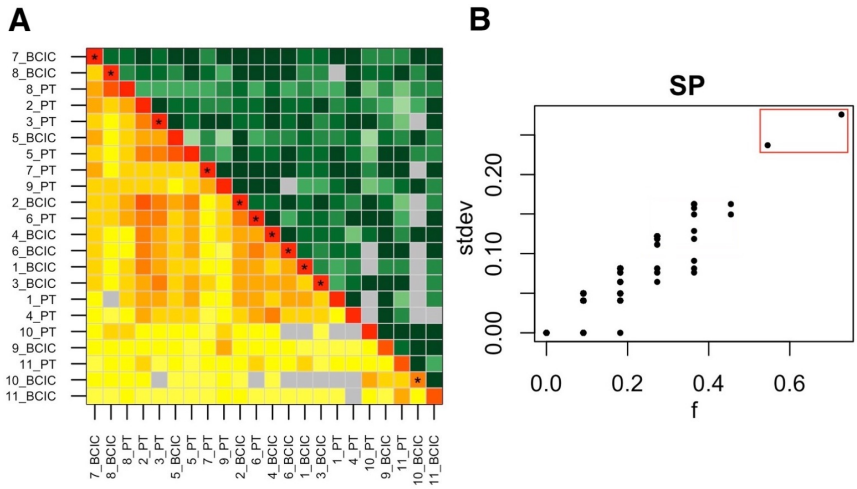


Figure 5.6: Co-occurrence of mutation genes across subjects and gene mutation frequency variability. (**A**) Total number of mutated genes (diagonal), mutation co-occurence across subjctes (below diagonal) and corresponding coefficients of variations across pipelines (above diagonal). Asterisks indicate a coefficients of variations greater than 1. (**B**) Standard deviation (stdev) of gene mutation frequency (f) across pipelines, red box indicates genes with a variation of f greater than 0.20.

5.4 Networks enriched in mutation detected in BCICs

Since BCICs plays an important role in development and progression of breast cancer [181,182,200,201], we focused on genes whose mutations were detected only in BCICs samples or in both in BCICs and their paired bulk PTs with an allelic frequency in BCICs 2-fold greater than bulks ("BCICs-ENRICHED" from here on). Therefore, to find gene modules of functionally related genes supported by one or more samples, we applied mND algorithmn (*Chapter 3*, paragraph 3.2) to BCICs mutation profiles of the 11 subjects.

Molecular interactions were collected from the database assembled by Ghiassian et al. [95], for a total of 13 244 genes and 138 045 links. A genes-by-subjects input matrix $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{11}]$ was defined, where each element x_{ij} was set to 1 if the gene i was member of "BCICs-ENRICHED" group in the subject j or otherwise to 0. ND was applied to $\boldsymbol{X}_0$ using the genome-wide interactome represented by the symmetric normalized adjacency matrix $\boldsymbol{W}$ (Equation 3.1) and α parameter was set to 0.7. mND score was calculated with the parameter k set equal to 3 (see paragraph 3.5.2) and statistical significance of gene scores was assessed by empirical p-values, calculated using 1000 permutations of the input matrix $\boldsymbol{X}_0$.

At the top of the resulting genome-wide ranking, a series of genes with significant scores (Figure 5.7 A) were found. In particular, we found 559 genes with $0.01 \leq p < 0.05$ and 181 genes with $p \geq 0.01$. From a topological point of view, mND prioritized both hubs (Figure 5.7 A, red box) and genes with a lower number of connections (Figure 5.7 A, blue box). In addition, genes prioritized by mND were mutated both in more than one sample and in none of them (Figure 5.7 B).

To assess the presence of a significant gene module, the gene list was analyzed with network resampling [42] and a dense integrative module of 189 genes (BCICs-MODULE) was identified (Figures 5.7 C-D). These genes form a largest connected component of 180 elements, while among the remaining genes, six are isolated and other three form a very small module.

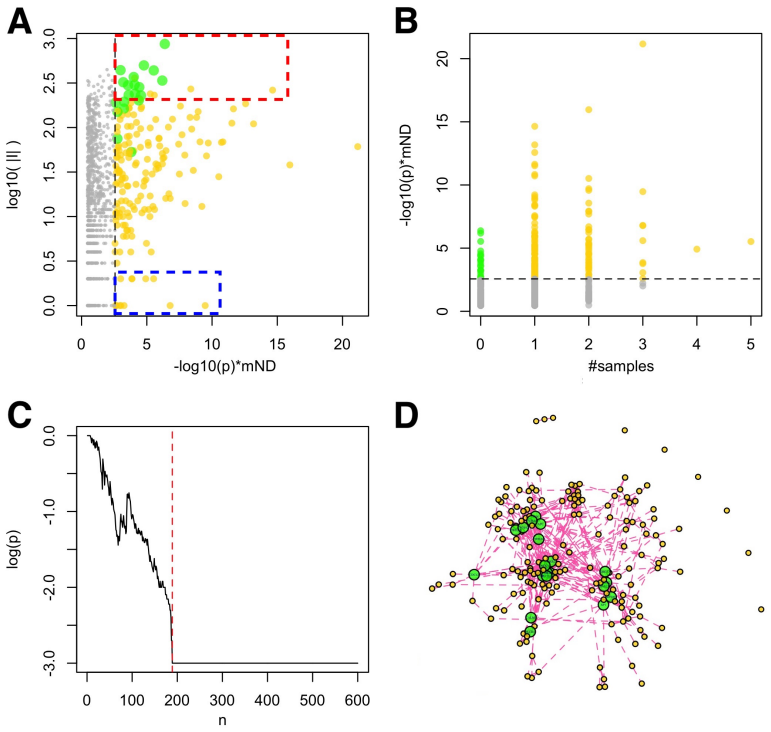


Figure 5.7: Integrative analysis of BCICs mutation profiles. (**A**) Global network diffusion scores (horizontal axis) and number of interactions (vertical axis) of the top ranking genes; (**B**) Global network diffusion score (vertical axis) and number of samples supported the gene (horizontal axis) of the top ranking genes; (**C**) Network resampling on the genes top ranked by the BCICs mutation profiles analysis; (**D**) Network of the top 189 genes (BCICs-MODULE). (**A-B**) Dashed line separates the top 189 genes belonging to the BCICs-MODULE; (**A-B, D**) Yellow circles: genes of BCICs-MODULE and mutated in at least one sample; Green circles: genes not mutated.

Interestingly, BCICs-MODULE includes several genes not mutated in any of BCICs samples, mutated in only one sample as well as genes mutated in more than one sample (Figure 5.7, green circles).

To functionally characterize the BCICs-MODULE, the largest connected component was partitioned in topological clusters and an over-representation analysiswas carried out to assess the enrichment of each cluster in terms of molecular pathways, using the hypergeometric test, gene-pathway associations from KEGG database [105], and Bonferroni-Hochberg correction of nominal p-values.

Topological community identification was performed using methods based on different rationales such as modularity/energy function optimization, edge removal, label propagation, leading eigenvector and random walks. Modularity was quantified using Newman definition [169], with functions implemented in igraph R package [170]. As in *Chapter 4*, several community detection strategies were explored and the highest modularity (Q) was found with a partition of 10 clusters ($Q = 0.5$) (Table 5.1, Figure 5.8). In particular, we found the three largest clusters composed of 43 (cluster #8), 37 (cluster #9) and 36 (cluster # 5) genes (Figure 5.8).

Algorithm	**Modularity**	n
fastgreedy	0.463	9
labprop	0.199	5
walktrap	0.434	22
eigen	0.431	8
multilev	0.503	10
infomap	0.489	19
spinglass	0.503	13
edge_betweenness	0.434	34
louvain	0.503	10

Table 5.1: Modularity of different partitions of the BCICs-MODULE. Modularity and number of clusters (n) obtained partitioning the BCICs-MODULE by means of 9 community detection algorithms.

We observed significant pathways ($q < 0.01$) in six of the ten clusters (Table 5.2). Cluster #8 is enriched in genes that are part of the Focal adhesion pathway ($q < 10^{-6}$), whose role in maintaining mammary cancer stem/progenitor cell population have been recently identified, playing a prominent role in breast cancer initiation, progression and relapse [202–204]. Cluster #9 (37 genes) contains member of Hippo signalling path-

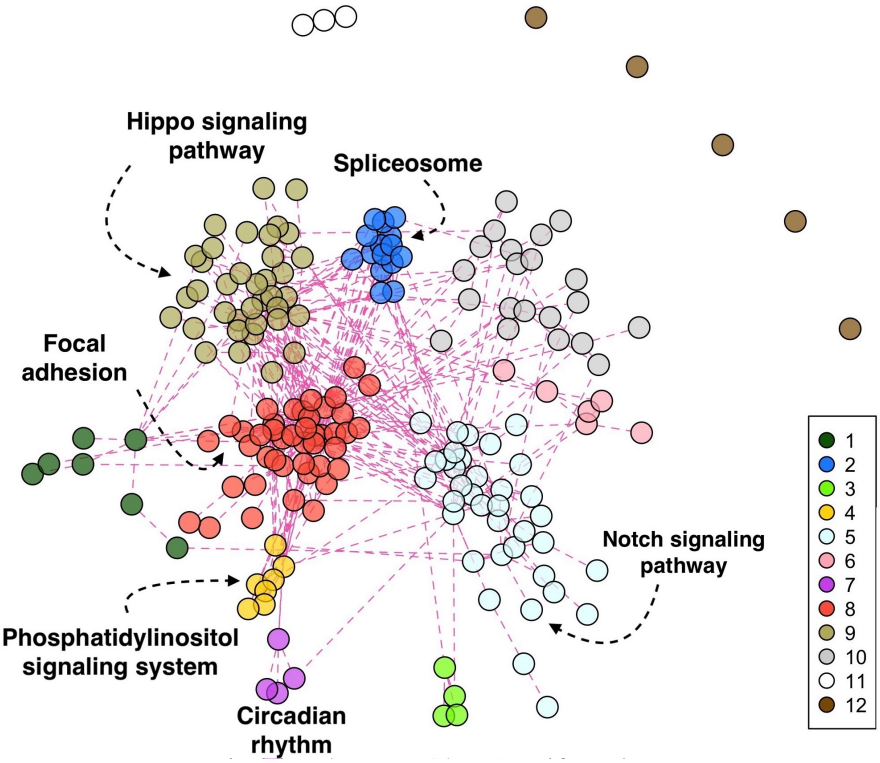


Figure 5.8: Functional characterization of the BCICs-MODULE. Topological clusters; #1-10: Clusters of the largest connected component; #11: Cluster of three genes; #12: The remaining six genes.

way ($q < 10^{-4}$), whose key components have been demonstrated to regulate breast tumor growth, metastasis, and drug resistance [205–207]. Cluster #5 (36 genes) is composed of genes associated with notch signaling pathway ($q < 10^{-2}$), which is vital to tumorigenicity of CSC and is one of the most intensively studied putative therapeutic targets in CSC [208–211]. Interestingly, genes of cluster #8 and cluster #5 are targeted by known drugs approved for breast cancer treatment [212–214] (Table 5.2): chemotherapy and targeted/biological therapies.

Cluster ID	#genes	Pathway ID	Description	q	Druggable pathway
2	15	hsa03040	Spliceosome	$1.49 \cdot 10^{-12}$	-
4	7	hsa04070	Phosphatidylinositol signaling system	$3.39 \cdot 10^{-10}$	-
5	36	hsa04330	Notch signaling pathway	$1.5 \cdot 10^{-3}$	yes
7	5	hsa04710	Circadian rhythm	$9.06 \cdot 10^{-9}$	-
8	43	hsa04510	Focal adhesion	$2.15 \cdot 10^{-7}$	yes
9	37	hsa04390	Hippo signaling pathway	$7.11 \cdot 10^{-5}$	-

Table 5.2: Significant pathways in BCICs-MODULE. Pathway ID: KEGG Identifier; Description: refers to the pathway name; q: adjusted p-values (Bonferroni-Hochberg method); Druggable pathways: "yes" if the pathway contains genes that are druggable according to the list "gene-drugs" generated by DGIdb database [215] (filters: "Anti-neoplastics" and "FDA-approved") and if the drugs have been approved for breast cancer treatment [212–214]; "-"' otherwise.

