



Network-Based Approaches To Rare Diseases

Doctoral thesis at the Medical University of Vienna for obtaining
the academic degree

Doctor of Philosophy

Submitted by

Pisanu Buphamalai

B.Sc. (Hons), M.Sc.

Supervisor:

Univ.Prof. Dr. Jörg Menche

Max Perutz Labs and Department of Mathematics, University of Vienna

CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences

Vienna, 12/2021

Declaration

This thesis was supervised by Univ.-Prof. Dr. Jörg Menche, and the present work was performed at CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (until October 2020) and the Max Perutz Labs, University of Vienna (October 2020 onwards). The author of this thesis contributed substantially to the original research and writing of the presented manuscripts. The manuscripts were published under licences permitting their reproduction in this thesis. Peer-reviewed funding was granted by the Vienna Science and Technology Fund (WWTF, project number VRG15-005).

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Contents

Declaration	i
Contents	iii
Abstract in English	v
Zusammenfassung auf Deutsch	vii
Publications arising from this thesis	ix
Abbreviations	xi
Acknowledgements	xiii
1 Introduction	1
1.1 Network medicine	3
1.2 The human interactome	49
1.3 The multiscale organization of molecular complexity	58
1.4 Rare diseases	63
2 Aims of the thesis	67
3 Results	69
3.1 Main publication	70
4 Discussion	103
4.1 Scale representation, network construction and characterization	103
4.2 Mapping rare disease genes and modularity quantification	106
4.3 Disease gene retrieval and prediction performance	108

4.4 Future Prospects	110
Bibliography	113
Appendix	127
Reporting summary for the main publication	127
Curriculum vitae	133

Abstract in English

At the turn of the millennium, the rise of network science together with the growing map of protein-protein interactions – the interactome - have allowed the novel discipline of network biology and medicine to be founded. Early network medicine studies on the human interactome successfully demonstrated the concept of disease modularity and thereby the quantification of disease relationships from the molecular roots. However, the interactome as a map is largely incomplete, and has failed to capture underlying functional relationships beyond physical interactions. In other words, it only captures one side of the multifaceted biological complexity. More recent studies have attempted to reconstruct networks from the increasing abundance of high throughput experimental data to gain additional insights into the system through co-expression, genetic, pathway, or regulatory networks. Nevertheless, these networks have primarily been considered separately, and in the context of individual diseases. How to integrate networks derived from various types of data to elucidate cross-scale biological organization and investigate the impact of different networks to disease aetiology and mechanisms remained an open question. This doctoral thesis aims to apply network biology principles to rare disease studies where the scarcity of relevant data has long hindered the diagnostic and therapeutic efforts. Building on what we have learned from specific diseases, we develop a general network-based framework to systematically investigate rare diseases. In particular, mechanisms for how severe genetic defects affect the various levels of biological organization were elucidated. To this end, we compiled a multi-layer gene network of over 20 million relationships across various scales, from interactions at the genetic level to phenotypic similarities, as well as a comprehensive set of over 3700 rare diseases with known genetic basis. To our knowledge, this represents the first attempt to systematically map all known rare diseases to a cross-scale network. We demonstrated that this massive dataset can be leveraged to address several practical and conceptual challenges in rare disease research, particularly the prioritization of disease causal genes where the framework developed in this doctoral thesis have shown to outperform previously established methods.

Zusammenfassung auf Deutsch

Um die Jahrtausendwende hat der Aufstieg der Netzwerkwissenschaft zusammen mit der wachsenden Karte der Protein-Protein-Interaktionen - dem Interaktom - die Gründung der neuen Disziplin der Netzwerkbiologie und -medizin ermöglicht. Frühe netzwerkmedizinische Studien zum menschlichen Interaktom haben das Konzept der Modularität von Krankheiten und damit die Quantifizierung von Krankheitszusammenhängen von den molekularen Wurzeln her erfolgreich demonstriert. Das Interaktom als Karte ist jedoch weitgehend unvollständig und kann, die zugrunde liegenden funktionellen Beziehungen jenseits der physikalischen Interaktionen nur unzureichend erfassen. Mit anderen Worten: Es erfasst nur eine Seite der vielschichtigen biologischen Komplexität. In neueren Studien wurde versucht, aus der zunehmenden Fülle von experimentellen Hochdurchsatzdaten Netzwerke zu rekonstruieren, um durch Koexpressions-, Gen-, Signalweg- oder regulatorische Netzwerke zusätzliche Erkenntnisse über das System zu gewinnen. Dennoch wurden diese Netzwerke in erster Linie separat und im Zusammenhang mit einzelnen Krankheiten betrachtet. Die Frage, wie Netzwerke, die aus verschiedenen Datentypen abgeleitet wurden, integriert werden können, um die skalenübergreifende biologische Organisation aufzuklären und die Auswirkungen der verschiedenen Netzwerke auf die Ätiologie und die Mechanismen von Krankheiten zu untersuchen, blieb offen. Diese Doktorarbeit zielt darauf ab, die Prinzipien der Netzwerkbiologie auf die Erforschung seltener Krankheiten anzuwenden, bei denen der Mangel an relevanten Daten lange Zeit die diagnostischen und therapeutischen Bemühungen behindert hat. Aufbauend auf den Erkenntnissen über spezifische Krankheiten entwickeln wir einen allgemeinen, netzwerkbasierten Rahmen für die systematische Erforschung seltener Krankheiten. Insbesondere wurden die Mechanismen aufgeklärt, wie sich schwere Gendefekte auf die verschiedenen Ebenen der biologischen Organisation auswirken. Zu diesem Zweck haben wir ein mehrschichtiges Netzwerk zusammengestellt, das aus über 20 Millionen Genbeziehungen auf verschiedenen Ebenen besteht, von genetischen Interaktionen bis hin zu phänotypischen Ähnlichkeiten, sowie aus einem umfassenden Satz von über 3700 seltenen Krankheiten mit bekannter genetischer Grundlage. Nach unserem Kenntnisstand ist dies der erste Versuch, alle bekannten seltenen Krankheiten systematisch in einem skalenübergreifenden Netzwerk abzubilden. Wir haben gezeigt, dass dieser riesige Datensatz genutzt werden kann, um verschiedene praktische und konzeptionelle Herausforderungen in der Forschung zu seltenen Krankheiten anzugehen, insbesondere die Priorisierung von krankheitsverursachenden Genen, bei der das in dieser Doktorarbeit entwickelte System nachweislich besser abschneidet als die bisher etablierten Methoden.

Publications arising from this thesis

* indicates equal contributions

1. Pisanu Buphamalai^{*}, Michael Caldera^{*}, Felix Muller, and Jörg Menche
Book Chapter: **Network Medicine: Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists**, edited by Nataša Pržulj, 414–58. *Cambridge University Press*, 2019
DOI: 10.1017/9781108377706
© 2019, Cambridge University Press
Reprinted with permission.
2. Michael Caldera^{*}, Pisanu Buphamalai^{*}, Felix Muller, and Jörg Menche
Review article: **Current Opinion in Systems Biology**, **3**: 88–94, 2017.
DOI: 10.1016/j.coisb.2017.04.015
© 2017 The Authors. Published by Elsevier Ltd.
This article is available under the Creative Commons CC-BY-NC-ND license and permits non-commercial use of the work as published, without adaptation or alteration provided the work is fully attributed. The publication was reprinted in its entirety.
3. Pisanu Buphamalai, Tomislav Kokotovic, Vanja Nagy, and Jörg Menche
Research article: **Nature Communications**, volume 12, Article number: 6306 (2021)
DOI: 10.1038/s41467-021-26674-1
© 2021, The Authors. Published by Springer Nature Limited.
This article is available under the Creative Commons Attribution 4.0 International License and permits non-commercial use of the work as published, without adaptation or alteration provided the work is fully attributed. The publication was reprinted in its entirety.

Abbreviations

AF Allele Frequency

AP-MS Affinity-Purification Mass Spectrometry

BP Biological Process

GO Gene Ontology

GTE_x Genotype-Tissue Expression

HPO Human Phenotype Ontology

HuRI Human Reference Interactome

ID Intellectual Disability

LCC Largest Connected Component

MF Molecular Function

MPO Mammalian Phenotype Ontology

ORDO Orphanet Rare Disease Ontology

PPI Protein-Protein Interaction

RD Rare Diseases

RWR Random Walk with Restart

Y2H Yeast-Two-Hybrid

Acknowledgements

First of all, I would like to thank CeMM for giving me the opportunity to work in a unique, supportive, interdisciplinary, and, above all, fun and creative environment. From scientific seminars to the occasional parties, CeMM always ensures the perfect balance for everyone to be professional, and yet playful and kind to one another. CeMM has become a home, and some of the peers I have met here have become the best friends that I ever had.

I would like to thank the Max Perutz Labs and the University of Vienna for providing us a new home after the group moved. It's the coffee at the Perutz that fueled my final PhD year, and it is where my main paper and this thesis were born as a result.

I am deeply grateful to my closest collaborator and biggest supporter, the Nagy lab, for their kindness and their patience in introducing me to the fascinating world of neurobiology. I learned so much and had lots of fun in the process. It has truly been my pleasure to see how the two very different disciplines merged and formed beautiful and impactful projects.

I would like to thank my thesis committee, Prof. Stefan Thurner, Prof. Roded Sharan and Dr. Vanja Nagy for their valuable advice in the conceptualization phase of this doctoral thesis.

After all, the degree that I seek may not be of any less importance than the friendship that I gained. I would therefore like to thank these groups of people: the Macrophages for being the best PhD cohort CeMM has ever seen; Zozo & Friends for being a part of this journey from the beginning in all aspects of my life; the Olos for the proud, cozy and vibrant space that we built and shared; the Asian Amigos for being the best friends (sometimes mothers), for keeping me fed, cared and alive; and my flatmate Andi for preventing me from going to work on Sundays, for his encouragements, supports and optimism.

None of this would have been possible if it was not for my family respecting my decision to freely choose my own path from the very beginning, for allowing me to give up being one kind of doctor so that I can become another. I know it has never been easy for my mother to have to send her only son away across the globe for so long. Her daily virtual hugs have been as warm as caring as ever, and definitely helped me weather through all the good and bad days when I needed them most.

I would like to thank the Menche lab for the companion over all these years, and for all the evenings we delved into science and not-so-science discussion. I learned so much from each of your unique characteristics, and from all of us together as a group. I learned to lead and to follow, to speak up and to listen, to give and to get - thanks for giving me the stage to perform all these roles and to grow together as a team.

Lastly, I would like to thank my ‘Doktorvater’ Jörg who is not only my dedicated supervisor, but also a fine mentor and a really cool friend. Thanks for constantly reinforcing the *Ph* part so that I am worth the *PhD* title that I hope to soon carry. Your approach to science and the world has truly fostered me into not only a better scientist, but also a better human being. Thanks for your wholehearted trust, confidence and support you had in me, and for showing that kindness prevails in the busy and competitive world that we live in.

Thank you.

1

Introduction

The essence of reductionism is the assumption of the dissectible nature of a system, and that it can be explained in terms of its constituents and their individual direct interactions. For centuries, scientific research has been conducted and advanced based upon such a reductionist approach. In an attempt to understand a system, its constituent parts first have to be defined, and then comprehensively studied, although often in isolation. On the one hand, such an approach enables the understanding of a phenomenon or system of interest through detailed inspection of its components. On the other hand, as a system rarely operates in isolation, a reductionist approach can only yield limited insights to the entire system (van Riel & Van Gulick, 2019). Furthermore, constituents that make up the system are often incomplete, and their relationships that may give rise to additional, emergent properties are often neglected. Holistic perspectives were urgently needed to tackle this increasingly apparent challenge (Gallagher et al., 1999). The science of complexity, a discipline that considers collective behaviours of the system in its essence, was rapidly advancing throughout the past decades in all domains, from statistical physics to sociology, and from information theory to economics (Turner et al., 2018; Auyang, 1998; Newman, 2011).

Modern network theory is at the forefront of the rapid advances of complexity science. Properties including scale-free power-law degree distribution (Barabasi & Albert, 1999; Barabási, 2009) and the small-world effect (Watts & Strogatz, 1998) have been observed to be universal in most real-world systems from the Internet (Albert et al., 1999) to protein-protein interactions (Wagner, 2003). This universality allows methodologies developed in one discipline to be adopted and applied in the others. These methodologies, combined with comprehensive studies of biological interactions in health and diseases gave rise to the novel discipline of network biology (Barabási & Oltvai, 2004) and later network medicine (Barabási et al., 2011). Mapping disease genes onto a network of physical interactions among proteins (*the Interactome*) reveals the

existence of *disease modules*, *i.e.*, topological neighbourhood where disease genes are strongly localized (Barabási et al., 2011; Vidal et al., 2011). The identification of disease modules based on known genes has enabled scientists to look for potential novel genes associated with a disease or certain molecular functions and quantify disease relationships (Menche et al., 2015). Elegant mathematical formulation combined with the powerful visualization capability of the networks have rapidly matured the discipline (Vespignani, 2018). Network-based methodologies and visualizations have become standard practices in various studies from module detection in cancer (Leiserson et al., 2015; Wang et al., 2015), identifying molecular aetiology in asthma (Sharma et al., 2015), neurological (Stam, 2014; Li et al., 2017), and immunological diseases (Rieckmann et al., 2017). Beyond disease contexts, further applications such as gene-drug interactions (Guney et al., 2016; Subramanian et al., 2017; Keenan et al., 2017; Caldera et al., 2019) have also been explored. The explosion of network medicine unprecedentedly shifted our perception of biological systems to a holistic view.

Despite the success of high-throughput technologies, linking genotypes to phenotypes for complex diseases remains a challenging task that requires both further implementation of technology and thorough study design to yield sufficient statistical power and the dissection of genetic and environmental factors of the diseases (Visscher et al., 2012; Boyle et al., 2017; McCarthy & Birney, 2021). In comparison, rare diseases are often early onset and characterized by much clearer genotype-phenotype relationships. This makes them ideal models to establish general principles to understand disease causality. Studies incorporated the protein-protein interaction networks in patient gene prioritization (Köhler et al., 2008; Smedley et al., 2015), as well as constructing tissue-specific interactome in an attempt to understand mechanistic processes of disease development (Kitsak et al., 2016; Luck et al., 2020). However, despite advances in sequencing technologies, the proportion of undiagnosed rare disease patients remain high, let alone the rate of patients with successful treatments (Graessner et al., 2021). Efficient tools to integrate scarce information of rare diseases into actionable insights remain urgently needed.

This chapter aims to introduce four major concepts central to this doctoral thesis: (i) *network medicine* - this section is accompanied by a comprehensive book chapter that I co-authored on the historical and current methodologies, along with future prospects of the discipline; (ii) *the interactome* as the first and most studied biological networks in the context of diseases as well as its strengths and limitations. This section is accompanied by a scholarly review that I co-authored; (iii) *the network-based data integration* to capture the multifaceted complexity of biological systems; Lastly, (iv) *rare diseases* as a perfect opportunity to apply state-of-the-art network data and methodologies to overcome its bottlenecks in both therapeutic and diagnostic settings.

1.1 Network medicine

BOOK CHAPTER

Pisanu Buphamalai^{*}, Michael Caldera^{*}, Felix Muller, and Jörg Menche

Published in **Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists**, edited by Nataša Pržulj, 414–58. *Cambridge University Press*, 2019
DOI: 10.1017/9781108377706

^{*} Authors contributed equally

© 2019, Cambridge University Press
Reprinted with permission.

The book chapter introduces various aspects of the emerging field of network medicine, where fundamental ideas and principles in this doctoral thesis were built upon. The book chapter is organized as follows. First, it introduces different types of interactions which gives rise to biological organization across scales, e.g., from gene regulatory and protein-protein interaction networks at the molecular scale to the social networks at the population scale. Second, methods for extraction of relationships to construct network are outlined. This process is subject to available data types where respective biological scales are represented. For example, co-expression networks derived from gene expression data are constructed by means of correlation, while pathway and disease networks derived from annotation data are constructed by means of bi-partite association. Furthermore, the book chapter outlines key quantification of network characteristics including the measurement of localization, distances, and randomization. Network construction and quantification measurements presented in this book chapter have been fundamental to the methods and findings in this thesis.

10 Network Medicine

Pisanu Buphamalai, Michael Caldera*, Felix Müller,
and Jörg Menche*

10.1 Introduction

Since the publication of the first draft of the human genome less than two decades ago [1, 2], rapid technological progress has revolutionized biomedical research. Thanks to a diverse array of “omics” technologies (e.g., genome sequencing, transcriptome mapping, proteomics, metabolomics, and others), we can now quantify both healthy and disease states at molecular resolution. At the same time, it has become clear that the detailed characterization of the individual molecular components alone (genes, proteins, metabolites, etc.) does not suffice to truly understand the nature of (patho-) physiological states and how to modulate them. Indeed, biomolecules do not act in isolation, but within an intricate and tightly coordinated machinery of complex interactions, such as protein–protein, gene regulatory, or signaling interactions. Network medicine is an emerging field that aims to apply tools and concepts from network theory to elucidate this machinery. Network approaches have helped unravel the molecular mechanisms of a broad range of diseases, from rare Mendelian disorders [3], cancer [4] or metabolic diseases [5], to basic attack strategies of viruses [6], to name but a few examples. While the molecular networks that underly biological processes may be the most natural candidate for applying network concepts in biomedical research, they are certainly not the only one. Networks are used across the full spectrum of medicine, from biomarker [7] to drug discovery [8], from the spread of obesity [9] to global outbreaks of infectious diseases [10], and from characterizing the relationships among diseases [11] to those among physicians within the health care system [12].

This chapter aims to give a general introduction to the dynamic field of network medicine. We start with a broad overview of major network types that are relevant to medicine. We then discuss with more detail the cellular network of molecular interactions among proteins and other biomolecules, the perhaps most widely used network in biomedical research. In the last section, we introduce disease module analysis, an important application of network tools to elucidate the molecular mechanisms of a particular disease.

* equal contribution

10.2 Networks in Medicine

10.2.1 Overview

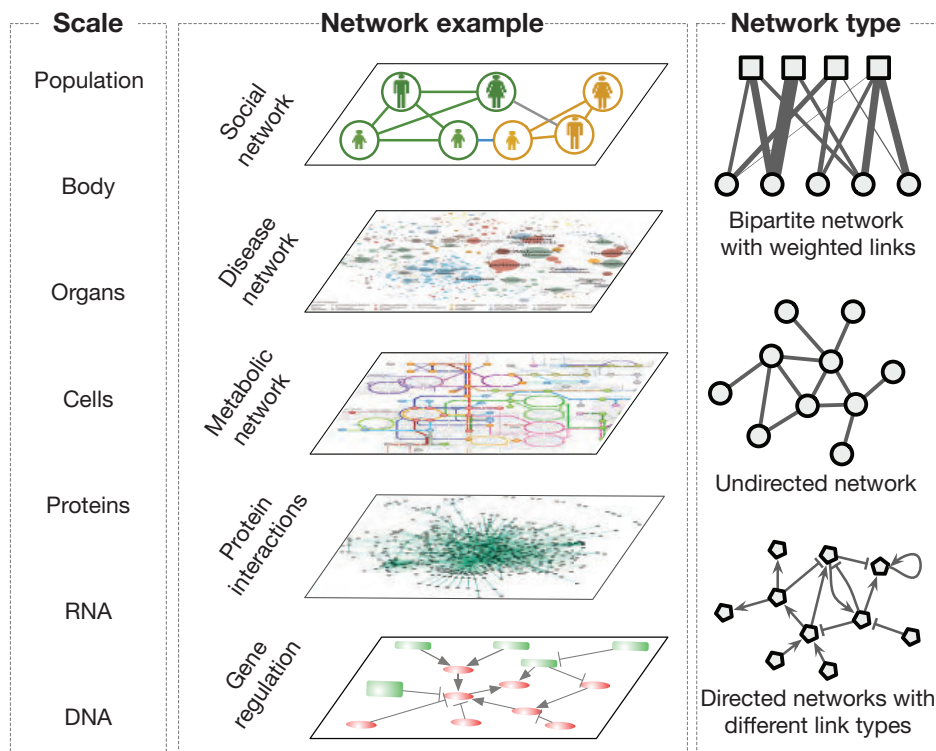
One can distinguish three basic network types that cover different disease-relevant relationships: (i) Molecular networks describing the relationships between the molecular constituents of living organisms, for example, maps of all protein–protein interactions or metabolic reactions in a cell. The observation that such molecular maps share certain universal topological features with vastly different systems, e.g., the World Wide Web, collaboration networks, power grids, and many others, was instrumental for the development of network science. Today, it seems almost trivial that networks provide the most natural way of describing and analyzing the large-scale organization of biomolecules and their interactions. (ii) Disease networks are a powerful tool to investigate the diverse relationships between diseases. For example, two diseases can be linked if they share genetic associations or if they have similar clinical manifestations. In contrast to molecular networks, in which links often represent direct physical interactions, disease–disease networks represent more abstract relationships. They therefore serve as beautiful examples for the power of networks as a general tool for the analysis, integration, and intuitive visualization of large and complex data. (iii) Population-scale networks, i.e., networks describing the complex interactions among humans have been very successful in modeling and predicting the spread of contagious diseases, for example, global swine flu or ebola pandemics. These studies show the enormous potential of networks to serve as a platform for translating exact analytical results from physics and mathematics and translating them to concrete applications in medicine. (See Box 10.1.)

10.2.2 Molecular Networks

There are a plethora of molecular networks describing different aspects of the molecular and cellular organization of living organisms. A broad distinction can be made between physical and functional interaction networks. Physical interactions involve actual physical contact between the participating biomolecules, for example, proteins that assemble in a complex or receptor–ligand binding. Functional interaction, on the other hand, can refer to any kind of biologically relevant relationship. In co-expression networks, for example, genes are connected if their expression patterns are strongly correlated [13]. In the following we introduce the main types of molecular networks that are used to elucidate diverse disease mechanisms. Some of them were introduced in previous chapters, but we also summarize them here for completeness.

10.2.2.1 Protein–Protein Interaction Networks

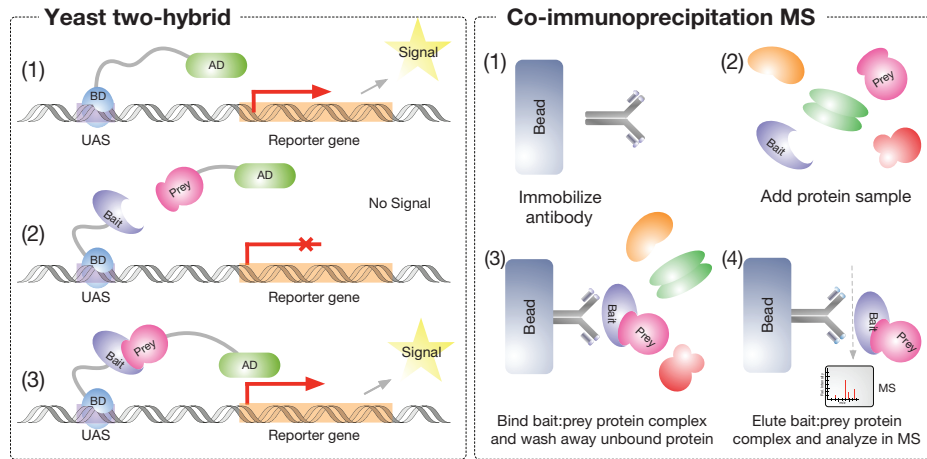
Many molecular processes within a cell are performed by molecular machines consisting of a large number of protein components organized by their protein–protein interactions (PPIs). PPIs result from biochemical events steered by electrostatic forces leading to physical contacts of high specificity between two or more proteins [14]. Perturbed PPIs are involved in the pathobiology of many diseases, ranging from diabetes and obesity to Crohn’s disease or cancer [15]. In analogy to the “genome” representing

Box 10.1: Networks in medicine

The diverse networks that are studied in network medicine reflect the different levels of organization that are relevant to human disease. From the molecular level, e.g., networks of interacting biomolecules that form the basis of all cellular processes, to the level of social interactions that are involved in the transmission of infectious diseases. Depending on the particular system, different network types are used for their description. Undirected and unweighted networks represent the most basic network type. More complex types may include a link directionality, link weights or use different types of nodes, for example in bipartite networks.

the collection of all genes in an organism, the collection of all molecular interactions is often referred to as the “interactome.” The interactome can be represented by a network in which the nodes are proteins and the edges correspond to physical interaction between them. Over the last decade, significant experimental efforts have been made to map out the complete human interactome. High-throughput techniques such as yeast two-hybrid (Y2H) and immunoprecipitation linked to mass spectrometry are capable of mapping thousands of interactions in parallel (see Box 10.2). There has also been substantial work in curating interactions that were identified in small-scale experiments, as well as using computational tools to predict interactions [15].

Box 10.2: Mapping the human interactome



There are two major high-throughput techniques for the identification of protein interactions:

Yeast two-hybrid: (1) the system uses a protein consisting of a DNA binding domain (BD) and an activation domain (AD) that is responsible for activating transcription of DNA. (2) In Y2H, the two domains are separated and fused to proteins whose interaction is investigated. The BD is fused to the so-called bait, the AD to the prey. (3) Upon interaction between the two proteins of interest, the AD comes in close proximity to the reporter gene and the transcription leads to a signal.

Co-immunoprecipitation coupled to mass spectrometry: (1) In a first step, a target (bait) protein-specific antibody is immobilized on beads (e.g., agarose). (2) When the cell lysate is added, the antibody will specifically bind the target protein and indirectly capture proteins (prey) that are capable of binding to it. (3) After washing away unbound proteins, (4) the proteins of interest are eluted and analyzed using mass spectrometry. In short, the sample (the proteins) is first ionized and fragmented into smaller molecules, e.g., amino acids and peptides. Their mass-to-charge ratios can then be determined by accelerating the ions and subjecting them to an electric and/or magnetic field. Finally, the proteins in the sample can be identified by comparing with databases of known masses and characteristic fragmentation patterns.

Despite these promising first steps, our knowledge of the human interactome map remains far from complete, estimates indicate that only 10–30% of the full interactome has been revealed currently [16]. Nevertheless, interactome-based studies have contributed substantially to our understanding of biological processes both in homeostasis and in disease states, see Section 10.3.

10.2.2.2 *Metabolic Networks*

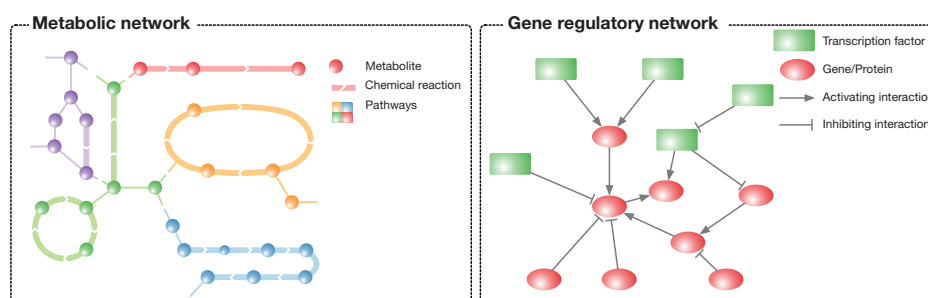
Metabolism (from Greek *μεταβολή* for “change”) refers to the sum of all processes that are involved in assembling and disassembling the basic building blocks of cells, in particular the biochemical reactions for energy conversion. Traditionally, these reactions have been organized into specific pathways, for example the tricarboxylic acid (TCA) cycle, which corresponds to the sequence of chemical reactions in the cell that produces energy (also known as citric acid – or Krebs cycle, named after Hans Krebs, a Nobel Laureate in 1953). Metabolic networks represent collections of such pathways that connect chemical compounds (metabolites), biochemical reactions, enzymes, and genes. The relationships between the individual components of a given metabolic system can be inferred using comparative genomics combined with metabolomic data [17]. Metabolic networks are the most complete among the different biological networks, i.e., they reflect a near exhaustive knowledge of the involved biochemical processes [18]. They are available for a wide range of species and can be accessed through databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] or Reactome [20]. The currently most comprehensive human metabolic network, Recon 2.2 [21], includes 5,324 metabolites, 7,785 reactions, and 1,675 associated genes. Such metabolic networks do not only offer deep insights into the basic machinery of cells, but can also be used for *in silico* simulations to study how different parameters (e.g., metabolite concentrations) affect local and global properties of the biochemical network. The two most commonly used methods employ either (1) deterministic approaches (e.g., systems of ordinary differential equations) or (2) stochastic models (e.g., effect probabilities upon network perturbation) [22]. Metabolic network analyses can yield profound insights into the evolutionary emergence of complex life forms [23, 24], help understand the molecular mechanisms that drive the response to vaccination [25], or elucidate the interplay between metabolism and gene regulation [26]. (See Box 10.3.)

10.2.2.3 *Regulatory Networks*

Regulatory networks describe the complex machinery of genes and their corresponding proteins and RNAs, as well as the interactions between them that control the level of gene expression across the genome under specific conditions. Of particular importance for expression regulation are transcription factors (TFs), i.e., DNA-binding proteins that modulate the first step in gene expression [27]. In the most common representation of regulatory networks, nodes correspond to genes and links to the regulation of the expression of one gene by the product of the other. The links are typically directed and have either an activating (i.e., an increase in the concentration of one leads to an increase in the expression of the other) or inhibitory effect (increase in the concentration of one leads to decrease in the other) [28, 29]. Several experimental techniques exist to create large-scale data for building genome-wide regulatory networks, such as Chromatin-Immunoprecipitation Chip (ChIP-on-chip) [30] and ChIP-Sequencing [31]. Comprehensive databases include the Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE) [32] or JASPAR [33].

Gene regulatory networks provide powerful tools to identify key transcription factors that control cell fate, for example in early blood development [34, 35].

Box 10.3: Metabolic and regulatory networks



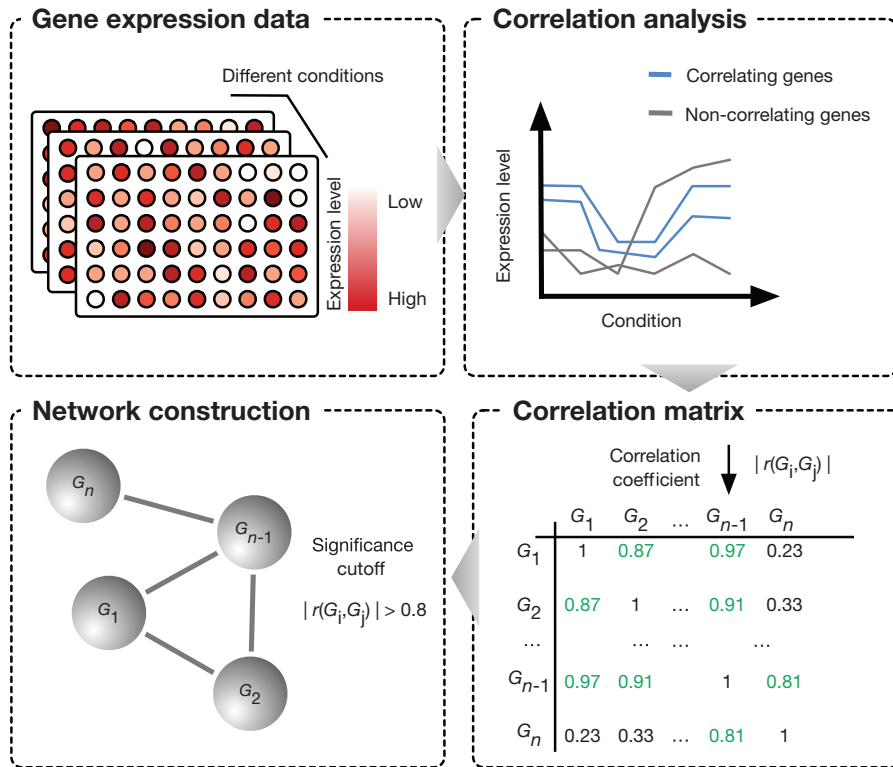
Metabolic networks describe the conversion/transformation of chemicals (metabolites) within a cell, organ, or whole organism. The nodes represent specific molecules while the edges describe the chemical reactions that take place between the nodes. Often these reactions are catalyzed by enzymes. Specific routes/compartments that are known to perform a particular function are called pathways.

Gene regulatory networks consist of genes that regulate each other. Often these genes are transcription factors that are capable of binding to DNA. The type of interactions can be either positive leading to an increase of protein concentration of the regulated gene, or negative, which leads to a decrease in protein concentration.

They can also be used to interpret variants identified in genome-wide association studies (GWAS), as they often perturb regulatory modules that are highly specific to disease-relevant cell types or tissues [36]. Lastly, gene regulatory networks also shed light on evolutionary conditions and pathways by which new regulatory functions emerge [37]. (See Box 10.3.)

10.2.2.4 Co-Expression Networks

In co-expression networks, genes are linked if their expression levels are significantly correlated under different experimental conditions, for example over time, across different tissues or cell types, or across a patient population (see Box 10.4 for an overview of the construction process) [13, 39]. In contrast to regulatory networks, co-expression networks do not offer an immediate causal relationship between genes. They can be used, however, to identify groups of genes that are more broadly functionally related, for example, controlled by the same transcriptional regulatory program, or members of the same pathway or protein complex [40]. Network analyses have been used to identify commonly affected pathways in heterogeneous diseases like autism spectrum disorder [41] or inflammatory bowel disease [42], predict causal GWAS genes associated with bone mineral density [43], or help explain the mechanism of breast cancer development [44].

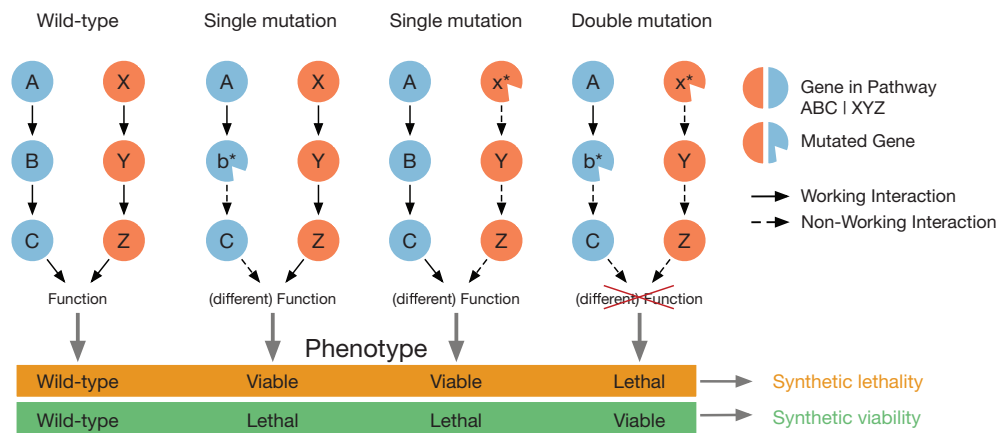
Box 10.4: Co-expression networks

Construction of a co-expression network: Creating a co-expression network requires gene expression data over several conditions, for example different treatments, across several tissues or patients. For each gene pair one can then calculate a correlation coefficient for their respective expression values across the different conditions, resulting in a correlation matrix. Extracting biologically meaningful correlations can be quite challenging, as true signals are often masked by noise that can arise, for example, from experimental confounding factors, batch effects, or sample heterogeneity. A widely used alternative to somewhat arbitrary global thresholds preserves the continuous nature of correlation scores and instead applies soft thresholding to identify network subclusters [13]. With recent large-scale resources, such as GTEx [38], noise from sample heterogeneity can be reduced and co-expression networks can be constructed in a tissue-specific manner, thus providing deeper and more robust insights onto the regulatory system in diseases.

10.2.2.5 Genetic Interactions

Two genes are linked by a genetic interaction if the effect of a simultaneous alteration (e.g., a mutation or the complete knock-down) of both genes differs from the

Box 10.5: Genetic interactions



Genetic interactions occur when the phenotype of two combined mutations differs significantly from the expectation based on the individual mutations. These interactions can be either positive (combined effect stronger than expected) or negative (combined effect weaker). The two most extreme outcomes are called “synthetic lethality” and “synthetic viability.” In **synthetic lethality** the two individual mutations often occur in two independent, yet redundant pathways, so that the loss of one can be compensated for by the second. Only when targeting both pathways the system fails. In **synthetic viability** the mutation in one pathway often leads to a toxic gene product. Only by also affecting another pathway the production of this toxic product is stopped and the resulting phenotype is again viable.