In conclusion, the integrative analysis allowed the definition of gene networks enriched in mutations supported by one or more samples and the identification of meaningful pathways that could help in the development of new potential therapeutic strategies for cancer treatment that target specifically the BCICs component of the tumor bulk.

Chapter 6

Conclusion

The availability of multiple omics data from the same biological system allows gaining insights into molecular events associated with human diseases, but, at the same time, it poses challenges in data analysis and interpretation. Network-based methods for the analysis of multi-omics leverage the complex web of macromolecular interactions occurring within cells to extract significant patterns of molecular alterations involving molecular systems. However, existing network-based approaches typically address specific combinations of omics and are limited in terms of the number of layers that can be jointly analysed.

The aim of this work consisted in the design and in the development of a network-based method that enables the integration of multi-omics data. Specifically, a new algorithm, named mND, relying on the mathematical machinery of network diffusion, was developed. mND assesses the relevance of a gene considering its own importance within each layer, its network proximity of the gene and its first neighbours to other relevant genes; the importance of a gene is represented by the combination of its mND score and its statistical significance assessed by the dataset permutations. In addition, the developed framework allows a layer-specific gene classification to suggest and clarify the genes role in relation to the datasets studied.

Beyond the novel ways of using diffusion score for gene ranking and gene classification, the methodology proposed in this research activity introduces an important advance in the class of multi-omics methods: mND can be applied to any disease for which omics data and molecular network information are available, overcoming limitations in terms of layer number and data types. The results generated by mND can be further processed with other existing tools, for example to characterize the top ranking genes using current annotations (e.g. pathways) or network theory (e.g. centrality measures).

To support the choice of the proposed method, performance of mND has been evaluated under different scenarios: considering the general problems of locating high scoring genes in network proximity across multiple layers and recovering known cancer genes in four cancer types. In both problems, mND reported better performances than existing methods. The first analysis was repeated using two different interactomes and similar results were obtained. Collectively, these results support the usefulness of mND for global ranking of genes considering multiple evidences.

Moreover, mND was applied in three different integrative problems: integration of *2*-omics dataset collected from TCGA for breast invasive carcinoma; integration of *3*-omics dataset collected in studies on ASDs; integration of a single type of omics considering a series of subjects' molecular profiles observed in breast cancer initiating cells.

In the first problem (*2*-omics), the integrative analysis with mND of mutations and differential expression data - two types of omics with relevant differences for data analysis in terms of distribution and sparsity - allowed to spot meaningful pathways underlying the disease and genes that are important in both layers or in one layer only, as well as genes with marginal signal but relevant topological role. In principle, if this is not the case, a simple solution could be to weight each layer in relation to the research questions under investigation, adding an appropriate coefficient in the two sums of Equation 3.4.

In the second problem (*3*-omics), the integrative analysis of genomics, epigenomics and transcriptomics data highlighted a series of genes associated with one or more types of molecular alteration and identified key

molecular pathways that could provide a more comprehensive view of the disease.

In the last problem (single type of omics), the integration of subjects' mutation profiles led to the identification of gene networks enriched in mutation and molecular pathways containing potentially valuable targets for breast cancer treatment.
These results further stressed the utility and ability of mND in the integrative analysis of different types of molecular data. A future working direction regards the application of mND to find active gene networks from single-cell data.

The developed method is currently applied to analyze multi-omics data in projects where our institution is involved: "Single-cell analysis of lymphocytes that infiltrate autoimmunity sites: dissecting immunological mechanisms of rheumatoid arthritis" (LYRA-2015-0010) supported by Fondazione Regionale per la Ricerca Biomedica (Regione Lombardia); "Genome, Environment, Microbiome & Metabolome in Autism: an integrated multi-omics systems biology approach to identify biomarkers for personalized treatment and primary prevention of Autism Spectr" supported by European Union's Horizon 2020 research and innovation programme; "An integrated approach to predict disease activity in the early phases of multiple sclerosis" supported by Fondazione Regionale per la Ricerca Biomedica (Regione Lombardia) and "Integration of clinical and multi-omics multiple sclerosis data into a predictive algorithm of disease activity to accelerate personalized medicine" supported by Italian Ministry of Health.

At present, mND applies to an interactome with a fixed topology and without edge directions. In future work, it would be interesting to explore the sensitivity of mND in relation to low confidence interactions. This analysis could be performed by evaluating the ability of the method to prioritize disease-genes against perturbations of the network (e.g. adding or removing interactions). Moreover, the generalization of mND pipeline to include layers with different topologies, as well as the inclusion of edge directions, are interesting opportunities for future developments. However, the latter information is currently lacking for most PPIs and would imply a significant reduction of coverage in terms of the genes studied. As all

network-based methods, the performance of mND is bounded by the reliability of current models that describe intracellular circuits. As the data about macromolecular interactions will become more and more available and reliable, network-based analyses will be less affected by the lack of a reference human interactome.

In this context, the impact of tools like mND in molecular biology will presumably increase.

Appendix A

Appendix A

A.1 Tables

	#P	Cores	L	Interactome	#V	#E	ND [hh:mm:ss]	mND [hh:mm:ss]	Total [hh:mm:ss]
BC-T1	1000	4	2	WU	6016	128 150	0:33:39	0:06:11	0:39:50
LUSC-T1	1000	4	2	WU	6016	128 150	0:33:49	0:06:18	0:40:07
THCA-T1	1000	4	2	WU	6016	128 150	0:19:57	0:06:18	0:26:15
PRAD-T1	1000	4	2	WU	6016	128 150	0:32:53	0:05:56	0:38:49
BC-T1	1000	4	2	STRING	11 796	309 850	1:35:05	0:13:35	1:48:40
BC-T2	1000	4	35	WU	6016	128 150	2:27:34	1:36:31	4:04:05
LUSC-T2	1000	4	23	WU	6016	128 150	2:45:38	1:02:34	3:48:12
THCA-T2	1000	4	17	WU	6016	128 150	1:45:34	0:45:42	2:31:16
PRAD-T2	1000	4	27	WU	6016	128150	1:48:30	1:13:49	3:02:19

Table A.1: Runtimes Total run times are split in "ND", the time required up to and including network diffusion, and "mND", the following part of the pipeline. "-T1" refers to the analysis of mutations and expression change, "-T2" refers to the analysis of mutation profiles of multiple patients; #P: number of permutations; L: number of layers; #V: number of genes; #E: number of interactions.

A.2 Figures

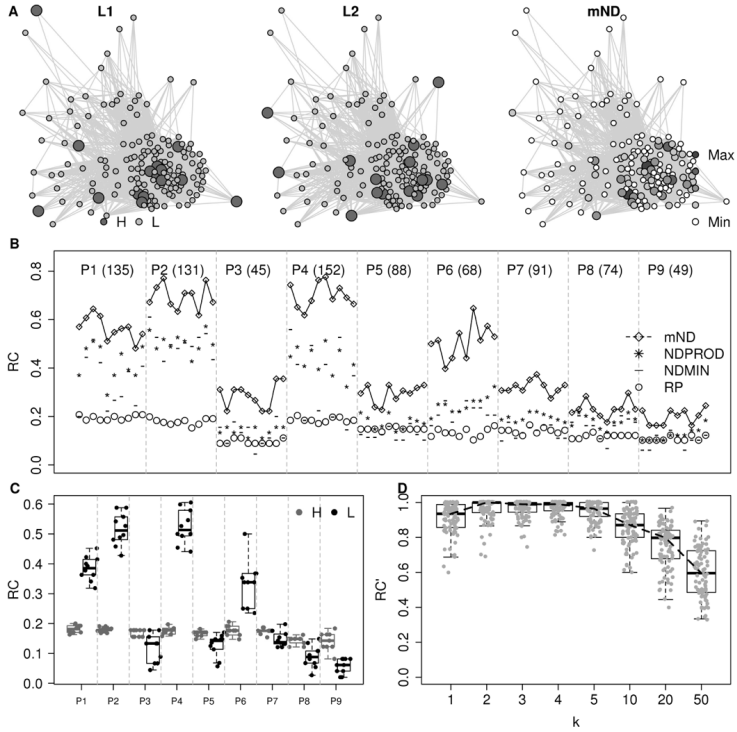


Figure A.1: Performance in ranking high scoring genes in network proximity in STRING. (**A**) Example of a gene module with its high scoring genes (H, black) in each of the two layers and the resulting mND score; only genes belonging to the module and links occurring among such genes are reported. (**B**) Recall values for 10 signal permutations for each of the 9 modules (P1, P2, ..., P9), using mND score and other methods; the number between parentheses after module id is module size. (**C**) Recall values, shown separately for high scoring genes and other genes in each module. (**D**) Recall values normalized by the highest recall found for each input configuration at varying number of neighbors (k). (**A-D**) These results were obtained using interactome STRING.

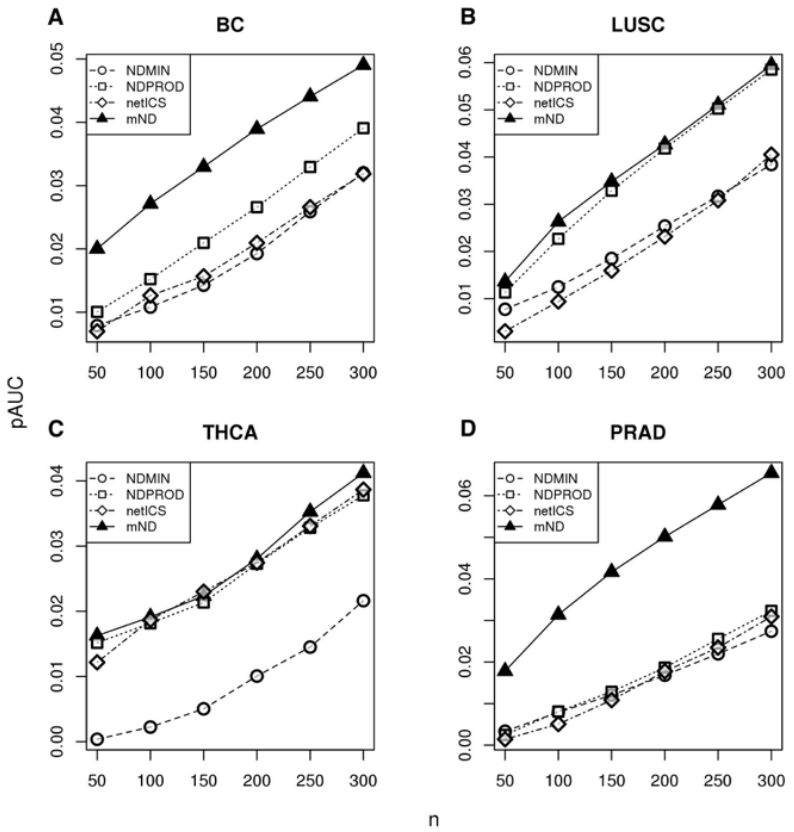


Figure A.2: Performance in recovering known cancer genes. Partial AUC (pAUC) at varying number of top false positive ranking genes (n) in integration of mutation profiles of subjects. The reference gene set was composed of both mutated genes and differentially expressed genes. (**A-D**) These results were generated using interactome WU.

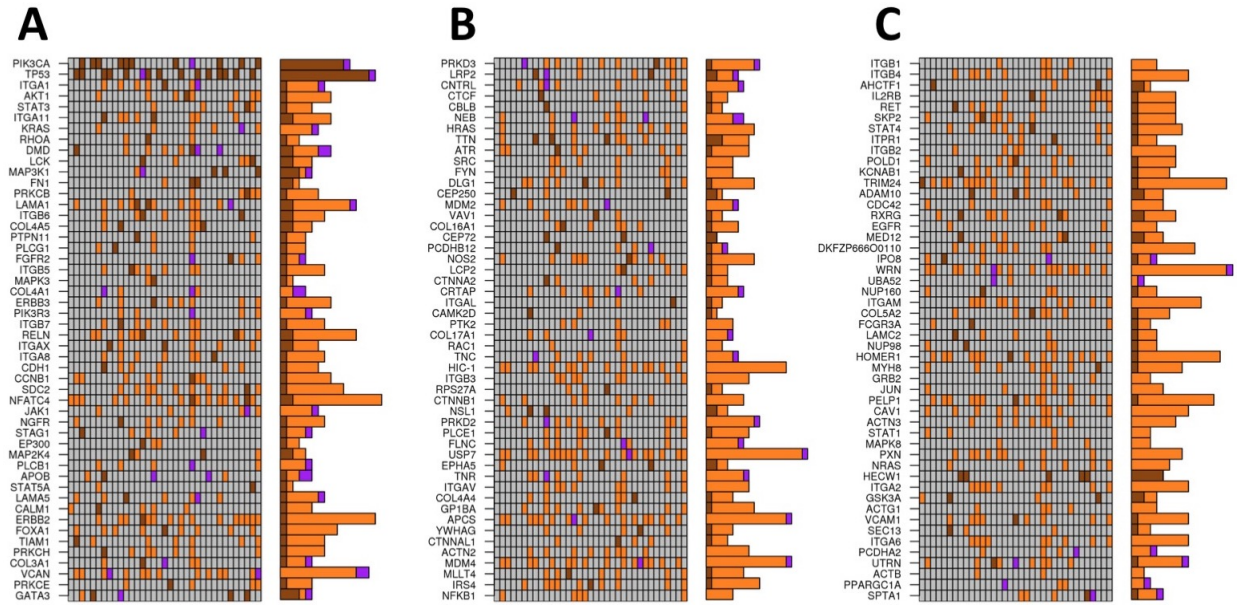


Figure A.3: Classification of genes across layers in the integration of mutation profiles of subjects. Top 150 ranked genes by mND: (**A**) 1-50; (**B**) 51-100; (**C**) 101-150. (**A-C**) Brown: isolated; orange: linker; purple: module; grey: not significant. Barplots reflect the occurrence of the different labels - isolated, linker, module - a gene may assume across layers. These results were generated using interactome WU and BC data from TCGA.

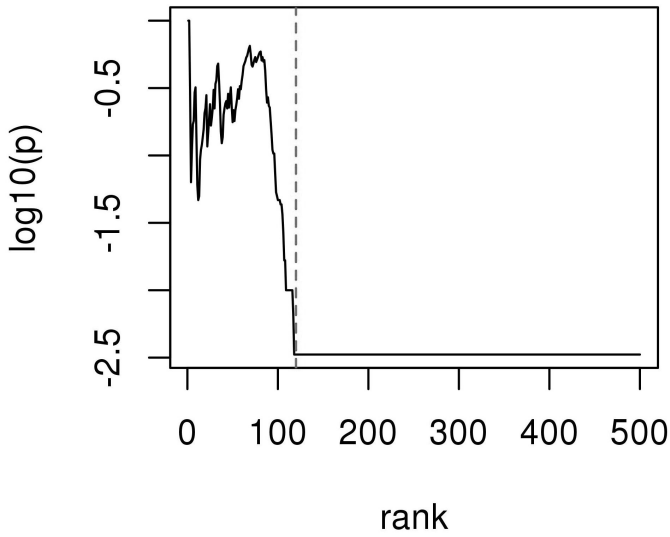


Figure A.4: Network resampling on breast cancer data from TCGA. Logarithm of p-value (y-axis) calculated for each rank of a gene list (x-axis) ordered by decreasing values of mNDp.

Appendix B

Appendix B

Recent comparative studies have brought to our attention how somatic mutation detection from next-generation sequencing data is still an open issue in bioinformatics, because different pipelines result in a low consensus. In this context, it is suggested to integrate results from multiple calling tools, but this operation is not trivial and the burden of merging, comparing, filtering and explaining the results demands appropriate software.

We developed isma (integrative somatic mutation analysis), an R package for the integrative analysis of somatic mutations detected by multiple pipelines for matched tumor-normal samples. The package provides a series of functions to quantify the consensus, estimate the variability, underline outliers, integrate evidences from publicly available mutation catalogues and filter sites. In this chapter, we illustrate the capabilities of isma analysing breast cancer somatic mutations generated by The Cancer Genome Atlas (TCGA) using four pipelines.[1]

[1]The contents of this chapter are published in: *N. Di Nanni, M. Moscatelli, M. Gnocchi, L. Milanesi, E. Mosca E. (2019) "isma: an R package for the integrative analysis of mutations detected by multiple pipelines". BMC Bioinformatics, 20:107.* https://doi.org/10.1186/s12859-019-2701-0.

The package is available for non-commercial users at the URL: https://www.itb.cnr.it/isma.

B.1 isma: an R package for the integrative analysis of mutations detected by multiple pipelines

The identification of somatic mutations from Next Generation sequencing (NGS) data is a challenging task. Several studies compared the single nucleotide variations (SNVs) [192–194] and insertions/deletions (INDELs) [195, 196] detected by different computational tools and underlined relevant discrepancies. Therefore, it is recommended to analyse the same NGS data using multiple callers, like Mutect [102], SomaticSniper [103] and Varscan [104], which generate lists of mutations encoded in Variant Call Format (VCF) [199]. This way of facing conflicting predictions demands appropriate tools that harmonize different outputs and enable comparative analyses [195]. Indeed, for instance, mutation callers encode the same information in multiple ways (Table B.1) and generate outputs with relevant qualitative (e.g. germline/somatic/loss-of-heterozygousity, SNVs/INDELs) and quantitative (number of site found) differences. More generally if, in principle, the use of multiple callers is expected to reduce false positive findings, in practice, the resulting large and heterogeneous lists of mutation sites increase the complexity of the subsequent interpretations. Existing tools like myVCF [216], NGS-pipe [217], VariantTools [218], vcfR [219] and VCFTools [199], implement functions and pipelines to work with VCF files, but do not specifically address the problem of integrating and comparing the results of different mutation callers. A few tools exist to address this problem: Cake [220] (a bioinformatics pipeline implemented in perl) offers the opportunity to run multiple callers and apply customizable filtering steps to obtain a final unique list of single nucleotide variations (SNVs); BAYSIC [221] (implemented in perl) provides a bayesian method for combining SNVs from different variant calling programs. Here, we describe isma (integrative somatic mutation analysis), an R package that provides functions for the joint analysis of VCF files generated by somatic mutation

callers from NGS data (Figure B.1). Differently from existing tools, beyond site integration and filtering, isma provides functions for a more in-depth analysis of mutation sites occurrence across subjects and tools, considering both SNVs and INDELs. The results generated by isma underline common patterns (e.g. recurrent calls, tool consensus in each subject), specificities (e.g. outlier samples, pipeline specific sites, genes enriched in calls from a single pipeline), as well as sites already catalogued by other studies (e.g. The Cancer Genome Atlas (TCGA) [99]), so as to design and apply filtering strategies to screen more reliable sites.

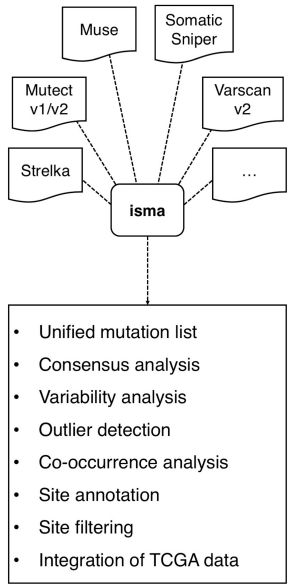


Figure B.1: Overview of isma. Integrative analysis of somatic mutations detected by multiple pipelines.

	Variant type	Mutation inheritance	Model	Implementation	Allelic depth	
					Fields	Values
Mutect [102]	SNV	Somatic	Bayesian	Java	AD	2 comma separated numbers
Mutect (v2) [102]	SNV, INDELs	Somatic	Bayesian	Java	AD	2 comma separated numbers
Muse [101]	SNV	Somatic	Bayesian Markov	C/C++	AD	2 comma separated numbers
SomaticSniper [103]	SNV	Germline, somatic, LOH	Bayesian	C	BCOUNT	4 comma separated numbers
Strelka [222]	SNV, INDELs	Somatic	Bayesian	Perl	AU:CU:GU:TU	4 comma separated numbers
Varscan (v2) [104]	SNV, INDELs	Germline, somatic, LOH	Fisher's exact statistics	Java	AD and RD	2 numbers

Table B.1: Pipelines for somatic mutation call from matched tumor-normal samples. (*) The way in which the allelic depth (number of reads supporting an allele) is encoded in VCF files is reported as an example of heterogeneity among pipeline outputs.

B.2 Implementation

The software isma is implemented in R. The package takes in input mutation sites encoded in VCF files or tab-delimited text files. isma extracts mutation site information from the output of multiple mutation callers by means of specific parsers and integrates sites into a unique data structure:

```
mut_sites <- pre_process (‘‘config.txt’’)
```

Most of the analyses can be easily carried out through a few wrapper functions, like "site_analysis" and "gene_analysis" for site- and gene-level analyses respectively. Nevertheless, many routines are available as part of the user interface to carry out custom analyses (Table B.2).

Gene-level analyses require mutation site annotation, for which isma relies on the R package VariantAnnotation [110] or, alternatively, on user-provided files. Computationally demanding analyses (e.g. the comparison among all-pairs of hundreds of subjects) are implemented in parallel, using the support provided by the R package parallel.

The package isma contains a tutorial available as R vignettes:

```
vignette(‘isma’)
```

Function name	Description
pre_process	Read and integrate input files; generate unique identifiers
site_analysis	Perform site-level analyses, calling get_sites_statistics, overlap_Tools, overlap_Subjects
gene_analysis	Perform gene-level analyses, calling get_sites_statistics, overlap_Tools, overlap_Subjects, gene_mutation
site_annotation	Perform site annotation
integrete_TCGA	Integrate mutation evidence from TCGA
consensus_Tools	Calculate the consensus among tools
get_sites_statistics*	Calculate the co-occurrence of mutation sites/genes across callers and subjects
overlap_Subjects*	Calculate subject-by-subject site/gene co-occurrence matrix
overlap_Tools*	Calculate tool-by-tool site/gene co-occurrence matrix
ese_allsubj*	Calculate the variation of site/genes amount and show the results for each tool
ese_tool_subj*	Calculates the variation of site/genes amount, considering separately each tool and returns the results for each subject
ese_subj_tool*	Calculates the variation of site amount, considering separately each subject and returns the results for each caller
calculate_dist_to_exon	Calculate the site distance from the nearest exons
gene_mutation	Calculate the gene-by-subject mutation matrix and the gene mutation frequency vectors
filtering_sites	Filter sites

Table B.2: isma user interface. The asterisk (*) indicates functions that work both at site- and gene-level.