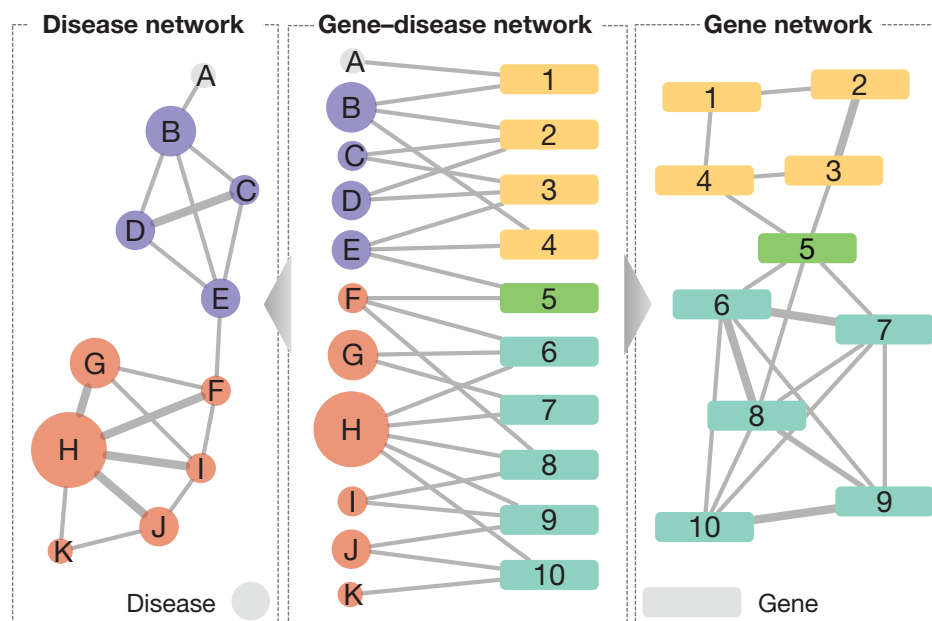
expectation based on the individual alterations [45] (see Box 10.5). The most extreme negative genetic interaction, often called “synthetic lethality,” occurs when the simultaneous mutation of two genes is lethal, while individually both mutations are viable. Conversely, the most extreme positive genetic interaction (“synthetic viability”) occurs, when a combination of two mutations is viable, while both individual mutations are lethal. Genetic interactions imply a functional relationship between the two genes, for example involvement in a common biological process or pathway, or conversely involvement in compensatory pathways with unrelated apparent function [46]. Hence, genetic interactions are an effective tool for biological discovery, e.g., for dissecting signaling pathways. They may also explain a considerable component of undiscovered genetic associations with human diseases and might help identify potential therapeutic targets. Over the last decade, genetic interactions have been investigated using mainly synthetic genetic array technology and RNA interference in yeast and *Caenorhabditis elegans*. A recent yeast based high-throughput screen [47], for example, tested all pairwise combinations of 6,000 genes resulting in almost 1 million interactions. Such maps can be used to study the large-scale organization of functions

in a cell [47], identify the hierarchical organization of specific biological processes [48], or generate hypotheses on the function of uncharacterized genes [49].

10.2.3 Disease Networks

Disease networks are a powerful framework for systematically investigating the diverse relationships among diseases. Such relationships exist on the molecular level (e.g., common genetic origin), on the phenotypic level (e.g., similar clinical manifestations) and on the population level (e.g., frequent co-occurrence in patients). A first comprehensive map of the human “diseaseome” was presented in [11], where 1,377 diseases were linked by shared genetic associations reported in the OMIM database [50] (see Box 10.6). The resulting network showed clearly that diseases can rarely be viewed as isolated quantities, each with a distinct genetic origin, but fall into highly connected clusters of disease groups with overlapping molecular roots. It was also found that diseases that are more central within the disease network tend to be more prevalent and have higher mortality rates [51]. The genetic overlap

Box 10.6: Disease networks



Disease networks in which diseases are linked if they share a genetic association are based on gene–disease association data that can be represented as a bipartite network (middle panel). This bipartite network can then be projected either onto the diseases, resulting in a disease–disease network (left) or onto a gene–gene network, in which links represent a common disease association.

among diseases also extends towards physical interactions among the respective gene products, as well as similar gene expression profiles.

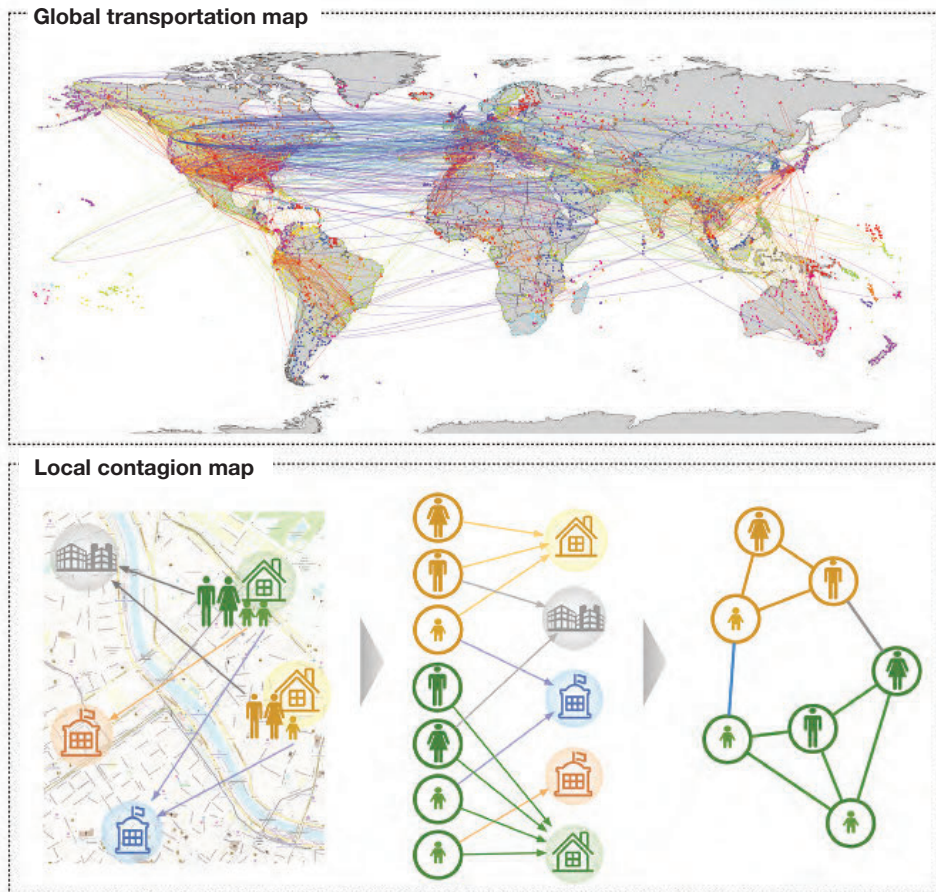
Similar results were obtained in a disease network in which diseases were linked by the similarity of their clinical manifestations [52] that were extracted from a large-scale screen of the biomedical literature and the annotated Medical Subject Headings (MeSH) metadata [53]. Confirming the strong correlation between the similarity of the symptoms of two diseases, the number of shared genetic associations and the extent to which their corresponding proteins interact, the study further revealed that the diversity of the clinical manifestations of a disease can be related to the degree of localization of the associated genes on the underlying protein interaction network. More detailed analyses that compared disease networks of different disease classes (e.g., complex diseases, Mendelian diseases, or cancer) and protein interaction networks identified interesting differences between diseases with different inheritance modes [54, 55, 56].

Networks can also be used to study comorbidity, i.e., the tendency of certain diseases to co-occur in the same patient. A disease network extracted from over 30 million patient records revealed that disease progression patterns of individual patients can be related to topological properties of the respective diseases within the co-morbidity network, for example, peripheral diseases tend to precede more central diseases [57]. These central, highly connected diseases are in turn associated with a higher mortality rate. More recently, differences in disease progression patterns that are related to age and sex have been characterized [58]. Co-morbidity networks have been used to address a wide range of further biomedical challenges, from drug repurposing [59] to the identification of potential drug side-effects [60], from biomarker identification [61] to approaches how to disentangle genetic and environmental factors of diseases [62].

10.2.4 Social Networks

A third important application of networks in medicine addresses the spread of contagious diseases, such as viral or bacterial infections (Box 10.7). Mathematical models of disease spreading go back as far as the year 1760, when Daniel Bernoulli formulated the first analytical method for quantifying the effectivity of inoculation against smallpox [64] (see Box 10.8 for an overview of important epidemiological models). Some 240 years later, the rise of complex networks made it possible to add a key ingredient to such models, namely realistic topologies of the networks on which diseases propagate, in particular global transportation maps and networks of social interactions [63] (see Box 10.7). Detailed information on interactions between humans on a local scale and on worldwide travel patterns is crucial for accurate predictions of the spatio-temporal spread of infectious diseases. Historically, the mobility of humans was largely confined by geography, such as rivers or mountains that could not be crossed easily. Such geographical borders naturally confined the propagation of epidemics. In present day, however, where both humans and goods can easily and quickly travel worldwide via air traffic, not even oceans can limit contagions [10]. As a consequence, an infection that started in a remote rural region may quickly propagate all over the world once

Box 10.7: Networks of disease spread



Global air traffic plays a major role in the spread of epidemic disease across the world. Locally, infectious diseases, but also personal traits like happiness or habits like smoking, are transmitted through social interactions. These interactions can occur, for example at home or at work, which can be represented as a bipartite network that can be mapped to a person-to-person network (illustration adapted from [63]).

it has reached an airport, leading to much faster, much wider, and seemingly more erratic patterns of global epidemics.

10.2.4.1 Transportation Networks

Network-based epidemiological models that incorporate the structure of worldwide transportation networks can shed light on the complicated propagation patterns observed in recent pandemic outbreaks, help identify the source of an outbreak, predict future highly affected areas, or design most effective immunization or prevention strategies [67, 68]. Examples for recent outbreaks of infectious diseases that were

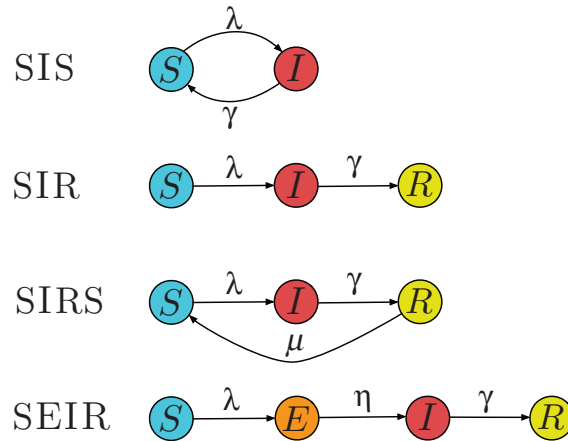
studied with the help of network models include the SARS pandemic in 2003 [69], the H1N1 outbreak in 2009 [70], the ebola crisis of 2015, or the spread of HIV in the Philippines [71].

Like many other real world networks, air-traffic networks have been found to be approximately scale free [72]. Scale-free networks are therefore the prime model for analytical studies of epidemic outbreaks and for the analysis of real data from past and current epidemics [73]. Important global properties of a pandemic are directly linked to the structure of the underlying networks. For example, the characteristic (super-) hubs of scale-free networks can often be identified with large airports that play an important role in the spread of a disease, both through the large number of people gathering at such airports and through the large number of destinations that they serve. Indeed, scale-free networks are generally more prone to global infections than more regular network structures that do not exhibit the “small world effect.” The critical spreading rate at which an infection is likely to propagate through the entire network is given by the ratio between the average degree and its variance. In large scale-free networks with degree distribution $P(k) \sim k^{-\gamma}$, the variance goes to infinity for power coefficients $\gamma < 3$. The critical spreading drops to zero in this case, meaning that a local infection is likely to become global, even for small infection rates [74].

10.2.4.2 *Social Contagion*

Approaches used to elucidate large-scale properties of infectious disease outbreaks can also be used to study the dynamics of social interactions, such as the spread of ideas, attitudes, and behaviors [75]. Reflecting the complexity of social relationships, links in social networks may represent, for example, friendship, family relationships, common work-place, shared political preferences, and many more. Collectively, these relationships not only define and shape our social relationships, but may also have concrete medical impact as shown in a seminal work on the spread of obesity [9]: The authors quantified how changes in body-mass index correlated among members of a social network of friends and family. Surprisingly, they found that obesity preferentially spreads through close social relationships. This effect is strong between men and between women, but almost negligible between man and woman. Similar studies were carried out to dissect the social component of starting to smoke [76] or of general happiness in life [77]. The results suggest that people surrounded by many happy people and those who are central in the network are more likely to become happy in the future. This effect was not observed among co-workers [77].

Recently there have also been efforts to combine global disease dynamics of transportation networks with contagion occurring on social networks. Multiplex or multilayer networks provide the analytical platform for combining several networks [78, 79, 80]. In such multilayer networks, different types of contact (at work, in the supermarket, at the airport) can be represented by distinct layers. It has been shown that the epidemic threshold is determined by the largest eigenvalue of the contact probability matrices of the different layers [78]. A powerful tool to study the full dynamics of spreading phenomena on networks, both simple or multilayered, are reaction diffusion processes [81].

Box 10.8: Basic mathematical models of disease spread

Classical epidemic models aim to determine the fraction of a population affected by a contagious disease over time. Most models represent the disease-status of an individual by one of three basic states [65]:

The **susceptible** (S) state, in which an individual can contract a disease. The **infected** (I) state, in which the individual carries the disease and can transmit it. The **recovered** (R) state, in which an individual is immune to repeated infections. More advanced models may also include further states, such as the **exposed** (E) state, in which an individual is already infected, but cannot yet transmit the disease. The microscopic dynamics of epidemiological models is given by transitions between the different states, macroscopic properties emerge from the interaction of many individuals. The most widely studied models are the following:

The **SIS model**, in which the recovery of a disease does not convey immunization, but renders an individual susceptible again, for example the common cold. The dynamics of the system are completely determined by the two rates of infection λ and recovery γ , respectively.

In the **SIR model** [66] susceptible individuals become infected with rate λ and recover with rate γ . This system exhibits an epidemic threshold $\alpha = \frac{\lambda}{\gamma}$, such that for $\alpha \leq 1$ a disease will die out in the long run, whereas for $\alpha > 1$ it will persist in the population.

The **SIRS model** contains an additional temporary immunity state, so that recovered individuals become susceptible again with rate μ . The impact of the incubation periods can be modeled by adding an exposed state (E), in which an individual has been infected, but is not yet infectious.

In network-based generalizations of these models, the individuals are identified with nodes and diseases spread along the connections of the network. In the simplest case this can be done by substituting the infection rate λ with a degree-dependent rate $\lambda = \lambda(k)$, so that the likelihood of becoming infected grows with the number of infected neighbors.

10.3 Interactome Analysis

As we have seen above, there exists a great variety of molecular interaction networks that can yield important insights into disease mechanisms. In the following, we will focus on “interactome networks” containing only physical interactions. The basic tools and concepts apply readily to other types of networks, however.

10.3.1 Interactome Construction

A large number of publicly available databases provide comprehensive collections of interactions between proteins and other relevant biomolecules (e.g. protein–DNA, protein–RNA, enzyme–metabolite interactions) in human, but also in other species, see [82] for a compendium of available resources. Among the most comprehensive, actively maintained and widely used databases are STRING [83], BioGRID [84], and MIntACT [85]. Note that they may also contain interactions that are not strictly physical, for example co-expression or other types of functional relationships among genes and their products. A well curated collection of only physical interactions has recently been published in the HIPPIE database [86]. Each interaction in HIPPIE is annotated with the original publication(s), details on the experimental protocol and an aggregated confidentiality score, thus allowing the user to adapt the final interactome network to specific requirements and preferences.

Generally, one can distinguish between three main sources of PPIs: (1) **interactions curated from the scientific literature** and typically derived from small-scale experiments, for example using co-immunoprecipitation, X-ray crystallography, or nuclear magnetic resonance. (2) **Interactions from systematic, proteome-scale mapping efforts**. The two main techniques are yeast two-hybrid (Y2H) assays [87] and binding affinity purifications coupled to mass spectrometry (MS) [88, 89], which produce rather different, yet complementary results (see Box 10.2). Y2H can map out precise, binary protein interactions, yet without biological context. It is not guaranteed, for example, that an experimentally observed interaction is biologically relevant, or whether the two respective proteins are in fact never expressed at the same time in the same cell. Co-complexes observed in MS experiments, on the other hand, are derived from a specific biological sample, yet are more difficult to translate into precise pairwise interactions [14]. (3) **Interactions from computational predictions**, for example based on protein structure [90] or other genomic data [91]. All three sources of PPIs have strengths and limitations in terms of comprehensiveness, noise and biases [92], such as biases in the selection of protein pairs [93] or experimental biases, for example towards highly expressed genes [87].

10.3.2 Basic Interactome Properties

Figure 10.1 gives a visual impression of a manually curated interactome from [16] and summarizes its global topological properties. In total, it contains 13,460 proteins connected via 141,296 physical interactions, so on average each protein has about 21 interaction partners. Characteristic not only to this, but also to many other complex

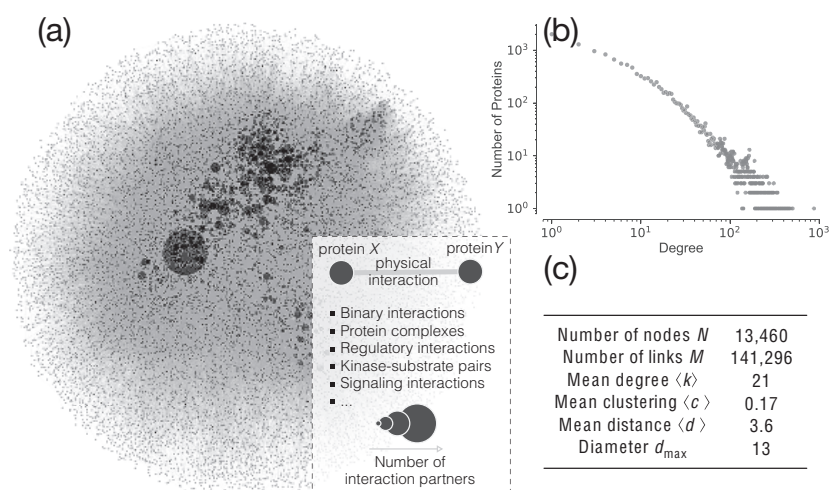
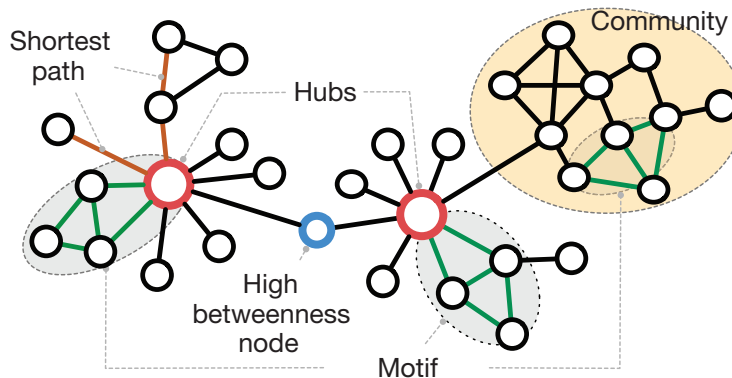


Figure 10.1: (a) A global picture of the interactome (original data curated by [16], figure adapted from [94]). The network consists of 13,460 proteins and 141,296 interactions that have been collected from different sources with various kinds of physical interactions, including binary interactions from systematic yeast two-hybrid screens, protein complexes, kinase-substrate pairs and others. (b) The overall topology is characterized by a highly heterogeneous degree distribution that follows approximately a power-law. (c) Other important structural properties of the interactome.

networks, is the high heterogeneity among the degrees of the nodes, i.e., in the number of connections they have to other nodes differs widely (see Box 10.9 for an overview of important terms in network science). While the vast majority of proteins have only few neighbors (more than 2,000 have only a single link), there is also a considerable number of nodes with hundreds of connections, such as *GRB2* (degree $k = 872$), *YWHAZ* ($k = 502$) and *TP53* ($k = 450$), so-called “hubs.” The histogram of all nodes’ degrees shows “scale-free” properties,¹ i.e., $P(k)$ follows approximately a power-law $P(k) \sim k^{-\gamma}$. As laid out in more detail in Chapter 3, the broad degree distribution and, as a consequence, the presence of hubs have a profound impact on many network properties. Hubs serve as shortcuts that connect distinct parts of the network, resulting in a network property often referred to as the “small world effect” [96] (in some cases of scale-free networks even “ultra-small” [97]). In the interactome, for example, it takes on average less than four steps ($\langle d \rangle = 3.6$) to reach any other protein from any given starting point. This high degree of connectedness is also associated with a remarkable resilience of the overall network structure against random failure of individual nodes and/or edges. Scale-free networks can maintain global connectedness even upon removal of a considerable fraction of nodes and edges [98, 99, 100, 101]. The flipside of this robustness towards random failure, however, is a particular vulnerability towards targeted attack against the hubs [102]. For the interactome, for example,

¹ How accurately this and other networks can be described by a power-law is subject to some debate, see [95] for a thorough discussion. For our purposes, however, the precise mathematical nature of the degree distribution plays only a secondary role.

Box 10.9: Basic topological characteristics of networks



- The degree of a node is the number its direct neighbors. The degree distribution across all nodes is an important global network characteristic.
- Scale free networks are characterized by a degree distribution that follows a power law: While most nodes have few neighbors, there are also a few highly connected hubs with a large number of neighbors.
- A path between two nodes is a sequence of links connecting the two. The minimum number of links needed to connect the two is called shortest path length and represents their network distance.
- Centrality measures quantify the topological importance of a node within the network. There are different types of centrality measures, the betweenness centrality, for example, quantifies how many shortest paths of the full network cross through a certain node.
- Clustering describes a tendency observed in many biological (and other) networks that two neighbors of a node are often also connected to each other, thus forming a triangle.
- Motifs are small recurrent subgraphs in a network that occur particularly frequently.
- Network communities are groups of tightly interconnected nodes that have more connections among themselves than to the rest of the network.

the removal of $\sim 30\%$ of the most highly connected nodes is sufficient to completely destroy the network, leaving only disconnected fragments.

10.3.3 Interactome Topology and Biological Function

The degree of connectedness of a protein is directly related to its biological importance: As first shown for the yeast *Saccharomyces cerevisiae* [103], and later confirmed also in

human cell lines [49], the products of essential genes, i.e., genes that are critical for the survival of an organism, tend to have a high number of interaction partners and take on central positions in the interactome. In contrast, genes whose loss of function can be more easily compensated for tend to have fewer interactions and are situated at the periphery of the interactome.

Interactome networks have also important structural features that go beyond the degree (or other measures of centrality) of individual nodes: “Network modules,” i.e. groups of nodes that are densely interconnected among themselves, but sparsely connected to the rest of the network, can often be identified with proteins that jointly perform a certain function [104, 105, 106]. This relation between functional similarity of genes (see ahead to Box 10.14) and their closeness in interactome networks has also been found for shared pathway membership, co-localization in the same cellular component or co-expression [87, 89]. The local aggregation of cellular function within interactome networks represents a fundamental biological organization principle that forms the basis for many important applications, ranging from the prediction of protein function to disease gene identification and drug target prioritization.

10.3.4 Diseases in the Interactome

The observation that functionally similar proteins are often densely interconnected can be generalized also to other relationships among genes, in particular to shared disease associations. Genes that are implicated in the same disease tend to have more interactions among each other than expected for completely randomly distributed genes [107]. Note, however, that this does not necessarily imply particularly densely interconnected network patterns as those observed for genes involved in the same function. Indeed, *dys*function is typically distributed among several, often only loosely connected functional modules within the interactome [108]. A systematic study on ~ 300 complex diseases showed that currently available interactome networks offer sufficient coverage to identify these “disease modules,” thereby confirming a fundamental hypothesis of interactome-based approaches to human disease [16]. The specific topological properties of disease modules differ between classes of diseases (e.g., complex diseases, Mendelian diseases, or cancer) and inheritance modes (autosomal dominant or recessive). Cancer driver genes are often highly central, while recessive disease genes tend to be more isolated at the periphery of the interactome [56].

10.3.5 Localization in Networks

As shown above, network-based localization of (dys)function is a central part of many interactome-based studies. In network science, the identification of densely connected groups of nodes is known as “community detection” [109]. While numerous algorithms exist for this task, they are usually not well suited for the identification of only weakly connected local network neighborhoods such as disease modules [108]. In order to quantify the tendency of a given set of disease genes to be localized in a certain neighborhood, we first need to inspect different possibilities for **measuring distances among a set of nodes in a network**. The simplest way to summarize the

localization of a set \mathbf{S} consisting of s nodes into a single quantity is to compute the network distance d_{ij} for all $\binom{S}{2} = \frac{s(s-1)}{2}$ pairs of nodes i and j and take the average:

$$d_{\text{av}}(\mathbf{S}) = \frac{2}{s(s-1)} \sum_{ij} d_{ij}, \quad (10.1)$$

which can be interpreted as a diameter of the set \mathbf{S} . As a consequence of the “small-world” nature of many relevant networks, differences in the absolute values of d_{av} for different gene sets are often relatively small. Several variations and extensions of Equation 10.1 have therefore been proposed [110]. For example, instead of taking the average over all possible node pairs, one can consider only the distance to the next closest node, respectively:

$$d_{\text{close}}(\mathbf{S}) = \frac{1}{s} \sum_i \min_{j \in \{\mathbf{S} \setminus i\}} (d_{ij}). \quad (10.2)$$

This gives different results as d_{av} in situations where a module is split into several “islands,” for example due to network incompleteness. Whereas d_{close} correctly reflects the high degree of localization within the individual islands, it is diluted when the distances of all pairs are averaged. Other variations include adding weights to different path lengths d_{ij} , see Box 10.10 for more examples. Complementary to such distance-based measures, one can also use **connectivity-based measures** to determine the degree of connectedness among a set of nodes. The simplest way is to consider the number of links between them. A perhaps more intuitive measure is given by the size of the largest connected component, i.e., the highest number of nodes that are directly connected to one another. We can apply tools from statistical physics to understand many of its properties analytically [111]. It is, however, relatively sensitive to data incompleteness. In extreme cases, a single missing link in the network or a missing node from the set \mathbf{S} , e.g., a protein, whose disease association is yet unknown, can fragment the connected component into isolated nodes.

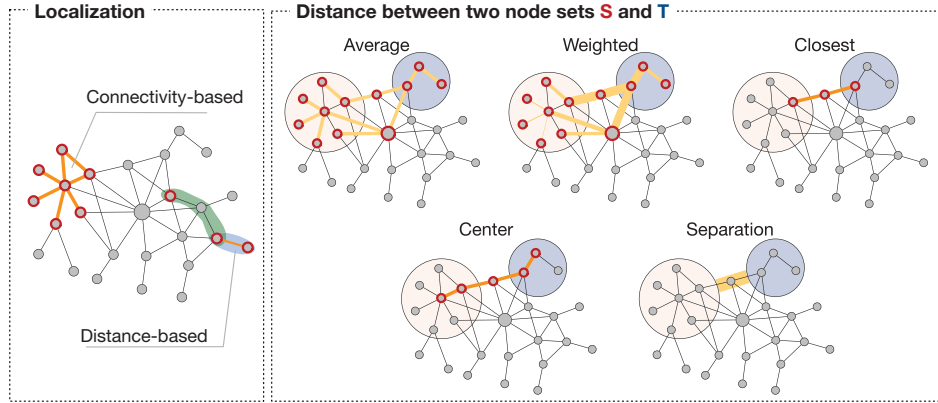
The concepts introduced above can be readily extended to measure distances between two node sets \mathbf{S} and \mathbf{T} , for example, for quantifying the interactome-based similarity between two diseases [16]. The equivalent of Equation 10.1, i.e., the average over all possible pairs of nodes between two node sets is given by

$$d_{\text{av}}(\mathbf{S}, \mathbf{T}) = \frac{1}{s} \sum_{i \in \mathbf{S}} \frac{1}{t} \sum_{j \in \mathbf{T}} d_{ij}. \quad (10.3)$$

Similarly to different linkage methods in hierarchical clustering algorithms, there are different ways to compute the distance between two sets of nodes, see Box 10.10 for a number of frequently used options.

10.3.6 Randomization of Network Properties

By themselves, the absolute values of localization or distance as introduced above bring few insights. To judge whether an observed clustering of a particular node set is significant, we need to compare it to suitable random models. Many quantities that

Box 10.10: Distance measures in networks

There are different ways to quantify the degree of “localization” of a given set of nodes S , i.e., whether or not they aggregate in a certain network neighborhood. Distance-based localization measures are based on different averages over pairwise distances d_{ij} between all nodes in the set, e.g.:

$$d_{\text{av}}(S) = \frac{2}{s(s-1)} \sum_{ij} d_{ij} \quad (10.4)$$

$$d_{\text{close}}(S) = \frac{1}{s} \sum_i \min_{j \in S \setminus i} (d_{ij}) \quad (10.5)$$

$$d_{\text{exp}}(S) = -\frac{2}{s(s-1)} \ln \sum_{ij} \exp(-d_{ij}) \quad (10.6)$$

These measures can be generalized to two node sets S and T :

$$d_{\text{av}}(S, T) = \frac{1}{s} \sum_{i \in S} \frac{1}{t} \sum_{j \in T} d_{ij} \quad (10.7)$$

$$d_{\text{close}}(S, T) = \frac{1}{s+t} \left[\sum_{i \in S} \min_{j \in T} (d_{ij}) + \sum_{i \in T} \min_{j \in S} (d_{ij}) \right] \quad (10.8)$$

$$d_{\text{exp}}(S, T) = -\frac{1}{s} \sum_{i \in S} \frac{1}{t} \ln \sum_{j \in T} \exp(-d_{ij}) \quad (10.9)$$

Nodes that are common to both sets S and T are usually taken to contribute with $d_{ij} = 0$ in the above formula. Instead of averaging over all pairs of nodes between S and T one can also define a center for each and use the distance between them:

$$d_{\text{center}}(S, T) = d(\text{center}(S), \text{center}(T)) \quad (10.10)$$

Another option is the separation parameter introduced in [16]:

$$\text{sep}(S, T) = d_{\text{close}}(S, T) - \frac{1}{2}(d_{\text{close}}(S) + d_{\text{close}}(T)) \quad (10.11)$$

Negative values $\text{sep}(S, T) < 0$ suggest overlapping network modules, while $\text{sep}(S, T) > 0$ indicates separated modules. Note, however, that the separation parameter is not an intensive quantity, i.e., its magnitude depends on the number of nodes in the respective sets.

occur in the context of network analyses do not follow normal (Gaussian) distributions, such as the scale-free degree distribution, and therefore require particular care when choosing statistical tests. Comparisons with ensembles of randomized networks obtained from simulations are often the best choice. In general, we can distinguish two types of randomizations: (1) **Randomizing the network topology**, for example the interaction partners of a particular protein, and (2) **randomizing node attributes**, such as the disease associations of a group of genes.

10.3.6.1 *Randomizing the Network Topology*

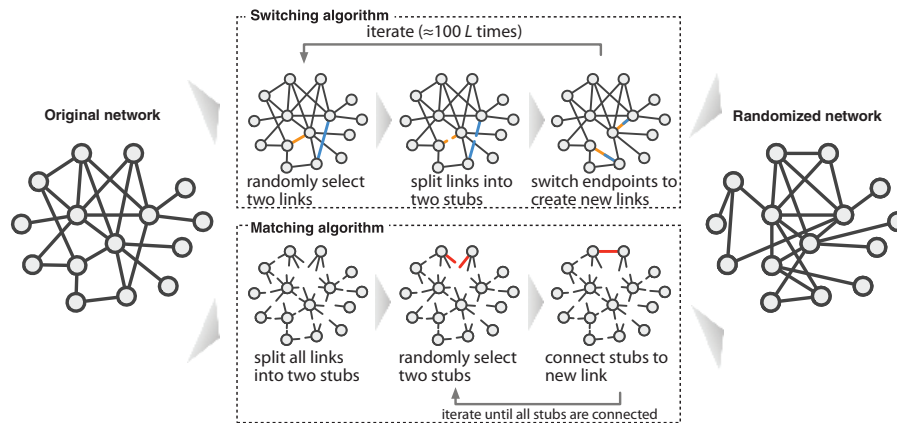
To exclude that a seemingly interesting observation, for example, the local aggregation of disease genes in the interactome, could be a generic consequence of the overall topology of the underlying network, we need to compare our results from the original network with those obtained from networks with randomized topology. There are numerous randomization procedures. Which one is most suited, depends on the particular reference that is needed for a specific observation. The simplest method is to fix only the number of nodes N and the number of links L of the original network and to redistribute the links completely at random among the nodes. As shown in Chapter 3, this procedure results in an Erdős-Rényi network. Many properties of Erdős-Rényi networks can be calculated analytically and without extensive computer simulations, for example the expected clustering or the size of the largest connected component. However, the topology of most real world networks differs substantially from the one of a corresponding complete random graph, for instance hubs are completely absent in the latter. Hence, comparisons between the two are rarely meaningful and can in fact be rather misleading.

A more adequate reference that is suitable for most applications is given by networks in which the number of neighbors of every node are kept constant, but the specific interaction partners are completely randomized. This ensures that important structural features, in particular the degree distribution and presence of hubs, are preserved in the ensemble of randomized networks. Box 10.11 introduces the two main algorithms that are used to generate such randomized networks: The “switching algorithm” [112], is an iterative method, where at each step two links are selected at random and their endpoints are swapped. For example, the links connecting the nodes $n_1 \leftrightarrow n_2$ and $n_3 \leftrightarrow n_4$, respectively, can be reconnected to $n_1 \leftrightarrow n_3$ and $n_2 \leftrightarrow n_4$. Note that this may result in multiple links between two nodes or self-loops. In an application where such links are not meaningful, the original link pairs should be restored. As we repeatedly apply this procedure, the interactions of the network become more and more randomized, without altering the degree of each node. A drawback of this simple method is that no precise criteria exist as to how many switches should be performed to ensure a good mixing. Empirical results suggest $100L$ switching attempts, which can be computationally rather expensive for large networks [113].

A more efficient method for generating random networks with a prescribed degree sequence is to apply a variation of the “configuration model” [114, 115]. The second algorithm introduced in Box 10.11 is the “matching algorithm,” in which all links of a given network are broken at once and then randomly reassembled one by one. As in the switching algorithm, the potential creation of self-loops and multiple links may

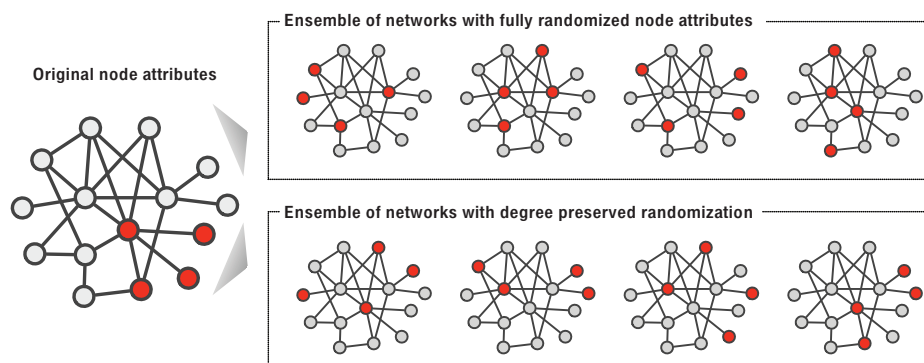
Box 10.11: Network randomization

Randomizing the network topology



There are two frequently used algorithms to generate an ensemble of randomized networks with fixed degree distribution. In the *switching algorithm*, two links are chosen at random and their endpoints switched. Repeating this procedure will eventually lead to a fully randomized version of the original network. In the **matching algorithm**, all links of the given network are broken at once and then one by one reconnected at random.

Randomizing node attributes



The most basic procedure to randomize node attributes (e.g., disease associations of genes) is to redistribute them completely at random on the network. For more restricted random controls, one can also keep specific topological properties of a node attribute constant, in particular the degree of the annotated node. In this case, only nodes with the same (or at least similar) properties are allowed choices.