B.3 Results

In this section, isma will be described considering breast cancer (BC) mutations from TCGA, collected using the function "get_TCGA_sites". In particular, we considered mutation profiles of 975 subjects detected by four variant callers: Mutect2, Varscan2, Muse and SomaticSniper.

```
mut_sites <- get_TCGA_sites (tools = c('muse','
   mutect2','varscan2','somaticsniper'), n_subjects
   = 975)
```

Note that these sites were already filtered by TCGA and are therefore less noisy than the corresponding initial variant caller outputs that would constitute the input of isma in a typical use scenario. Nevertheless, the exploratory analyses made possible by isma underlined interesting patterns even among such filtered calls from TCGA. The analyses presented below can be easily run by means of "site_analysis and "gene_analysis" wrapper functions and include quantification of site/gene occurrence across callers and subject, consensus among tools, detection of outlier subjects and tools, variation of detected sites at different cut-offs on alignment results (e.g. read depth) and integration of information from TCGA.

B.3.1 Site occurrence across callers and subjects

The co-occurrence of sites across tools and subjects is quantified by "get_sites_statistics". This operation allows the user to quantify the fraction of tool-specific calls, the distribution of the sites across tools in each subject and tool consensus on sites. These results are used to detect and mark outlier features (subjects and tools), defined by the inter-quartile range (Tukey's fences) (Table B.3). The amount of shared sites between each pair of callers and subjects is calculated and organized, respectively, in callers-by-callers and subjects-by-subjects site co-occurrence matrices by the functions "overlap_Tools" and "overlap_Subjects". Site co-occurrence matrices are used to summarize consensus and dispersion. Caller consensus relative to a subject is quantified by means of the consensus score (CS), defined as the sum of ratios between the amount of co-occurring sites (off-diagonal elements of the tools-by-tools site co-occurrence matrix) and tool-specific calls (diagonal elements) normalized by the total number of possible tool pairs:

$$CS = \frac{\sum_i^n \left(\frac{1}{x_{i,j}} \sum_{j \neq i}^n x_{i,j} \right)}{P(n,2)} \tag{B.1}$$

where n is the number of tools, $x_{i,j}$ are the sites shared between tools i and j, and $P(n,2)$ is the number of permutations of tools in pairs.

Subjects	Hypermutated	Imbalance in the number of sites across tools	Imbalance in consensus among tools	Tool consensus scores(CS)
A0JC	NO	YES	YES	YES
A1G6	NO	YES	YES	YES
A1L1	NO	YES	NO	YES
A0U0	YES	YES	YES	YES

Table B.3: Outlier subjects report. Examples of subjects recognized as outliers according to site number, imbalance in site number across tools, imbalance in consensus among tools and tool consensus score.

The results of these analyses are summarized into consensus plots, co-occurrence matrices plot and a series of text files, like the summary table of outlier subjects. The overall consensus plot (Fig. B.2) reports the total number of sites found by each tool and the fraction of calls shared among tools. Note how mutect2 found the highest number of sites, the 50% of which was not reported by other callers (Fig. B.2). The consensus plot per subject shows the total number of unique sites, the fraction of sites found by each tool, the distribution of the consensus across subjects and the CS (Fig. B.3). Note the presence of a few hypermutated subjects (i.e. A1XQ, A0U0, A08H, A1J5, A1NC and A25A) (Fig. B.3 A). Several subjects display an imbalance of calls among the pipelines (Fig. B.3 B). Further, there are subjects with a relevant (e.g. A1J5 and A0XR) or poor (e.g. AIKO and A0JC) proportion of sites supported by more than one caller (Fig. B.3 C). Lastly, note how CS underlines, by means of a unique score, subjects with issues in tool consensus, including imbalances in site number or consensus among tools (Fig. B.3 D and Table B.3).

Site co-occurrence between callers revealed that mutect2 detected up to 3 times more sites than other tools, while muse and varscan shared approximately the 60% of their sites (Fig. B.4 A). The mutation co-occurrence in each pair of subjects underlines similarities between mutation profiles; this information is completed with an estimation of the variability (coefficient of variation) of such co-occurrences due to the use of different callers (Fig. B.4 B). The package provides the possibility of calculating, for every gene, the fraction of subjects with at least one mutation, i.e. the gene mutation frequency across subjects (f), and its dispersion across callers. The cor-

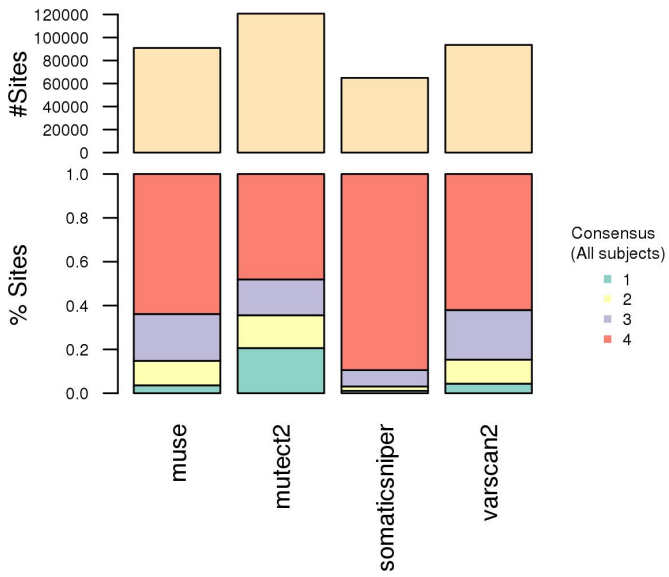


Figure B.2: Global consensus plot. Overall consensus among pipelines; results obtained on BC mutations detected by TCGA in 975 subjects.

responding plot, obtained on BC TCGA sites, underlined the presence of some genes, including known BC genes as GATA3 and CDH1, with a particularly higher variation of f (Fig. B.4 C): indeed, mutect2 and varscan2 detected much more sites than other callers in GATA3 and CDH1 (Fig. B.4 D).

B.3.2 Called sites and sequencing results

The variation of caller output at different cut-offs on site-level quantities (e.g. minimum number of reads, allele frequency) is informative of caller performance and samples (subjects) specificities. This analysis can be done by the function:

```
ese1 <- ese_allsubj(mut_sites$sites,type ='Site')
```

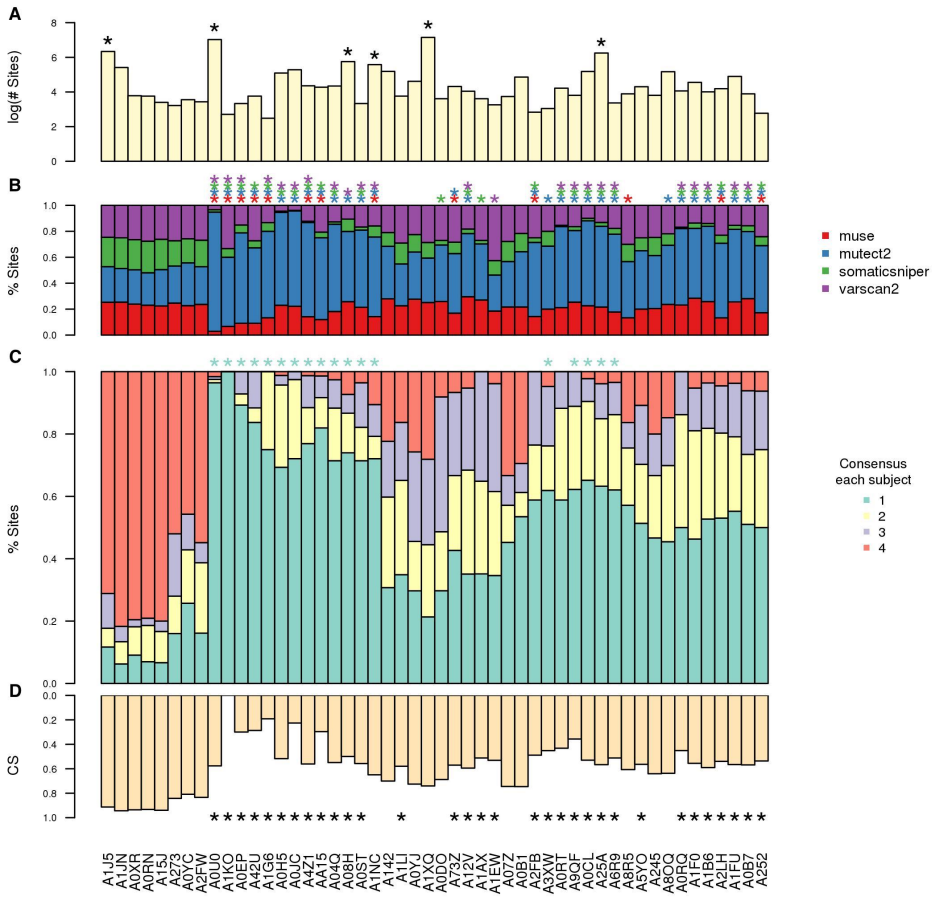


Figure B.3: Detailed consensus plot. (**A**) Number of mutation sites. (**B**) Fraction of sites called by different pipelines. (**C**) Tool Consensus across subjects. (**D**) Consensus score (CS). (**A-D**) Asterisks indicate outliers. Results shown only for 50 subjects (out of 975), selected to include different types of outliers as well as samples without abnormal behaviours.

The pipelines used to call mutations in TCGA BC data show a different

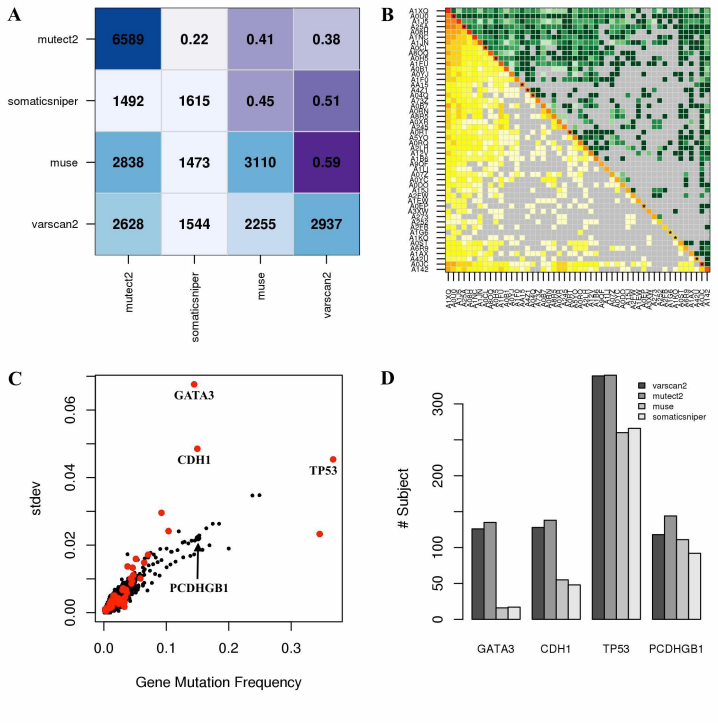


Figure B.4: Site co-occurrence plots and gene mutation frequency variability. (**A**) Total number of sites (diagonal), site co-occurrence among mutation callers (below diagonal) and corresponding similarity between callers (Jaccard index, above diagonal). (**B**) Total number of mutated genes (diagonal), mutation co-occurrence across subjects (below diagonal) and corresponding coefficients of variations (CVs) across pipelines (above diagonal). Asterisks indicate CVs greater than 1; grey colour indicates no mutation co-occurrence between two subjects. (**C**) Standard deviation of gene mutation frequency across pipelines; red: genes associated with BC [110–112]. (**D**) Number of subjects with mutations detected by each tool. (**A-B**) Results obtained on BC mutations from 50 subjects; (**C-D**) Results obtained on all 975 subjects.

behaviour, especially at low tumor variant allele frequency (VAF). In fact, in this range, mutect2 calls more sites than other tools, SomaticSniper

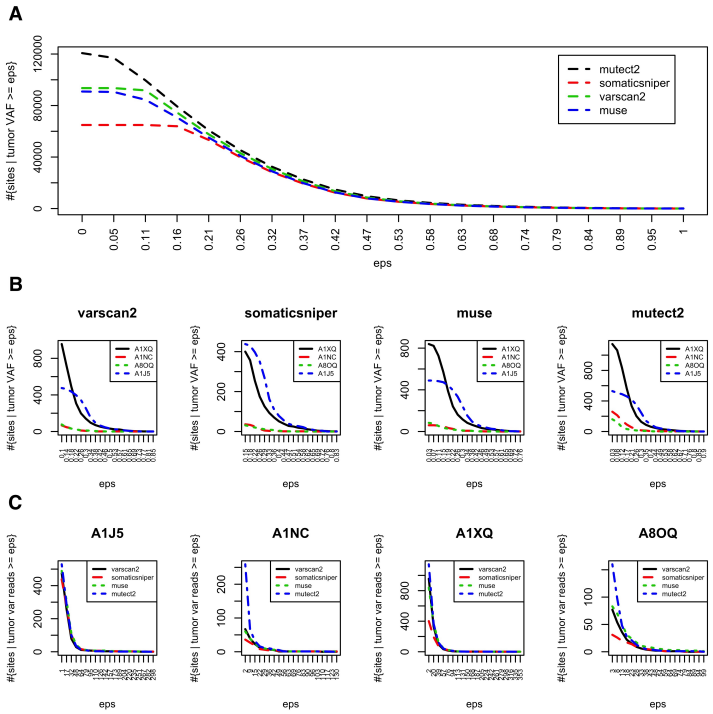


Figure B.5: Number of called sites at various filtering criteria. Number of mutation sites at varying tumor VAF for (**A**) the whole dataset (975 subjects) and (**B**) in single subjects. (**C**) Number of sites at varying number of reads supporting the alternative allele in four subjects. (**A-C**) Results obtained on BC mutations detected by TCGA.

detects almost half of mutect2 sites, while muse and varscan2 show similar trend and are halfway between mutect2 and SomaticSniper (Fig. B.5 A). This global pattern is particularly relevant in some subjects (Fig. B.5 B-C).

B.3.3 Collecting data from the TCGA

The function "integrate_TCGA" uses the R package TCGAbiolinks [100] to collected data from the TCGA. These data are used to support the

mutation sites under analysis with the possible evidence of availability of the same sites among those already catalogued at TCGA, which would be an additional evidence of site reliability.

B.3 Conclusion

The R package isma provides functions for the integrative analysis of mutation sites detected by multiple pipelines. It quantifies the consensus between somatic mutation call pipelines, estimates pipeline variability and biological variability, and underlines outlier features (subject/tools) that may require further investigation. Indeed, an outlier subject may reflect a biological phenomenon (e.g. due to tumor genetic heterogeneity) and/or an experimental problem (e.g. poor biological sample, sequencing performance). The application of isma on BC mutations from TCGA underlined relevant variations among pipelines across subjects, with extreme cases characterized by a very poor consensus. Relevant imbalances among pipelines were also spotted at gene level, which implies a significant variability in the estimation of gene mutation frequency according to the pipeline used. In general, mutect2 reported a higher number of sites at low VAF in comparison to other callers.

In conclusion, the knowledge emerging from the analyses made possible by isma is useful to screen more reliable mutation sites, carry out comparative analysis among pipelines and, lastly, may suggest novel biological insights.

Bibliography

[1] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302, 2016.

[2] Rui Chen and Michael Snyder. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):73–82, 2013.

[3] Maria Gallo Cantafio, Katia Grillone, Daniele Caracciolo, Francesca Scionti, Mariamena Arbitrio, Vito Barbieri, Licia Pensabene, Pietro Guzzi, and Maria Di Martino. From single level analysis to multi-omics integrative approaches: a powerful strategy towards the precision oncology. *High-throughput*, 7(4):33, 2018.

[4] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, 8(1):33, 2015.

[5] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299, 2018.

[6] Noël Malod-Dognin, Julia Petschnigg, and Nataša Pržulj. Precision medicine-a promising, yet challenging road lies ahead. *Current Opinion in Systems Biology*, 7:1–7, 2018.

[7] FDA. `www.fda.gov/medical-devices/vitro-diagnostics/precision-medicine/`.

[8] Roger Higdon, Rachel K Earl, Larissa Stanberry, Caitlin M Hudac, Elizabeth Montague, Elizabeth Stewart, Imre Janko, John Choiniere, William Broomall, Natali Kolker, et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *Omics: a journal of integrative biology*, 19(4):197–208, 2015.

[9] Eva Loth, Declan G Murphy, and Will Spooren. Defining precision medicine approaches to autism spectrum disorders: concepts and challenges. *Frontiers in psychiatry*, 7:188, 2016.

[10] David Q Beversdorf and MISSOURI AUTISM SUMMIT CONSORTIUM. Phenotyping, etiological factors, and biomarkers: Toward precision medicine in autism spectrum disorders. *Journal of Developmental and Behavioral Pediatrics*, 37(8):659, 2016.

[11] Seung Ho Shin, Ann M Bode, and Zigang Dong. Precision medicine: the foundation of future cancer therapeutics. *NPJ precision oncology*, 1(1):12, 2017.

[12] Andrea Garofalo, Lynette Sholl, Brendan Reardon, Amaro Taylor-Weiner, Ali Amin-Mansour, Diana Miao, David Liu, Nelly Oliver, Laura MacConaill, Matthew Ducar, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome medicine*, 8(1):79, 2016.

[13] Adam A Friedman, Anthony Letai, David E Fisher, and Keith T Flaherty. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, 15(12):747, 2015.

[14] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):S15, 2016.

[15] Jaeyun Sung, Yuliang Wang, Sriram Chandrasekaran, Daniela M Witten, and Nathan D Price. Molecular signatures from omics data: from chaos to consensus. *Biotechnology journal*, 7(8):946–957, 2012.

[16] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.

[17] Biswapriya B Misra, Carl D Langefeld, Michael Olivier, and Laura A Cox. Integrated omics: tools, advances, and future approaches. *Journal of molecular endocrinology*, 1(aop), 2018.

[18] Jingwen Yan, Shannon L Risacher, Li Shen, and Andrew J Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6):1370–1381, 2017.

[19] Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. Network embedding in biomedical data science. *Brief. Bioinform*, pages 1–16, 2018.

[20] Xingyi Li, Wenkai Li, Min Zeng, Ruiqing Zheng, and Min Li. Network-based methods for predicting essential genes or proteins: a survey. *Briefings in bioinformatics*, 2019.

[21] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[22] Noemi Di Nanni, Matteo Gnocchi, Marco Moscatelli, Luciano Milanesi, and Ettore Mosca. Network diffusion on multiple-layers: cur-

rent approaches and integrative analysis of rheumatoid arthritis data. *PeerJ Preprints*, 5:e3310v1, 2017.

[23] Noemi Di Nanni, Matteo Gnocchi, Marco Moscatelli, Luciano Milanesi, and Ettore Mosca. Gene relevance based on multiple evidences in complex networks. *Bioinformatics*, 2019.

[24] Noemi Di Nanni, Matteo Bersanelli, Francesca Anna Cupaioli, Luciano Milanesi, Alessandra Mezzelani, and Ettore Mosca. Network-based integrative analysis of genomics, epigenomics and transcriptomics in autism spectrum disorders. *International journal of molecular sciences*, 20(13):3363, 2019.

[25] Noemi Di Nanni, Marco Moscatelli, Matteo Gnocchi, Luciano Milanesi, and Ettore Mosca. isma: an r package for the integrative analysis of mutations detected by multiple pipelines. *BMC bioinformatics*, 20(1):107, 2019.

[26] Youjin Hu, Qin An, Katherine Sheu, Brandon Trejo, Shuxin Fan, and Ying Guo. Single cell multi-omics technology: methodology and application. *Frontiers in cell and developmental biology*, 6:28, 2018.

[27] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85, 2015.

[28] Ashar Ahmad and Holger Fröhlich. Integrating heterogeneous omics data via statistical inference and learning techniques. *Genomics and Computational Biology*, 2(1):e32–e32, 2016.

[29] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299, 2014.

[30] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011.

[31] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[32] Peggy I Wang and Edward M Marcotte. It's the machine that matters: predicting gene function and phenotype from protein networks. *Journal of proteomics*, 73(11):2277–2289, 2010.

[33] Michael Caldera, Pisanu Buphamalai, Felix Müller, and Jörg Menche. Interactome-based approaches to human disease. *Current Opinion in Systems Biology*, 3:88–94, 2017.

[34] Katja Luck, Gloria M Sheynkman, Ivy Zhang, and Marc Vidal. Proteome-scale human interactomics. *Trends in biochemical sciences*, 42(5):342–354, 2017.

[35] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[36] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, et al. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014.

[37] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[38] Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti,

Fiona SL Brinkman, Gianni Cesareni, et al. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nature methods*, 9(4):345, 2012.

[39] Javier De Las Rivas and Celia Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, 2010.

[40] Giulia Menichetti, Daniel Remondini, Pietro Panzarasa, Raúl J Mondragón, and Ginestra Bianconi. Weighted multiplex networks. *PloS one*, 9(6):e97857, 2014.

[41] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[42] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Gastone Castellani, and Luciano Milanesi. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific reports*, 6:34841, 2016.

[43] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[44] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106, 2015.

[45] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

[46] Yupeng Cun and Holger Fröhlich. Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS one*, 8(9):e73074, 2013.

[47] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108, 2013.

[48] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, and Joshua M Stuart. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 2013.

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

[50] Xiaoke Ma, Zaiyi Liu, Zhongyuan Zhang, Xiaotai Huang, and Wanxin Tang. Multiple network algorithm for epigenetic modules via the integration of genome-wide dna methylation and gene expression data. *BMC bioinformatics*, 18(1):72, 2017.

[51] Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble positive unlabeled learning for disease gene identification. *PloS one*, 9(5):e97079, 2014.

[52] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, 9(1):S4, 2008.

[53] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell systems*, 3(6):540–548, 2016.

[54] Xiaoke Ma, Long Gao, and Kai Tan. Modeling disease progression using dynamics of pathway connectivity. *Bioinformatics*, 30(16):2343–2350, 2014.

[55] Leihong Wu, Zhichao Liu, Joshua Xu, Minjun Chen, Hong Fang, Weida Tong, and Wenming Xiao. Netbags: a network-based clustering approach with gene signatures for cancer subtyping analysis. *Biomarkers in medicine*, 9(11):1053–1065, 2015.

[56] Christos Dimitrakopoulos, Sravanth Kumar Hindupur, Luca Häfliger, Jonas Behr, Hesam Montazeri, Michael N Hall, and Niko Beerenwinkel. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34(14):2441–2448, 2018.

[57] Wei Zhang, Jianzhu Ma, and Trey Ideker. Classifying tumors by supervised network propagation. *Bioinformatics*, 34(13):i484–i493, 2018.

[58] Michael Seifert and Andreas Beyer. regnet: An r package for network-based propagation of gene expression alterations. *Bioinformatics*, 34(2):308–311, 2017.

[59] Matthew Ruffalo, Mehmet Koyutürk, and Roded Sharan. Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS computational biology*, 11(12):e1004595, 2015.

[60] Kai Shi, Lin Gao, and Bingbo Wang. Discovering potential cancer driver genes by an integrated network-based approach. *Molecular BioSystems*, 12(9):2921–2931, 2016.

[61] Thanh Le Van, Matthijs Van Leeuwen, Ana Carolina Fierro, Dries De Maeyer, Jimmy Van den Eynden, Lieven Verbeke, Luc De Raedt, Kathleen Marchal, and Siegfried Nijssen. Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*, 32(17):i445–i454, 2016.