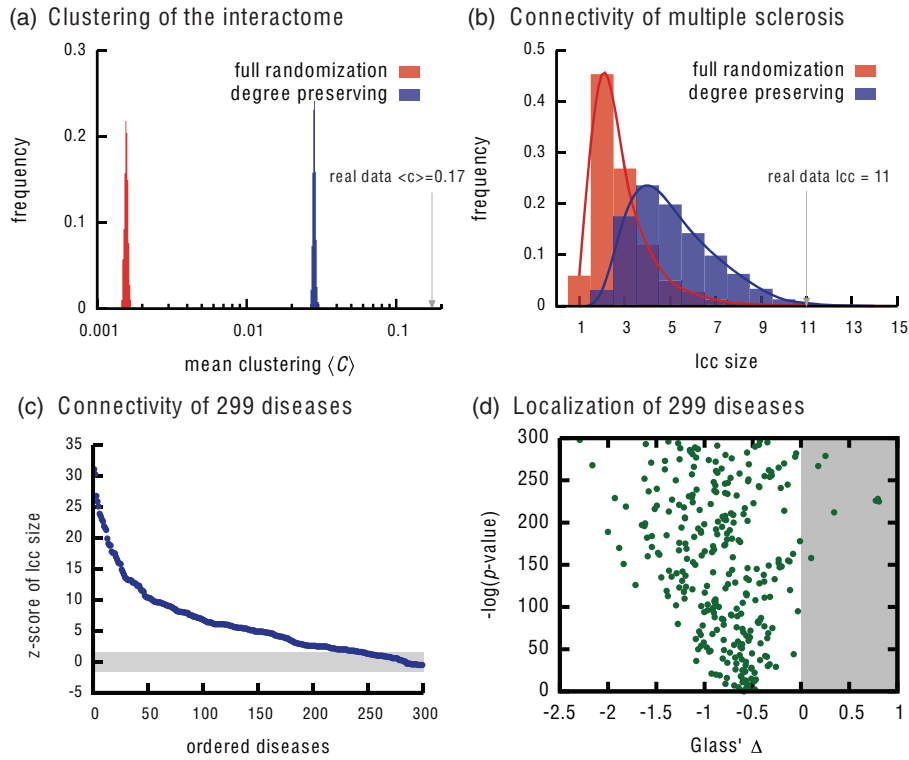


Figure 10.2: Network randomization. (a) Comparison of the clustering coefficient of the interactome (see Figure 10.1) with distributions obtained from complete randomization and degree-preserving randomization. (b) Comparison of the size of the largest connected component (lcc) of proteins associated with multiple sclerosis in the interactome with two distributions obtained from full and degree preserving randomization, respectively. (c) Sorted z-scores of the lcc size of 299 diseases in the interactome. (d) Significance and effect size of the observed localization $d_{av}(\mathbf{S})$ of 299 diseases compared to randomized gene sets. (Data from [16].)

need to be prevented in certain applications. Note that in this case the ensemble of the generated networks is no longer completely unbiased, but the effects are usually small and can often be neglected for large networks [113].

Figure 10.2a shows an application of the two randomization strategies to evaluate the observed mean clustering coefficient $\langle C \rangle = 0.17$ of the interactome. As expected, we find excellent agreement between the values observed in 10,000 simulations of a full random model corresponding to an Erdős-Rényi network and the respective analytical value $\langle C \rangle = p = \frac{2L}{N(N-1)} = 0.0016$. Simulations of the degree preserving matching algorithm yield the considerably higher mean value $\langle C \rangle = 0.03$, which is still significantly smaller than the originally observed clustering, indicating that the clustering of the interactome could not have emerged by chance.

10.3.6.2 Randomizing Node Properties

Instead of rewiring the structure of the network itself, it is often useful to consider randomizing certain node attributes, for example disease associations of individual genes in the interactome. In the simplest case of **random label permutation**, we detach the attribute of interest from their original nodes and redistribute them completely at

random among all nodes of the network. For example, to investigate the connectivity of N_d disease proteins in terms of their largest connected component (lcc), we select the same number of proteins randomly from the network and measure their lcc. Repeating this procedure yields a random control distribution that can then be used to determine the statistical significance of the original lcc. According to data from [16], multiple sclerosis has $N_d = 69$ known associated proteins in the interactome that form an lcc of size $S = 11$. Figure 10.2 (b) shows the lcc distribution for 69 randomly picked proteins from 10,000 simulations. The distribution has a mean of $\langle S_{\text{rand}}^{\text{full}} \rangle = 2.9$ and a standard deviation of $\sigma = 1.4$. The statistical significance of the observed lcc size can be quantified using the z-score

$$\text{z-score} = \frac{S - \langle S_{\text{rand}}^{\text{full}} \rangle}{\sigma}, \quad (10.12)$$

yielding $\text{z-score} = 5.8$. For normal distributions, z-scores > 1.65 correspond to a p -value < 0.05 (corresponding to a right-sided test, left- or two-sided tests are also possible) and are considered to be statistically significant. The empirical p -value, i.e., the fraction of all random simulations with $S_{\text{rand}}^{\text{full}} \geq S$ was found to be $p\text{-value} = 0.003$. Taken together, we conclude that the connected component for multiple sclerosis is unlikely to have emerged by chance or as a trivial consequence of the network topology, indicating the potential presence of a disease module.

10.3.6.3 Degree Preserving Label Permutation

There are also stricter attribute randomization procedures that impose certain constraints on the allowed set of nodes among which an attribute can be distributed. Prominent cancer genes, for example, tend to have a large number of interactions in literature-curated interactome networks, simply because they have been investigated more intensively than other genes. To test whether the high connectivity among such genes can be explained by their high degree alone, we need to generate random distributions of node attributes that maintain the degree of the individual nodes carrying the original annotation. Note that swapping only between nodes of exactly the same degree will be problematic for high-degree nodes, as there may be only few, or even a single node in the entire network that have a certain degree. It is therefore useful to relax the requirement of having exactly the same degree and work with bins of nodes with comparable degree instead. Figure 10.2 (b) shows the distribution $S_{\text{rand}}^{\text{degree}}$ obtained using such an approach. The mean value $\langle S_{\text{rand}}^{\text{degree}} \rangle = 5.1$ is larger than the one obtained from the full randomization, but still significantly smaller than the value $S = 11$ from the original data ($\text{z-score} = 3.1$, empirical p -value = 0.009), indicating that the high degree of the disease proteins alone does not explain their observed high connectivity.

These randomization procedures can also be applied to evaluate the distance-based localization measures introduced above, for example $d_{\text{av}}(\mathbf{S})$. From each random simulation we can extract $d_{\text{av}}^{\text{rand}}$ and then compute the mean $\langle d_{\text{av}}^{\text{rand}} \rangle$ and corresponding standard deviation $\sigma(d_{\text{av}}^{\text{rand}})$. In analogy to the z-score introduced above, we can use Glass' Δ to quantify the effect size of any difference observed between

the true value $d_{av}(\mathbf{S})$ and the values obtained in the respective randomization simulations:

$$\Delta = \frac{d_{av}(\mathbf{S}) - \langle d^{rand} \rangle}{\sigma(d^{rand})}. \quad (10.13)$$

The statistical significance of an observed difference in the respective means $d_{av}(\mathbf{S})$ and $\langle d^{rand} \rangle$ can be obtained from a Mann–Whitney U test, for example. Figure 10.2 (c–d) shows the results for the randomization valuation of the localization observed among 299 diseases on the interactome.

Numerous more advanced randomization procedures exist that can preserve topological features beyond the degree distribution. For example, there are algorithms to generate randomized networks that maintain the mean clustering coefficient of the original network [116] or the correlation structure between the degrees of adjacent nodes [117, 118]. Another level of sophistication needs to be applied when randomizing metabolic networks, where simple link rewiring would likely generate reactions that are biochemically impossible [119, 120].

10.4 Disease Module Analysis

10.4.1 Overview

Sequencing technology has accelerated the discovery of disease associated genetic variations significantly. For most diseases, however, we are still far from a complete understanding of the underlying molecular mechanisms. Most complex diseases, such as cardiovascular diseases, cancer, or diabetes mellitus (the three most frequent causes of death worldwide), involve hundreds of genes and their complex interactions. It has been estimated, for example, that more than 2,000 genes are involved in intellectual disabilities, yet our current knowledge includes only around 800 genes [121]. The situation is similar for rare Mendelian disorders. Estimates for the total number of rare genetic disorders range from 6,000 to 8,000, a majority of which likely to be caused by a single genetic aberration. Despite this simple genetic architecture, less than half of all suspected diseases and corresponding disease genes are currently known.

Network-based **disease modules** offer a general framework for investigating how the pathobiology of a particular disease may arise from a combination of many genetic (but also epigenetic, environmental, behavioral etc.) variations. Successful applications range from rare Mendelian disorders [3], to cancer [4] and other complex disorders, like metabolic [5], inflammatory [42], or developmental diseases [122]. A disease module is loosely defined as the comprehensive set of cellular components associated with a certain disease and their interactions. More specifically, the term refers to a connected subgraph of the interactome, whose perturbation causes the disease [18]. Figure 10.3 gives an overview of the disease module analysis process. The first step is to construct an interaction network and collect genes known to be associated with the particular disease of interest. These “seed genes” will serve as starting point for network-based gene prioritization algorithms. The resulting network module can then be validated and enriched with various additional datasets that will also be used in the biological interpretation of the final disease module.

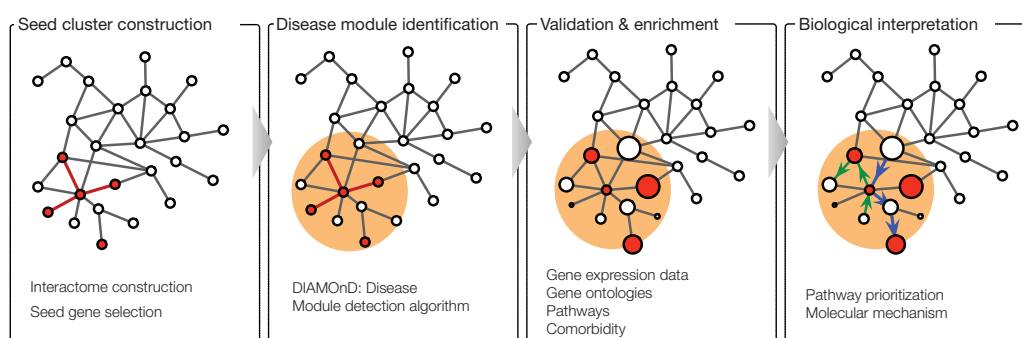


Figure 10.3: The basic steps of a disease module analysis process: First, interactome and seed gene data are collected. Next, a network-based disease gene prioritization method is employed. The performance of the predictions is then validated through comparison and enrichment with independent external data. In the last step, the module is explored for important biological pathways, overlap with other disease modules etc. (Figure adapted from [123].)

10.4.2 Seed Cluster Construction

The first step of the disease module analysis is the construction of a seed cluster, i.e., the curation of a suitable molecular interaction network and a set of genes known to be associated with the particular disease of interest. Box 10.12 lists a number of resources that may serve as a starting point.

10.4.2.1 Interactome Construction

As introduced above, one can make a broad distinction between physical interactions, e.g., protein co-complexes or binary protein–protein interactions, and functional interactions, e.g., genetic interactions or co-expression. By definition, physical interactions represent a direct molecular relationship, thus facilitating the identification of causal molecular mechanisms. Functional interactions, on the other hand, offer a much broader spectrum of potentially relevant associations between genes and gene products and can often be more easily adapted to a particular diseases, for example by incorporating tissue-specific expression data. Incorporating such information can considerably improve disease gene prioritization [124, 125, 126], see also Chapter 11. The choice of interaction type and used data sources will affect coverage (number of contained genes/proteins and their interactions), biases (for example, towards well-studied genes) and signal to noise ratio (number of false positive interactions) of the final interactome. Physical interactions offer more control over biases and signal to noise ratio, but often at the cost of lower coverage. Biases can be reduced by relying only on data obtained from systematic high-throughput studies, e.g., from [87, 89]. False positive interactions can be reduced by filtering for interactions that have been reported by several studies and by different experimental techniques. Several databases, such as HIPPIE [86] or STRING [83] offer integrated interaction scores for this purpose.

Box 10.12: Resources for disease module analyses**Interactome databases:**

BIOGRID	thebiogrid.org
BioPlex	bioplex.hms.harvard.edu
HIPPIE	cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/
IntAct	www.ebi.ac.uk/intact
MatrixDB	matrixdb.univ-lyon1.fr
MINT	mint.bio.uniroma2.it
STRING	string-db.org

A more comprehensive list can be found on EBI's PSICQUIC view that also offers programmatic access, see www.ebi.ac.uk/Tools/webservices/psicquic/view/

Disease genes:

DGA	dga.nubic.northwestern.edu
GWAS Catalog	www.ebi.ac.uk/gwas
Gene2Mesh	gene2mesh.ncibi.org
HGMD	hgmd.cf.ac.uk
OMIM	omim.org
OrphaNet	www.orpha.net

Integrated and functional web-based services:

DisGeNet	disgenet.org
GeneMANIA	genemania.org
HumanBase	hb.flatironinstitute.org

Ontologies:

Disease ontology (DO)	disease-ontology.org
Gene ontology (GO)	www.geneontology.org
Human phenotype ontology (HPO)	human-phenotype-ontology.github.io
Mammalian phenotype ontology (MPO)	www.informatics.jax.org/vocab/mp_ontology

A comprehensive list of biological ontologies can be accessed from EBI's Ontology Lookup Service under <https://www.ebi.ac.uk/ols/ontologies>

10.4.2.2 *Seed Gene Selection*

There are numerous resources that collect genes associated with diseases (see Box 10.12). Note that the term “disease associated gene” itself is only loosely defined and covers a wide spectrum from high penetrance dominant mutations to GWAS variants of rather small effect size or genes observed to be differentially regulated in patient subgroups. Similarly, the level of evidence for reported disease associations may differ greatly, from rare gene variants with a known and experimentally validated functional mechanism, to genes with unknown mechanism, yet repeatedly confirmed in multiple patient cohorts, to rather speculative associations inferred solely from text mining.

10.4.2.3 *Evaluation of the Seed Cluster*

Both the interactome construction and the seed gene selection involve a certain trade-off between using only highest-confidence data and achieving the highest possible coverage. There is no simple and universally applicable solution to this challenging problem that requires a certain amount of experimentation, ideally guided by a domain expert for the specific disease under study. From a network perspective, however, localization measures introduced above can be used as a rough indicator whether a particular combination of interactome and seed gene data meets the minimal criteria for a meaningful disease module analysis. Figure 10.4 shows the seed cluster for an asthma disease module from [123]. From a total of 129 seed genes that could be mapped to the interactome, 37 form the largest connected component, indicating a highly significant (z -score = 10.7) network localization. This suggests that the seed cluster has sufficient “signal” pinpointing the network neighborhood of the complete asthma module that can then be identified through a network-based expansion algorithm.

10.4.3 **Network-Based Disease Gene Prioritization**

Network-based disease gene prioritization methods build on the observation that genes associated with the same disease tend to be localized in the same interactome neighborhood. We can therefore use the network topology to extrapolate from a given set of seed genes to identify other genes that are likely to be also involved in the disease or at least strongly affected by the local interactome perturbation. Over the last years, numerous algorithms have been developed for this purpose. They can be broadly classified into three major categories: (1) connectivity based methods (2) path-based methods and (3) diffusion-based method (see Box 10.13).

10.4.3.1 *Connectivity-Based Methods*

Connectivity-based methods exploit the observed propensity among disease genes to interact with each other. Early pioneering approaches considered all direct neighbors of seed genes as potential candidate genes [127]. As more and more interactome and seed gene data become available, such approaches tend to generate an increasing number of false positives. More recent algorithms therefore utilize more advanced connectivity patterns, such as graphlets [128], or take the degree heterogeneity of the interactome explicitly into account [129]. Indeed, hubs in the network are expected to

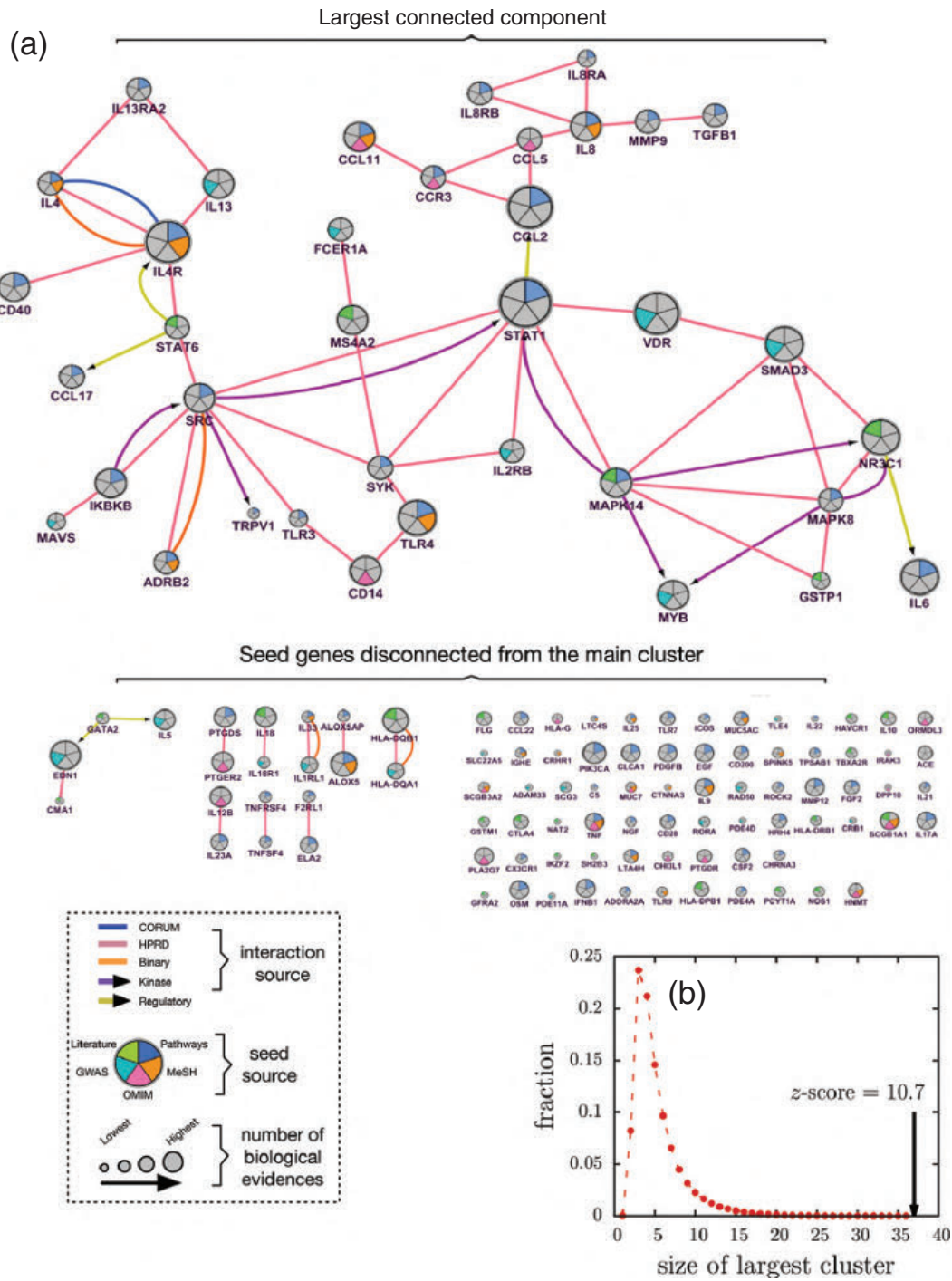


Figure 10.4: Seed cluster of an asthma disease module analysis from [123]. (a) Of the 129 expert curated seed gene, 37 form the largest connected component, the rest are scattered throughout the interactome. (b) The size of the largest connected component is highly significant (z -score = 10.7) compared to random expectation.

Box 10.13: Network-based disease gene prioritization

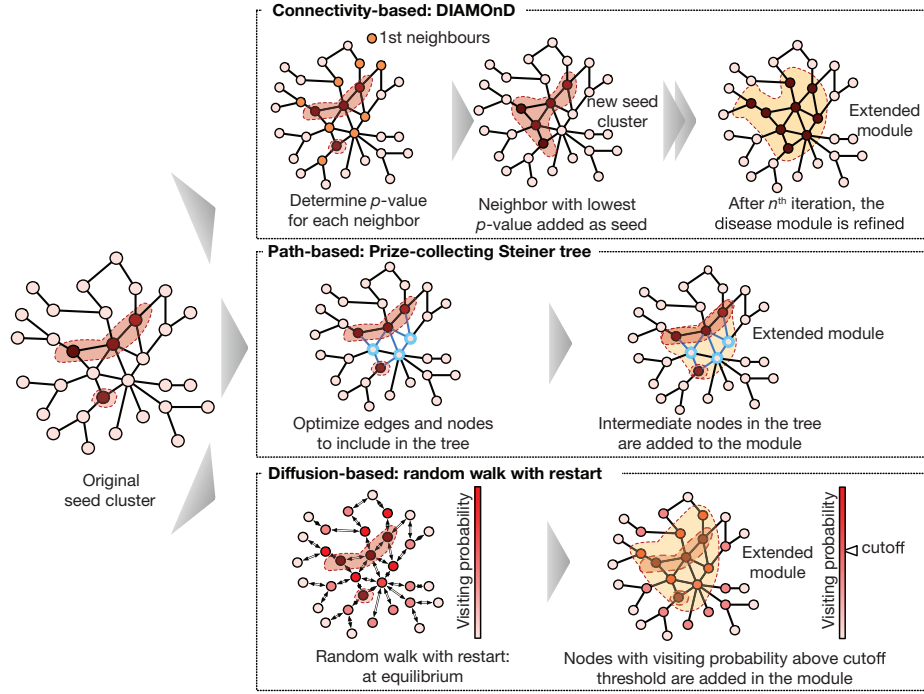


Illustration of three different methodologies for network-based disease gene prioritization: (1) **Connectivity-based** methods evaluate the direct neighbors of seed genes. (2) **Path-based** methods evaluate candidate genes based on their network distance to seed genes. (3) **Diffusion-based** methods use a dynamical process to rank candidate gene according to how strongly they are influenced by the seed genes.

also interact with a large number of seed genes without necessarily implying a disease-association. To correct for these effects, the DIAMOnD algorithm [108, 123] evaluates the *significance* of a given number of connections k_s to s seed genes with respect to the total degree k of a given candidate gene. In a network of size N , with s randomly distributed seed genes, the probability that a gene with degree k connects to exactly k_s seed genes is given by the hypergeometric distribution

$$P(X = k_s) = \frac{\binom{s}{k_s} \binom{N-s}{k-k_s}}{\binom{N}{k}}. \quad (10.14)$$

The significance of a given number of connections is therefore given by the p -value

$$p\text{-value} = \sum_{n=k_s}^k P(X = n), \quad (10.15)$$

which can then be used to iteratively rank all genes in the network. Note that the resulting disease module may consist of genes without direct connectivity to the initial seed genes.

10.4.3.2 Path-Based Methods

Instead of using the direct connectivity to seed genes, candidate genes can also be ranked according to their network distance to the set of seed genes (compare also with Box 10.10). A versatile set of algorithms that combines different distance measures for prioritizing candidate genes has been proposed in [130]. Instead of ranking the genes iteratively, it is also possible to search for an optimal set of candidate genes that collectively minimize the path lengths between the seed genes. Such approaches often implement variations of minimum spanning tree (or “Steiner tree”) search algorithms [131, 132, 133]. Basically, the algorithm will construct a tree consisting of a minimum amount of edges while connecting all the seeds into a single cluster.

10.4.3.3 Diffusion-Based Methods

The methods described above rely only on the static topology of the network. It is also possible to use dynamical models to explore the network neighborhood around the seed genes for gene prioritization [3, 4, 134, 135, 136, 137]. Among the most widely used dynamical models are diffusion processes, such as the random walk with restart (RWR) [138]: Here, the seed genes serve as starting points for a random walk process along the links of the network. At every time step, the walker either proceeds to a randomly picked neighboring gene, or returns with restart probability r to one of the seed genes. The restart ensures that the local neighborhood around the seed genes is emphasized by the walker, otherwise all seed gene information would be lost in the long run of the process. The frequencies with which the individual nodes in the network are visited will eventually converge to a steady state and can then be used to rank all genes in the network according to their “dynamical closeness” to the seed genes. The process can be formalized as follows: Consider the vector \mathbf{p}_t whose elements $p_i \dots p_N$ represent the probability of the walker visiting node i at time t . The visiting probability at time t can be derived from the visiting probability at time $t - 1$ via

$$\mathbf{p}_t = \mathbf{W}\mathbf{p}_{t-1}, \quad (10.16)$$

where \mathbf{W} is the so-called transition matrix and defined as the column normalized adjacency matrix \mathbf{A} with $W_{ij} = \frac{A_{ij}}{\sum_i k_i}$. At time t_0 , only seed genes have (uniform) non-zero probability p , as well after each restart, which happens at a rate r . Equation 10.16 then becomes

$$\mathbf{p}_t = (1 - r)\mathbf{W}\mathbf{p}_{t-1} + r\mathbf{p}_0. \quad (10.17)$$

The steady-state solution for Equation 10.17 is given by

$$\mathbf{p}_\infty = r(\mathbf{I} - (1 - r)\mathbf{W})^{-1}\mathbf{p}_0. \quad (10.18)$$

The genes in the network can then be ranked according to the visiting probability p_∞ . The restarting probability r can be used to adjust the influence of the seed genes on the

diffusive process, from free diffusion (walker is not restricted by seed genes, $r = 0$) to no diffusion at all (walker remains at seeds, $r = 1$).

10.4.4 Validation and Enrichment

After completion of the preferred candidate gene ranking procedure, we first need to evaluate its performance. A second, closely related task is to determine a sensible cutoff, i.e. how many ranked genes should be considered for the final disease module, as most prioritization methods rank all genes in the network without offering an intrinsic stopping criterion. There are two complementary approaches: (1) Estimating the predictive power of the disease gene predictions using cross-validation methods. (2) Comparison with independent biological data.

10.4.4.1 *Cross-validation of Prediction Performance*

In principle, cross-validation of disease gene prioritization algorithms works in the same way as with other classification tasks (compare also with Chapters 6–8): For a basic k -fold cross-validation, the set of seed genes is first randomly divided into k groups (the special case where k equals the number of seed genes is often referred to as “leave-one-out” cross-validation). One of the groups can then serve as the “test-set” of true positives, while the remaining $k - 1$ groups are used as modified seed gene pool. The gene prioritization algorithm is then run on this modified pool to test how well the method is able to retrieve the left out genes in the test set. Repeating this procedure k times with each of the k groups serving as test set yields a statistic on the expected average performance of the method. The choice of k determines the trade-off between high bias (large k) and high variance (small k). An important difference to many other classification tasks is the lack of clear true negatives, i.e., genes that we know not to be involved in the disease. Several proxies have been proposed, for example essential genes, genes of high genetic variability or manually curated genes that are unlikely to be involved in a particular disease according to their expression patterns. These gene sets can only offer approximations and remain necessarily incomplete, making the interpretation of standard performance measures difficult, such as receiver operating characteristic curves.

10.4.4.2 *Enrichment with Independent Biological Data*

A complementary approach for estimating the performance is to test for enrichment of the ranked genes with independent biological data (see Box 10.12). Figure 10.5 shows the biological enrichment of the top 400 ranked genes from an asthma disease module analysis [123]. To compare the biological signal of the ranked genes with the one of the manually curated seed genes, the authors chose a sliding window of ranked genes with the same size of the seed genes and within each window computed the enrichment with five different datasets: (1) Genes differentially expressed in a relevant case/control study, (2) genes participating in expert curated relevant biological pathways, (3) genes contained in general pathways that were found enriched in the seed genes, (4) genes annotated to similar biological processes as the seed genes according to the gene ontology (GO, see Box 10.14) and (5) genes that are known to be implicated in

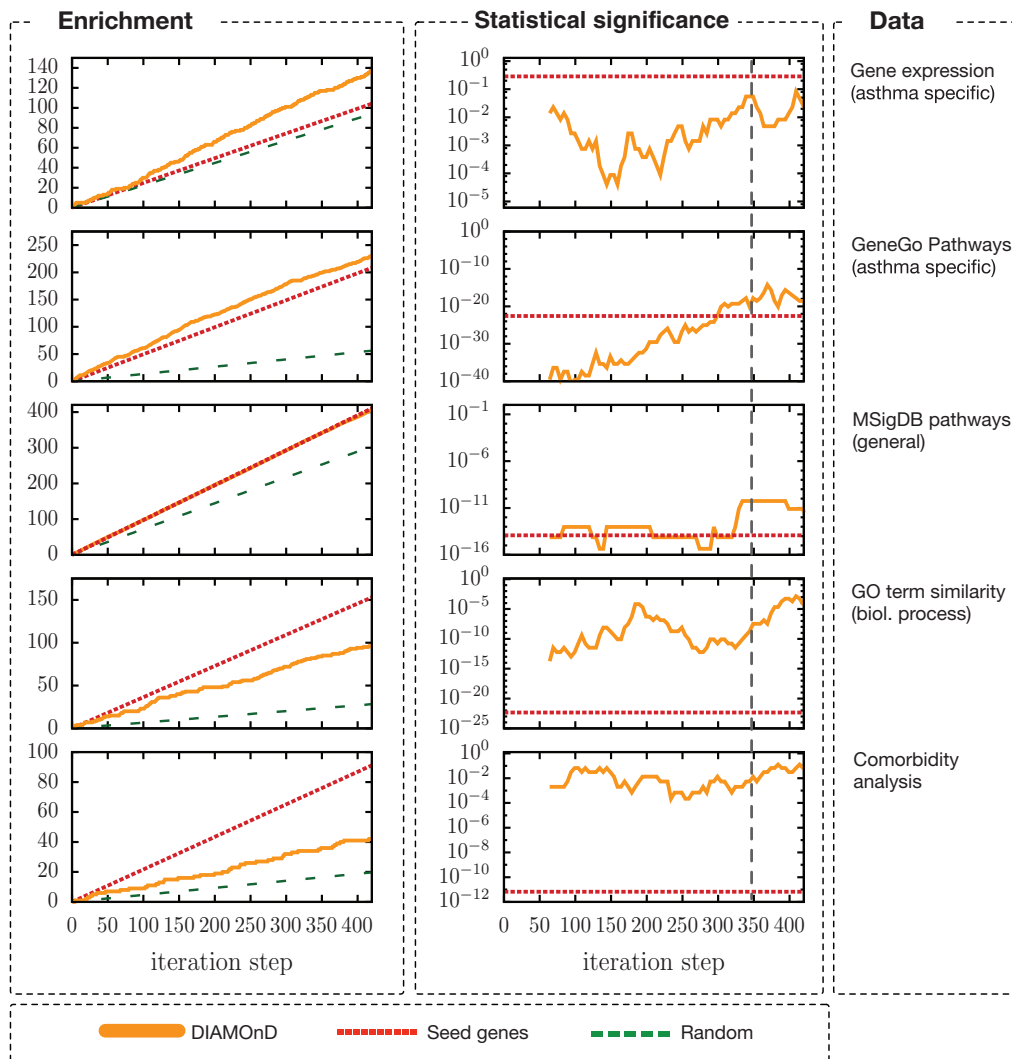
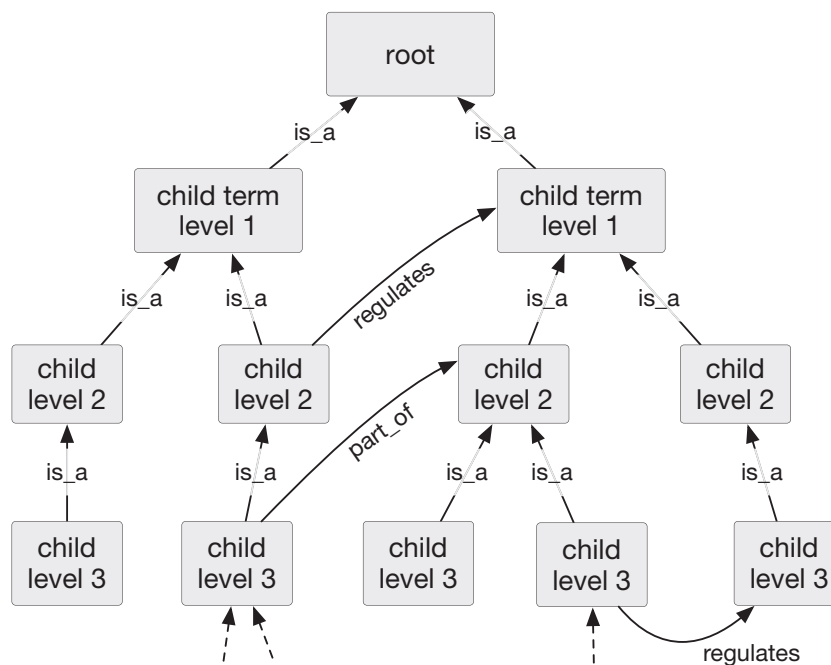


Figure 10.5: Biological enrichment of the asthma disease module in [123]. The first two columns show the number (and the corresponding statistical significance, respectively) of the identified candidate genes that were found in the different validation datasets indicated in the third column. The values for the candidate genes are shown in orange, the values for seed genes and random expectation in red and green, respectively.

diseases that show high co-morbidity with asthma. A comparison of the enrichments across different datasets allows for an evaluation of the general plausibility of the ranked genes, but also for an estimation of the border of the disease module.

10.4.5 Biological Interpretation

The data collected for the performance evaluation can further be used for an integrated analysis of the biological mechanisms represented in the disease module. Figure 10.6

Box 10.14: Ontologies

Ontologies are controlled vocabularies to organize the knowledge of a specific field, for example biological pathways, diseases or phenotypes (see Box 10.12 for a list of biomedical ontologies). These vocabularies are usually manually curated by an authoritative consortium of domain experts. An important vocabulary is the gene ontology (GO). It consists of three separate branches: (1) “cellular component” (4,195 terms), (2) “molecular function” (11,120 terms), and (3) “biological process” (29,682 terms), each forming a hierarchical, acyclic tree. The root term at the top is the most general, increasingly specific terms are connected by either **is_a**, **part_of** or **regulates** links that describe the particular relationship between the respectively linked terms.

Ontologies are not only useful for systematic annotation and collection of knowledge, but can also be used to assess the “semantic similarity” among different terms according to their relative position in the tree [139]. A common approach relates the specificity (tree depth) of a term to its information content (IC). The similarity between two terms can then be calculated from the IC of their most informative (i.e., highest IC) common ancestor. Note that most biological entities, such as gene products, are usually annotated with several terms and different strategies can be used to aggregate the similarity among several terms, see [139] for a detailed discussion.

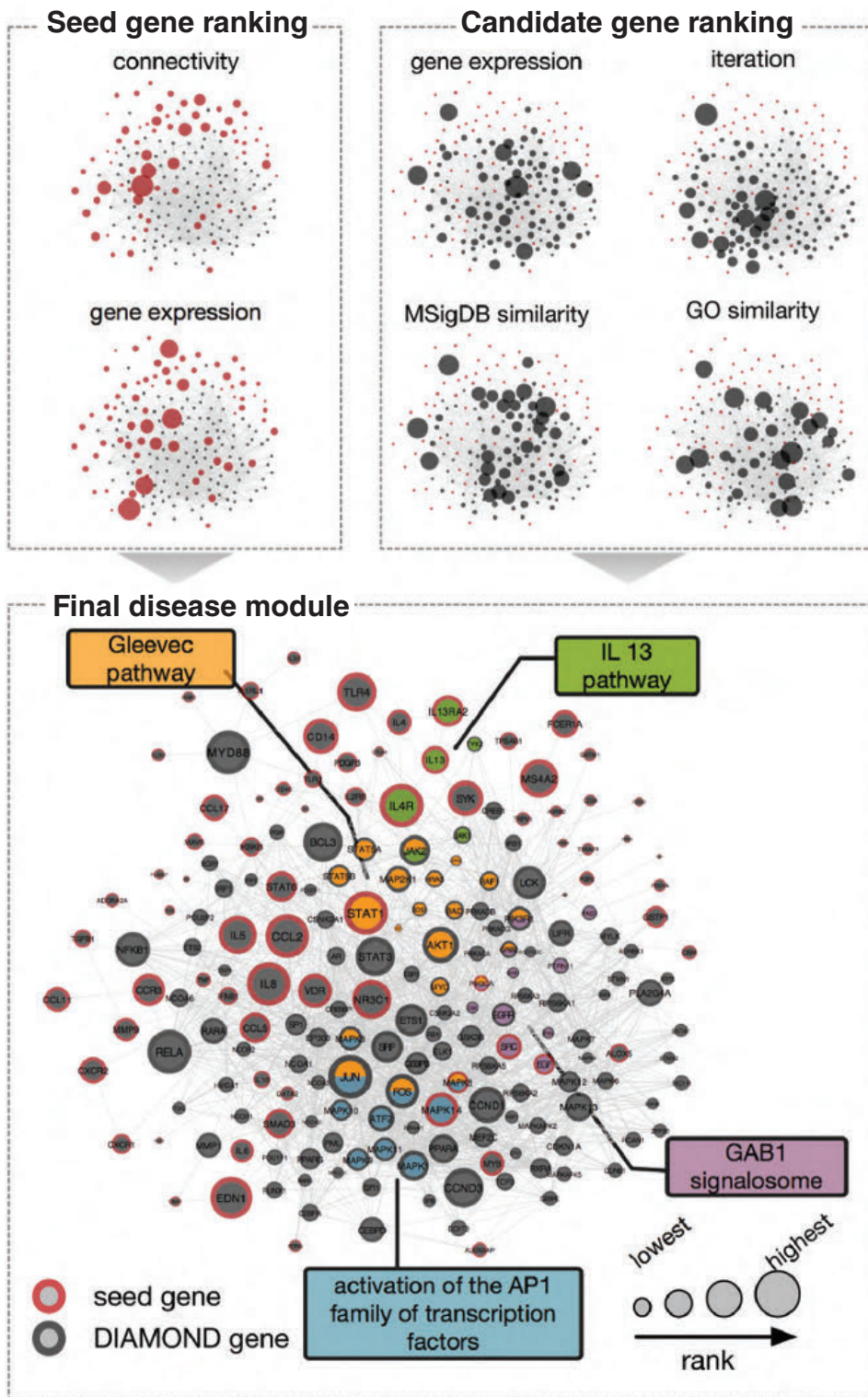


Figure 10.6: Illustration of the ranking procedure (top) and the final asthma disease module (bottom) from [123]. Seed genes and candidate genes are first ranked separately according to their enrichment with different biological datasets. The individual rankings are then combined into a final score for each gene in the disease module, which can then be used to prioritize pathways within the module.

illustrates how the different data are combined into a final score for each gene in the asthma disease module, which in turn can be used to prioritize pathways within the module. The first step is to create a ranking of all genes for each individual data source. Seed genes and candidate genes are often examined separately, which has the advantage that they can be given different weights when they are combined later on. Depending on the particular data type, the ranking can be based on fold-change for differential expression data, GWAS p -value or functional similarity with known processes (compare with Box 10.14), for example. The individual rankings can then be combined into a single score, e.g., using the so-called Borda-count [140]: The score of a gene is taken to correspond to its inverted rank and the scores of different rankings are simply added. Finally, the integrated gene score can be used to prioritize pathways within the module, thus complementing commonly used measures, such as coverage of genes in the pathway. The integrated biological relevance of a pathway within the module can be quantified by the average score of its genes. Additional potentially interesting network-based analyses that can be performed with the disease module include identifying overlaps with other diseases or with network modules known to be modulated by drugs, for example using the distance measures above, or applying community detection to identify potential submodules, for example for patient stratification.

10.5 Summary and Outlook

Network medicine is a highly dynamic and rapidly expanding field covering virtually all areas of biomedical research. This brief introduction can therefore only provide a necessarily incomplete and highly subjective selection. We hope that the references we provide may serve as a starting point for further reading and also recommend a recently published textbook focusing exclusively on this subject [141].

An important challenge in current biomedical research is to integrate the ever growing amount of “omics” data (e.g., genomics, epigenomics, proteomics, metabolomics, lipidomics). Network approaches are inherently holistic and integrative, and particularly multilayer networks are very promising candidates for addressing this challenge [79]. First analytical analyses of multilayer networks highlight the importance of a detailed, context-aware mapping of different types of interactions to fully understand the interplay between structure and dynamics of such complex networks [142]. So far, most studies on biomolecular networks focus on structural network properties and a thorough understanding of their dynamical properties remains an important issue. The concept of dynamic controllability, for example, is well established in network theory [143, 144] and could in principle be applied to driving a cell from a disease state to a healthy state [143]. We expect that such network approaches will be key to designing advanced therapeutics for complex diseases that cannot be understood, nor treated, by a simple mono-causal molecular mechanism. The ultimate goal of network medicine is of course to contribute not only to basic research, but to the translation to benefit patients. Based on the pace at which network medicine is progressing, we are confident that this exciting and challenging goal will be reached rather sooner than later.

10.6 Exercises

To familiarize yourself with some basic network-based approaches to human diseases we will perform a rudimentary disease module analysis. The exemplary solution we provide is based on the programming language python and utilizes heavily the excellent `networkx` module, but of course other programming languages offer similar functionalities.

10.1 Constructing the interactome

- (a) Use one of the databases listed in Box 10.12 to construct an interactome network. We suggest using HIPPIE, as it allows for both programmatic access via an API or download of the entire dataset in an easy to parse text format.
- (b) Construct different networks with different parameters, such as different confidence scores or different experimental sources.
- (c) Perform a basic characterization of the overall topology of each network, e.g., overall coverage, degree distribution, number of isolated components, distribution of shortest pathlengths, clustering coefficient, etc.

10.2 Constructing a seed cluster for a particular disease

- (a) Use one of the databases listed in Box 10.12 to assemble a set of seed genes for a specific disease.
- (b) Place the seed genes on the interactome and determine the degree of localization using different measures from Box 10.10.
- (c) Assess the statistical significance of the measured localization using different randomization schemes, both for the network topology and the seed genes (see Box 10.11).

10.3 Constructing a disease module

- (a) Implement two different network-based gene prioritization algorithms introduced in Box 10.13.
- (b) Rank all genes in the interactome using both methods and with varying parameters of the respective algorithms.
- (c) Evaluate how the results change when removing various fractions of the seed genes.

10.4 Perform an enrichment analysis of the disease module

- (a) Use the databases listed in Box 10.12 to assemble an independent set of genes with potential relevance to the disease, e.g., genes found to be differentially expressed in a patient cohort.
- (b) Test whether the ranked candidate genes are enriched for the genes of the independent validation set.
- (c) Perform a gene set enrichment analysis of the disease module using gene ontology to identify prominent biological processes within the module.

Note: Solutions are available to instructors at www.cambridge.org/bionetworks.

References

- [1] Craig Venter J, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*, 2001;291(5507):1304–1351.
- [2] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001;409(6822):860–921.
- [3] Smedley D, Köhler S, Czeschik JC, et al. Walking the interactome for candidate prioritization in exome sequencing studies of mendelian diseases. *Bioinformatics*, 2014;30(22):3215–3222.
- [4] Leiserson MDM, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 2015;47(2):106–114.
- [5] Chen Y, Zhu J, Lum PK, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 2008;452(7186):429–435.
- [6] Pichlmair A, Kandasamy K, Alvisi G, et al. *Nature*, 2012;487:486–490.
- [7] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 2007;3:140.
- [8] Csermely P, Korcsmfiaros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics*, 2013;138:333–408.
- [9] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 2007;357:370–379.
- [10] Colizza V, Barrat A, Barthélemy M, Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences USA*, 2006;103:2015–2020.
- [11] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proceedings of the National Academy of Sciences USA*, 2007;104(21):8685–8690.
- [12] Landon BE, Keating NL, Barnett ML, et al. Variation in patient-sharing networks of physicians across the United States. *Journal of the American Medical Association*, 2012;308(3):265–273.
- [13] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 2005;4:Article 17.
- [14] De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLOS Computational Biology*, 2010;6:e1000807.
- [15] Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell*, 2011;144:986–998.
- [16] Menche J, Sharma A, Kitsak M, et al. Disease networks. Uncovering disease–disease relationships through the incomplete interactome. *Science*, 2015;347(6224):1257601.

- [17] Thiele I, Palsson, BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 2010;5(1):93–121.
- [18] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 2011;12:56–68.
- [19] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 2017;45:D353–D361.
- [20] Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 2016;44:D481–D487.
- [21] Swainston N, Smallbone K, Hefzi H, et al. Recon 2.2: From reconstruction to model of human metabolism. *Metabolomics*, 2016;12:109.
- [22] Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circulation Research*, 2012;111:359–374.
- [23] Goldford JE, Hartman H, Smith TF, Segre D. Remnants of an ancient metabolism without phosphate. *Cell*, 2017;168:1126–1134.e9.
- [24] Josephides C, Swain PS. Predicting metabolic adaptation from networks of mutational paths. *Nature Communications*, 2017;8:685.
- [25] Li S, Sullivan NL, Roupahel N, et al. Metabolic phenotypes of response to vaccination in humans. *Cell*, 2017;169:862–877.e17.
- [26] Klosik DF, Grimbs A, Bornholdt S, Hutt MT. The interdependent network of gene regulation and metabolism is robust where it needs to be. *Nature Communications*, 2017;8:534.
- [27] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. *Science* 2005;309(5740):1559–1563.
- [28] Zhang Y. Gene regulatory networks: Real data sources and their analysis. In Iba H, Noman N, eds., *Evolutionary Computation in Gene Regulatory Network Research*. John Wiley & Sons, Inc.;2016, pp. 49–65.
- [29] Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 2008;9:770–780.
- [30] Blat Y, Kleckner N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, 1999;98:249–259.
- [31] Furey TS. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 2012;13:840–852.
- [32] Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 2015;43: D117–D122.
- [33] Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: A major expansion and update of the open access database of transcription factor binding profiles. *Nucleic Acids Research*, 2016;44:D110–D115.

- [34] Moignard V, Woodhouse S, Haghverdi L, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 2015;33:269–276.
- [35] Goode DK, Obier N, Vijayabaskar MS, et al. Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Developmental Cell*, 2016;36:572–587.
- [36] Marbach D, Lamparter D, Quon G, et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 2016;13:366–370.
- [37] Friedlander T, Prizak R, Barton NH, Tkačik G. Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nature Communications* 2017;8:216.
- [38] GTEx Consortium. Human genomics: The Genotype-Tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 2015;348:648–660.
- [39] De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* 2010;8:717–729.
- [40] Weirauch MT. Gene coexpression networks for the analysis of DNA microarray data. In Dehmer M, Emmert-Streib F, Graber A, Salvador A, eds., *Applied Statistics for Network Biology*. Wiley-VCH Verlag GmbH & Co. KGaA;2011, pp. 215–250.
- [41] Parikshak NN, Swarup V, Belgard TG, et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 2016;540:423–427.
- [42] Peters LA, Perriogues J, Mortha A, et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics*, 2017;49:1437–1449.
- [43] Calabrese GM, Mesner LD, Stains JP, et al. Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Systems*, 2017;4:46–59.e4.
- [44] Guo X, Xiao H, Guo S, Dong L, Chen J. Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer Gene Therapy*, 2017;24:333–341.
- [45] Boucher B, Jenna S. Genetic interaction networks: Better understand to better predict. *Frontiers in Genetics*, 2013;4:290.
- [46] Srivas R, Shen JP, Yang CC, et al. A network of conserved synthetic lethal interactions for exploration of precision cancer therapy. *Molecular Cell*, 2016;63:514–525.
- [47] Costanzo M, VanderSluis B, Koch EN, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 2016;353:aaf1420.
- [48] Kramer MH, Farré JC, Mitra K, et al. Active interaction mapping reveals the hierarchical organization of autophagy. *Molecular Cell*, 2017;65:761–774.e5.

- [49] Blomen VA, Májek P, Jae LT, et al. Gene essentiality and synthetic lethality in haploid human cells. *Science*, 2015;350:1092–1096.
- [50] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online mendelian inheritance in man (OMIM R), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 2015;43:D789–798.
- [51] Lee DS, Park J, Kay KA, et al. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences USA*, 2008;105:9880–9885.
- [52] Zhou X, Menche J, Barabasi, AL, Sharma A. Human symptoms–disease network. *Nature Communications*, 2014;5:4212.
- [53] NIH: US National Library of Medicine. Medical subject headings. Available online at www.nlm.nih.gov/mesh/.
- [54] Barrenas F, Chavali S, Holme P, Mobini R, Benson M. Network properties of complex human disease genes identified through genome-wide association studies. *PLOS ONE*, 2009;4:e8090.
- [55] Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. The orphan disease networks. *American Journal of Human Genetics*, 2011;88:755–766.
- [56] Pinero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of next generation sequencing. *Science Reports*, 2016;6:24570.
- [57] Hidalgo, CA, Blumm, N, Barabasi, AL, Christakis, NA. A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, 2009;5:e1000353.
- [58] Chmiel A, Klimek P, Thurner S. Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 2014;16:115013.
- [59] Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 2016;17: 615–629.
- [60] Duran-Frigola, M, Rossell, D, Aloy, P. A chemo-centric view of human health and disease. *Nature Communications*, 2014;5:5676.
- [61] Gomez-Cabrero, D. Menche J, Vargas C, et al. From comorbidities of chronic obstructive pulmonary disease to identification of shared molecular mechanisms by data integration. *BMC Bioinformatics*, 2016;17:441.
- [62] Klimek, P, Aichberger, S, Thurner, S. Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks. *Science Reports*, 2016;6:39658.
- [63] Pastor-Satorras, R, Castellano, C, Van Mieghem, P, Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics*, 2015;87: 925–979.
- [64] Bernoulli, D. Essai d’une nouvelle analyse de la mortalité causée par la petite verole et des avantages de l’inoculation pour la prevenir. *Histoire de l’Academie Royale des Sciences (Paris) Avec les Mémoires de Mathématique & de Physique*, 1760;1:1–45.

- [65] Hethcote HW. Three basic epidemiological models. *Applied Mathematical Ecology*, 1989;18:119–144.
- [66] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1927;115:700–721.
- [67] Longini Jr IM, Nizam A, Xu S, et al. Containing pandemic influenza at the source. *Science*, 2005;309:1083–1087.
- [68] Granell C, Gomez S, Arenas A. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical Review Letters*, 2013;111:128701.
- [69] Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences USA*, 2004;101:15124–15129.
- [70] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 2013;342:1337–1342.
- [71] Verdery AM, Siripong N, Pence BW. Social network clustering and the spread of HIV / AIDS among persons who inject drugs in two cities in the Philippines. *Journal of Acquired Immune Deficiency Syndromes*, 2017;76:26–32.
- [72] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*, 1999;286:509–512.
- [73] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. *Physical Review Letters*, 2001;86:3200–3203.
- [74] Pastor-Satorras R, Vespignani A. Immunization of complex networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2002;65:036104.
- [75] Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Reviews of Modern Physics*, 2009;81:591–646.
- [76] Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 2008;358:2249–2258.
- [77] Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. *BMJ*, 2008;337: a2338.
- [78] Cozzo E, Banos RA, Meloni S, Moreno Y. Contact-based social contagion in multiplex networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2013;88:050801.
- [79] Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. *Physics Reports*, 2014;544:1–122.
- [80] Kivelä M, Arenas A, Barthelemy M, et al. Multilayer networks. *Journal of Complex Networks*, 2014;2:203–271.
- [81] Noh JD, Rieger H. Random walks on complex networks. *Physical Review Letters*, 2004;92:118701.

- [82] Bader GD, Cary MP, Sander C. Pathguide: A pathway resource list. *Nucleic Acids Research*, 2006;34:D504–D506.
- [83] Szklarczyk, D. et al. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 2017;45:D362–D368.
- [84] Chatr-Aryamontri A, Oughtred R, Boucher L. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 2017;45:D369–D379.
- [85] Orchard S, Ammari M, Aranda B, et al. The MIntAct project: IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 2014;42:D358–63.
- [86] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 2017;45:D408–D414.
- [87] Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*, 2014;159:1212–1226.
- [88] Huttlin EL, Ting L, Bruckner RJ, et al. The BioPlex network: A systematic exploration of the human interactome. *Cell*, 2015;162:425–440.
- [89] Huttlin EL, Bruckner RJ, Paulo JA, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 2017;545:505–509.
- [90] Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 2012;490:556–560.
- [91] Jansen, R. Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, 2003;302:449–453.
- [92] Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein–protein interaction networks and biology: What’s the connection? *Nature Biotechnology*, 2008;26:69–72.
- [93] Gillis J, Ballouz S, Pavlidis P. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of Proteomics*, 2014;100:44–54.
- [94] Caldera M, Buphamalai P, Muller F, Menche J. Interactome-based approaches to human disease. *Current Opinion in Systems Biology*, 2017;3:88–94.
- [95] Clauset A, Shalizi C, Newman M. Power-law distributions in empirical data. *SIAM Review*, 2009;51:661–703.
- [96] Watts DJ, Strogatz SH. Collective dynamics of “small-world” networks. *Nature*, 1998;393:440–442.
- [97] Cohen R, Havlin S. Scale-free networks are ultrasmall. *Physical Review Letters*, 2003;90:058701.
- [98] Callaway DS, Newman ME, Strogatz SH, Watts DJ. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 2000;85:5468.
- [99] Newman ME, Strogatz SH, Watts, DJ. Random graphs with arbitrary degree distributions and their applications. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2001;64:026118.

- [100] Cohen R, Erez K, Ben-Avraham D, Havlin S. Resilience of the internet to random breakdowns. *Physical Review Letters*, 2000;85:4626.
- [101] Dorogovtsev SN, Mendes JF. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press;2003.
- [102] Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature*, 2000;406:378–382.
- [103] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*, 2001;411:41–42.
- [104] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*, 1999;402:C47–52.
- [105] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences USA*, 2003;100:12123–12128.
- [106] Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 2004;5:101–113.
- [107] Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences USA*, 2008;105:4323–4328.
- [108] Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLOS Computational Biology*, 2015;11:e1004120.
- [109] Fortunato S. Community detection in graphs. *Physics Reports*, 2010;486:75–174.
- [110] Guney E, Menche J, Vidal M, Barabasi AL. Network-based in silico drug efficacy screening. *Nature Communications*, 2016;7:10331.
- [111] Newman ME. The structure and function of complex networks. *SIAM Review*, 2003;45:167–256.
- [112] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*, 2002;296, 910–913.
- [113] Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U. On the uniform generation of random graphs with prescribed degree sequences. arXiv preprint available at <https://arxiv.org/pdf/cond-mat/0312028.pdf>. 2004.
- [114] Bender EA, Canfield ER. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 1978;24:296–307.
- [115] Bollobás B. Random graphs. In *Graph Theory*, 123–145 (Springer, 1979).
- [116] Serrano MA, Boguna M. Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E* 72, 036133 (2005).
- [117] Boguna, M, Pastor-Satorras, R. Class of correlated random networks with hidden variables. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2003;68:036112.
- [118] Weber S, Porto M. Generation of arbitrarily two-point-correlated random networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2007;76:046111.

- [119] Samal A, Martin OC. Randomizing genome-scale metabolic networks. *PLOS ONE*, 2011;6:e22295.
- [120] Basler G, Ebenhoh O, Selbig J, Nikoloski Z. Mass-balanced randomization of metabolic networks. *Bioinformatics*, 2011;27:1397–1403.
- [121] Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, 2016;17:9–18.
- [122] Krishnan A, Zhang R, Yao V, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 2016;19: 1454–1462.
- [123] Sharma A, Menche J, Chris Huang C, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Human Molecular Genetics*, 2015;24: 3005–3020.
- [124] Barshir R, Shwartz O, Smoly IY, Yeager-Lotem E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLOS Computational Biology*, 2014;10:e1003632.
- [125] Magger O, Waldman YY, Ruppin E, Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLOS Computational Biology*, 2012;8:e1002690.
- [126] Li M, Zhang J, Liu Q, Wang J, Wu FX. Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. *BMC Medical Genomics*, 2014;7(Suppl 2):S4.
- [127] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 2006;43:691–698.
- [128] Wang XD, Huang JL, Yang L, et al. Identification of human disease genes from interactome network using graphlet interaction. *PLOS ONE* 2014;9:e86142.
- [129] Erten S, Bebek G, Ewing RM, Koyuturk M, et al. DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 2011;4:19.
- [130] Guney E, Oliva B. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLOS ONE*, 2012;7:e43557.
- [131] Bailly-Bechet M, Borgs C, Braunstein A, et al. Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences USA*, 2011;108:882–887.
- [132] Tuncbag N, McCallum S, Huang SSC, Fraenkel E. SteinerNet: A web server for integrating ‘omic’ data to discover hidden components of response pathways. *Nucleic Acids Research*, 2012;40:W505–W509.
- [133] Tuncbag N, Gosline SJC, Kedaigle A, et al. Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLOS Computational Biology*, 2016;12:e1004879.
- [134] Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: Bridging linkage and molecular-network information for

- identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences USA*, 2004;101:15148–15153.
- [135] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 2010;6:e1000641.
 - [136] Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 2011;18:507–522.
 - [137] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: A universal amplifier of genetic associations. *Nature Reviews Genetics*, 2017;18:551–562.
 - [138] Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics* 82, 949–958 (2008).
 - [139] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. *PLOS Computational Biology*, 2009;5:e1000443.
 - [140] Van Erp M, Schomaker, L. Variants of the borda count method for combining ranked classifier hypotheses. In Schomaker L, Vuurpijl L, eds., *Proceedings 7th International Workshop on Frontiers in Handwriting Recognition*. International Unipen Foundation;2000, pp. 443–452.
 - [141] Loscalzo J, Barabasi AL, Silverman EK, eds. *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Harvard University Press;2017.
 - [142] De Domenico M, Granell C, Porter MA, Arenas A. The physics of spreading processes in multilayer networks. *Nature Physics*, 2016;12:901–906.
 - [143] Liu YY, Slotine JJ, Barabasi AL. Controllability of complex networks. *Nature*, 2011;473:167–173.
 - [144] Liu YY, Slotine JJ, Barabasi AL. Observability of complex systems. *Proceedings of the National Academy of Sciences USA*, 2013;110:2460–2465.

1.2 The human interactome

REVIEW ARTICLE

Interactome-Based Approaches to Human Disease

Michael Caldera^{*}, Pisanu Buphamalai^{*}, Felix Müller, and Jörg Menche

Published in **Current Opinion in Systems Biology**, **3**: 88–94, 2017.

DOI: 10.1016/j.coisb.2017.04.015

^{*} Authors contributed equally

© 2017 The Authors. Published by Elsevier Ltd.

This article is available under the Creative Commons CC-BY-NC-ND license and permits non-commercial use of the work as published, without adaptation or alteration provided the work is fully attributed.

The human interactome, the complete set of physical interactions among macromolecules inside the cell, is a central resource to the development of network medicine as a discipline as well as a key network that this doctoral thesis has been developed based upon. This review article was written to target readers who may have not have been familiarized with the concept of the interactome and its usages. It discusses various definitions of the interactome, provides an overview of experimental methods used to discover novel interactions, and describes its current topological characteristics. The review further introduces various application aspects of the interactome including (i) the identification of disease modules, (ii) the network-based prioritization of disease genes, (iii) the quantification of disease relationships, (iv) the construction of context-specific interactome *i.e.*, incorporation of tissue specificity, and (v) the interactome for drug target prioritization.



Interactome-based approaches to human disease

Michael Caldera¹, Pisanu Buphamalai¹, Felix Müller and Jörg Menche

Abstract

Recent advances in high-throughput technologies have created exciting opportunities for systematically investigating the molecular basis of human disease. In addition to a growing catalog of disease-associated genetic variations, we can now map out an increasingly detailed network diagram of the complex machinery of interacting molecules that constitutes the basis of (patho-) physiological states. The emerging field of ‘network medicine’ applies tools and concepts from network theory to interpret this diagram and elucidate the relation between perturbations on the molecular level and phenotypic disease manifestations. The interactome, i.e. the integrated network of all physical interactions within the cell, can be interpreted as a map and diseases as local perturbations. Network-based approaches can aid in identifying the specific interactome neighborhood that is perturbed in a certain disease, guide the search for therapeutic targets and reveal common molecular mechanisms between seemingly unrelated diseases.

Addresses

CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

Corresponding author: Menche, Jörg (jmenche@cemm.oew.ac.at)

¹ These authors contributed equally to this work.

Current Opinion in Systems Biology 2017, 3:88–94

This review comes from a themed issue on **Clinical and translational systems biology (2017)**

Edited by **Jesper Tegnér** and **David Gomez-Cabrero**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online 4 May 2017

<http://dx.doi.org/10.1016/j.coisb.2017.04.015>

2452-3100/© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

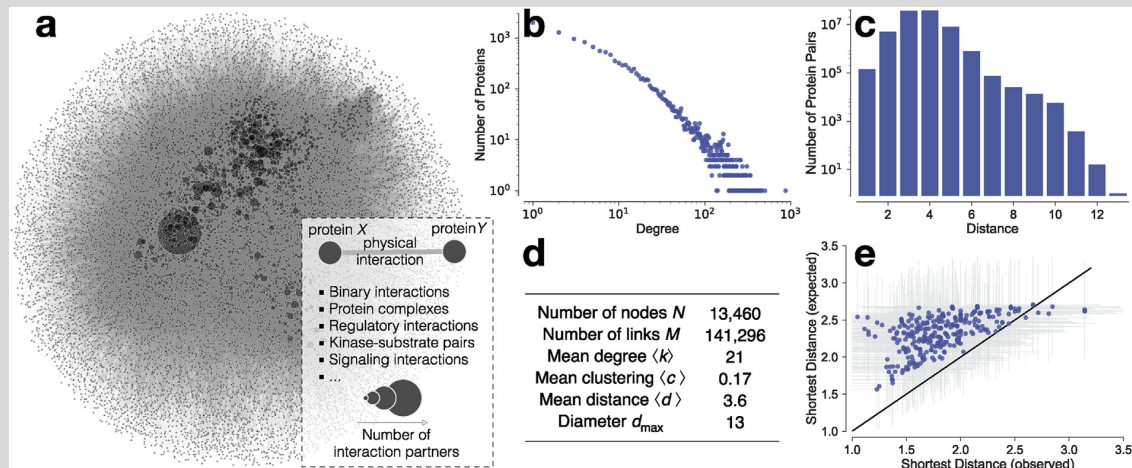
Introduction

The Online Mendelian Inheritance in Man (OMIM) database [1] currently lists over 3700 genes with mutations that are known to have a phenotypic impact, e.g. sequence alterations that are causal for Mendelian diseases or variants that increase the susceptibility to complex diseases or cancer. Yet, despite this ever growing wealth of data, many details of how exactly genetic alterations contribute to the disease pathobiology remain in the dark. A crucial roadblock for

translating gene-level discoveries into a mechanistic understanding of disease pathogenesis and concrete strategies for prevention, diagnosis, and treatment is that gene products do not act in isolation, but in the context of other genes and proteins. Biological processes are ultimately the result of a highly dynamic and regulated interplay of macromolecules, such as interactions between proteins or between proteins and DNA or RNA. The entirety of all such biologically relevant interactions form a large and highly connected network, often referred to as the ‘*interactome*’ (Box 1). The interactome can therefore be understood as a map to investigate how individual (or several) genetic alterations propagate throughout the network and perturb the system as a whole. The emerging field of ‘network medicine’ applies tools and concepts from network theory (Box 2) to interpret this map and elucidate the relation between perturbations on the molecular level and phenotypic disease manifestations [2]. In the last decade, network-based approaches have been successfully applied to a broad range of diseases, with examples ranging from rare Mendelian disorders [3], cancer [4] or metabolic diseases [5], to identifying basic strategies by which viruses hijack the host interactome [6], to name but a few. In the following we will review the basic ideas that underly interactome-based approaches to human disease and highlight important recent conceptual advances.

The interactome

The term ‘interactome’ is only loosely defined and may refer to networks that contain rather different types of interactions. It is instructive to distinguish between *physical* and *functional* interactions. Physical interactions involve actual physical contact between the participating biomolecules, for example proteins that assemble in a complex or receptor-ligand binding. Functional interaction, on the other hand, can refer to any kind of biologically relevant relationship. In co-expression networks, for example, genes are connected if their expression patterns are strongly correlated [7]. Another important functional relationship are ‘genetic interactions’, where two genes are linked if the effect of a simultaneous alteration of both genes differs from the expectation based on the individual alterations. An extreme form is *synthetic lethality*, where a combined loss of two genes leads to cell death, while the loss of each individual gene does not [8]. *Synthetic viability*, conversely, occurs when the lethal effect of a mutation in one gene is rescued by a simultaneous mutation in a

Box 1. The interactome.

(a) A global picture of the interactome (as used in [16]) showing its highly complex and interconnected nature. It contains 13,460 proteins and 141,296 interactions that have been curated from different sources with various kinds of physical interactions, including binary interactions from systematic yeast two-hybrid screens, protein complexes, kinase–substrate pairs and others. (b) The overall topology is characterized by a highly heterogeneous degree distribution that follows a power-law. The vast majority of proteins have only few connections, but there is also a considerable number of extremely highly connected proteins, so-called hubs (33 proteins have more than 300 interactions). (c) These hubs serve as shortcuts, so that on average, all proteins are directly connected to each other with less than four intermediate steps, a phenomenon often called the ‘small-world’ effect. The maximum distance between any two proteins in the interactome is 13. (d) Other important structural properties of the interactome. (e) A comparison of the distances observed among genes associated with the same disease and the respective random expectation reveals that disease genes are not scattered randomly in the interactome, but aggregate in local, disease-specific neighborhoods, so-called disease modules.

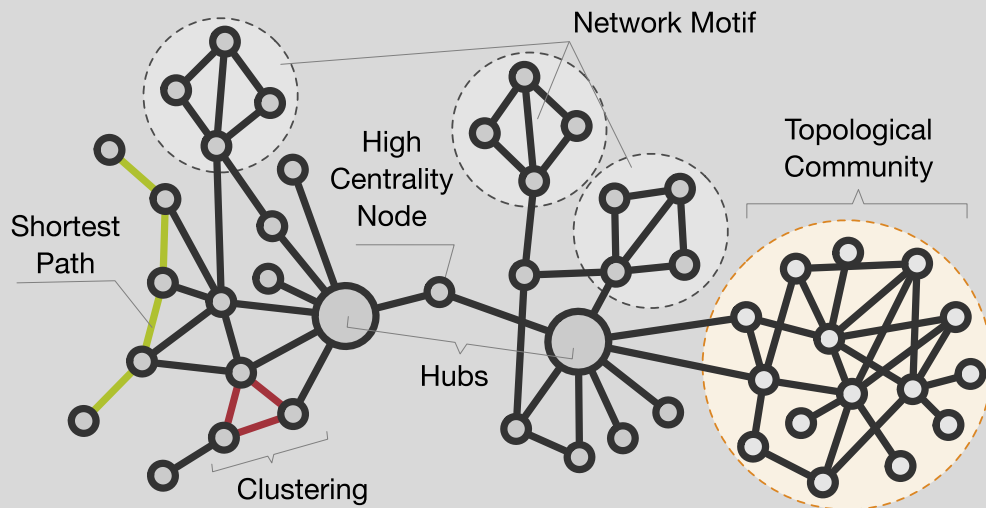
second gene [9]. While both functional and physical interaction networks can yield important insights into disease mechanisms, we will focus mostly on the more narrowly defined physical interactions in the following.

A number of publicly available databases provide comprehensive lists of physical protein–protein interactions (PPIs), as well as other relevant interactions (e.g. protein–DNA, protein–RNA, enzyme–metabolite) in human, but also in other species [10]. There are three main sources for the PPIs reported therein: (i) interactions curated from the scientific literature and typically derived from small-scale experiments. (ii) Interactions from systematic, proteome-scale mapping efforts, the two main techniques being yeast two-hybrid assays for binary interactions [11] and binding affinity purifications coupled to mass spectrometry for co-complexes [12]. (iii) Interactions from computational predictions, for example based on protein structure [13]. It is important to note that each of these sources may introduce different kinds of noise and biases [14], such as biases in the selection of which protein pairs

have been tested [15] or experimental biases, for example towards highly expressed genes [11]. Another important consideration for interactome-based analyses is the considerable incompleteness of currently available data. It is estimated, for example, that high-throughput methods cover less than 20% of all potential pairwise protein interactions in the human cell [11]. It is therefore imperative to carefully evaluate both the effect of potential biases, as well as the influence of missing interactions, when analyzing and interpreting interactome data. Box 1 summarizes the main topological properties of a manually curated interactome from [16].

Disease modules in the interactome

Among the first evidence for a direct correspondence between the biological importance of a gene and the interactome position of its product was the observation that the phenotypic impact of deleting a gene in the yeast *Saccharomyces cerevisiae* correlates with the number of interaction partners of the corresponding protein [17]. This trend was later confirmed also for genes that

Box 2. Basic topological characteristics of networks.

The **degree** of a node is the number of links attached to it, i.e. the number of direct neighbors. The distribution of the degrees across all nodes is an important global characteristic of a network.

Scale free networks are characterized by a heterogeneous degree distribution that follows a power-law: while most nodes have only few neighbors, there are also a few highly connected 'hubs' with a large number of neighbors.

A **path** between two nodes is a sequence of links connecting the two. The minimum number of links needed to connect the two is called 'shortest path length' and represents their 'network distance'.

Centrality measures exist for both nodes and for links and quantify their topological importance within the network. There are different types of centrality measures, e.g. the 'degree centrality' (simply given by the degree) or 'betweenness centrality' (quantifying how many shortest paths of the full network cross through a certain node).

Clustering describes a tendency observed in many biological (and other) networks that two neighbors of a node are often also connected to each other, thus forming a triangle.

Motifs are small recurrent subgraphs in a network that occur particularly frequently.

Network communities are groups of tightly interconnected nodes that have more connections among themselves than to the rest of the network.

are essential for the viability of human cell lines [18]. The topological properties of disease-associated genes are generally more diverse and may differ between disease classes (e.g. complex diseases, Mendelian diseases or cancer), as well as inheritance modes (autosomal dominant or recessive): cancer driver genes generally show a strong tendency towards high network centrality (Box 2), while recessive disease genes are often more isolated and located at the periphery of the interactome [19].

To further elucidate the detailed mechanisms, by which a disease-associated perturbation contributes to the pathobiological phenotype, it is important not only to understand the network properties of individual associated genes, but also their interactome environment and emerging collective properties. This is particularly evident for complex diseases that involve potentially hundreds of genes. Similar to the functional coherence of interactome neighbors (i.e., interacting proteins are often involved in the same biological process [20]),

genes associated with the same disease have been found to interact with each other more frequently than expected by chance [21]. This observation has been verified systematically for a large number of diseases [16], thus confirming a fundamental hypothesis of interactome-based approaches to human disease, namely that disease genes tend to cluster within so-called *disease modules*. Such disease modules are connected subgraphs of the interactome that contain all molecular determinants of a certain disease. The first step towards elucidating the biological mechanisms of a disease in a network-based framework is therefore to identify the respective disease module.

Interactome-based gene prioritization

In recent years, a plethora of disease-module identification methods have been proposed that explore the local network neighborhood around known disease-associated genes ('seed genes') to infer likely new disease gene candidates [22]. They can roughly be classified into three main categories: (i) *Path-based approaches* consider the genes along the shortest paths between the known disease genes as potential candidate genes. These candidate genes can then be further ranked, for example according to the number [23] or significance [24] of paths they participate in, or filtered such that they form a minimal connected subgraph, a so-called Steiner-tree [25]. (ii) *Dynamical approaches* aim to identify candidate genes by propagating known disease associations using dynamical models, for example diffusive processes, where the network neighborhood around seed genes is scanned by simulating random walks along the links [26–29]. Genes that are visited more frequently are considered dynamically closer to the seed genes and therefore ranked higher. (iii) *Connectivity-based approaches* algorithms rank candidate genes according to their number of links to seed genes [30–32].