[62] Taosheng Xu, Thuc Duy Le, Lin Liu, Rujing Wang, Bingyu Sun, and Jiuyong Li. Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data. *PloS one*, 11(4):e0152792, 2016.

[63] Xue Zhong, Hushan Yang, Shuyang Zhao, Yu Shyr, and Bingshan Li. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC genomics*, 16(7):S7, 2015.

[64] Akihiro Fujimoto, Mayuko Furuta, Yasushi Totoki, Tatsuhiko Tsunoda, Mamoru Kato, Yuichi Shiraishi, Hiroko Tanaka, Hiroaki Taniguchi, Yoshiiku Kawakami, Masaki Ueno, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nature genetics*, 48(5):500, 2016.

[65] Zhaoqi Liu and Shihua Zhang. Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. *BMC genomics*, 16(1):503, 2015.

[66] Nishant Agrawal, Rehan Akbani, B Arman Aksoy, Adrian Ally, Harindra Arachchi, Sylvia L Asa, J Todd Auman, Miruna Balasundaram, Saianand Balu, Stephen B Baylin, et al. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, 2014.

[67] Justin M Drake, Evan O Paull, Nicholas A Graham, John K Lee, Bryan A Smith, Bjoern Titz, Tanya Stoyanova, Claire M Faltermeier, Vladislav Uzunangelov, Daniel E Carlin, et al. Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell*, 166(4):1041–1054, 2016.

[68] Liang Huang, Dan Liu, Na Wang, Shaoping Ling, Yuting Tang, Jun Wu, Lingtong Hao, Hui Luo, Xuelian Hu, Lingshuang Sheng, et al. Integrated genomic analysis identifies deregulated jak/stat-myc-biosynthesis axis in aggressive nk-cell leukemia. *Cell research*, 28(2):172, 2018.

[69] Vésteinn Thorsson, David L Gibbs, Scott D Brown, Denise Wolf, Dante S Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F Gao, Christopher L Plaisier, James A Eddy, et al. The immune landscape of cancer. *Immunity*, 48(4):812–830, 2018.

[70] Xiaoke Ma, Wanxin Tang, Peizhuo Wang, Xingli Guo, and Lin Gao. Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(2):647–658, 2016.

[71] Xiaoke Ma, Penggang Sun, and Guimin Qin. Identifying condition-specific modules by clustering multiple networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1636–1648, 2017.

[72] Xiaoke Ma, Di Dong, and Quan Wang. Community detection in multi-layer networks using joint nonnegative matrix factorization. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):273–286, 2018.

[73] Cui Chen, Feng-Wei Ma, Cui-Yun Du, and Ping Wang. Multiple differential networks strategy reveals carboplatin and melphalan-induced dynamic module changes in retinoblastoma. *Medical science monitor: international medical journal of experimental and clinical research*, 22:1508, 2016.

[74] Li Han, Cui Chen, Chang-Hui Liu, Min Zhang, and Ling Liang. Revealing differential modules in uveal melanoma by analyzing differential networks. *Molecular medicine reports*, 15(4):2261–2266, 2017.

[75] Jia Zhou, Chao Chen, Hua-Feng Li, Yu-Jie Hu, and Hong-Ling Xie. Revealing radiotherapy-and chemoradiation-induced pathway dynamics in glioblastoma by analyzing multiple differential networks. *Molecular medicine reports*, 16(1):696–702, 2017.

[76] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[77] Josef Gladitz, Barbara Klink, and Michael Seifert. Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. *Acta neuropathologica communications*, 6(1):49, 2018.

[78] Su-Ping Deng, Shaolong Cao, De-Shuang Huang, and Yu-Ping Wang. Identifying stages of kidney renal cell carcinoma by combining gene expression and dna methylation data. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(5):1147–1153, 2016.

[79] Florence MG Cavalli, Marc Remke, Ladislav Rampasek, John Peacock, David JH Shih, Betty Luu, Livia Garzia, Jonathon Torchia, Carolina Nor, A Sorana Morrissy, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer cell*, 31(6):737–754, 2017.

[80] Sean P Pitroda, Nikolai N Khodarev, Lei Huang, Abhineet Uppal, Sean C Wightman, Sabha Ganai, Nora Joseph, Jason Pitt, Miguel Brown, Martin Forde, et al. Integrated molecular subtyping defines a curable oligometastatic state in colorectal liver metastasis. *Nature communications*, 9(1):1793, 2018.

[81] Di Zhang, Peng Chen, Chun-Hou Zheng, and Junfeng Xia. Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. *Oncotarget*, 7(4):4298, 2016.

[82] Benjamin J Raphael, Ralph H Hruban, Andrew J Aguirre, Richard A Moffitt, Jen Jen Yeh, Chip Stewart, A Gordon Robertson, Andrew D Cherniack, Manaswi Gupta, Gad Getz, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32(2):185–203, 2017.

[83] Alec M Chiu, Mithun Mitra, Lari Boymoushakian, and Hilary A Coller. Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer. *Scientific reports*, 8(1):11807, 2018.

[84] Loris De Cecco, Marco Giannoccaro, Edoardo Marchesi, Paolo Bossi, Federica Favales, Laura Locati, Lisa Licitra, Silvana Pilotti, and Silvana Canevari. Integrative mirna-gene expression analysis enables refinement of associated biology and prediction of response to cetuximab in head and neck squamous cell cancer. *Genes*, 8(1):35, 2017.

[85] Johnny M Tkach, Askar Yimit, Anna Y Lee, Michael Riffle, Michael Costanzo, Daniel Jaschob, Jason A Hendry, Jiongwen Ou, Jason Moffat, Charles Boone, et al. Dissecting dna damage response pathways by analysing protein localization and abundance changes during dna replication stress. *Nature cell biology*, 14(9):966, 2012.

[86] Jimena Giudice, Zheng Xia, Eric T Wang, Marissa A Scavuzzo, Amanda J Ward, Auinash Kalsotra, Wei Wang, Xander HT Wehrens, Christopher B Burge, Wei Li, et al. Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature communications*, 5:3603, 2014.

[87] Rajendra Karki, David Place, Parimal Samir, Jayadev Mavuluri, Bhesh Raj Sharma, Arjun Balakrishnan, RK Subbarao Malireddi, Rechel Geiger, Qifan Zhu, Geoffrey Neale, et al. Irf8 regulates transcription of naips for nlrc4 inflammasome activation. *Cell*, 173(4):920–933, 2018.

[88] Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10:45–62, 2019.

[89] Hadas Biran, Martin Kupiec, and Roded Sharan. Comparative analysis of normalization methods for network propagation. *Frontiers in genetics*, 10:4, 2019.

[90] Frederik Gwinner, Gwénola Boulday, Claire Vandiedonck, Minh Arnould, Cécile Cardoso, Iryna Nikolayeva, Oriol Guitart-Pla, Cécile V Denis, Olivier D Christophe, Johann Beghain, et al. Network-based analysis of omics data: the lean method. *Bioinformatics*, 33(5):701–709, 2016.

[91] Heiko Horn, Michael S Lawrence, Candace R Chouinard, Yashaswi Shrestha, Jessica Xin Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Eejung Kim, Atanas Kamburov, et al. Netsig: network-based discovery from cancer genomes. *Nature methods*, 15(1):61, 2018.

[92] Dezső Módos, Krishna C Bulusu, Dávid Fazekas, János Kubisch, Johanne Brooks, István Marczell, Péter M Szabó, Tibor Vellai, Péter Csermely, Katalin Lenti, et al. Neighbours of cancer-related proteins have key influence on pathogenesis and could increase the drug target space for anticancer therapies. *NPJ systems biology and applications*, 3(1):2, 2017.

[93] Yufei Xiao, Tzu-Hung Hsiao, Uthra Suresh, Hung-I Harry Chen, Xiaowu Wu, Steven E Wolf, and Yidong Chen. A novel significance score for gene selection and ranking. *Bioinformatics*, 30(6):801–807, 2012.

[94] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[95] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology*, 11(4):e1004120, 2015.

[96] Guanming Wu, Xin Feng, and Lincoln Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5):R53, 2010.

[97] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, 43(D1):D36–D42, 2014.

[98] Marc Carlson, S Falcon, H Pages, and N Li. org. hs. eg. db: Genome wide annotation for human. *R package version*, 3(1), 2013.

[99] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[100] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2015.

[101] Yu Fan, Liu Xi, Daniel ST Hughes, Jianjun Zhang, Jianhua Zhang, P Andrew Futreal, David A Wheeler, and Wenyi Wang. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*, 17(1):178, 2016.

[102] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213, 2013.

[103] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis,

Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2011.

[104] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[105] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[106] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[107] Ettore Mosca, Roberta Alfieri, and Luciano Milanesi. Diffusion of information throughout the host interactome reveals gene expression variations in network proximity to target proteins of hepatitis c virus. *PloS one*, 9(12):e113660, 2014.

[108] Ettore Mosca, Matteo Bersanelli, Matteo Gnocchi, Marco Moscatelli, Gastone Castellani, Luciano Milanesi, and Alessandra Mezzelani. Network diffusion-based prioritization of autism risk genes identifies significantly connected gene modules. *Frontiers in Genetics*, 8:129, 2017.

[109] Michelle S Scott and Geoffrey J Barton. Probabilistic prediction and ranking of human protein-protein interactions. *BMC bioinformatics*, 8(1):239, 2007.

[110] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue

of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2018.

[111] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333, 2013.

[112] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495, 2014.

[113] Hayley M Dingerdissen, John Torcivia-Rodriguez, Yu Hu, Ting-Chia Chang, Raja Mazumder, and Robel Kahsay. Biomuta and bioxpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic acids research*, 46(D1):D1128–D1136, 2017.

[114] Guangcun Cheng, Xueqing Sun, Jinglong Wang, Gang Xiao, Xiumin Wang, Xuemei Fan, Lidong Zu, Mingang Hao, Qing Qu, Yan Mao, et al. Hic1 silencing in triple-negative breast cancer drives progression through misregulation of lcn2. *Cancer research*, 74(3):862–872, 2014.

[115] Yingying Wang, Xiaoling Weng, Luoyang Wang, Mingang Hao, Yue Li, Lidan Hou, Yu Liang, Tianqi Wu, Mengfei Yao, Guowen Lin, et al. Hic1 deletion promotes breast cancer progression by activating tumor cell/fibroblast crosstalk. *Journal of Clinical Investigation*, 128(12):5235–5250, 2018.

[116] Daniel Elias and Henrik J Ditzel. Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacological research*, 100:250–254, 2015.

[117] Ga-Hang Lee, Ki-Chun Yoo, Yoojeong An, Hae-June Lee, Minyoung Lee, Nizam Uddin, Min-Jung Kim, In-Gyu Kim, Yongjoon Suh, and

Su-Jae Lee. Fyn promotes mesenchymal phenotypes of basal type breast cancer cells through stat5/notch2 signaling node. *Oncogene*, 37(14):1857, 2018.

[118] Eleni Chrysanthou, Kylie L Gorringe, Chitra Joseph, Madeleine Craze, Christopher C Nolan, Maria Diez-Rodriguez, Andrew R Green, Emad A Rakha, Ian O Ellis, and Abhik Mukherjee. Phenotypic characterisation of breast cancer: the role of cdc42. *Breast cancer research and treatment*, 164(2):317–325, 2017.

[119] Toru Mukohara. Pi3k mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer: Targets and Therapy*, 7:111, 2015.

[120] Nam Nhut Phan, Chih-Yang Wang, Kuan-Lun Li, Chien-Fu Chen, Chung-Chieh Chiao, Han-Gang Yu, Pung-Ling Huang, and Yen-Chang Lin. Distinct expression of cdca3, cdca5, and cdca8 leads to shorter relapse free survival in breast cancer patient. *Oncotarget*, 9(6):6977, 2018.