Relationship between diseases

Considering the highly connected interactome, it is apparent that diseases can rarely be understood as independent entities. Uncovering such relationships between diseases systematically can help us understand how different pathological phenotypes are linked together at the molecular level and shed light on disease comorbidity, i.e. the observation that certain groups of diseases frequently arise together [33]. Indeed, a large-scale evaluation of shared gene associations revealed a highly connected 'diseaseome', in which more than 500 diseases form a giant component and more than 800 diseases have at least one link to another disease [34]. Other disease–disease networks have been constructed based on shared metabolic pathways [35], phenotype similarity [36,37], the structure of disease ontologies [38] or comorbidity extracted from patient records [39,40]. In an interactome-based framework, the relationship between two diseases is represented by

overlapping disease modules, indicating that perturbations causing one disease are likely to also affect the other disease. A systematic study of over 44,000 disease pairs revealed that the degree of this overlap is highly predictive for the pathobiological similarity of diseases, such that diseases with overlapping modules show significant co-expression patterns, symptom similarity, and comorbidity, while those that reside in separated interactome neighborhoods are pathobiologically and clinically distinct [16].

The considerable molecular-level overlap that has been observed for many diseases pinpoints a limitation of canonical disease classifications that, historically, are largely based on clinicopathological evidence and often categorized according to the organ system that the disease primarily affects. Interactome-based methodologies could provide a more holistic framework for disease classification based on molecular mechanism [41].

Tissue-specific interactomes

The studies discussed above considered an integrated interactome containing interactions that have been identified using various techniques and were observed under different experimental and biological conditions. While such a global interactome provides invaluable information for discovering general principles of disease-associated network perturbations, it cannot account for the cell-type or tissue-specific manifestations that characterize many diseases. Directly measured context-specific interactome networks are scarce, but can be approximated by integrating more widely available transcriptome or proteome information [42,43]. The main idea is to use tissue-specific expression information to filter the global interactome for interactions that are feasible in a given tissue, i.e. both interaction partners are present [44]. Consequently, the resulting tissue-specific interactomes are generally smaller and sparser. In line with the observation that essential genes are more central in the global interactome, genes that are expressed across many tissues (such as 'house-keeping' genes) were found to form a core interactome to which the more tissue-specific genes then attach, thus forming tissue-specific peripheries [45–47]. A comparison between the global and tissue-specific interactomes further revealed that diseases typically manifest in those tissues, in which the corresponding disease-module is least fragmented [48]. Tissue-specific interactome networks can therefore shed light onto the detailed disease-associated rewiring events [49,50] and considerably improve disease gene prioritization [47,51,52].

Drugs in the interactome

From a network-based perspective, the action of drugs can be interpreted similarly to the effect of disease-

associated genetic variants, i.e. as a local perturbation of the interactome. Many of the concepts and tools introduced above can be therefore immediately applied in the context of network pharmacology [53,54]. Several studies of drug-target networks have shown that most currently used drugs are less selective than previously assumed and instead target multiple proteins [55,56]. These target proteins tend to be more highly connected than random proteins, but less so than essential proteins. Most drugs do not target the corresponding disease module as a whole, but only a small subset or adjacent interactome neighborhood [57]. It was further found that drugs whose affected interactome neighborhood is closer to the disease module tend to be more effective in the clinic. These insights could help in selecting the most promising drug targets, for example by prioritizing targets according to their topological properties [58], as well as in designing multitarget drugs that act specifically and directly on the respective disease module [54]. Another promising application of interactome-based drug–disease relationships are approaches to drug repurposing, for example by systematically identifying diseases with shared molecular mechanism that may be modulated by the same therapeutic intervention [59].

Conclusion

Interactome-based approaches to human disease have matured considerably in the past few years, now possessing both a firm theoretical fundament, as well as a broad range of successful applications across all major areas of human disease research. At the same time, the interactome represents only one layer of relevant information. A pressing challenge on the way towards the next generation of (network) medicine is to integrate the ever growing amount of omics data (e.g., genomics, epigenomics, proteomics, metabolomics, lipidomics). Interactome-based, and more generally, network-based approaches are inherently holistic and integrative, thus offering unique opportunities in this endeavor.

Acknowledgements

J.M. is supported by the Vienna Science and Technology Fund, WWTF [grant number WWTF-VRG005].

Glossary

- Interactome** A global network representing all molecular interactions in a cell. In most cases, the term specifically refers to *physical* interaction networks consisting mostly of protein–protein interactions, but also of protein–DNA or protein–RNA interactions. More generally, the term interactome may also be used to describe *functional* interactions, such as genetic interactions.
- Disease Gene** Gene with a known disease association. Sometimes the term is reserved to genes with a known mutant genotype that causes an inherited disorder. More generally, the term is used also for genes containing a risk variant for complex diseases or other, more indirect associations to a particular disease.

- Candidate gene** Gene with suspected role in the pathobiology of a disease based on prior evidence. The goal of disease gene prioritization methods is to identify the most likely candidates.
- Disease module** The comprehensive set of cellular components associated with a certain disease and their interactions. More specifically, the term refers to a connected subgraph of the interactome, whose perturbation causes the disease. Network-based disease module detection methods aim to identify this subgraph, in analogy to gene prioritization methods.
- Context-specific interactomes** Contain only interactions that occur in a given biological context, such as cell-type, tissue, or a specific disease condition. Such interactomes are most commonly obtained by filtering out proteins that are not expressed in the respective context.
- Comorbidity** The tendency of certain diseases to co-occur in the same patient, suggesting shared underlying molecular mechanisms.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- * of special interest
- ** of outstanding interest

- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A: **OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders.** *Nucleic Acids Res* 2015, **43**:D789–D798.
- Loscalzo J, Barabási A-L, Silverman EK: **Network medicine: complex systems in human disease and therapeutics.** Harvard University Press; 2017.
- Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, *et al.*: **A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease.** *Am J Hum Genet* 2016, **99**:595–606, <http://dx.doi.org/10.1016/j.ajhg.2016.07.005>.
- Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, *et al.*: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.** *Nat Genet* 2015, **47**:106–114, <http://dx.doi.org/10.1038/ng.3168>.
- Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, *et al.*: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452**:429–435, <http://dx.doi.org/10.1038/nature06757>.
- Pichlmair A, Kandasamy K, Alvisi G, Mulhern O, Sacco R, Habjan M, *et al.*: **Viral immune modulators perturb the human molecular network by common and unique strategies.** *Nature* 2012, **487**:486–490, <http://dx.doi.org/10.1038/nature11289>.
- Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4**:17, <http://dx.doi.org/10.2202/1544-6115.1128>.
- Srivastava R, Shen JP, Yang CC, Sun SM, Li J, Gross AM, *et al.*: **A network of conserved synthetic lethal interactions for exploration of precision cancer therapy.** *Mol Cell* 2016, **63**:514–525, <http://dx.doi.org/10.1016/j.molcel.2016.06.022>.
- Motter AE, Gulbahce N, Almaas E, Barabási A-L: **Predicting synthetic rescues in metabolic networks.** *Mol Syst Biol* 2008, **4**:168, <http://dx.doi.org/10.1038/msb.2008.1>.

10. De Las Rivas J, Fontanillo C: **Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.** *PLoS Comput Biol* 2010, **6**:e1000807, <http://dx.doi.org/10.1371/journal.pcbi.1000807>.
11. Rolland T, Taşan M, Charleatoux B, Pevzner SJ, Zhong Q, Sahni N, *et al.*: **A proteome-scale map of the human interactome network.** *Cell* 2014, **159**:1212–1226, <http://dx.doi.org/10.1016/j.cell.2014.10.050>.
This paper introduces the largest currently available binary interactome map obtained from a systematic yeast two-hybrid screen. Particular emphasis is given to a quantification of the influence of biases in literature-curated interaction maps.
12. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, *et al.*: **The BioPlex network: A systematic exploration of the human interactome.** *Cell* 2015, **162**:425–440, <http://dx.doi.org/10.1016/j.cell.2015.06.043>.
The largest currently available interaction map based on the affinity purification – mass spectrometry approach. A detailed topological analysis reveals that the network architecture reflects biological organisation principles.
13. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, *et al.*: **Structure-based prediction of protein-protein interactions on a genome-wide scale.** *Nature* 2012, **490**:556–560, <http://dx.doi.org/10.1038/nature11503>.
14. Hakes L, Pinney JW, Robertson DL, Lovell SC: **Protein-protein interaction networks and biology—what's the connection?** *Nat Biotechnol* 2008, **26**:69–72, <http://dx.doi.org/10.1038/nbt10108-69>.
15. Gillis J, Ballouz S, Pavlidis P: **Bias tradeoffs in the creation and analysis of protein–protein interaction networks.** *J Proteom* 2014, **100**:44–54, <http://dx.doi.org/10.1016/j.jprot.2014.01.020>.
16. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, *et al.*: **Uncovering disease-disease relationships through the incomplete interactome.** *Science* 2015, **347**:1257601, <http://dx.doi.org/10.1126/science.1257601>.
A systematic study of 299 diseases revealing that current interactome maps have reached sufficient coverage to show that genes associated with the same disease tend to cluster in the same interactome neighborhood. Disease pairs with overlapping disease modules show significant molecular similarity, elevated coexpression of their associated genes, similar symptoms and high comorbidity.
17. Jeong H, Mason SP, Barabási A-L, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41–42, <http://dx.doi.org/10.1038/35075138>.
18. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, *et al.*: **Gene essentiality and synthetic lethality in haploid human cells.** *Science* 2015, **350**:1092–1096, <http://dx.doi.org/10.1126/science.aac7557>.
A first large-scale investigation of essential genes in human cell lines, confirming their central position in interactome networks.
19. Piñero J, Berenstein A, Gonzalez-Perez A, Chermomoretz A, Furlong LI: **Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing.** *Sci Rep* 2016, **6**:24570, <http://dx.doi.org/10.1038/srep24570>.
A thorough investigation of the topological interactome properties of disease genes for different classes of diseases and inheritance modes, offering a much more diverse picture than previously appreciated that can also explain apparent contradictions in the literature.
20. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47–C52, <http://dx.doi.org/10.1038/35011540>.
21. Feldman I, Rzhetsky A, Vitkup D: **Network properties of genes harboring inherited disease mutations.** *Proc Natl Acad Sci U. S. A* 2008, **105**:4323–4328, <http://dx.doi.org/10.1073/pnas.0701722105>.
22. Wang X, Gulbahce N, Yu H: **Network-based methods for human disease gene prediction.** *Brief Funct Gen* 2011, **10**:280–293, <http://dx.doi.org/10.1093/bfgp/eln024>.
23. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction.** *Nucleic Acids Res* 2006, **34**:e130, <http://dx.doi.org/10.1093/nar/gkl707>.
24. Dezső Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, *et al.*: **Identifying disease-specific genes based on their topological significance in protein networks.** *BMC Syst Biol* 2009, **3**:36, <http://dx.doi.org/10.1186/1752-0509-3-36>.
25. Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, François J-M, *et al.*: **Finding undetected protein associations in cell signaling by belief propagation.** *Proc Natl Acad Sci U. S. A* 2011, **108**:882–887, <http://dx.doi.org/10.1073/pnas.1004751108>.
26. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A: **Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease.** *Proc Natl Acad Sci U. S. A* 2004, **101**:15148–15153, <http://dx.doi.org/10.1073/pnas.0404315101>.
27. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R: **Associating genes and protein complexes with disease via network propagation.** *PLoS Comput Biol* 2010, **6**:e1000641, <http://dx.doi.org/10.1371/journal.pcbi.1000641>.
28. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol* 2011, **18**:507–522, <http://dx.doi.org/10.1089/cmb.2010.0265>.
29. Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, *et al.*: **Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases.** *Bioinformatics* 2014, **30**:3215–3222, <http://dx.doi.org/10.1093/bioinformatics/btu508>.
30. Guney E, Oliva B: **Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization.** *PLoS One* 2012, **7**:e43557, <http://dx.doi.org/10.1371/journal.pone.0043557>.
31. Wang X-D, Huang J-L, Yang L, Wei D-Q, Qi Y-X, Jiang Z-L: **Identification of human disease genes from interactome network using graphlet interaction.** *PLoS One* 2014, **9**:e86142, <http://dx.doi.org/10.1371/journal.pone.0086142>.
32. Ghiassian SD, Menche J, Barabási A-L: **A Disease Module Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome.** *PLoS Comput Biol* 2015, **11**:e1004120, <http://dx.doi.org/10.1371/journal.pcbi.1004120>.
33. Hu JX, Thomas CE, Brunak S: **Network biology concepts in complex disease comorbidities.** *Nat Rev Genet* 2016, **17**:615–629, <http://dx.doi.org/10.1038/nrg.2016.87>.
34. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L: **The human disease network.** *Proc Natl Acad Sci U. S. A* 2007, **104**:8685–8690, <http://dx.doi.org/10.1073/pnas.0701361104>.
35. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L: **The implications of human metabolic network topology for disease comorbidity.** *Proc Natl Acad Sci U. S. A* 2008, **105**:9880–9885, <http://dx.doi.org/10.1073/pnas.0802208105>.
36. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM: **A text-mining analysis of the human phenotype.** *Eur J Hum Genet* 2006, **14**:535–542.
37. Zhou X, Menche J, Barabási A-L, Sharma A: **Human symptoms–disease network.** *Nat Commun* 2014, **5**:4212, <http://dx.doi.org/10.1038/ncomms5212>.
38. Caniza H, Romero AE, Paccanaro A: **A network medicine approach to quantify distance between hereditary disease modules on the interactome.** *Sci Rep* 2015, **5**:17658, <http://dx.doi.org/10.1038/srep17658>.
39. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA: **A dynamic network approach for the study of human phenotypes.** *PLoS Comput Biol* 2009, **5**:e1000353, <http://dx.doi.org/10.1371/journal.pcbi.1000353>.
40. Klimek P, Aichberger S, Thurner S: **Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks.** *Sci Rep* 2016, **6**:39658, <http://dx.doi.org/10.1038/srep39658>.
A systematic study of over 300 diseases that integrates comorbidity networks and molecular networks in order to dissect the role of environmental and genetic factors in the pathogenesis of each individual disease.

41. Chan SY, Loscalzo J: **The emerging paradigm of network medicine in the study of human disease.** *Circ Res* 2012, **111**: 359–374, <http://dx.doi.org/10.1161/CIRCRESAHA.111.258541>.
 42. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, *et al.*: **Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.** *Mol Cell Proteom* 2014, **13**: 397–406, <http://dx.doi.org/10.1074/mcp.M113.035600>.
 43. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, *et al.*: **The human transcriptome across tissues and individuals.** *Science* 2015, **348**:660–665, <http://dx.doi.org/10.1126/science.aaa0355>.
 44. Yeger-Lotem E, Sharan R: **Human protein interaction networks across tissues and diseases.** *Front Genet* 2015, **6**:257, <http://dx.doi.org/10.3389/fgene.2015.00257>.
 45. Bossi A, Lehner B: **Tissue specificity and the human protein interaction network.** *Mol Syst Biol* 2009, **5**:260, <http://dx.doi.org/10.1038/msb.2009.17>.
 46. Liu W, Wang J, Wang T, Xie H: **Construction and analyses of human large-scale tissue specific networks.** *PLoS One* 2014, **9**:e115074, <http://dx.doi.org/10.1371/journal.pone.0115074>.
 47. Barshir R, Shwartz O, Smoly IY, Yeger-Lotem E: **Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases.** *PLoS Comput Biol* 2014, **10**:e1003632, <http://dx.doi.org/10.1371/journal.pcbi.1003632>.
- The authors examine the topological features of over 300 diseases in tissue-specific interactomes and identify an increased number of interactions as a major determinant for tissue-specific disease manifestation.
48. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, *et al.*: **Tissue specificity of human disease module.** *Sci Rep* 2016, **6**:35241, <http://dx.doi.org/10.1038/srep35241>.
 49. Ideker T, Krogan NJ: **Differential network biology.** *Mol Syst Biol* 2012, **8**:565, <http://dx.doi.org/10.1038/msb.2011.99>.
 50. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, *et al.*: **Understanding multicellular function and disease with human tissue-specific networks.** *Nat Genet* 2015, **47**:569–576, <http://dx.doi.org/10.1038/ng.3259>.
 51. Magger O, Waldman YY, Ruppin E, Sharan R: **Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks.** *PLoS Comput Biol* 2012, **8**: e1002690, <http://dx.doi.org/10.1371/journal.pcbi.1002690>.
 52. Li M, Zhang J, Liu Q, Wang J, Wu F-X: **Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation.** *BMC Med Gen* 2014, **7**(Suppl 2):S4, <http://dx.doi.org/10.1186/1755-8794-7-S2-S4>.
 53. Hopkins AL: **Network pharmacology: the next paradigm in drug discovery.** *Nat Chem Biol* 2008, **4**:682–690, <http://dx.doi.org/10.1038/nchembio.118>.
 54. Csérmely P, Korcsmáros T, Kiss HJM, London G, Nussinov R: **Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review.** *Pharmacol Ther* 2013, **138**:333–408, <http://dx.doi.org/10.1016/j.pharmthera.2013.01.016>.
 55. Yöldöröm MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M: **Drug—target network.** *Nat Biotechnol* 2007, **25**:1119–1126, <http://dx.doi.org/10.1038/nbt1338>.
 56. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, *et al.*: **Predicting new molecular targets for known drugs.** *Nature* 2009, **462**:175–181, <http://dx.doi.org/10.1038/nature08506>.
 57. Guney E, Menche J, Vidal M, Barabási A-L: **Network-based in silico drug efficacy screening.** *Nat Commun* 2016, **7**:10331, <http://dx.doi.org/10.1038/ncomms10331>.
- An analysis of the interactome relation between drug targets and the respective disease modules showed that the therapeutic effect of drugs is localized in a small network neighborhood of the disease genes.
58. Li Z-C, Huang M-H, Zhong W-Q, Liu Z-Q, Xie Y, Dai Z, *et al.*: **Identification of drug-target interaction from interactome network with “guilt-by-association” principle and topology features.** *Bioinformatics* 2016, **32**:1057–1064, <http://dx.doi.org/10.1093/bioinformatics/btv695>.
 59. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z: **A survey of current trends in computational drug repositioning.** *Brief Bioinform* 2016, **17**:2–12, <http://dx.doi.org/10.1093/bib/bbv020>.

1.3 The multiscale organization of molecular complexity

I think the [21st] century will be
the century of complexity

Stephen Hawking (1942-2018)

Driven by recent technological advances, biomedical research is currently undergoing a profound transformation towards a data-driven science. As we map out both healthy and disease states in larger scale, higher resolution, and with greater affordability, biology is moving rapidly into the era of big data. As shown in Section 1.2, network medicine primarily driven by the interactome has achieved milestones in elucidating underlying molecular mechanisms for various diseases. Over the past decades, novel PPI have continued to be mapped out through large- and small-scale experiments, generating detailed maps of physical interactions among proteins (Huttlin et al., 2017, 2021; Drew et al., 2021; Rolland et al., 2014; Luck et al., 2020). The latest milestones in large-scale PPI mapping include the recent publications of two complementary datasets, the Bioplex 3.0¹ (Huttlin et al., 2021) and the Human Reference Interactome (HuRI)² (Luck et al., 2020), each resulting in over 118k and 90k interactions, respectively (Figure 1.1a). Nevertheless, the PPI remains largely incomplete (Menche et al., 2015) and known large-scale interactions are predisposed to biases based on methods of identification (Figure 1.1b). In addition, literature-curated PPIs collected from independent studies, which represent the majority of the human interactome are heavily biased towards highly studied genes (Figure 1.1c) and known disease causal genes (Figure 1.1d). With the multifaceted and dedicatedly orchestrated nature of biological information transfer, the reliance of capturing molecular relationships on a single layer network that is both incomplete and biased is therefore no longer adequate to advance our understanding of health and disease states at the molecular level.

To overcome this issue, increasingly large volumes of data representing different types of relationships can be leveraged. Naturally, biological phenomena can be observed across both spatial and temporal scales. These phenomena can be perceived as networks where relationships among entities may occur within and across various scales, reflecting the many levels of organization. They also play different roles in disease phenomena - from the social network among individuals at the epidemiological level, the inter-tissue and -organ crosstalk in keeping metabolic balances, to signalling and physical interactions at the molecular levels (Figure 1.2) (Sin & Menche, 2021). In the context of Mendelian diseases where this thesis primarily focuses on, disease phenotypes can be regarded as a failure in genetic information transfer in which a single mutation could lead to a cascade of effects that results in devastating and diverse set of phenotypes. The heterogeneity of diseases with the same genetic basis cannot be explained via a ‘one gene, one

¹<https://bioplex.hms.harvard.edu/>

²<http://www.interactome-atlas.org/>

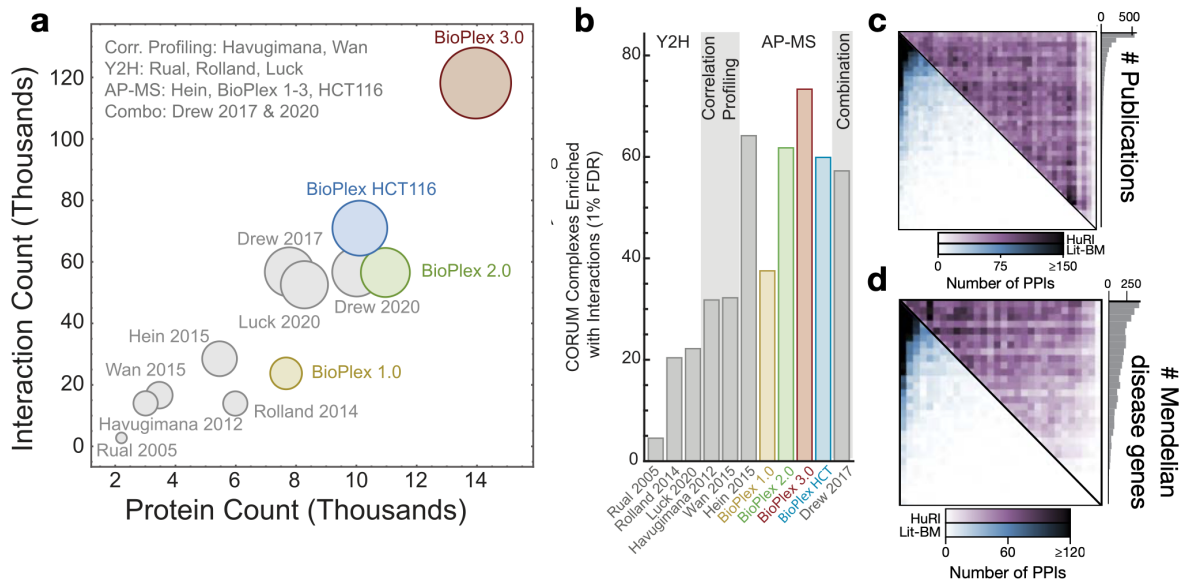


Figure 1.1: The development of protein-protein interaction databases. **a** Technological advancements have enabled larger scale discovery of protein-protein interactions. However, the largest PPI database to date only account for half of the proteome. Circle size indicates the number of interactions. **b** Different methods for identifying PPIs (Y2H: yeast-two-hybrid, AP-MS: affinity-purification mass spectrometry) and their characteristics and biases. The figure shows that affinity-purification mass spectrometry methods (BioPlex database) are more capable of retrieving protein complexes aggregated in the CORUM database. **c** Heat map showing the number of interactors of proteins ordered by the number of publications associated with the proteins. **d** A subset of **c** focusing on genes involved in Mendelian diseases. Both plots reveal that literature-curated PPIs (indicated in blue), representing the majority of the reported PPIs, show a strong bias towards highly studied genes. This effect is significantly reduced with large-scale interaction assays, such as the Human Reference Interactome (HuRI, labelled in purple).

Figure **a** and **b** are reprinted from (Huttlin et al., 2021) © Elsevier. Figure **c** and **d** are modified from (Luck et al., 2020), © Springer Nature. The reuse of the figures is in accordance with STM permission guidelines.

disease’ archetype that has previously been established (Beadle & Tatum, 1941; Cerrone et al., 2019), suggesting that the intermediary layers of information (mis)transfer may play substantial roles in modulating the disease phenotypes. This poorly understood process, together with the mounting number of undiagnosed rare disease patients, have posed significant challenges in both research and diagnostics (details discussed in Section 1.4). Consequently, efforts were made to capture molecular interaction beyond physical interactions. Emerging *-omics* technologies have enabled quantitative observation of events taking place from regulatory to metabolic levels, which allow scientists to further dissect the interplay between molecular components and their roles in disease aetiology at unprecedented details (Wong et al., 2021). Examples of comprehensive interactions resulting from such technologies include: (i) the measurements of genetic dependencies which resulted in observable genetic interactions in both small and larger scales (Costanzo et al., 2019; van Leeuwen et al., 2016; Kuzmin et al., 2018; van Leeuwen et al., 2017; Costanzo et al., 2016); and (ii) the inference of regulatory activities modulating gene expression

via co-expression networks (Seyfried et al., 2017; Pierson et al., 2015; Saha et al., 2017).

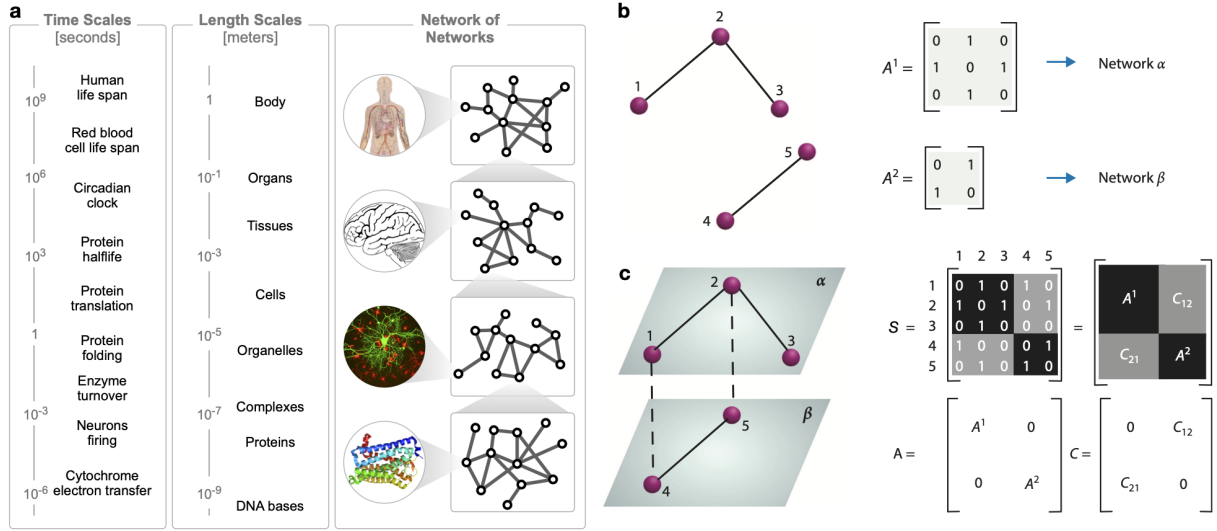


Figure 1.2: **a** Biological events naturally occur in a multiscale fashion, temporally and spatially. These events can be measured through various readouts, which their relationships can be inferred. **b** A mathematical representation of two individual simple networks. **c** Two interconnected networks, represented in the supraadjacency matrix

Figure **a** was adapted from (Sin & Menche, 2021) © Cambridge University Press. Figure **b** was modified from (Aleta & Moreno, 2019), © Annual Reviews. The reuse of the figures are according to STM permission guidelines.

Not all knowledge has been or can be generated through high-throughput experiments. For example, only 13.2% of all PPI data from the Human Integrated Protein–Protein Interaction rEference database (HIPPIE)³ were validated by two largest proteome-wide PPI assays, i.e., the Bioplex or HuRI databases. The majority of information about genes and proteins characterized throughout the past decades were conducted in independent studies and published as scientific literatures. Efforts have been made to transform this knowledge computer-readable via annotation in the form of ontologies and pathway databases (Mungall et al., 2017; Köhler et al., 2014; Kibbe et al., 2015; Smith & Eppig, 2009; Ashburner et al., 2000). Definitions and examples of such biological ontologies are overviewed in Section 1.1. These structured data have enabled the quantification of similarity between entities based on their annotation, i.e., *semantic similarity* (Pesquita, 2017; Žitnik et al., 2013). Functional relationships inferred from such computation has enabled the construction of genome-wide networks based on annotation similarity such as gene functions (Pesquita et al., 2008), pathways (Aguirre-Plans et al., 2019; Fabregat et al., 2018) and phenotypes (Goh et al., 2007; Haendel et al., 2015a; Yu, 2020). In addition, population-level health records have also been used as a resource to redefine disease relationships and comorbidity (Strauss et al., 2021; Hu et al., 2016; Jensen et al., 2014). These relationships have been established as additional layers to investigate disease phenomena.

³<http://cbdm-01.zdv.uni-mainz.de/mschaefer/hippie/>

The multiscale nature of complex systems does not only introduce additional complexity through the increased number of nodes in consideration, but also through their cross-layer interconnectivity via additional edges. Further mathematical formulation was developed to express such data. Fortunately, the inherent universality of such formulation to many complex systems have attracted scientists across all domains to develop various representations (Kivelä et al., 2013; Aleta & Moreno, 2019). As a result, ‘*multilayer networks*’ quickly became one of the fastest growing sub-discipline in network science - owing to synergistic methodological developments and characterization (Liu et al., 2020; Nicosia & Latora, 2015; De Domenico et al., 2016), and also a tool for data integration (Valdeolivas et al., 2019; Cho et al., 2016; Himmelstein et al., 2017; Cava et al., 2018; Huang et al., 2018). A framework of representation employed in this thesis, which is outlined in the Methods section of the main manuscript (Section 3.1) is based on the following formulation (also illustrated in Figure 1.2 b):

Consider two networks α and β , represented by tuples $G_\alpha = (V_\alpha, E_\alpha)$ and $G_\beta = (V_\beta, E_\beta)$ with number of vertices $|V_\alpha| = m$ and $|V_\beta| = n$, and number of edges $|E_\alpha|$ and $|E_\beta|$ respectively. Mathematically, the relationship *within* each network is stored in adjacency matrices $A_\alpha \in \mathbb{R}^{m \times m}$, and $A_\beta \in \mathbb{R}^{n \times n}$, respectively. To describe *interlayer* adjacency, additional matrices $C_{\alpha\beta} \in \mathbb{R}^{m \times n}$ and $C_{\beta\alpha} \in \mathbb{R}^{n \times m}$ are introduced to store inter-connectivity of vertices for $\alpha \rightarrow \beta$ and $\beta \rightarrow \alpha$, respectively, i.e. $c_{\alpha\beta}^{ij} = 1$ if there is a link connecting node i in layer α to node j in layer β . Several representations exist to depict both intra- and interlayer adjacencies (Kivelä et al., 2013). A convenient and yet powerful representation is the supra-adjacency representation where existing tools and frameworks described for matrices that represent monolayer networks can be immediately applied. In such representation, the supra-adjacency ($S \in \mathbb{R}^{(m+n) \times (m+n)}$) of the network layers introduced above is given by:

$$S = \begin{pmatrix} A_\alpha & C_{\beta\alpha} \\ C_{\alpha\beta} & A_\beta \end{pmatrix}$$

In the context of this thesis, where different layers reflect gene-centric relationships and all layers therefore represent the same amount of nodes, i.e., genes, such networks are conventionally known as ‘multiplex networks’. The supra-adjacency implementation allows multilayer or multiplex networks to be viewed as one ‘supra-graph’. This enables methodologies for network characterization and propagation introduced in Section 1.1 to be readily applied. In such a setting, cross-layer propagation allows network-based approaches to be established as a data integration tool for tasks such as gene prioritization (Valdeolivas et al., 2019) or drug repurposing (Ruiz et al., 2021). However, there remains open questions such as whether integrating more information immediately translates into better mechanistic understanding of the system or improve the accuracy of the prioritization. A crucial part of the thesis is the evaluation of significance of biomolecular network layers as well as how to integrate such data to recapitulate the information transfer in the system. In Section 3.1, a methodological framework was

constructed to address such issues. We applied the concept of disease modules to quantify the relevance of different network layers. This allows us to incorporate context awareness into a multilayer network propagation algorithm which was demonstrated to significantly improve the accuracy of disease gene identification in a large rare cohort of rare disease patients.

1.4 Rare diseases

When you hear hoofbeats behind
you, don't expect to see a zebra

Theodore E. Woodward
(1914-2005)

The analogy (along with its numerous variations) accredited to a medical professor Theodore Woodward in the 1940s has widely become common practices in medical diagnosis (Dickinson, 2016). For a medical practitioner to diagnose a patient presented with a set of symptoms, this means that they first think of common diseases. Albeit sensibly so, this practice has neglected a group of patients who are affected by rare diseases. It can take years, or in unfortunate cases, the patients' entire lives to receive correct diagnoses and treatment, even with recent advancements in clinical diagnostics (Graessner et al., 2021). And unlike the name suggests, the collective number of rare disease patients can pose a tremendous challenge to the healthcare sector. Clearly too many zebras were mistaken as horses.

Rare diseases, as defined by the EU, are diseases with a prevalence of less than 1 in 2000 (Rode, 2005). To date, over 7,000 rare diseases are known (Figure 1.3), and collectively they affect up to 8% of the population. Approximately 36 million people in the EU are or will be affected by one of these diseases (Julkowska et al., 2017). To put into perspective, this is an equivalent number to the population suffering from diabetes (Tamayo et al., 2014). The difference is that rare diseases are extremely diverse in terms of phenotypes and causalities, many of them have severe or multi-organ symptoms, half of them affect children, and many are left undiagnosed.

Throughout the development of molecular biology, techniques developed in laboratories have quickly turned into diagnostic tools (Figure 1.3) (Hartley et al., 2020; Lalonde et al., 2020). This started at the era of optogenetics in the discovery of trisomy of chromosome 21 as a cause of Down syndrome (Jacobs et al., 1959) during the 1950s, following the karyotyping of human chromosomes (Ford & Hamerton, 1956). A few decades later, with the discovery of restriction enzymes, known as 'molecular scissors' (Nathans & Smith, 1975), scientists were able to manipulate the genetic molecules directly, the era of DNA-based diagnostics has begun. These discoveries have contributed to the recent development of high-throughput sequencing used in modern day practices. Many Mendelian diseases are caused by a single mutation, the single nucleotide variation (SNV), and the vast majority (85%) of them are believed to lie in the exome - the protein coding regions that contribute to 1-2% of the genome (Botstein & Risch, 2003). With higher cost efficiency compared to whole genome sequencing, technologies such as panel sequencing and later exome sequencing have promptly been integrated into current diagnostic pipelines (Figure 1.3b). This has enabled the screening of all rare and functional

variants that individuals may possess, and identified the causality in many diseases (Yang et al., 2013; de Ligt et al., 2012). Despite this technological leap, it is estimated that the diagnostic yield of rare disease patients undergoing exome sequencing is only around 40% (Wright et al., 2018). This number falls in the same range as the information from our local statistics from the Ludwig Boltzmann Institute for Rare and Diagnosed Diseases (LBI-RUD), Austria's leading research institute for rare disease diagnostics and therapeutics (Figure 1.3b). With nearly half of the patients undiagnosed, efforts have been made to reanalyze the data via pooling of patients from multiple cohorts, use of different variant thresholding, apply practices of data sharing and structured data storage and analyses (Zurek et al., 2021; Matalonga et al., 2021).

With around 4 million genetic variations in a genome (or about 20 thousands in an exome) (Wright et al., 2018), it is needed to further narrow down the list is to identify rare and putative pathogenic variants, a process known as variant prioritization. Various tools are typically used in combination to evaluate the likelihood of variant pathogenicity. Table 1.1 lists tools employed in the Genome-Phenome Analysis Platform (GPAP) of the RD-Connect Project to identify rare and likely pathogenic variants. This is a two-stage approach: First, identifying rare variants by comparing variants detected in a patient with a pool of references, typically aggregated from exome or whole genome sequencing of healthy individuals (1000 Genomes Project Consortium et al., 2015; Karczewski et al., 2020). Variants with allele frequency (AF) of less than 1% is typically classified as rare. Second, identified rare variants undergo pathogenicity prediction. This usually involves information on protein structure and stability as well as conservation of the homologous sequences across multiple species, or a combination of different metrics into a single score (Adzhubei et al., 2013; Rentzsch et al., 2019; Schwarz et al., 2014; Ng & Henikoff, 2003). These criteria are not standardized and different diagnostic laboratories adopt different cutoffs in their routines. Figure 1.3c illustrates the diagnostic pipelines as well as the estimated number of variants remaining for each step.

After prioritizing for rare and potentially pathogenic variants, there remain on average 400 variants that met the stringent criteria (Wright et al., 2018). At this level, additional information such as mode of inheritance, family history, as well as phenotype information and other evidences are incorporated to further narrow down the list of potential causal variants. In addition, resources and platforms such as RD-Connect⁴ (Thompson et al., 2014; Zurek et al., 2021) for storing and sharing genetic information and PhenomeCentral⁵ (Buske et al., 2015) for matchmaking patients with similar phenotypes, have enhanced global data sharing and standardizing clinical annotation (Köhler et al., 2017; Mungall et al., 2017). Furthermore, screening platforms have been developed on cellular and model organism level to facilitate experimental validation and accelerate the discovery of novel treatments. The International Mouse Phenotyping Consortium (IMPC)⁶ aims to address the lack of knowledge regarding gene functions and

⁴www.rd-connect.eu

⁵<https://www.phenomecentral.org>

⁶www.mousephenotype.org

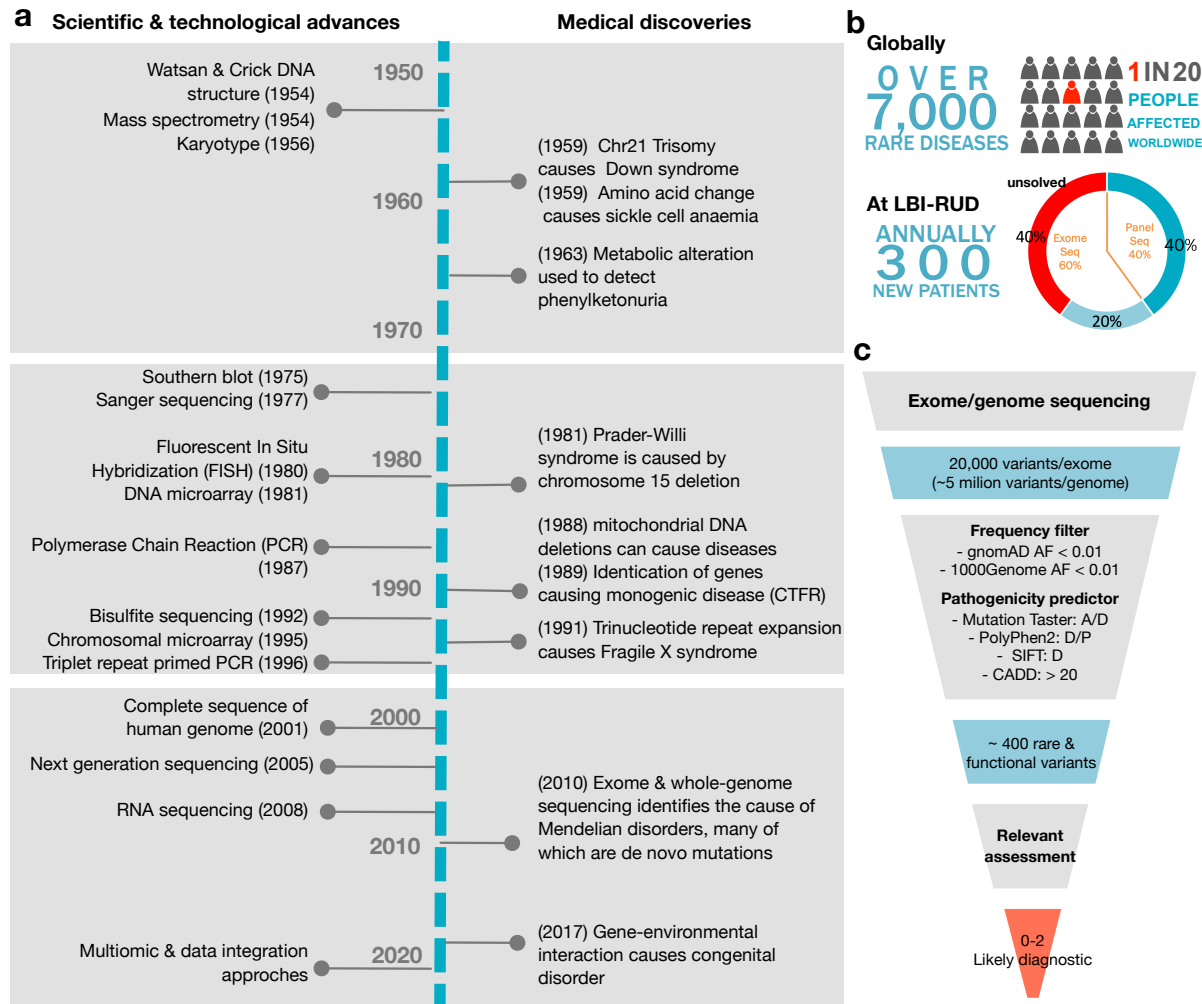


Figure 1.3: Rare disease timelines, statistics, and variant prioritization. **a** The timeline of scientific, technological and translational advancement in genetic diagnostics. The left column shows basic discoveries and techniques that have later been translated into biomedical diagnostics. The grey boxes mark major epochs based on diagnostic technologies at the time, from chromosomal and metabolic-based (before 1970s) to low-throughput DNA-based molecular diagnostics (1970s-2000s), and high-throughput sequencing (present), respectively. The timeline was compiled based on (Hartley et al., 2020; Lalonde et al., 2020). **b** Infographics of rare disease statistics. Top: there are over 7,000 rare diseases, resulting in a collective prevalence of 1 in 20 globally. Bottom: statistics from Austria's rare disease research institute, the Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases (LBI-RUD). There are 300 patients who underwent clinical diagnosis. Notably, 40% of the patients remained undiagnosed. **c** An example variant prioritization pipeline. Grey boxes mark techniques or filtering methods. Coloured boxes represent number of variants at each step. Data were taken from (Wright et al., 2018). The example criteria for frequency and pathogenicity filtering are based on high stringent threshold in the RD-Connect Genome-Phenome Analysis Platform. It is notable that even after stringent filtering for rare and likely pathogenic variants, there remain 400 variants on average where further contextualization and prioritization are required to pinpoint the true causal variant.

Resource	Description	Score range	Deleterious cutoff
Allele Frequency			
GnomAD	Reference exomes and genomes of over 100k individuals	0-1	<0.01
1000Genome	Aggregated common genetic variation from healthy individuals	0-1	<0.01
Single Nucleotide Variant (SNV) Effect predictor			
Mutation Taster	Conservation and structure	-	Disease causing (D,A)
PolyPhen2	Conservation and structure	0 to 1	>0.453 (P,D)
SIFT	Conservation	1 to 0	<0.05 (Damaging)
CaDD	Combines multiple resources	0 to 35+	>20

Table 1.1: Two main steps in variant annotation for rare diseases. Databases and threshold criteria based on the RD-Connect Genome-Phenome Analysis Platform: (1) Allele frequency: to identify rare variants compared to reference population in the databases. (2) Pathogenicity prediction: to estimate the degree of damaging effects of variants based on sequence conservation, and predicted effect of variants in structural and functional domains of resulting proteins. Combinations of different tools and thresholds are used in different diagnostic laboratories. Abbreviations: SNV = Single Nucleotide Variation; Mutation Taster classes: D = disease causing, A = annotated and disease causing; PolyPhen2 classes: P = possibly damaging, D = damaging.

pathogenicity by creating a genome- and phenome-wide ‘disease model’ catalogue of knockout mice (Meehan et al., 2017; Perry, 2017). Consortia such as Solve-RD also developed platforms where previously undiagnosed cases are catalogued and re-analyzed in light of new knowledge and data (Zurek et al., 2021; Wright et al., 2018).

Several tools have been developed to exploit the success of genomic diagnostics and the increasing amount of genotype-to-phenotype resources. This includes the utilization of cross-species phenotype and protein interaction information (Smedley & Robinson, 2015; Haendel et al., 2015b; Robinson et al., 2014), the elevated gene expression level in relevant tissues (Feiglin et al., 2017; Frésard et al., 2019), or the utilization of text mining to search the scientific literature for relevant information (Birgmeier et al., 2020). However, as data grow in both variety and volume, the questions that follows are (i) how to justify which data to be included in the discovery process, and (ii) with data being heterogeneously stored and represented, and often with varying quality, how to unify and integrate such data while (iii) enabling interpretability of the results.

Aims of the thesis

The rise of multifaceted molecular and phenotypic data (Section 1.3) opens up new opportunities to complement the physical interactome (Section 1.2) with additional maps of molecular connectivity for navigating disease relationships and understanding their underlying pathogenicity. The aims of this doctoral thesis are to apply such data to the challenges of prioritization and interpretation of rare disease causality (Section 1.4) through the following steps: First, we construct networks based on heterogeneous data types that represent different molecular layers and characterize their topological properties. This is to assess whether the resulting networks are redundant or complementary as well as whether they reflect intrinsic organizational principles at different biological scales. Second, we quantify whether the disease modularities of groups of rare diseases with similar phenotypical characteristics resemble those observed in common diseases, and whether they are observed in other molecular scales beyond the interactome. This will allow us to generalize a core concept of network medicine, disease modules, to be applied across multiple scales of biological organization. Thirdly, with different disease groups showing various levels of modularities across scales, we aim to incorporate this information into a network-based disease gene prioritization algorithm where the level of propagation is varied based on the modularity level of the disease in a particular network. Additionally, this information also allows us to elucidate the pathobiological mechanisms that are encoded in the modules of a particular rare disease group. Next, we validate the predictive power of cross-scale disease modules through our implemented informed propagation algorithm for disease gene identification and test whether incorporating cross-scale network modularity leads to better performance than using individual networks or all networks in an uninformed manner. This enables new network-based strategies for identifying and integrating the most relevant datasets for a particular biological application. Finally, we apply the framework to aid in the genetic diagnosis of individual patients using a cohort of patients suffering from rare neurological diseases.

3

Results

3.1 Main publication

PUBLISHED ARTICLE

Network Analysis Reveals Rare Disease Signatures Across Multiple Levels of Biological Organization

Pisanu Buphamalai*, Tomislav Kokotovic, Vanja Nagy, and Jörg Menche

Published in **Nature Communications**, volume 12, Article number: 6306 (2021)

© 2021, The Author(s)

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format.



To achieve the goals of using networks both as a tool for data integration and interpretation, we constructed multiplex networks from public resources, representing six major biological scales from genetic interactions to phenotypic similarities, and consisting of over 20 million gene relationships. A comprehensive comparison between the networks from different biological scales reveals that they offer distinct information that can help compensate for technological and experimental biases, as well as for incompleteness of individual layers. Moreover, we show that the so-called ‘disease modules’, the tightly connected subnetwork among genes causing the same disease initially identified on the interactome, are differentially clustered across different network scales. Disease genes appear to cluster more strongly, which results in more apparent modules, on the network layers that reflect pathobiological mechanisms of the disease. Next, we quantified this network modularity score and implemented a mathematical framework to predict potential causal genes in rare diseases, the task that have been proven challenging due to the data scarcity. The ‘informed multiplex network propagation’ developed in this manuscript has been extensively tested on data from independent cohorts of patients characterized by intellectual disability to prioritize their causal genes. The framework correctly identified true causal genes in the patients more accurately compared to methods utilizing only one network layer, or all networks layers in an uninformed way. Overall, the results presented in this manuscript not only further the fundamental understanding of how genomic aberrations impact various levels of organization, but also offer a novel platform to systematically explore the molecular origins of rare diseases.

ARTICLE

<https://doi.org/10.1038/s41467-021-26674-1>

OPEN

Network analysis reveals rare disease signatures across multiple levels of biological organization

Pisanu Buphamalai ^{1,2}, Tomislav Kokotovic^{1,3,4}, Vanja Nagy^{1,3,4} & Jörg Menche ^{1,2,5}✉

Rare genetic diseases are typically caused by a single gene defect. Despite this clear causal relationship between genotype and phenotype, identifying the pathobiological mechanisms at various levels of biological organization remains a practical and conceptual challenge. Here, we introduce a network approach for evaluating the impact of rare gene defects across biological scales. We construct a multiplex network consisting of over 20 million gene relationships that are organized into 46 network layers spanning six major biological scales between genotype and phenotype. A comprehensive analysis of 3,771 rare diseases reveals distinct phenotypic modules within individual layers. These modules can be exploited to mechanistically dissect the impact of gene defects and accurately predict rare disease gene candidates. Our results show that the disease module formalism can be applied to rare diseases and generalized beyond physical interaction networks. These findings open up new venues to apply network-based tools for cross-scale data integration.

¹CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria. ²Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna BioCenter 5, 1030 Vienna, Austria. ³Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria. ⁴Department of Neurology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria. ⁵Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria. ✉email: joerg.menche@univie.ac.at

Over the past 2 decades, rapid advances in DNA sequencing technology allowed us to uncover the genetic basis of over 6000 rare diseases^{1–3}. In contrast to common diseases, which are typically characterized by a complex interplay between multiple genetic and environmental factors, rare diseases can often be pinpointed to a single genetic lesion. Rare diseases thus offer unique opportunities to mechanistically dissect the relationship between genetic aberrations and their phenotypic consequences, which can then inform targeted treatment strategies. For individual rare diseases, this potential for a molecularly rooted, personalized medicine could already be demonstrated, for example in rare immunodeficiencies^{4–6}, neurodevelopmental^{7,8}, and metabolic disorders^{9,10}. At the same time, the costs and extended timelines of these individual efforts also highlight the need for novel, systematic approaches for investigating the large number of rare diseases that still remain uncharacterized. To this end, several practical and conceptual challenges need to be overcome:

First, rare disease phenomena cover a wide spectrum, from highly cell-type or organ-specific phenotypes to heterogeneous, syndromic diseases that affect the whole body. Our understanding of how a genetic aberration impacts various scales of biological organization between genotype and clinical phenotype is very limited. Second, the enormous complexity within and between different organizational scales, such as the transcriptome, proteome, intra- or intercellular communication, also poses important technical challenges: How can we identify and integrate the most relevant data? Third, the rarity of many conditions with monogenic origins implies that data are usually scarce. Traditionally, rare diseases have been studied following a one-gene, one-pathway, one-disease paradigm. A systematic approach for transferring knowledge from one rare disease to another, and for investigating differences and commonalities between different diseases, is still missing.

In this work, we propose a network-based framework for systematically investigating rare diseases that addresses these challenges, and, in turn, use the large number of rare diseases with a well-described genetic origin to deepen our understanding of disease-associated perturbations of molecular networks. Specifically, we introduce a multiplex network approach for integrating different network layers that represent different scales of biological organization ranging from the genome to the transcriptome and the phenotype. A systematic characterization of the network signatures of all rare diseases with known genetic causes allowed us to identify the connectivity patterns that determine the importance of a particular scale of biological organization for a given rare disease. Finally, we explored how these systems-level insights may help contextualize individual genetic lesions, investigate the impact of disease heterogeneity, and be translated into clinically actionable tools for the genetic diagnosis of rare disease patients with unknown gene defects.

Results

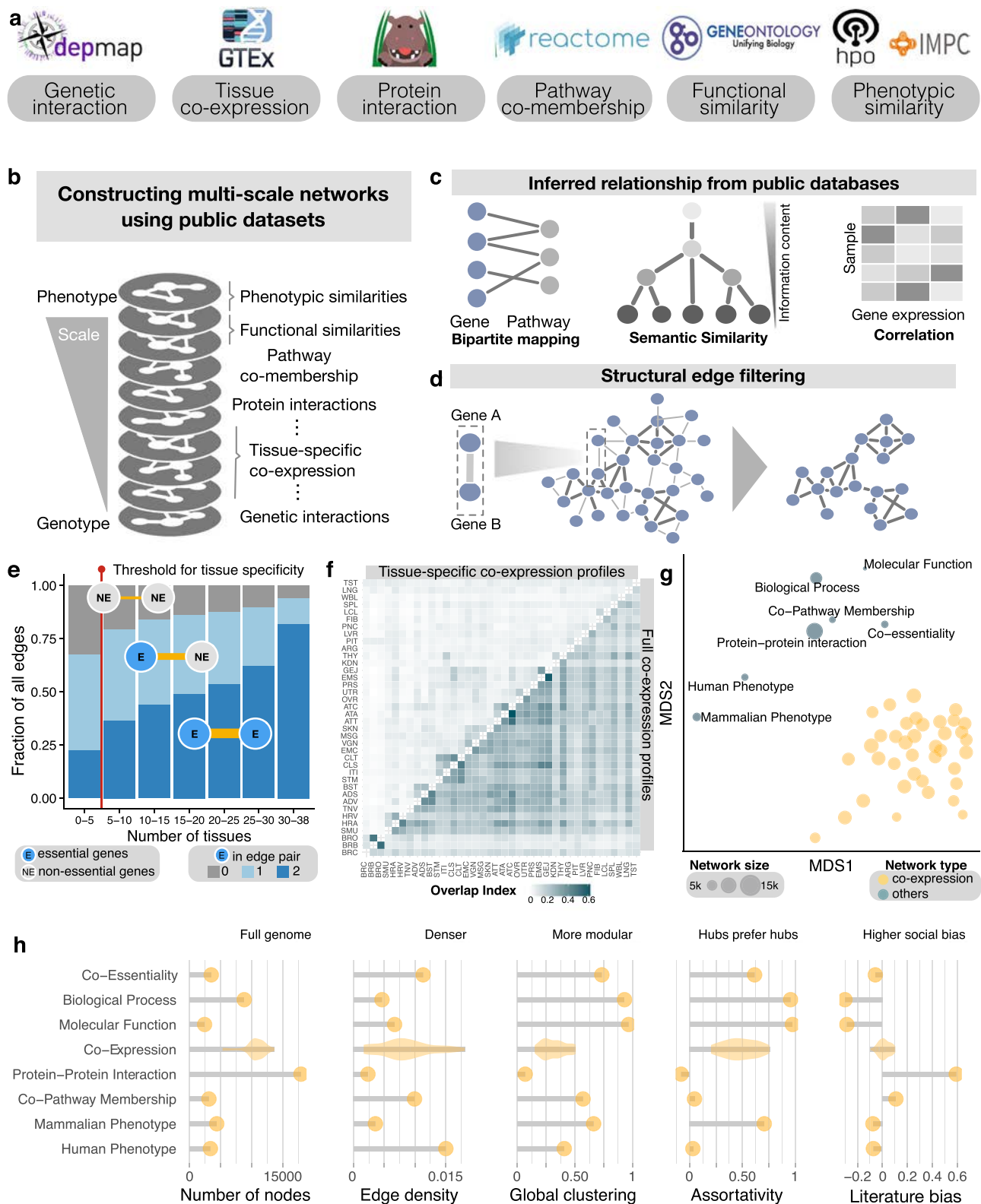
Constructing a gene network bridging molecular and phenotypic scales. Rare diseases affect many scales of biological organization which, conversely, may provide valuable information for elucidating a particular gene defect. At the genetic level, for example, interplay between genetic variants can modulate phenotypic outcomes¹¹ or even completely rescue disease-associated variants¹². At the protein level, members of the same complex or pathway are often implicated in similar phenotypes^{13,14} and expression patterns of a particular gene can reveal affected cell types and tissues^{15–17}. Finally, phenotypic similarities with known human or animal model gene defects can guide the annotation of genetic variants with unknown consequences¹⁸.

To integrate these diverse relationships into a unifying, gene-centric framework, we constructed a multiplex network

comprised of several layers: The nodes in each layer represent genes, the links represent their respective relationship at a particular scale of biological organization, ranging from direct interactions between gene products at the molecular level to phenotypic similarity of associated diseases at the phenotype level (Fig. 1a, b). We compiled information from seven databases and, where appropriate, applied a range of techniques for extracting gene relationship, such as bipartite mapping, ontology-based semantic similarity metrics and correlation-based relationship quantification, as well as filtering based on both statistical and network structural criteria¹⁹ (Fig. 1c, d and Supplementary Figs. 1 and 2, see Methods for details). The resulting multiplex network consisted of 46 layers containing over 20 million relationships between 20,354 genes (Supplementary Data 1 and 2). The relationships represent six major biological scales: (i) The genome scale, where links represent genetic interactions derived from CRISPR screening in 276 cancer cell lines²⁰. (ii) The transcriptome scale, where interactions represent co-expression, i.e., co-variability of gene transcription levels indicative of higher-level regulatory mechanisms. We included both pan-tissue and tissue specific networks derived from RNA-seq data across 53 tissues in the GTEx database¹⁷. (iii) The proteome scale, where links represent physical interactions between gene products obtained from the HIPPIE database²¹. (iv) The pathway scale, where links represent pathway co-membership derived from the REACTOME database²². (v) The scale of biological processes and molecular functions, where links represent similar functional annotations derived from the Gene Ontology²³. (vi) The phenotypic scale, where links represent similarity in annotated phenotypes derived from the Mammalian and Human Phenotype Ontologies (MPO and HPO)^{24,25}.

Characterizing the network architectures across biological scales.

To characterize the resulting cross-scale gene relationships, we first quantified the global similarity between all pairs of network layers *A* and *B* by the min overlap of their respective sets of edges *E*: $S_{AB} = |E_A \cap E_B| / \min(|E_A|, |E_B|)$. The highest similarities were found within the transcriptomic scale: co-expression networks of different tissues have an overlap of up to $S = 0.49$ (between brain tissues), compared to an average similarity of $S = 0.05$ between networks of other scales. A major contribution to this elevated similarity is given by a core of links that is preserved across multiple tissues. We found that the proportion of links that connect essential genes increases with the number of tissues in which a particular link is present (Fig. 1e). This suggests that the common core is related to essential housekeeping activities. To represent pan-tissue and tissue-specific interactions separately, we extracted broadly preserved co-expression edges and considered them as a separate core transcription network layer, consisting of 12,364 nodes and 1,062,924 edges (Supplementary Fig. 1c, d, Methods). We further combined redundant tissue types, resulting in a final set of 38 tissue-specific networks used in the downstream analyses (Fig. 1f, Methods, Supplementary Data 3). These tissue-specific networks still form a recognizable cluster within the multidimensional scaling (MDS) projection of the relative similarities between all networks (Fig. 1g). The differences between tissues, however, are comparable with differences to networks of other scales (median similarity among tissues: $S = 0.043$; similarity to other scales: $S = 0.018$). The clear separation between most network layers (median similarity $S = 0.033$) indicates that each layer contains unique information (Supplementary Fig. 3b). At the same time, a comparison with randomized networks reveals that a significant amount of interactions are preserved across levels of organization (Supplementary Fig. 3c, d), as shown by a significant similarity for 96.5% of



all network pairs (empirical p -value < 0.05 , see Methods). Finally, we noticed that the relative position of all network layers in Fig. 1g suggests a representative role for the protein-protein interaction (PPI) layer, which is located in a central position and close to the layers that directly encode phenotypic similarities.

We next compared the networks at the different biological scales in terms of five structural characteristics: genome coverage,

overall connectivity, clustering, assortativity and literature bias (Fig. 1h, Methods). The results revealed a wide structural diversity: The network layer with the highest genome coverage is the PPI scale, covering 17,944 proteins. This is due to the combination of a large number of literature curated small-scale experiments and several large-scale screening efforts. Such systematic, genome-wide measurements also underlie the high

Fig. 1 Construction and characterization of the cross-scale multiplex network. **a** Data resources for the major biological levels of organization represented in the multiplex network. **b** The multiplex network consists of 46 network layers, each representing a particular type of gene relationships, ranging from genetic interactions to phenotypic similarity. **c** Methods used for inferring networks: bipartite mapping was used to build gene relationships based on common annotations, e.g., pathways; semantic similarity was used to define relationships based on annotation similarity; correlation analyses were used to identify co-expression. **d** Weighted and dense networks were subsequently filtered based on structural network criteria for extracting the most relevant interactions. **e** Co-expressed gene pairs found in a higher number of tissues tend to be essential, reflecting core cellular functions. Edges found in five or fewer tissues were considered tissue-specific. **f** Full co-expression profiles are highly similar between tissues and thus redundant (lower triangle). The removal of core transcription profiles reveals tissue-specific patterns of the co-expression networks (upper triangle). **g** The multi-dimensional scaling (MDS) plot based on edge overlap similarity of all networks shows a clear distinction between major types and subtypes of the network layers. **h** Major network characteristics for all considered network layers: number of nodes edge density, global clustering, assortativity and social bias, as measured by the correlation between node degree and number of associated publications. The values of the 38 individual co-expression networks are shown in the form of a distribution.

coverage of the transcriptomic layers (with a total number of $N = 17,432$ genes across all tissues, and an average number of 10,527 genes per tissue, Supplementary Fig. 1d, see Methods for the filtering processes). Our incomplete understanding of how these molecular interactions translate into biological processes, however, is indicated by the low coverages observed among the functional and phenotypic levels ($N = 2407$ and 3342 for the molecular function and HPO networks, respectively). The high connectivity and clustering among these functional layers, in turn, is the basis for their predictive power for transferring gene annotations within functional clusters^{11,20} (e.g., edge density = 1.13×10^{-2} and clustering = 0.73 for the co-essentiality network). The PPI represents the sparsest network (edge density = 2.359×10^{-3} ; average density across all layers = 7.76×10^{-3}), which, in part, reflects the incompleteness of currently available data²⁶. Curiously, the PPI is the only network in our collection that exhibits a (modest) level of disassortativity ($r = -0.08$), i.e., a tendency of hubs to connect preferentially to low-degree nodes, a property that was previously suggested to be a universal feature of biological networks²⁷. Disassortativity may arise when the neighbors of high-interest nodes are mapped out more extensively than the interaction partners of these neighbors. For the PPI, this is likely to be the case in network data curated from hypothesis-driven, small-scale experiments, but can also occur in unbiased large-scale efforts (Supplementary Fig. 3e, f, Methods). A further characterization of curated and unbiased subsets of the PPI (Supplementary Fig. 4a, b, Methods) revealed that the relatively high literature bias present in the PPI, as measured by the correlation between the degree of a protein and the number of associated publications (Spearman's $\rho = 0.59$, Supplementary Fig. 4b), is largely driven by its curated subset, which represents 87% of the full PPI. This emphasizes that despite recent efforts towards unbiased, high-throughput protein–protein interaction screening, a large fraction of the currently available PPI network information still reflects the reductionist, hypothesis driven research paradigm where new knowledge preferentially accumulates around proteins with an already known important function. This literature bias is notably absent in all other network layers.

In summary, the structural diversity observed among the individual network layers reflects both organizational principles intrinsic to a particular biological scale, as well as technical or historical details pertaining to the curation process of the underlying database (Supplementary Fig. 4c–f). We expect that this diversity further corresponds to complementary pieces of information contained in the different biological scales, collectively increasing their potential to drive novel insights into the relationships between rare disease genes.

Identifying cross-scale network signatures of rare diseases. To investigate the connectivity patterns among rare disease genes, we collected 3953 genes associated with 3771 rare disease terms from

the Orphanet database, the largest rare disease ontology and resource for genetic associations (Supplementary Data 4). Collectively, rare diseases represent an extraordinarily rich resource of causative genetic aberrations and their phenotypic consequences. For individual rare diseases, however, the situation is the opposite: Over 3501 diseases in the Orphanet database (~93%) are associated with fewer than five genes. This represents a major challenge for systematic, comparative rare disease research in general, and for network-based approaches in particular: Network approaches are based on the fundamental observation that genes associated with the same disease are not scattered randomly in molecular networks, but aggregate in disease-specific neighborhoods or “disease modules”^{26,28}. However, the incompleteness of currently available network maps sets a lower bound for the number of genes that can be recognized as a connected module. This minimal number was estimated to be around 20 for the PPI network²⁶, so that individually, only few rare diseases have a sufficiently large number of associated genes.

We hypothesized that the disease module concept can be generalized to groups of rare diseases with closely related phenotypes. Collectively, these related rare diseases could thus reach the required minimum number of genes to form a recognizable disease module (Fig. 2a). To test this hypothesis, we used the hierarchical classification of rare diseases within the Orphanet Disease Ontology to aggregate rare diseases with similar phenotypes and collect all genes associated with their corresponding descendant terms. We identified a total of 26 rare genetic disease groups that are sufficiently broad or well-studied, respectively, to result in a number of associated genes required for network module approaches (i.e., more than 20), while retaining the pathophysiological specificity of rare disease phenotypes (Supplementary Fig. 5a). The disease groups range from smaller groups, such as RASopathy (ORPHA:536391) or rare genetic vascular diseases (ORPHA:233655) (with 20 and 22 associated genes, respectively), to large groups with over 1000 associated genes, such as rare genetic neurological disorder (ORPHA:71859) or rare developmental defect during embryogenesis (with 1649 and 1598 associated genes each). The average number of genes per disease group was 339 (Fig. 2b and Supplementary Data 5). Despite the wide range in the total number of associated genes per disease group, the average number of genes per disease term remains comparable across all disease groups, thus ensuring similar levels of disease specificity across the disease domain. In addition, there is only little overlap between the disease terms contained in the different groups, with 90.5% of all disease pairs being distinct (Jaccard Index < 0.1), indicating that the groups provide non-redundant disease definitions (Supplementary Fig. 5b).

We first inspected the network localization of the aggregated rare disease groups within two-dimensional network embeddings obtained from the node2vec algorithm²⁹, which aims to preserve

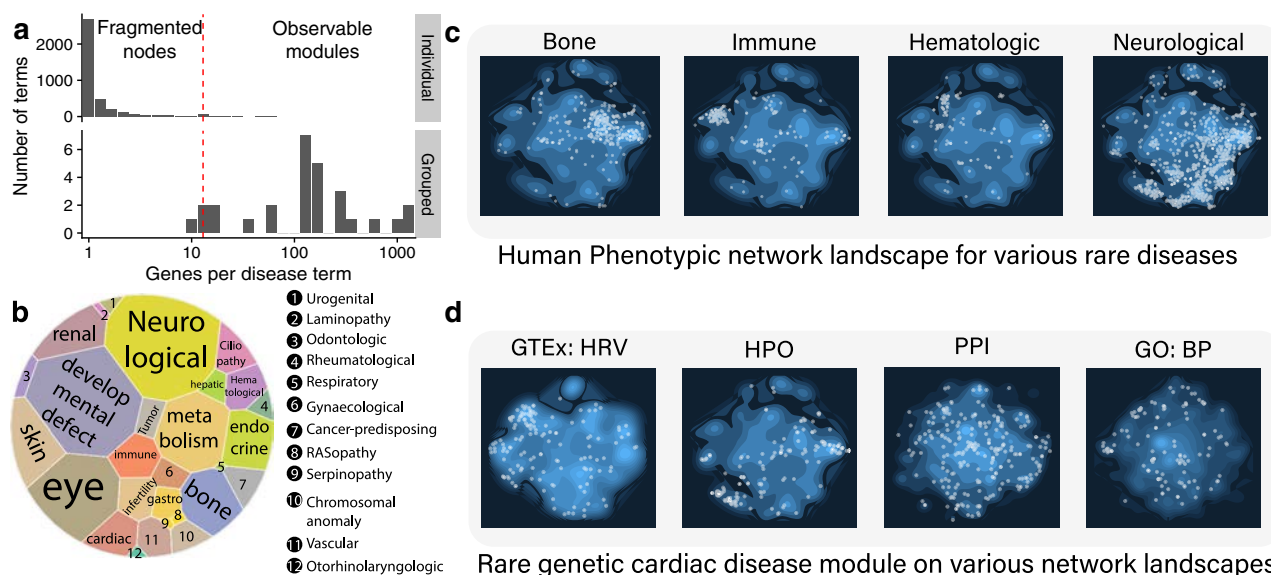


Fig. 2 Rare disease grouping and network mapping reveal network- and disease-specific connectivity patterns. **a** Rare genetic diseases are typically associated with only a few genes and therefore remain fragmented on molecular networks. Grouping rare diseases by phenotypic similarity can overcome data scarcity and result in identifiable disease modules, thus allowing for further network-based inspection. **b** Voronoi treemap showing the 26 rare genetic disease groups used in this study. The size of each disease group is proportional to the number of associated genes. **c** Network landscape obtained using the node2vec embedding algorithm. Network distances between genes are preserved in the embedding and illustrate differential modularity of different rare disease groups on the Human phenotype similarity network layer. The bright dots represent disease associated genes and the blue contour map represents all genes in a network. **d** Localization of the rare cardiac disease group on different network layers.

network distances between nodes (Supplementary Fig. 5c, Methods). Fig. 2c shows the resulting network landscape of the human phenotypic network with four rare disease groups highlighted: rare bone, immune, hematologic and neurological diseases (the complete landscape of all diseases in all networks can be explored via our MultiOmeExplorer web app: www.menchelab.com/MultiOmeExplorer, see also Supplementary Fig. 6). We found that all four disease groups localize within specific network neighborhoods. Given that the HPO network is based on phenotypic similarity of individual gene defects, this localization confirms that aggregating diseases based on disease ontology relationships indeed leads to groups of phenotypically related diseases. We further noticed that the different disease groups cover network areas of varying size, from highly localized immune diseases to more broadly spread neurological diseases. In part, this spread can be attributed to the larger number of genes associated with the latter disease group. More generally, it may reflect varying degrees of coherence and specificity among the phenotypic manifestations of the diseases represented within a particular group. The close relationship between the spread of disease-associated perturbations within molecular networks and the heterogeneity of clinical symptoms has previously been shown for complex diseases³⁰. Similarly, the spread of the rare neurological disease cluster recapitulates the high level of comorbidities observed among affected patients. Finally, we noted that the proximity between neighborhoods is indicative of disease similarity, e.g., between rare immune and hematologic diseases, where the interplay between blood and immune system often leads to similar phenotypes.

We next inspected the network signatures of rare disease groups across different network layers. Figure 2d shows that rare genetic cardiac diseases are strongly localized on a heart-specific co-expression network (heart right ventricle; HRV) and the human phenotypic similarity network (HPO). The more dispersed signals on the PPI network and the network of shared biological processes (GO:BP), on the other hand, suggest that the

respective genes might be involved in a broad range of molecular processes that cannot be adequately depicted in a two-dimensional projection.