[121] Jiangang Liu, Andrew Campen, Shuguang Huang, Sheng-Bin Peng, Xiang Ye, Mathew Palakal, A Keith Dunker, Yuni Xia, and Shuyu Li. Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. *BMC medical genomics*, 1(1):39, 2008.

[122] Changran Wei, Ying Wang, and Xiangqi Li. The role of hippo signal pathway in breast cancer metastasis. *OncoTargets and therapy*, 11:2185, 2018.

[123] Mohd Farhan, Haitao Wang, Uma Gaur, Peter J Little, Jiangping Xu, and Wenhua Zheng. Foxo signaling pathways as therapeutic targets in cancer. *International journal of biological sciences*, 13(7):815, 2017.

[124] Milena Gasco, Shukri Shami, and Tim Crook. The p53 pathway in breast cancer. *Breast cancer research*, 4(2):70, 2002.

[125] Justin Cidado and Ben Ho Park. Targeting the pi3k/akt/mtor pathway for breast cancer therapy. *Journal of mammary gland biology and neoplasia*, 17(3-4):205–216, 2012.

[126] Aparna Mani and Edward P Gelmann. The ubiquitin-proteasome pathway and its role in cancer. *Journal of clinical oncology*, 23(21):4776–4789, 2005.

[127] Christian P Schaaf and Huda Y Zoghbi. Solving the autism puzzle a few pieces at a time. *Neuron*, 70(5):806–808, 2011.

[128] Bernie Devlin and Stephen W Scherer. Genetic architecture in autism spectrum disorder. *Current opinion in genetics & development*, 22(3):229–237, 2012.

[129] Brett S Abrahams, Dan E Arking, Daniel B Campbell, Heather C Mefford, Eric M Morrow, Lauren A Weiss, Idan Menashe, Tim Wadkins, Sharmila Banerjee-Basu, and Alan Packer. Sfari gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (asds). *Molecular autism*, 4(1):36, 2013.

[130] Barbara Wiśniowiecka-Kowalnik and Beata Anna Nowakowska. Genetics and epigenetics of autism spectrum disorder-current evidence in the field. *Journal of applied genetics*, 60(1):37–47, 2019.

[131] Shan V Andrews, Brooke Sheppard, Gayle C Windham, Laura A Schieve, Diana E Schendel, Lisa A Croen, Pankaj Chopra, Reid S Alisch, Craig J Newschaffer, Stephen T Warren, et al. Case-control meta-analysis of blood dna methylation and autism spectrum disorder. *Molecular autism*, 9(1):40, 2018.

[132] Rui Luo, Stephan J Sanders, Yuan Tian, Irina Voineagu, Ni Huang, Su H Chu, Lambertus Klei, Chaochao Cai, Jing Ou, Jennifer K Lowe, et al. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent cnvs in autism spectrum disorders. *The American Journal of Human Genetics*, 91(1):38–55, 2012.

[133] Marta Codina-Solà, Benjamín Rodríguez-Santiago, Aïda Homs, Javier Santoyo, Maria Rigau, Gemma Aznar-Laín, Miguel Del Campo, Blanca Gener, Elisabeth Gabau, María Pilar Botella, et al. Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. *Molecular autism*, 6(1):21, 2015.

[134] Sek Won Kong, Christin D Collins, Yuko Shimizu-Motohashi, Ingrid A Holm, Malcolm G Campbell, In-Hee Lee, Stephanie J Brewster, Ellen Hanson, Heather K Harris, Kathryn R Lowe, et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PloS one*, 7(12):e49475, 2012.

[135] Daniel S Tylee, Daniel M Kawaguchi, and Stephen J Glatt. On the outside, looking in: A review and evaluation of the comparability of blood and brain "-omes". *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162(7):595–603, 2013.

[136] Shan V Andrews, Shannon E Ellis, Kelly M Bakulski, Brooke Sheppard, Lisa A Croen, Irva Hertz-Picciotto, Craig J Newschaffer, Andrew P Feinberg, Dan E Arking, Christine Ladd-Acosta, et al. Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder. *Nature communications*, 8(1):1011, 2017.

[137] Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular autism*, 8:1–17, 2017.

[138] Silvia De Rubeis, Xin He, Arthur P Goldberg, Christopher S Poultney, Kaitlin Samocha, A Ercument Cicek, Yan Kou, Li Liu, Menachem Fromer, Susan Walker, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209, 2014.

[139] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[140] Valerie W Hu, Tewarit Sarachana, Kyung Soon Kim, AnhThu Nguyen, Shreya Kulkarni, Mara E Steinberg, Truong Luu, Yinglei Lai, and Norman H Lee. Gene expression profiling differentiates autism case–controls and phenotypic variants of autism spectrum disorders: Evidence for circadian rhythm dysfunction in severe autism. *Autism research*, 2(2):78–97, 2009.

[141] Tiziano Pramparo, Michael V Lombardo, Kathleen Campbell, Cynthia Carter Barnes, Steven Marinero, Stephanie Solso, Julia Young, Maisi Mayo, Anders Dale, Clelia Ahrens-Barbeau, et al. Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers. *Molecular systems biology*, 11(12), 2015.

[142] Jeffrey P Gregg, Lisa Lit, Colin A Baron, Irva Hertz-Picciotto, Wynn Walker, Ryan A Davis, Lisa A Croen, Sally Ozonoff, Robin Hansen, Isaac N Pessah, et al. Gene expression changes in children with autism. *Genomics*, 91(1):22–29, 2008.

[143] Thanit Saeliw, Chayanin Tangsuwansri, Surangrat Thongkorn, Weerasak Chonchaiya, Kanya Suphapeetiporn, Apiwat Mutirangura, Tewin Tencomnao, Valerie W Hu, and Tewarit Sarachana. Integrated genome-wide alu methylation and transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. *Molecular autism*, 9(1):27, 2018.

[144] Jingjing Li, Minyi Shi, Zhihai Ma, Shuchun Zhao, Ghia Euskirchen, Jennifer Ziskin, Alexander Urban, Joachim Hallmayer, and Michael Snyder. Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Molecular systems biology*, 10(12), 2014.

[145] Neelroop N Parikshak, Rui Luo, Alice Zhang, Hyejung Won, Jennifer K Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H Geschwind. Integrative functional genomic analyses implicate specific

molecular pathways and circuits in autism. *Cell*, 155(5):1008–1021, 2013.

[146] Dalila Pinto, Elsa Delaby, Daniele Merico, Mafalda Barbosa, Alison Merikangas, Lambertus Klei, Bhooma Thiruvahindrapuram, Xiao Xu, Robert Ziman, Zhuozhi Wang, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics*, 94(5):677–694, 2014.

[147] Fereydoun Hormozdiari, Osnat Penn, Elhanan Borenstein, and Evan E Eichler. The discovery of integrated gene networks for autism and related disorders. *Genome research*, 25(1):142–154, 2015.

[148] Fei Yi, Tamas Danko, Salome Calado Botelho, Christopher Patzke, ChangHui Pak, Marius Wernig, and Thomas C Südhof. Autism-associated shank3 haploinsufficiency causes ih channelopathy in human neurons. *Science*, 352(6286):aaf2669, 2016.

[149] Mengye Zhu, Vinay Kumar Idikuda, Jianbing Wang, Fusheng Wei, Virang Kumar, Nikhil Shah, Christopher B Waite, Qinglian Liu, and Lei Zhou. Shank3-deficient thalamocortical neurons show hcn channelopathy and alterations in intrinsic electrical properties. *The Journal of physiology*, 596(7):1259–1276, 2018.

[150] Caroline Nava, Carine Dalle, Agnès Rastetter, Pasquale Striano, Carolien GF De Kovel, Rima Nabbout, Claude Cancès, Dorothée Ville, Eva H Brilstra, Giuseppe Gobbi, et al. De novo mutations in hcn1 cause early infantile epileptic encephalopathy. *Nature genetics*, 46(6):640, 2014.

[151] Jinong Feng, Richard Schroer, Jin Yan, Wenjia Song, Chunmei Yang, Anke Bockholt, Edwin H Cook Jr, Cindy Skinner, Charles E Schwartz, and Steve S Sommer. High frequency of neurexin 1β signal peptide structural variants in patients with autism. *Neuroscience letters*, 409(1):10–13, 2006.

[152] Julie Gauthier, Tabrez J Siddiqui, Peng Huashan, Daisaku Yokomaku, Fadi F Hamdan, Nathalie Champagne, Mathieu Lapointe, Dan Spiegelman, Anne Noreau, Ronald G Lafrenière, et al. Truncating mutations in nrxn2 and nrxn1 in autism spectrum disorders and schizophrenia. *Human genetics*, 130(4):563–573, 2011.

[153] Daniel J Parente, Caryn Garriga, Berivan Baskin, Ganka Douglas, Megan T Cho, Gabriel C Araujo, and Marwan Shinawi. Neuroligin 2 nonsense variant associated with anxiety, autism, intellectual disability, hyperphagia, and obesity. *American journal of medical genetics Part A*, 173(1):213–216, 2017.

[154] Stéphane Jamain, Hélène Quach, Catalina Betancur, Maria Råstam, Catherine Colineaux, I Carina Gillberg, Henrik Soderstrom, Bruno Giros, Marion Leboyer, Christopher Gillberg, et al. Mutations of the x-linked genes encoding neuroligins nlgn3 and nlgn4 are associated with autism. *Nature genetics*, 34(1):27, 2003.

[155] Li-Feng Jiang-Xie, Hsiao-Mei Liao, Chia-Hsiang Chen, Yuh-Tarng Chen, Shih-Yin Ho, Dai-Hua Lu, Li-Jen Lee, Horng-Huei Liou, Wen-Mei Fu, and Susan Shur-Fen Gau. Autism-associated gene dlgap2 mutant mice demonstrate exacerbated aggressive behaviors and orbitofrontal cortex deficits. *Molecular autism*, 5(1):32, 2014.

[156] Christian R Marshall, Abdul Noor, John B Vincent, Anath C Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, et al. Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488, 2008.

[157] Dalila Pinto, Alistair T Pagnamenta, Lambertus Klei, Richard Anney, Daniele Merico, Regina Regan, Judith Conroy, Tiago R Magalhaes, Catarina Correia, Brett S Abrahams, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368, 2010.

[158] H Poquet, L Faivre, SE Chehadeh, J Morton, D Mcmullan, and S Hamilton. Further evidence for dlgap2 as strong autism spectrum disorders/intellectual disability candidate gene. *Autism Open Access*, 6, 2016.

[159] Kazuhiko Nakamura, Ayyappan Anitha, Kazuo Yamada, Masatsugu Tsujii, Yoshimi Iwayama, Eiji Hattori, Tomoko Toyota, Shiro Suda, Noriyoshi Takei, Yasuhide Iwata, et al. Genetic and expression analyses reveal elevated expression of syntaxin 1a (stx1a) in high functioning autism. *International Journal of Neuropsychopharmacology*, 11(8):1073–1084, 2008.

[160] Kazuhiko Nakamura, Yasuhide Iwata, Ayyappan Anitha, Taishi Miyachi, Tomoko Toyota, Satoru Yamada, Masatsugu Tsujii, Kenji J Tsuchiya, Yoshimi Iwayama, Kazuo Yamada, et al. Replication study of japanese cohorts supports the role of stx1a in autism susceptibility. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):454–458, 2011.

[161] Takefumi Kofuji, Yuko Hayashi, Tomonori Fujiwara, Masumi Sanada, Masao Tamaru, and Kimio Akagawa. A part of patients with autism spectrum disorder has haploidy of hpc-1/syntaxin1a gene that possibly causes behavioral disturbance as in experimentally gene ablated mice. *Neuroscience letters*, 644:5–9, 2017.

[162] Jaroslava Durdiaková, Varun Warrier, Sharmila Banerjee-Basu, Simon Baron-Cohen, and Bhismadev Chakrabarti. Stx1a and asperger syndrome: a replication study. *Molecular autism*, 5(1):14, 2014.

[163] S Hossein Fatemi, Teri J Reutiman, Timothy D Folsom, Robert J Rooney, Diven H Patel, and Paul D Thuras. mrna and protein levels for gaba a α4, α5, β1 and gaba b r1 receptors are altered in brains from subjects with autism. *Journal of autism and developmental disorders*, 40(6):743–750, 2010.

[164] Shuhan Yang, Xuan Guo, Xiaopeng Dong, Yu Han, Lei Gao, Yuanyuan Su, Wei Dai, and Xin Zhang. Gaba a receptor subunit gene polymorphisms predict symptom-based and developmental deficits in chinese han children and adolescents with autistic spectrum disorders. *Scientific reports*, 7(1):3290, 2017.

[165] Bradley P Coe, Holly AF Stessman, Arvis Sulovari, Madeleine R Geisheker, Trygve E Bakken, Allison M Lake, Joseph D Dougherty, Ed S Lein, Fereydoun Hormozdiari, Raphael A Bernier, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature genetics*, 51(1):106, 2019.

[166] Anamaria Sudarov, Frank Gooden, Debbie Tseng, Wen-Biao Gan, and Margaret Elizabeth Ross. Lis1 controls dynamics of neuronal filopodia and spines to impact synaptogenesis and social behaviour. *EMBO molecular medicine*, 5(4):591–607, 2013.

[167] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-ha Lee, Giuseppe Narzisi, Anthony Leotta, et al. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299, 2012.

[168] Sarah R Gilman, Ivan Iossifov, Dan Levy, Michael Ronemus, Michael Wigler, and Dennis Vitkup. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70(5):898–907, 2011.

[169] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[170] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[171] Lewis Y Geer, Aron Marchler-Bauer, Renata C Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H

Bryant. The ncbi biosystems database. *Nucleic acids research*, 38(suppl_1):D492–D496, 2009.

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

[173] Karina Griesi-Oliveira, Allan Acab, Abha R Gupta, Daniele Yumi Sunaga, Thanathom Chailangkarn, Xavier Nicol, Yanelli Nunez, Michael F Walker, John D Murdoch, Stephan J Sanders, et al. Modeling non-syndromic autism and the impact of trpc6 disruption in human neurons. *Molecular psychiatry*, 20(11):1350, 2015.

[174] Patricia CB Beltrão-Braga and Alysson R Muotri. Modeling autism spectrum disorders with human neurons. *Brain research*, 1656:49–54, 2017.

[175] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338, 2013.

[176] E Melo Felipe De Sousa, Louis Vermeulen, Evelyn Fessler, and Jan Paul Medema. Cancer heterogeneity-a multifaceted view. *EMBO reports*, 14(8):686–695, 2013.

[177] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479, 2013.

[178] Jeffrey A Magee, Elena Piskounova, and Sean J Morrison. Cancer stem cells: impact, heterogeneity, and uncertainty. *Cancer cell*, 21(3):283–296, 2012.

[179] Michael F Clarke and Margaret Fuller. Stem cells and cancer: two faces of eve. *Cell*, 124(6):1111–1115, 2006.

[180] Jeffrey Koury, Li Zhong, and Jijun Hao. Targeting signaling pathways in cancer stem cells for cancer treatment. *Stem cells international*, 2017, 2017.

[181] Liheng Zhou, Yiwei Jiang, Tingting Yan, Genhong Di, Zhenzhou Shen, Zhimin Shao, and Jinsong Lu. The prognostic role of cancer stem cells in breast cancer: a meta-analysis of published literatures. *Breast cancer research and treatment*, 122(3):795–801, 2010.

[182] Judy Crabtree and Lucio Miele. Breast cancer stem cells. *Biomedicines*, 6(3):77, 2018.

[183] Komal Qureshi-Baig, Pit Ullmann, Serge Haan, and Elisabeth Letellier. T umor-i nitiating c ells: a cri tic al review of isolation approaches and new challenges in targeting strategies. *Molecular cancer*, 16(1):40, 2017.

[184] Dokyoon Kim, Ruowang Li, Scott M Dudek, John R Wallace, and Marylyn D Ritchie. Binning somatic mutations based on biological knowledge for predicting survival: an application in renal cell carcinoma. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 96–107. World Scientific, 2014.

[185] Junjun Zhang, Joachim Baran, Anthony Cros, Jonathan M Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, et al. International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

[186] Pau Creixell, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, et al. Pathway and network analysis of cancer genomes. *Nature methods*, 12(7):615, 2015.

[187] hg19. `www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/`.

[188] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[189] Peter N. Robinson, Rosario Michael Piro, Marten Jager. *Computational Exome and Genome Analysis*. CRC Press, 2017.

[190] PICARD. `http://broadinstitute.github.io/picard/`.

[191] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491, 2011.

[192] Lei Cai, Wei Yuan, Zhou Zhang, Lin He, and Kuo-Chen Chou. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific reports*, 6:36540, 2016.

[193] Nicola D Roberts, R Daniel Kortschak, Wendy T Parker, Andreas W Schreiber, Susan Branford, Hamish S Scott, Garique Glonek, and David L Adelson. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–2230, 2013.

[194] Qingguo Wang, Peilin Jia, Fei Li, Haiquan Chen, Hongbin Ji, Donald Hucks, Kimberly Brown Dahlman, William Pao, and Zhongming Zhao. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome medicine*, 5(10):91, 2013.

[195] Tyler S Alioto, Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D Eldridge, Eivind Hovig, Lawrence E Heisler, Timothy A Beck, Jared T Simpson, Laurie Tonon, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature communications*, 6:10001, 2015.

[196] Anne Bruun Krøigård, Mads Thomassen, Anne-Vibeke Lænkholm, Torben A Kruse, and Martin Jakob Larsen. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PloS one*, 11(3):e0151664, 2016.

[197] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 2016.

[198] Ensembl Variation - Calculated variant consequences. `www.ensembl.org/info/genome/variation/prediction/predicted_data.html`.

[199] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[200] Sònia Palomeras, Santiago Ruiz-Martínez, and Teresa Puig. Targeting breast cancer stem cells to overcome treatment resistance. *Molecules*, 23(9):2193, 2018.

[201] Xiaoxian Li, Michael T Lewis, Jian Huang, Carolina Gutierrez, C Kent Osborne, Meng-Fen Wu, Susan G Hilsenbeck, Anne Pavlick, Xiaomei Zhang, Gary C Chamness, et al. Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *Journal of the National Cancer Institute*, 100(9):672–679, 2008.

[202] Ming Luo and Jun-Lin Guan. Focal adhesion kinase: a prominent determinant in breast cancer initiation, progression and metastasis. *Cancer letters*, 289(2):127–139, 2010.

[203] Jun-Lin Guan. Integrin signaling through fak in the regulation of mammary stem cells and breast cancer. *IUBMB life*, 62(4):268–276, 2010.

[204] Mei-Ren Pan, Ming-Feng Hou, Fu Ou-Yang, Chun-Chieh Wu, Shu-Jyuan Chang, Wen-Chun Hung, Hon-Kan Yip, and Chi-Wen Luo. Fak is required for tumor metastasis-related fluid microenvironment in triple-negative breast cancer. *Journal of clinical medicine*, 8(1):38, 2019.

[205] Shihua Wang, Xiaodong Su, Meiqian Xu, Xian Xiao, Xiaoxia Li, Hongling Li, Armand Keating, and Robert Chunhua Zhao. Exosomes secreted by mesenchymal stromal/stem cell-derived adipocytes promote breast cancer cell growth via activation of hippo signaling pathway. *Stem cell research & therapy*, 10(1):117, 2019.

[206] Xin Zhou, Shuyang Wang, Zhen Wang, Xu Feng, Peng Liu, Xian-Bo Lv, Fulong Li, Fa-Xing Yu, Yiping Sun, Haixin Yuan, et al. Estrogen regulates hippo signaling via gper in breast cancer. *The Journal of clinical investigation*, 125(5):2123–2135, 2015.

[207] Kayla E Denson, Ashley L Mussell, He Shen, Alexander Truskinovsky, Nuo Yang, Natesh Parashurama, Yanmin Chen, Costa Frangou, Fajun Yang, and Jianmin Zhang. The hippo signaling transducer taz regulates mammary gland morphogenesis and carcinogen-induced mammary tumorigenesis. *Scientific reports*, 8(1):6449, 2018.

[208] Fouad Saeg and Muralidharan Anbalagan. Breast cancer stem cells and the challenges of eradication: a review of novel therapies. *Stem cell investigation*, 5, 2018.

[209] Vandana Venkatesh, Raghu Nataraj, Gopenath S Thangaraj, Murugesan Karthikeyan, Ashok Gnanasekaran, Shanmukhappa B Kaginelli, Gobianand Kuppanna, Chandrashekrappa Gowdru Kallappa, and Kanthesh M Basalingappa. Targeting notch signalling pathway of cancer stem cells. *Stem cell investigation*, 5, 2018.

[210] Fokhrul Hossain, Ayse D Ucar Bilyeu, Claudia Sorrentino, Judy Crabtree, Antonio Pannuti, Margarite Matossian, Matthew Burow, Todd

Golde, Barbara Osborne, and Lucio Miele. Targeting cancer stem-like cells in triple negative breast cancer, 2018.

[211] Fokhrul Hossain, Claudia Sorrentino, Deniz A Ucar, Yin Peng, Margarite Matossian, Dorota Wyczechowska, Judy Crabtree, Jovanny Zabaleta, Silvana Morello, Luis Del Valle, et al. Notch signaling regulates mitochondrial metabolism and nf-κb activity in triple-negative breast cancer cells via ikkα-dependent non-canonical pathways. *Frontiers in oncology*, 8:575, 2018.

[212] breastcancertrial. `https://www.breastcancertrials.org/BCTIncludes/Resources/BreastCancerDrugs.html`.

[213] Breastcancer.org. `https://www.breastcancer.org/treatment/druglist`.

[214] NCI -Drugs Approved for Breast Cancer. `https://www.cancer.gov/about-cancer/treatment/drugs/breast`.

[215] Kelsy C Cotto, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, and Malachi Griffith. Dgidb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic acids research*, 46(D1):D1068–D1073, 2017.

[216] Alessandro Pietrelli and Luca Valenti. myvcf: a desktop application for high-throughput mutations data management. *Bioinformatics*, 33(22):3676–3678, 2017.

[217] Jochen Singer, Hans-Joachim Ruscheweyh, Ariane L Hofmann, Thomas Thurnherr, Franziska Singer, Nora C Toussaint, Charlotte KY Ng, Salvatore Piscuoglio, Christian Beisel, Gerhard Christofori, et al. Ngs-pipe: a flexible, easily extendable and highly configurable framework for ngs analysis. *Bioinformatics*, 34(1):107–108, 2017.

[218] Michael Lawrence and Robert Gentleman. Varianttools: an extensible framework for developing and testing variant callers. *Bioinformatics*, 33(20):3311–3313, 2017.

[219] Brian J Knaus and Niklaus J Grünwald. vcfr: a package to manipulate and visualize variant call format data in r. *Molecular Ecology Resources*, 17(1):44–53, 2017.

[220] Mamunur Rashid, Carla Daniela Robles-Espinoza, Alistair G Rust, and David J Adams. Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics*, 29(17):2208–2210, 2013.

[221] Brandi L Cantarel, Daniel Weaver, Nathan McNeill, Jianhua Zhang, Aaron J Mackey, and Justin Reese. Baysic: a bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC bioinformatics*, 15(1):104, 2014.

[222] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.