Quantifying network modularity of rare diseases. The results so far indicate that the concept of disease modules, observed widely across complex diseases on PPI networks, can also be generalized to groups of rare diseases and to other network data representing relationships beyond the molecular scale of PPIs. Based on the heterogeneous degrees of modularity for different diseases and networks observed above, we further hypothesized that the degree of modularity can be related to the degree of relevance of the underlying information to a particular disease phenotype. To investigate this hypothesis and further dissect the characteristics of rare disease modules across biological scales, we systematically assessed all rare disease groups across all network layers. We quantified the level of modularity by the significance of the size of the largest connected component (LCC) of disease genes on a given network, as measured by the corresponding z-score compared to random gene sets (Fig. 3a, Methods). Figure 3b shows the module significance for all rare disease groups on all network layers summarized in one heatmap. We observed a high degree of differential modularity, i.e., the levels of localization vary greatly between disease groups and network layers. The largest number of significantly localized rare disease groups are found on the PPI network, the phenotypic networks (HP, MP), the core transcription network, and the network of shared biological processes. This consistent localization across a wide range of rare diseases confirms the existence of disease modules also for rare diseases. In contrast to the core transcription layer observed to be relevant across multiple disease groups, the tissue-specific co-expression networks provide a more disease-specific picture with unique signatures that reflect the molecular mechanisms that underlie a particular disease group on a given tissue. For example, the wider localization pattern of rare neurological disease genes in the

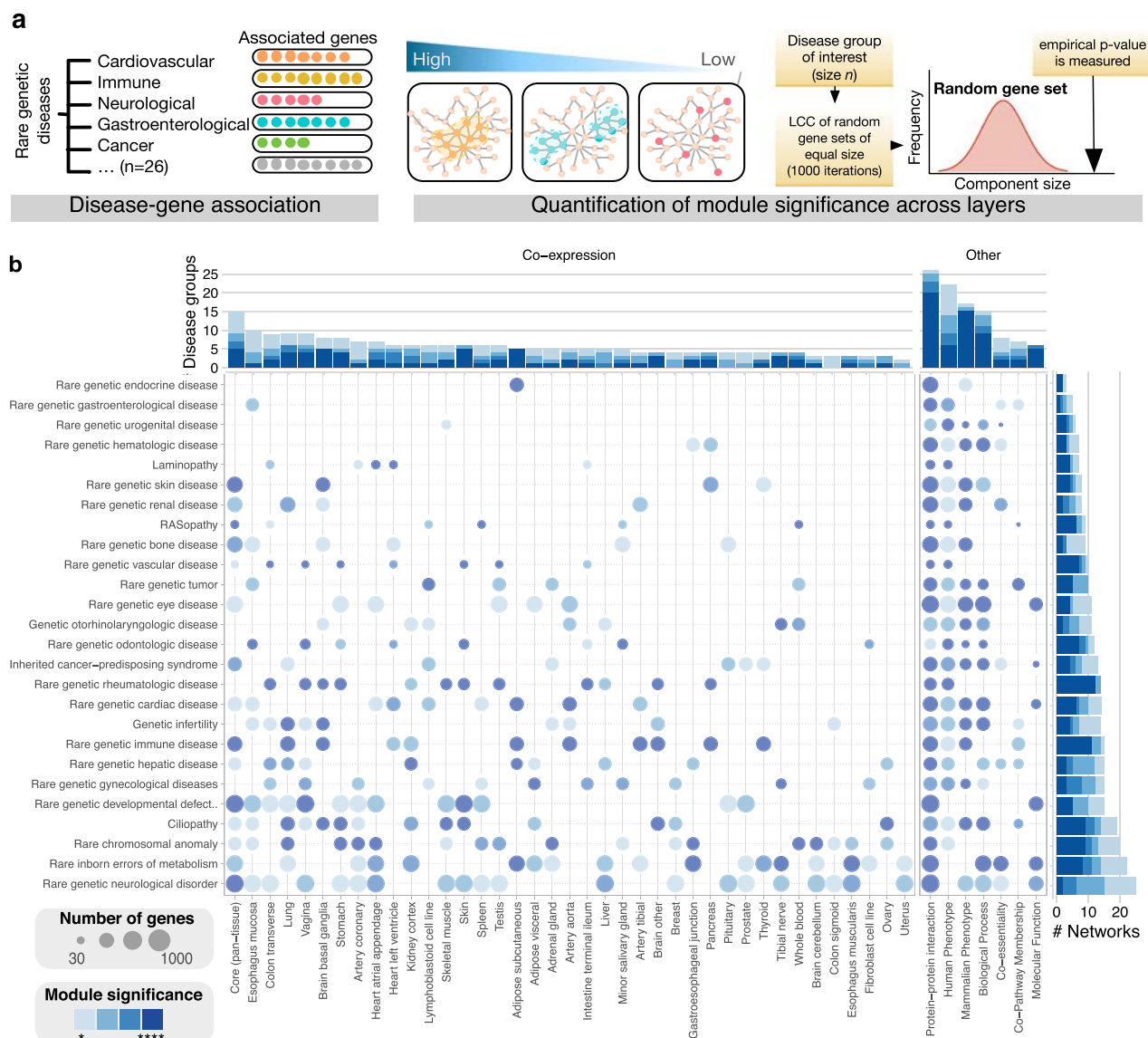


Fig. 3 Multiplex network modularity of rare disease groups. **a** Pipeline for disease module significance assessment. The size of the largest connected component (LCC) for genes associated with rare genetic disease groups collected from Fig. 2 were used to determine network relevance. **b** The heatmap shows the modularity of all rare disease groups across all network layers as measured by the respective module size significance ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***, $p < 0.0001$: ****, Benjamini-Hochberg corrected empirical p -values determined by node randomization, see Methods). In the tissue-specific network layers, only selected disease groups display pronounced modularity, often recapitulating known mechanisms and tissue specificities of particular rare diseases, but also revealing novel relationships. Network layers containing relationships that are relevant across biological levels of organization, such as protein-protein interaction, phenotypic and functional similarity networks, also display modularity across a wide range of disease groups.

phenotypic landscape observed in Fig. 2c corresponds to their significant modularity across co-expression networks in a wide range of tissues, which, in turn, reflects the often syndromic and heterogeneous phenotypes of these diseases.

Using differential modularity to contextualize rare disease gene clusters. The individual layers within the cross-scale network capture different pathobiological mechanisms. The observed differential modularities can thus offer insights into the disease etiology specific to a particular layer. For example, rare genetic gastroenterological diseases, a disease group consisting of 92 disease terms with 140 associated genes, were found to be significantly localized on five network layers (Fig. 4). Detailed inspection and enrichment of these submodules (Supplementary

Fig. 6a, b, Methods) enables us to interpret the disease characteristics within each layer: We found that genes causing the Bardet-Biedl syndrome (BBS) form pronounced clusters in the phenotypic and PPI layers. Together with the absence of modularity in other layers, this pinpoints that the emergence of this particular disease phenotype is mainly determined by interactions at the protein level, while co-essential, functional or pathway levels play less important roles. This observation is supported by our current knowledge of BBS pathological mechanisms: The proteins encoded by BBS genes form a complex crucial for transporting vesicles to cilia, a process whose defect is suspected to be a major cause of BBS³¹. At the same time, these proteins are of diverse functional character³² and involved in disparate pathways³³, explaining the lower modularity on the respective network layers.

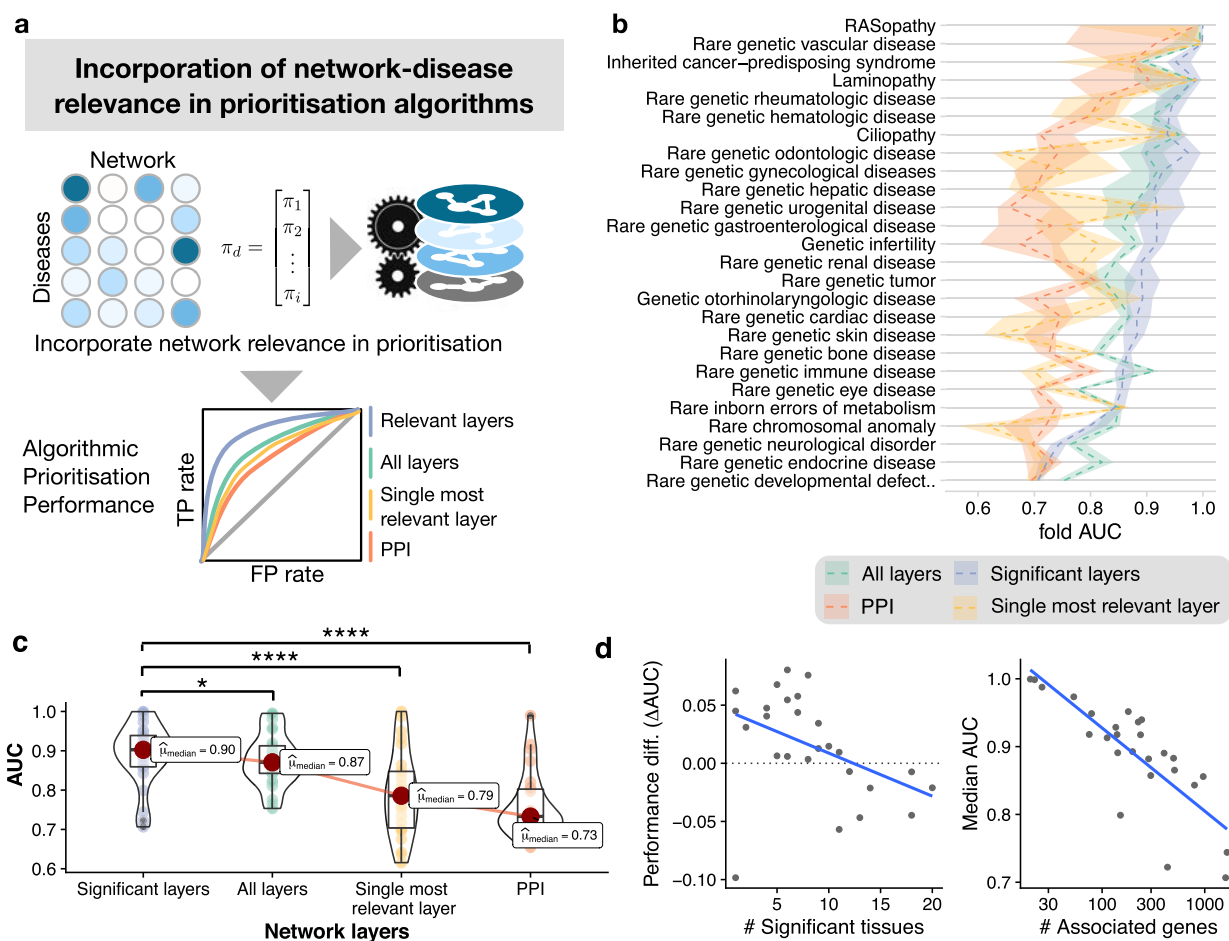


Fig. 5 Using network modularity as relevance prior in the informed propagation algorithm for gene prioritization. **a** Schematic overview of the informed multiplex network propagation algorithm that incorporates modularity as measure of relevance of a particular network level for a given disease group. **b** Comparison of 10-fold cross-validation performance in rare disease gene retrieval for different choices of included networks: Informed algorithm with most relevant network (blue), all networks (green), the PPI (red), and the single most relevant layer for each disease (yellow). Dashed lines show median value across all folds, shaded areas represent the interquartile range. The retrieval performance indicates that disease mechanisms are generally better recapitulated by incorporating relevant networks only. **c** Comparison of the AUROC from all four methods. Utilizing the significant networks lead to more accurate disease gene retrieval compared to all networks, the single most relevant layer, or the PPI. (Bonferroni-Holm corrected Durbin-Conover test p -value = 0.026, 1.22e-6, and 3e-16 respectively). Threshold for p -values: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****; $n = 26$ rare disease groups across all network sets. Bounds of box represent 25th and 75th percentiles, center the median, whiskers 10th and 90th percentiles, respectively. **d** Factors correlated with the retrieval performance. The algorithm that incorporates all networks can outperform the informed algorithm for diseases with high levels of syndromicity, i.e., disease that manifest in multiple physiological systems (left, Spearman's $\rho = -0.53$, corresponding p -value = 0.004). Decreasing functional relevance as the number of genes increases also led to lower predictive performance (right, Spearman's $\rho = -0.83$, p -value = 1.94e-6). The corresponding p -value of correlation was determined by Fisher z-transformation, two-sided.

0.95 (Fig. 5b), confirming the general applicability of multiplex network propagation to rare disease gene prediction. A comparison between the four methods revealed that incorporating only relevant network layers (median AUROC = 0.90) generally outperforms the PPI (median AUROC = 0.73), the most relevant single layer benchmark (median AUROC = 0.79), as well as the incorporation of all layers (median AUROC = 0.86), with corresponding Bonferroni-Holm corrected Durbin-Conover test p -value = 3e-16, 1.22e-6, and 0.002 respectively (Fig. 5c). We concluded that network modularity thus provides a network-based criterion to curate and integrate the most relevant data and levels of biological organization for a specific disease.

Interestingly, we also observed differences in retrieval performance related to characteristics of the diseases themselves: We found that syndromic disease groups, i.e., those with significant disease modules across multiple tissue types, tend to have lower

retrieval performance and benefit from incorporating all tissue co-expression networks (Spearman's $\rho = -0.53$, p -value = 0.004, Fig. 5d left panel, Supplementary Fig. 7b). On the one hand, this poses a challenge for disease groups that manifest various anatomical features such as rare genetic neurological disorders. On the other hand, this reflects limitations of broad ORDO disease group definitions such as rare genetic defects during embryological development. We further found that the retrieval performance correlated negatively with the number of genes associated with a particular disease group (Spearman's $\rho = -0.83$, p -value = 1.94e-6, Fig. 5d right panel). These two factors are closely related, as both the syndromicity level and overall heterogeneity tend to increase as more genes are involved in the disease group. Taken together, these findings indicate that well defined disease groups with low to moderate number of associated disease genes are more likely to capture molecular

disease characteristics at a level of specificity that results in better network-based predictions. This suggests that more fine-grained, mechanism-based disease definitions, together with high-resolution phenotyping will aid in further improving the predictive power of the introduced network methods.

To further dissect the contribution of individual networks and potential curation biases on the overall predictive power, we performed several additional benchmarks on different subsets of the multiplex network (Methods). Our comparisons between curated, unbiased and size-matched random subsets of the PPI indicate that the performance is largely driven by network size rather than potential literature biases in the interaction curation process (Supplementary Fig. 7c). We also evaluated the differences in performance upon removing individual layers, as well as groups of layers from the full multiplex network (Supplementary Fig. 7d). The results suggest that the performance is not driven by individual network layers and that the predictive power of the multiplex network can be best understood as a collective characteristic of all disease relevant layers.

Application to candidate gene prioritization in rare disease patients. Based on the performance of the informed multiplex propagation for retrieving genes across all rare disease groups we hypothesized that the method can also act as an additional evaluation metric for prioritizing genomic variants in individual rare disease patients. Starting point in a diagnostic setting is next-generation sequencing of a patient's genome, typically yielding rare genomic variants (allele frequency < 1%) in dozens to hundreds of different coding regions, and with unknown consequences. These variants may be further filtered down, for example based on frequency in the general population, deleteriousness scoring, or segregation analysis, resulting in up to a few dozen high confidence candidate genes^{38,39}. Identifying the one causal gene among them remains a critical challenge both in research and in clinical practice (Fig. 6a).

We tailored the informed propagation algorithm to individual patients by using seed genes associated with patient-specific phenotypes, combined with the network relevance scores from the corresponding Orphanet disease group (Supplementary Fig. 7e). Altogether, this enables us to perform patient-specific multiplex network propagation to prioritize candidate genes. We applied the method to filtered lists of genes with rare variants obtained from a cohort of 139 rare disease patients suffering from various neurological symptoms with intellectual disability as a predominant phenotype (Fig. 6b, Supplementary Fig. 7f, g, Supplementary Data 7, and Methods for details of the cohort). The causal variants of all patients were already confirmed and could thus be utilized for benchmarking. After standard methods of filtering for high confidence variants were exhausted, up to 997 candidate genes per patient remained (mean = 401.2). We found that our algorithm prioritizes causal genes with an overall AUROC of 0.95 (Fig. 6c). Furthermore, we benchmarked the performance of our method against predictions based on various gene-level properties, using the same data as in the network construction, specifically (1) pathway information, (2) expression level information, (3) literature counts and (4) phenotypic similarity (see Methods for details). Among these methods, phenotypic similarity was most predictive (AUC = 0.87), followed by literature counts (AUC = 0.72), expression information (AUC = 0.71) and pathway counts (AUC = 0.59). However, the informed multiplex propagation outperformed all gene-based methods (p -value = 8.39×10^{-5} , DeLong's test of ROCs between our approach and the best performing gene-level method, i.e., phenotypic similarity).

Note that for the specific use case of gene variant prioritization in rare disease patients, it is instructive to not only consider the global ROC-based performance, but also the exact ranking of the causal gene when assessing the utility of the framework in research and clinical settings. In our cohort, the multiplex propagation method placed the true causal gene among the top five ranked genes for 64 out of 131 patients (48.9%). For the purely gene-based methods, the causal gene was among the top five in only between 4 and 11 patients (3.1–8.4%, Fig. 6d).

We also performed an additional benchmark using a temporal-holdout setting to ensure that the performance of our method is not primarily driven by confirmatory biases (Supplementary Fig. 8a, Methods). To this end, we curated a set of 21 patients with causal genes that were unknown at the time of network construction, thus minimizing the likelihood that their disease association contributed to information in any of the curated databases (Supplementary Fig. 8a, Methods). We found that the overall performance as measured by the AUROC remained high for all tested prediction methods. The observed slight reductions were within the 10-fold interquartile range in most cases and may also be attributed to the smaller sample size. For example, the informed multiplex propagation AUROC was reduced from 0.90 to 0.86 (Supplementary Fig. 8b). A closer inspection of the ranking showed that our framework maintained its proportion of true causal genes being ranked in the top five gene list, whereas almost all gene-based approaches had difficulties in retrieving them at highly ranked positions (Supplementary Fig. 8c, d).

Discussion

In the context of complex diseases, numerous network-based studies have revealed an intimate relationship between genetic disease associations, their interaction patterns, and pathophysiological manifestations^{40,41}. Most importantly, it was found that disease genes are not scattered randomly in molecular networks, but instead agglomerate in disease-specific modules²⁶. Molecular networks can thus serve as maps to guide the search for new disease genes^{42–44}, suggest drug repurposing^{45–47} and combination strategies^{48,49}, or elucidate disease relationships^{26,50}, to name but a few important applications.

Our work expands this concept in several directions: First, we showed that by aggregating individual gene defects into groups of related phenotypes, we can apply tools originally developed for common, polygenic diseases also to rare, monogenic diseases. Our comprehensive analysis of over 3,584 individual gene defects revealed that as a group, they exhibit network signatures similar to those observed for complex diseases. This opens up a wide range of network medicine tools and concepts to be applied to rare diseases. Existing tools, for example for prioritizing rare variant genes, often augmented by additional clinical data^{51–55}, demonstrate the potential for network-based methods in this area.

We further showed that the central network medicine concept of disease modules can be generalized towards multiplex networks representing various levels of biological organization. Previous work relied prominently on physical protein–protein interactions, which have been mapped out systematically for nearly two decades^{13,14,56,57}. Our analysis of 46 network layers containing over 20 million interactions showed that disease modules can be identified across a wide range of relevant gene relationships. We further found that the degree of modularity is indicative of the impact of disease-associated perturbations on a particular level of biological organization, and thereby determines the disease relevance of datasets from the respective level. The performance of the informed propagation algorithm for rare

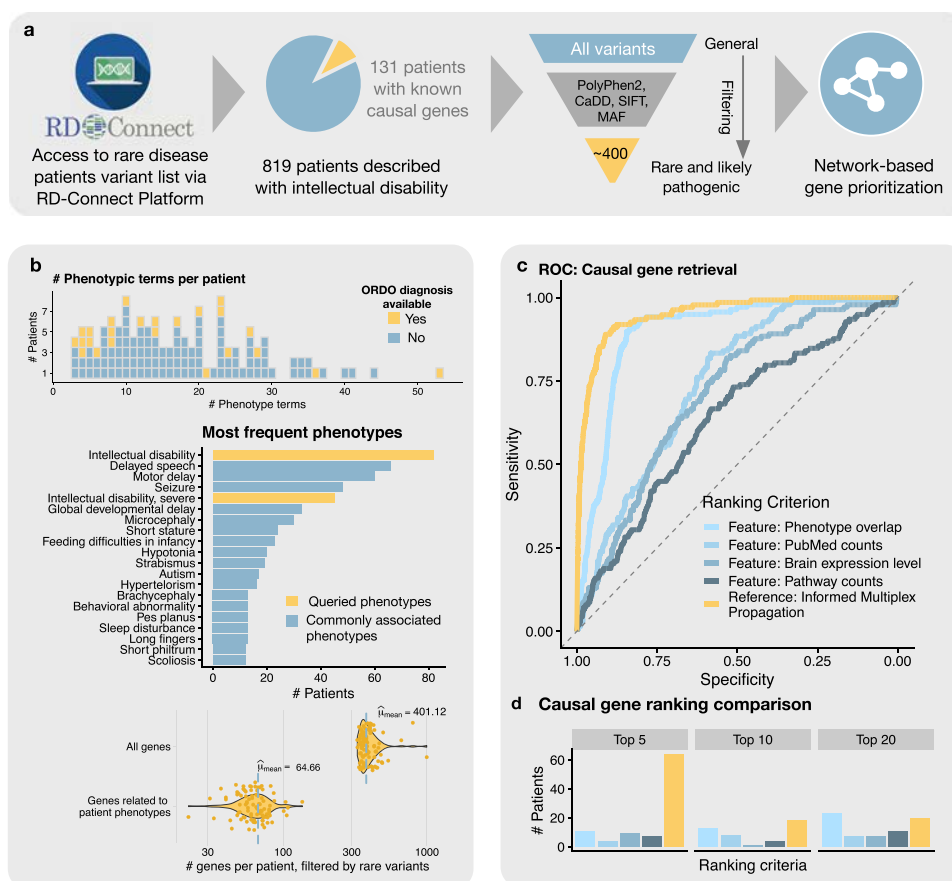


Fig. 6 Patient cohort and gene prioritization performance. **a** Data access and filtering: Querying for intellectual disability phenotypes resulted in 819 patients, 131 of which were solved cases with rare and pathogenic variants in an average of over 400 genes. **b** Basic characteristics of patient variants, associated phenotypes and diagnoses. **c** ROC curves for the performance of causal gene prioritization of our approach (yellow, AUROC = 0.95) and various gene level based benchmarks (AUROC between 0.59 and 0.87). **d** Number of patients for which the true causal gene was prioritized among the top five, 10, and 20 for all considered methods. The informed multiplex propagation placed the true causal gene among the top five ranked genes for 64 out of 131 patients (48.9%). For the purely gene-based methods, the causal gene was among the top five in only between 4 and 11 patients (3.1–8.4%).

disease gene prediction demonstrates the practical utility of this finding. We expect that the general principle for identifying the most relevant datasets will be applicable in other contexts as well, including studies on cancer and other complex diseases. Indeed, as biomedical research is becoming more data intensive in general, and more biological network maps become available in particular^{58,59}, new strategies for integrating diverse data are required. A range of methodologies have been developed for this purpose, including network-based strategies^{60–63} and advanced machine-learning approaches^{64–66}. Our results could potentially enhance these strategies by using disease modularity as a criterion for curating and excluding potentially uninformative datasets.

Finally, to enable a broad community of researchers in the areas of rare disease, network medicine or biomedical data integration to build on our work, all datasets and algorithms presented in this work are publicly available.

Methods

Resources and network construction. Resources used in the multiplex network construction are listed in Table 1. We incorporated seven major databases, each representing distinct biological layers.

Protein–protein Interactions. Protein–protein interaction data was taken from the HIPPIE database²¹ and filtered for interactions with supporting PubMed articles. To assess the impact of interactions collected from small-scale, hypothesis-driven experiments compared to those stemming from large-scale, unbiased screens, we further collected the most recent versions of the two largest systematic high-

throughput PPI studies: the Human Reference Interactome (HuRI) based on yeast two-hybrid (Y2H) screening¹³ (retrieved from <http://www.interactome-atlas.org/data/HuRI.tsv> on 31 May 2021), and the BioPlex interactome constructed from affinity-purification mass spectrometry profiling⁶⁷ (retrieved from https://bioplex.hms.harvard.edu/data/BioPlex_293T_Network_10K_Dec_2019.tsv on 31 May 2021). The PPI network layer can therefore be split into two categories: the unbiased PPI for interactions that are contained in any of these two resources, and the curated PPI for the remaining edges (Supplementary Fig. 4a).

Tissue Co-expression networks. Transcriptomic data is one of the most abundant publicly available high-throughput data. Differential expression profiles across tissues and cell types have been widely analyzed as a probe for disease specificity. In the context of network analyses, expression data has been used in two major ways: as a means to filter out genes from generic interactomes based on their expression level in a particular context of interest^{13,60}, and for constructing co-expression networks. Here, we follow the latter approach, and use co-expression as a proxy for tissue- or cell type-specific functional and regulatory relationships. As primary resource we used the Genotype-Tissue Expression (GTEx) data^{68,69}, which provides genome-scale expression profiles across 53 human tissues that have been used previously to construct co-expression networks^{15,16}. We used the following pipeline:

1. We downloaded the GTEx expression profiles in the format of transcripts per million (TPMs) from the Expression Atlas (<https://www.ebi.ac.uk/gxa/experiments/E-MTAB-5214/>). The data was subsequently processed using the bioconductor package SummarizedExperiment (<https://bioconductor.org/packages/release/bioc/html/SummarizedExperiment.html>).
2. Tissues with a number of samples less than a minimum quality threshold similar to the GTEx Portal ($n \geq 70$) were removed. These included fallopian tube ($n = 14$), ectocervix ($n = 12$), endocervix ($n = 10$) and urinary bladder ($n = 24$).

Table 1 Resources used for the multiplex network construction.

Layer representation	source
Biological process	GO (BP) release 2018-11-24, retrieved from http://purl.obolibrary.org/obo/go/go-basic.obo
Molecular function	GO (MF) release 2018-11-24, retrieved from http://purl.obolibrary.org/obo/go/go-basic.obo
Human phenotype	HPO release 2018-10-09, retrieved from http://purl.obolibrary.org/obo/hp.obo
Mammalian phenotype	MPO release 2018-11-23, retrieved from http://purl.obolibrary.org/obo/mp.obo
Co-essentiality	co-essential networks inferred from correlated fitness profile across diverse cancer cell lines from Kim et al. ²⁰ .
Protein-protein interaction	HIPPIE v2.2 (release 2019-02-14), retrieved from http://cbdm-01.zdv.uni-mainz.de/mschaefer/hippie/download.php
Co-pathway membership	Reactome Pathways Gene Set, retrieved on 30 January 2019 from https://reactome.org/download/current/ReactomePathways.gmt.zip
Tissue co-expression	RNA-seq from GTEx v7, retrieved from the Expression Atlas https://www.ebi.ac.uk/gxa/experiments/E-MTAB-5214/

- For the remaining 49 tissues, we further merged tissues with similar expression profiles to reduce redundancy and increase signals⁷⁰. Most notably, brain regions (13 tissues) were merged into three major groups and relabeled by their anatomical entities (Supplementary Fig. 1a). This process not only merged potentially redundant tissues, but also increased sample sizes for some tissue groups that would otherwise have been undersampled. The resulting 38 tissue groups along with the sample sizes are shown in Supplementary Data 3.
- The GTEx database contains an average of 29,779 ± 1972 transcripts per tissue. We next filtered for protein coding transcripts (e.g., exclusion of pseudogenes, long non-coding (lnc) RNAs, miRNAs, and other non-coding biotypes) by discarding transcripts without corresponding accession numbers in the UniProt Knowledgebase (www.uniprot.org) according to a query of MyGene (<https://mygene.info>, retrieved on 21 August 2019). Supplementary Data 10 lists the 21,310 transcripts in consideration, resulting in 17,716 ± 369 protein-coding genes per tissue (Supplementary Fig. 1b).
- For each tissue, Spearman's correlation coefficient (ρ) of all protein-coding gene pairs was used to determine the strength of their respective co-expression levels. Gene pairs with $|\rho| \leq 0.75$ were discarded, resulting in 11,161 ± 1082 genes per tissue.
- We applied a disparity filter¹⁹ to remove weak, structurally redundant edges and to extract the backbone of each network. Edges with a corresponding disparity filter p -value < 0.05 were selected. This process yielded 10,526 ± 1825 genes per tissue. Note that even though the number of nodes decreased only slightly, the disparity filter excluded a large amount of spurious correlations (median number of interactions before and after = 1.83e6 and 4.78e5, respectively, Supplementary Fig. 1b (right)). The disparity filter represents a dynamic cutoff, where lowly expressed genes tend to be removed and highly expressed genes tend to remain, while also allowing for the detection of lowly expressed genes that are strongly correlated with other genes (Supplementary Fig. 1e). As a reference, we also show the comparable reduction of remaining genes if they were filtered using a standard expression threshold of TPM > 1 (13,567 ± 874 genes, Supplementary Fig. 1b).

The resulting networks consist of edges that are shared across multiple tissues (core transcriptional modules), as well as edges that are only present in a small number of tissues (tissue-specific modules). We considered edges present in less than five tissues as tissue-specific, and edges present in at least five tissues as core transcriptional modules (Supplementary Fig. 1c, d).

Ontology-derived functional and phenotypic similarity network. To capture gene relationships on functional and phenotypic levels, we incorporated expert curated data and systematic ontologies. To transform ontological annotations into gene-centric networks, we defined that two genes are functionally or phenotypically connected if they are semantically similar based on the corresponding ontology^{71,72} as follows:

We first compared several widely used measures of semantic similarity to ensure that the scores are robust for our purposes:

- Information content (IC)-based similarity based on Resnik's method⁷³. The similarity of two terms is derived from their most informative common ancestor (MICA) in the ontology. Given ontology terms t_1 and t_2 , their pairwise similarity is given by $\text{sim}_{\text{Resnik}}(t_1, t_2) = \text{IC}(\text{MICA})$, where $\text{IC}(t) = -\log(p(t))$, $p(t)$ represents the frequency of term t defined by $p(t) = \frac{n_t}{N}$, n_t denotes the number of descendants of term t , and N the number of descendants of the root term of interest in the ontology tree.
- Information content (IC)-based similarity based on Lin's method⁷⁴. Unlike Resnik's method, Lin's similarity measure restricts the value to be in the range between zero and one, and is given by $\text{sim}_{\text{Lin}}(t_1, t_2) = \frac{2\text{IC}(\text{MICA})}{\text{IC}(t_1) + \text{IC}(t_2)} \in [0, 1]$.
- After collecting all pairwise term similarities for annotations of two genes, we next employed the Best-Match Average (BMA) strategy to combine them into a single gene similarity score. Their pairwise similarity of genes g_1 and g_2 with m and n annotated terms, respectively, is given by

$\text{sim}_{\text{BMA}}(g_1, g_2) = \frac{\sum_{i=1}^m \text{colmax}(S) + \text{rowmax}(S)}{m+n}$, where $S \in \mathbb{R}^{m \times n}$ is the matrix containing the pairwise similarity values of the ontology terms associated with the two genes, $\text{rowmax}(S)$ and $\text{colmax}(S)$ are vectors of length m and n , containing the maximum similarity values across all rows and columns of matrix S .

- Frequency-based similarity, where the similarity between two genes is given by the number of shared annotations, i.e., $\text{sim}_{\text{freq}}(g_1, g_2) = |T_{g_1} \cap T_{g_2}|$, where T_{g_k} is the set of ontology terms (including ancestor terms) associated with gene g_k .

We found that the respective similarity values are strongly correlated, indicating that the resulting networks are robust against details of the used methods (Supplementary Fig. 2a). We chose to proceed with the IC-based Resnik's method with the Best-Match Average (BMA) combination strategy, as it has been demonstrated to both be among the simplest methods, while also providing the most reliable performances across different tasks^{71,75}. We used the R packages GoSemSim⁷⁶ and OntologyX⁷⁷ to navigate and compute the similarity measurements.

Gene pairs with minimal similarity value, i.e., pairs whose only common annotation is the root term of the considered ontology branch (i.e., "Molecular Function" or "Biological Process") were considered as unrelated and therefore removed from further consideration. For example, there are over 21M gene pairs connected at this level in the GO (BP) branch (similar score = 0.447, Supplementary Fig. 2b). This led to the removal of 230 genes with no commonly associated MICA with other genes beyond the root term.

All ontology-based networks (GO:BP, GO:MF, MPO and HPO) were constructed according to the following procedure summarized in Supplementary Fig. 2c: Pairwise similarity scores given by the procedures above resulted in dense weighted networks. We further applied the disparity filter¹⁹ to extract the backbone of the network and discard structural redundant edges (gene pairs with corresponding disparity p -value > 0.05). The disparity filter provides a dynamic cutoff that considers the strength of the similarity scores of a gene in reference with all similarity values of its neighbors. Similar to using a hard cutoff, edges between gene pairs with low similarity scores (e.g., $0 < \text{sim}(g_1, g_2) < 3$ in GO:BP) are removed while those with high similarity scores ($\text{sim}(g_1, g_2) > 6$) are virtually unaffected. Edges with medium similarity scores ($3 < \text{sim}(g_1, g_2) < 6$) may either remain or be discarded based on their similarity score with respect to all other connected genes (Supplementary Fig. 2d).

Overall, networks derived from semantic similarity measures favor gene pairs that are similarly annotated over highly, but diversely annotated gene pairs (Supplementary Fig. 2e). Gene pairs with high similarity scores often belong to the same protein families such as the ER membrane protein complexes (EMC), olfactory receptors (OR), and membrane transporters, and tend to share a large fraction of annotated GO terms (Supplementary Fig. 2e, right). We further demonstrated this for the example of GO terms associated with TP53 (gene with highest number of publications) and TGFBI (gene with highest number of associated GO terms). While both genes are well characterized, with 87 and 176 annotated GO terms, respectively, only ten annotations are shared, indicating that they are involved in distinct biological processes (Supplementary Fig. 2f). As a result, the computed similarity score and subsequent disparity p -value failed to reach the significance threshold, meaning that the two genes are not connected (Supplementary Fig. 2e). This effect is observed across most well characterized genes, leading to the slightly negative literature bias of ontology-derived networks (Fig. 1h and Supplementary Fig. 4c–f). We found that densely connected clusters within the constructed networks recapitulate biological processes corresponding to shared terms on their respective ontologies (Supplementary Fig. 3a, clusters with Bonferroni-Holm corrected enrichment hypergeometric p -value < 1e-20 were labeled).

Pathway co-membership networks. Gene-pathway associations were downloaded from the Reactome website <https://reactome.org/download-data/> (accessed 25 January 2019) under the Reactome Pathways Gene Set section. For every gene pair, we collected the number of shared pathway annotations. In the pathway co-membership network construction, two genes were connected if they share at least five Reactome pathway annotations (to prevent associations due to common pathways).

Disparity filter. To extract the backbone of dense, weighted networks resulting from semantic and correlation-based construction, we applied a disparity filter¹⁹. For a given network, we computed a p -value for all edges between nodes i and j as $p_{ij} = (1 - w_{ij})^{k-1}$, where w_{ij} is the edge weight for node i normalized over all its edges, and k denotes its degree. We only kept edges for which both p_{ij} and p_{ji} reached a threshold significance level.

All network data and corresponding details are available in Supplementary Data 1, 2.

Measurements of network characteristics. The network characteristics shown in Fig. 1 h (number of nodes and edges, clustering and assortativity) were computed using the R package igraph⁷⁸ (<https://igraph.org>).

For a global assessment of the literature bias present in a particular network we used the Spearman's correlation coefficient between the network degree of a gene and the number of publications mentioning the gene. The number of publications was queried using the INDRA python module 78 (<http://www.indra.bio>, accessed on 12 April 2019), the resulting data is provided in Supplementary Data 8.

For a more local assessment of correlation structures within the connection patterns of a network, we used the local assortativity (ρ), a node-level property whose sum over all nodes is equal to the assortativity of the network⁷⁹. It is defined as $\rho = \frac{j(j+1)(\bar{k}-\mu_k)}{2M\sigma_k^2}$, where j is the excess degree, \bar{k} the average excess degree, and M the number of edges in the network. The excess degree follows the distribution $q(k) = \frac{(k+1)p(k+1)}{k}$. We employed the concept to demonstrate that the overall assortativity can also be present among interactions derived from high-throughput studies such as the BioPlex network (Supplementary Fig. 3f).

Network similarity computation and randomization. Given a pair of networks A and B with the set of edges E_A and E_B respectively, we quantified the network similarity using the edge overlap index (S_{AB}):

$$S_{AB} = \frac{|E_A \cap E_B|}{\min(|E_A|, |E_B|)}$$

We used a dissimilarity measure defined as $d_{AB} = 1 - S_{AB}$ to construct a 2D map that preserves network dissimilarities by employing Kruskal's non-metric multi-dimensional scaling (R package MASS). Finally, we compared the measured similarity of each network pair to random expectation: For each network, we performed 10 permutations of node indices, resulting in 100 permutations for a network pair, which we used as random reference distribution to assess the measured overlap similarity. We then computed the z -score and corresponding empirical p -value. A network pair with p -value < 0.05 is considered significantly similar (Supplementary Fig. 3c, d).

Characterization of co-expression network with essentiality data. We characterized our tissue-specific co-expression networks constructed based on GTEx expression data as follows: We hypothesized that genes that are highly co-expressed across several tissues are likely required for cellular development and survival, and should show a strong tendency of being essential genes. To test this hypothesis, we used the list of human essential genes from the OGEE database (v2, retrieved on 16 April 2019, Supplementary Data 9).

Rare genetic disease gene association data. The structure of the Orphanet Rare Disease Ontology was queried and processed using the R interface of the Ontology Lookup Service (<https://lgatto.github.io/rols/index.html>). We considered all descendant terms of "Rare genetic disease" (Orphanet:98053) that were associated with at least 20 genes, resulting in 26 rare genetic disease groups. The disease groups and all disease-gene associations can be found in Supplementary Data 5).

Disease-network landscapes via node2vec embedding algorithm. To visualize large (genome-scale) networks where the modularity can be difficult to observe, we employed the python3 implementation of the node2vec graph embedding algorithm²⁹ (<https://github.com/eliocr/node2vec>). Nodes were embedded into 64-dimensional Euclidean space and subsequently projected on a 2D plane using t-SNE⁸⁰ (Supplementary Fig. 6c). Note that the predictions in this work were performed on the original network space as the resulting coordinates in the embedded Euclidean space are subject to the parameterization in both the node embedding and the dimensionality reduction. Since different node embedding techniques and parameter sets may preserve different topological structures^{81–83}, their reliability may vary depending on the particular machine learning task⁸⁴.

Identification of the significance of a disease module. The size of the largest connected component of random subsets of m nodes in a network is expected to follow a normal distribution, provided that m is larger than the percolation threshold. We can therefore empirically estimate the significance of a given module size by the z -score and corresponding p -value compared to randomly selected nodes. Networks in which the size of the largest connected component of the genes

associated with a particular disease exceeded a threshold of p -value < 0.05 (after Benjamini–Hochberg correction) were considered significant.

Gene ID mapping, homolog conversion, and enrichment analysis. All human gene identifiers from different resources were mapped to NCBI standard symbols. For mouse to human gene mapping, we used the Mouse Genome Informatics homologs mapping <http://www.informatics.jax.org/downloads/reports/index.html>. Gene enrichment results were queried using EnrichR⁸⁵.

Informed multiplex network propagation algorithm. The standard multiplex network propagation is defined by an equal probability for the random walker to visit any neighbor from the current layer m or any other layer n ⁸⁶. For L network layers with N nodes each, this can be represented through the supra-adjacency matrix $S \in \mathbb{R}^{NL \times NL}$:

$$S = \begin{bmatrix} A_1 & I & \dots & I \\ I & A_2 & \dots & I \\ \vdots & \vdots & \ddots & \vdots \\ I & I & \dots & A_L \end{bmatrix}$$

where A_m is the adjacency matrix for network layer m ($m \in \{1 \dots L\}$) and I denotes the identity matrix.

We extended this standard algorithm towards an informed propagation method where the walker visits more relevant layers with higher probability. We quantify the relevance of a network m for a disease group d by the corresponding z -score z_{dm} of the largest component of associated genes. We considered all network layers with $z_{dm} \geq 1.645$ (corresponding to the 95% confidence level under normal distribution) as informative and defined the relevance score (π_{dm}) as the normalized z -score across all informative layers:

$$\pi_{dm} = z_{dm} / \sum_m z_{dm} \text{ and } \sum_m \pi_{dm} = 1$$

The relevance score π_{dm} was then used to determine the transition probability $p(m|n)$ between layers n and m , so that the walker visits more informative layers with a higher probability corresponding to their respective π_{dm} values. This is achieved by employing the concept of reversible Markov chain Monte Carlo that requires the following detailed balance condition:

$$\pi_n p(m|n) = \pi_m p(n|m)$$

To satisfy this condition, we define $p(m|n) = \frac{1}{L} \min(1, \frac{\pi_m}{\pi_n})$ and $p(m|m) = 1 - \sum_{n \neq m} p(m|n)$. The informed supra-adjacency matrix \tilde{S} can thus be written as

$$\tilde{S} = P \circ S = \begin{bmatrix} p_{11}A_1 & p_{12}I & \dots & p_{1L}I \\ p_{21}I & p_{22}A_2 & \dots & p_{2L}I \\ \vdots & \vdots & \ddots & \vdots \\ p_{L1}I & p_{L2}I & \dots & p_{LL}A_L \end{bmatrix}$$

Finally, we incorporate the informed supra-adjacency matrix into the random walk with restart algorithm:

$$p_{t+1} = (1 - r)\tilde{S}p_t + rp_0$$

where p_0 is the initial visiting probability vector with $p_0(i) = 1/k$ if node i is one of k seed nodes, and $p_0(i) = 0$ otherwise. p_t is the visiting probability at iteration step t , and $r \in [0, 1]$ is the restart probability. In this analysis, we chose $r = 0.7$.

The final visiting probability (p) can be obtained numerically when the convergence criteria are met ($|p_{t+1} - p_t| = 0$). The visiting probability of a node is the arithmetic mean of the visiting probability across all layers. In retrieval tasks, nodes are ranked based on this final visiting probability. Seed nodes are omitted from the ranking.

Cross-validation performance assessment. The prediction performance was assessed using 10-fold cross-validation for retrieving genes associated with individual rare disease groups. Area under the receiver operating characteristic curve (AUROC) computations and plots were performed using the cvAUC and pROC packages in R. Differences between ROCs were evaluated using the two-sided DeLong's test⁸⁷.

We first considered four different settings: (1) baseline single layer (the PPI), (2) the most relevant single layer for each disease according to the lowest LCC z -score, (3) all network layers, and (4) all relevant network layers, i.e., those with a significant LCC z -score for the disease (p -value < 0.05 , Benjamini–Hochberg correction for multiple hypotheses).

To further investigate the contribution of individual layers, as well as potential curation biases on the overall predictive power, we performed several additional benchmarks on different subsets of the multiplex network:

We first analyzed the impact of interactions curated from small-scale experiments on the prediction performance of the PPI network layer (Supplementary Fig. 4a). To this end, we considered two subsets of the full PPI, an unbiased subset consisting of interactions from systematic high-throughput

studies, and a curated subset consisting of all other interactions (see above). The unbiased PPI contributes to 13% of all interactions in the full PPI, and, as expected, shows a less pronounced literature bias (Supplementary Fig. 6b). While the curated PPI performs equally well as the full PPI in the disease gene prediction task, the performance of the unbiased PPI drops significantly (median AUROC = 0.62, p -value = 1.76×10^{-9} , FDR-corrected Durbin-Conover non-parametric test, Supplementary Fig. 7c). To assess the extent to which the reduced size of the unbiased PPI contributes to this drop, we repeated the analysis on ten random subsets of the curated PPI that are of the same size as the unbiased PPI subset. We found that these random subnetworks have a performance comparable to the one of the unbiased PPI (with a median AUROC of 0.58 even slightly reduced, Supplementary Fig. 7c). This indicates that the performance of the PPI network is mainly driven by its size, rather than details of the interaction curation. This, in turn, suggests that confirmatory biases that may result from including curated interaction data are likely to play only a minor role for the overall performance, at least for PPI data.

We next assessed the prediction performance of the multiplex network upon removing other network layers derived from curated databases, specifically the layers based on shared pathway membership, phenotypic similarity (HPO and MPO), and GO (BP and MF) similarity. We first computed the 10-fold cross-validation AUROC after removing each of these layers individually. For most layers, we observed only a slight drop in the performance (median AUROC between 0.87 and 0.88; Supplementary Fig. 7d), indicating that the core connectivity of disease genes across different layers is robust against the removal of individual layers. The only layer with a stronger impact is the HPO phenotype layers, whose removal resulted in a reduction of AUROC to 0.80 (p -value = 0.0003, FDR-corrected Durbin-Conover non-parametric test). This is not unexpected given the strong predictive power of phenotypes as close proxies to diseases which forms the basis for their usage in clinical settings and is documented in the literature^{1,2}, as well as in the gene-level benchmarks discussed in the patient candidate gene prioritization below.

Finally, we determined the predictive performance of the multiplex network after removing all layers that involve curated data (Reactome, GO, HP, MP, and PPI), leaving only relevant co-expression and co-essentiality networks for the propagation. While these high-throughput data alone do carry predictive power, their performance was significantly lower compared to using all available data sources (AUROC = 0.71, p -value = 1.17×10^{-11}). Interestingly, we also observed an occasional increase in performance, such as for rare genetic endocrine diseases, one of the worst performing disease groups in the reference setting (AUROC increased from 0.64 to 0.71). The propagation only took place on the adipose tissue co-expression network (ADS), which, in addition to its traditional role for excess lipid storage, has recently been recognized as an endocrine organ^{16,17}.

Taken together, these results suggest that the predictive power of the multiplex network can be best understood as a collective characteristic of all disease relevant layers, rather than being primarily driven by specific individual layers.

Cohort of patients with intellectual disability. We first developed and tested our method on a locally available, well-controlled cohort of patients with intellectual disability (ID), before applying it to a much larger cohort obtained from the RD-Connect Genome-Phenome Analysis Platform (GPAP)⁸⁸. To conduct a temporal-holdout benchmarking, we also curated a subset of the RD-Connect cohort containing patients with causal genes discovered after all data used in the network construction was retrieved. The details of the three cohorts are as follows:

Local cohort. We gained access to variant data from eight patients with confirmed causal gene (two females and six males aged between three to twenty years old; see Supplementary Data 11 for details). The recruitment was based on the referral by clinicians, with the purpose of genetic testing and there was no compensation involved. Informed consents were signed by the patients or their legal guardians and the processes were reviewed by Ethics Committee of the Medical University of Vienna; and/or Haunerschen Kinderspital, Munich, Germany; Servizio di Consulenza Genetica, Bolzano, Italy; University Hospital Zagreb, Zagreb, Croatia; General Hospital Varazdin, Varazdin, Croatia; and Tehran University of Medical Sciences, Tehran, Iran in accordance with the Declaration of Helsinki. All patient variant data were obtained from exome-sequencing performed at the Biomedical Sequencing Facility (BSF) at the CeMM Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM). Genomic DNA was extracted (QIAamp DNA Mini Kit, Qiagen) from whole blood from patients, parents and participating siblings. Quantity and quality of patient DNA were validated by Qubit 2.0 Fluorometric Quantitation system (Life Technologies). Exome libraries were prepared using the Nextera DNA Flex Exome Library Prep Kit (Illumina). Genomic DNA was tagged, size-selected and amplified followed by two rounds of hybridization with biotinylated baits and capture with streptavidin-conjugated magnetic beads. After enrichment, library fragments representing in total 45 Mb coding region were amplified and size-selected. Final library pools were quality controlled and sequenced on a HiSeq 3000 instrument (Illumina) using 75 bp paired-end chemistry. DNA sequences were mapped to GRCh37 (hg19) version of human reference genome using Burrows-Wheeler Aligner with default parameters.

Single nucleotide variants (SNVs) and indels were annotated with gnomAD⁸⁹, CADD-Phred⁹⁰, dbSNP⁹¹ and ClinVar⁹² data. Subsequent filtering of remaining variants of interest was based on the inheritance pattern, variant type (high or moderate impact as classified by Ensembl database), allele frequency (<1%) in gnomAD database, and gene lists of interest in relation to the patient's symptoms annotated by Human Phenotype Ontology (HPO).

Genes associated with HPO terms describing a patient's major symptoms were used as patient-specific seed genes (Supplementary Fig. 7f), weighted by the frequency of association, i.e., a gene will be given a higher weight if it is associated with more than one phenotype found in the patient. After standard methods of filtering for high confidence variants were exhausted, up to 46 candidate genes remained, with an average number of 16 candidate genes per patient. Our patient-specific multiplex network propagation ranked the validated causal gene first in four cases, in all cases it was among the top five predictions (Supplementary Fig. 7g). Strikingly, the algorithm correctly pinpointed the causal gene in the two most complex cases, where patients presented with high confidence variants from 46 and 33 genes, respectively.

RD-Connect cohort. To overcome the small number of patients available in our local cohort, we have gained access to RD-Connect Genome-Phenome Analysis Platform (GPAP), one of the largest global infrastructures for storing and sharing genotype and phenotype data of rare disease patients (<https://platform.rd-connect.eu/>)⁸⁸. To match our local cohort, we queried patients whose phenotypes are characterized by intellectual disability (HPO term HP:0001249). Of the resulting 819 patients, 131 were solved cases, i.e., patients with a confirmed causal variant that could thus be utilized for benchmarking (Fig. 6a). The inclusion of these patients expanded the original sample size by a factor of over 16. The variants were filtered for highly stringent pathogenicity include these following tools and criteria: (1) Variant type: SNV, (2) SNV effect prediction: Mutation Taster—A (Annotated and disease causing) and D (Disease causing); PolyPhen2—D (Possibly damaging) and P (Possibly damaging); SIFT—D (Damaging), CADD score ≥ 20 , (3) Minor Allele Frequency: gnomAD allele frequency < 0.01; 1000Genome Protect AF < 0.01.

Temporal-holdout benchmarking cohort. All curated databases (GO, MPO, HPO, and the PPI) were retrieved before March 2019, we thus sought to filter for patients with causal genes that were discovered only after that point in time (Supplementary Fig. 8a). To this end, we collected the list of confirmed intellectual disability (ID) causal genes from Genomics England PanelApp⁹³, a large expert reviewed platform for disease gene causality evaluation (<https://panelapp.genomicsengland.co.uk/panels/285>). We downloaded the ID panel v3.0, which has the signed off date of 10/12/2019 and consists of 1,085 confirmed ID genes. Within our cohort of 131 RDconnect patients, 21 had causal genes not included in this panel gene list. These genes can thus be considered to have been unknown to the expert community at the time of network curation. By restricting our validation analysis to these 21 causal genes, we can assume that their disease association is not implicitly contained in the data that we use in the prediction.

Gene-level ranking benchmark. As a benchmark for the network-based informed multiplex propagation for patient candidate gene prioritization, we also implemented several ranking methods relying solely on gene-based features. Specifically, we employed the same gene features that were used to construct the multiplex networks: (1) pathway information—ranking genes involved in more pathways higher; (2) expression level information—ranking genes with higher expression levels in brain tissues higher; (3) general literature counts—ranking genes linked to a higher number of publications higher; (4) phenotypic similarity—ranking genes higher that are associated with Human Phenotype Ontology (HPO) terms described in a patient.

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

Data availability

Data generated in this study are provided in the Supplementary Information/Source Data file. The RDconnect Genome-Phenome Analysis Platform (GPAP) data are available under restricted access, which can be obtained by validated users via the platform at <https://platform.rd-connect.eu/>.

Code availability

Source code and cache data is available at the <https://github.com/menchelab/MultiOme>⁹⁴. The supplementary Explorer app for detailed inspection of disease-network specificity is available at www.menchelab.com/MultiOmeExplorer.

Received: 30 April 2021; Accepted: 19 October 2021;

Published online: 09 November 2021

References

- Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
- Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14**, 681–691 (2013).
- Fernandez-Marmiesse, A., Gouveia, S. & Couce, M. L. NGS technologies as a turning point in rare disease research, diagnosis and treatment. *Curr. Med. Chem.* **25**, 404–432 (2018).
- Ozen, A. et al. CD55 deficiency, early-onset protein-losing enteropathy, and thrombosis. *N. Engl. J. Med.* **377**, 52–61 (2017).
- Dobbs, K. et al. Inherited DOCK2 deficiency in patients with early-onset invasive infections. *N. Engl. J. Med.* **372**, 2409–2422 (2015).
- Salzer, E. et al. RASGRP1 deficiency causes immunodeficiency with impaired cytoskeletal dynamics. *Nat. Immunol.* **17**, 1352–1360 (2016).
- Nagy, V. et al. HACE1 deficiency leads to structural and functional neurodevelopmental defects. *Neurol. Genet.* **5**, e330 (2019).
- Kochinke, K. et al. Systematic phenomics analysis deconvolutes genes mutated in intellectual disability into biologically coherent modules. *Am. J. Hum. Genet.* **98**, 149–164 (2016).
- Anikster, Y. et al. Biallelic mutations in DNAJC12 cause hyperphenylalaninemia, dystonia, and intellectual disability. *Am. J. Hum. Genet.* **100**, 257–266 (2017).
- Tarailo-Graovac, M. et al. Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **374**, 2246–2255 (2016).
- Costanzo, M. et al. Global genetic networks and the genotype-to-phenotype relationship. *Cell* **177**, 85–100 (2019).
- Velimezi, G. et al. Map of synthetic rescue interactions for the Fanconi anemia DNA repair pathway identifies USP48. *Nat. Commun.* **9**, 2280 (2018).
- Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 1–7 (2020).
- Huttlin, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* <https://doi.org/10.1038/nature22366> (2017).
- Pierson, E. et al. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.* **11**, e1004220 (2015).
- Saha, A. et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).
- GTEX Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Shefchek, K. A. et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **48**, D704–D715 (2020).
- Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl Acad. Sci. U. S. A.* **106**, 6483–6488 (2009).
- Kim, E. et al. A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Sci Alliance* **2**, 1–15 (2019).
- Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* **45**, D408–D414 (2017).
- Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
- Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).
- Köhler, S. et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
- Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
- Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 026126 (2003).
- Sharma, A. et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* **24**, 3005–3020 (2015).
- Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855–864 (2016).
- Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms-disease network. *Nat. Commun.* **5**, 4212 (2014).
- Nachury, M. V. et al. A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. *Cell* **129**, 1201–1213 (2007).
- Tayeh, M. K. et al. Genetic interaction between Bardet-Biedl syndrome genes and implications for limb patterning. *Hum. Mol. Genet.* **17**, 1956–1967 (2008).
- Tobin, J. L. & Beales, P. L. Bardet-Biedl syndrome: beyond the cilium. *Pediatr. Nephrol.* **22**, 926–936 (2007).
- Qiao, J.-G., Zhang, Y.-Q., Yin, Y.-C. & Tan, Z. Expression of Survivin in pancreatic cancer and its correlation to expression of Bcl-2. *World J. Gastroenterol.* **10**, 2759–2761 (2004).
- Yasuda, A. et al. The stem cell factor/c-kit receptor pathway enhances proliferation and invasion of pancreatic cancer cells. *Mol. Cancer* **5**, 46 (2006).
- Stepensky, P. et al. Mutations in EFL1, an SBDS partner, are associated with infantile pancytopenia, exocrine pancreatic insufficiency and skeletal anomalies in a Shwachman-Diamond like syndrome. *J. Med. Genet.* **54**, 558–566 (2017).
- Bezzeri, V. & Cipolli, M. Shwachman-diamond syndrome: molecular mechanisms and current perspectives. *Mol. Diagn. Ther.* **23**, 281–290 (2019).
- Frésard, L. & Montgomery, S. B. Diagnosing rare diseases after the exome. *Cold Spring Harb. Mol. Case Stud.* **4**, a003392 (2018).
- Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* <https://doi.org/10.1038/nrg.2017.116> (2018).
- Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
- Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
- Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DiSeAse MOdule Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* **11**, e1004120 (2015).
- Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
- Huang, J. K. et al. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6**, 484–495.e5 (2018).
- Guney, E., Menche, J., Vidal, M. & Barabási, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
- Nabirotkin, S. et al. Next-generation drug repurposing using human genetics and network biology. *Curr. Opin. Pharmacol.* <https://doi.org/10.1016/j.coph.2019.12.004> (2020).
- Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
- Caldera, M. et al. Mapping the perturbome network of cellular perturbations. *Nat. Commun.* **10**, 5140 (2019).
- Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **10**, 1197 (2019).
- Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **17**, 615–629 (2016).
- Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
- Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
- Robinson, S. et al. Incorporating interaction networks into the determination of functionally related hit genes in genomic experiments with Markov random fields. *Bioinformatics* **33**, i170–i179 (2017).
- Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
- Itan, Y. et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* **15**, 256 (2014).
- Stelzl, U. et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
- Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
- Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
- Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
- Basha, O. et al. Differential network analysis of human tissue interactomes highlights tissue-selective processes and genetic disorder genes. *bioRxiv* <https://doi.org/10.1101/612143> (2019).
- Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-019-0200-9> (2020).
- Gaudelet, T. et al. Unveiling new disease, pathway, and gene associations via multi-scale neural network. *PLoS ONE* **15**, e0231059 (2020).
- Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.* **12**, 1796 (2021).

64. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
65. Malod-Dognin, N. et al. Towards a data-integrated cell. *Nat. Commun.* **10**, 805 (2019).
66. Zanin, M. et al. Community effort endorsing multiscale modelling, multiscale data science and multiscale computing for systems medicine. *Brief. Bioinform.* **20**, 1057–1062 (2019).
67. Huttlin, E. L. et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28 (2021).
68. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.* <https://doi.org/10.1038/ng.3969> (2017).
69. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
70. Paulson, J. N. et al. Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinforma.* **18**, 437 (2017).
71. Pesquita, C. In *The Gene Ontology Handbook* (eds. Dessimoz, C. & Škunca, N.) 161–173 (Springer New York, 2017).
72. Pesquita, C., Faria, D., Falcão, A. O., Lord, P. & Couto, F. M. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **5**, e1000443 (2009).
73. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
74. Lin, D. An Information-Theoretic Definition of Similarity. in *Proceedings of the Fifteenth International Conference on Machine Learning* 296–304 (Morgan Kaufmann Publishers Inc., 1998).
75. Pesquita, C. et al. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinforma.* **9**(Suppl 5), S4 (2008).
76. Yu, G. In *Stem Cell Transcriptional Networks: Methods and Protocols* (Kidder, B. L. ed.) 207–215 (Springer US, 2020).
77. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).
78. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Interjournal Comp. Syst.* **1695**, 1–9 (2006).
79. Piraveenan, M., Prokopenko, M. & Zomaya, A. Y. Local assortativeness in scale-free networks. *EPL* **84**, 28002 (2008).
80. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
81. Gu, W., Tandon, A., Ahn, Y.-Y. & Radicchi, F. Principled approach to the selection of the embedding dimension of networks. *Nat. Commun.* **12**, 3772 (2021).
82. Yue, X. et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* **36**, 1241–1251 (2020).
83. Goyal, P. & Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *Knowl.-Based Syst.* **151**, 78–94 (2018).
84. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *J. Mach. Learn. Res.* **22**, 1–73 (2021).
85. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
86. Valdeolivas, A. et al. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **35**, 497–505 (2019).
87. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
88. Zurek, B. et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet.* <https://doi.org/10.1038/s41431-021-00859-0> (2021).
89. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
90. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
91. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
92. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
93. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
94. Buphalmai, P. et al. Network analysis reveals rare disease signatures across multiple levels of biological organization. <https://github.com/menchelab/MultiOme>. <https://doi.org/10.5281/zenodo.5562924> (2021).

Acknowledgements

This work was supported by the Vienna Science and Technology Fund (WWTF) through project VRG15-005 granted to J.M. The authors would like to thank Celine Sin for helping to host our MultiOme Explorer web app.

Author contributions

P.B. and J.M. developed the concept and designed the study. P.B. collected the data, implemented the algorithm, and conducted the analysis. T.K. and V.N. provided patient data and interpretation. P.B. and J.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-26674-1>.

Correspondence and requests for materials should be addressed to Jörg Menche.

Peer review information *Nature Communications* thanks Martin Schaefer, Lincoln Stein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

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

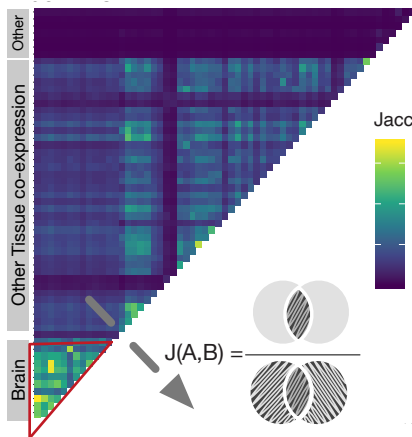


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

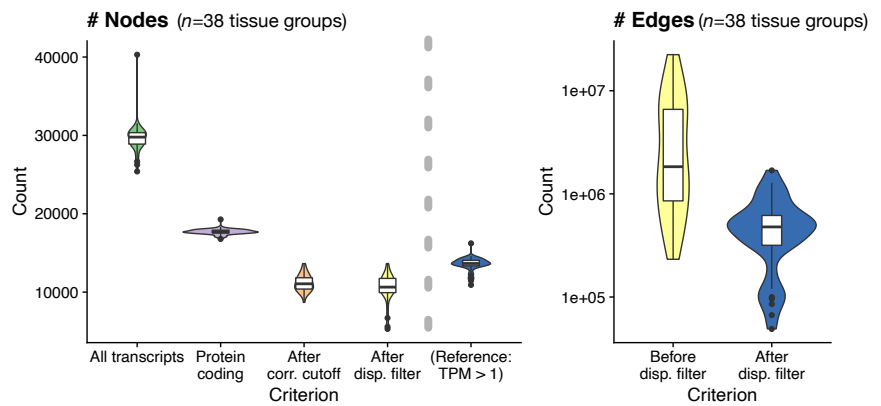
© The Author(s) 2021

Supplementary Figure 1

a Redundancy in original GTEx tissue definition

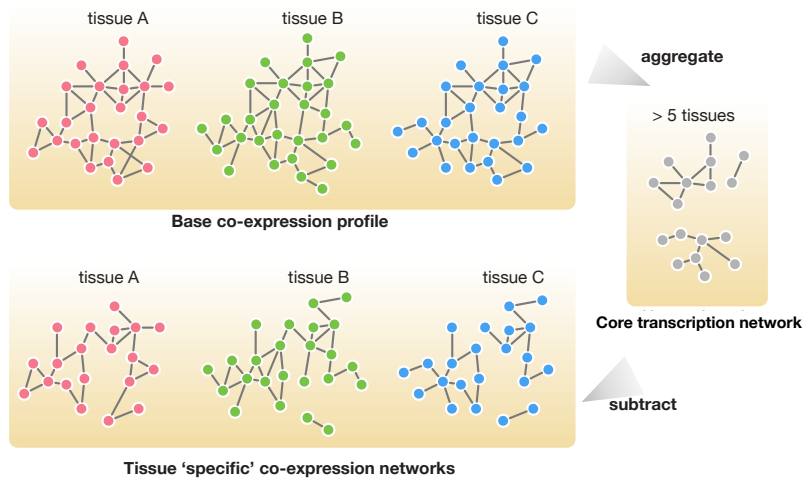
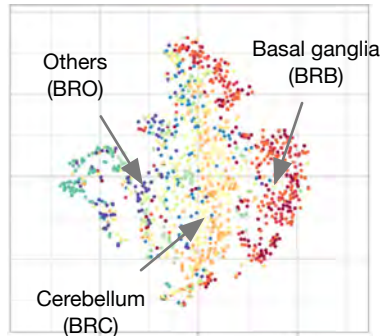


b # Nodes and edges at different stages of GTEx network construction

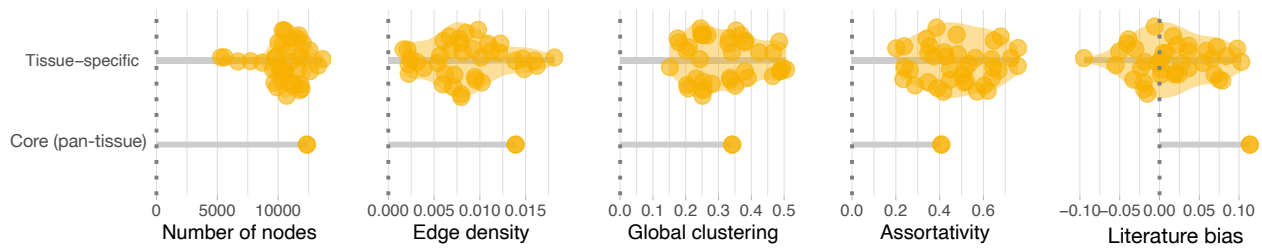


c Extraction of tissue-specific from core co-expression signals

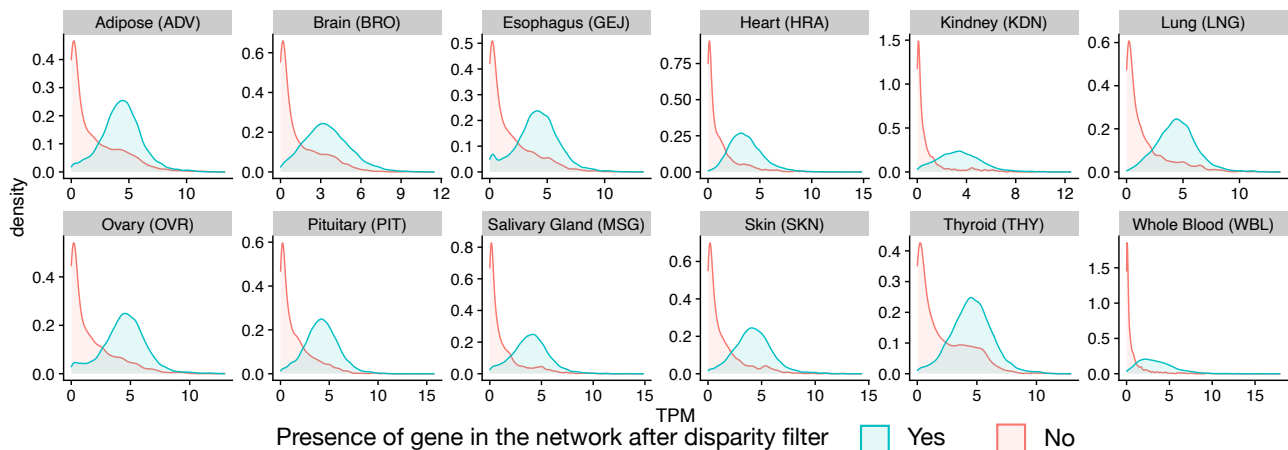
Brain (original = 13 sub tissues)



d Topological properties of core transcription network compared to its tissue-specific counterparts



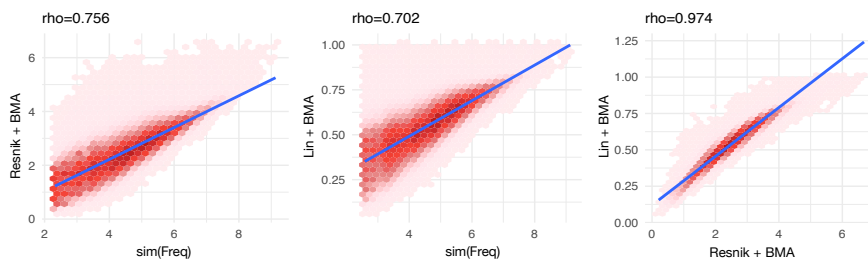
e Density plot of average tissue-specific gene expression, coloured by their presence in the networks



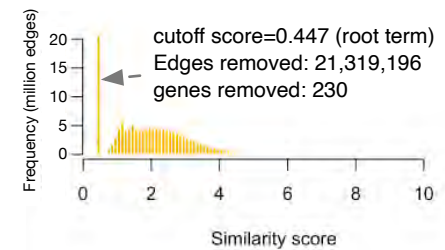
Supplementary Figure 1. Co-expression network construction. **(a)** Heatmap with pairwise Jaccard similarity between co-expression matrices of all GTEx tissues. Inset shows the t-SNE projection of expression profiles in 13 brain tissues, which were merged into three groups based on anatomical and expression similarity to reduce redundancy. In total, the process resulted in 38 tissue groups. **(b)** Left: number of genes for each step of the co-expression network construction process. Right: number of edges before and after applying the disparity filter ($n=38$ tissue groups). The network-based method removed a large number of spurious correlations (median number of interactions before and after = $1.83e6$ and $4.78e5$, respectively), while only slightly decreasing the number of genes. Bounds of box represent 25th and 75th percentiles, center the median, whiskers 10th and 90th percentiles, respectively. **(c)** Extracting tissue-specific co-expression networks by separating all edges observed in five or more tissues into an own core transcriptional network layer. **(d)** Topological properties of the core transcription layer, consisting of 12,364 nodes and 1,062,924 edges, compared to its tissue-specific counterparts. **(e)** Distribution of gene expression levels in the tissue-specific networks. The dynamic cutoff of the disparity filter generally tends to remove lowly expressed genes and keep highly expressed genes, while also allowing for the inclusion of genes that are lowly expressed, yet strongly correlated with other genes.

Supplementary Figure 2

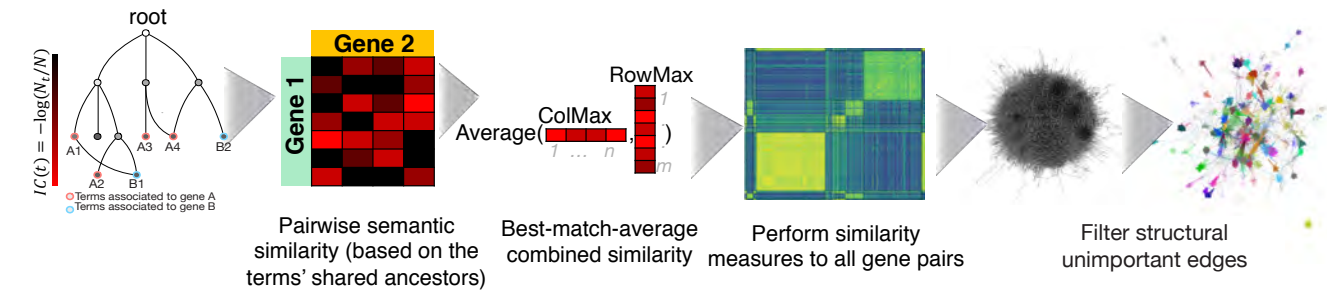
a Robustness of multiple methods in deriving ontology-based relationship



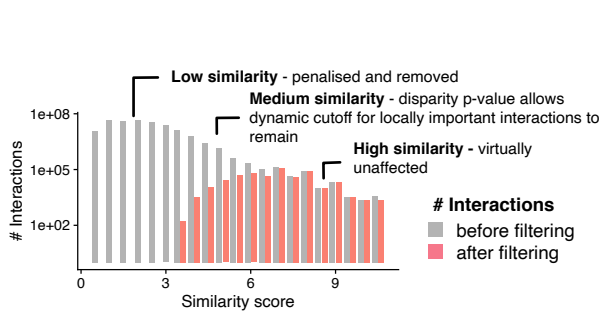
b Histogram of GO similarity score



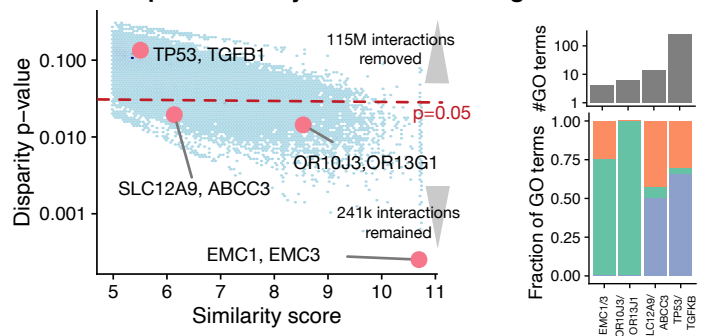
c Pipelines for extracting ontology-based relationships



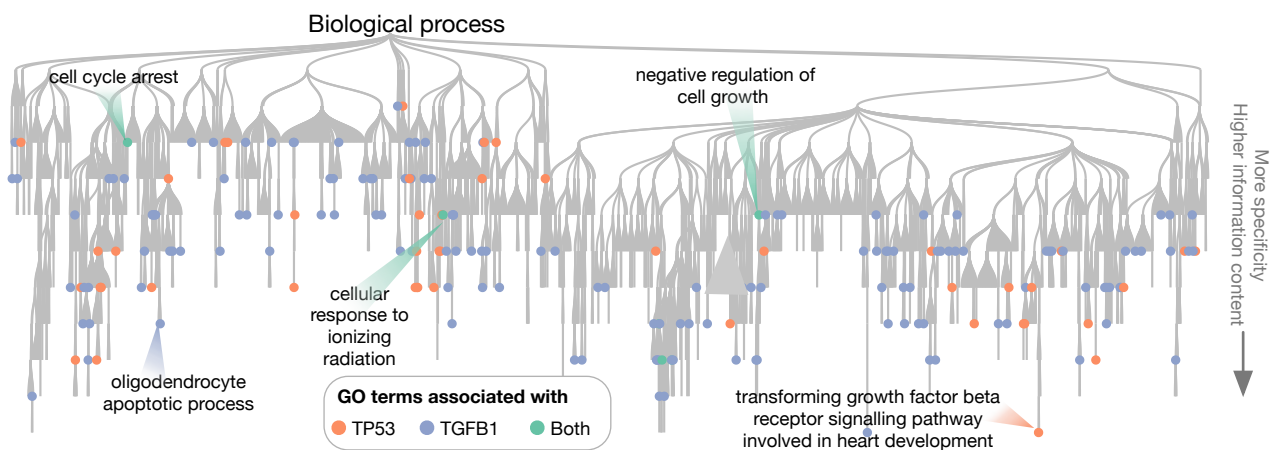
d Disparity filter allows dynamic scoring cutoff



e Gene pair similarity score and their significance



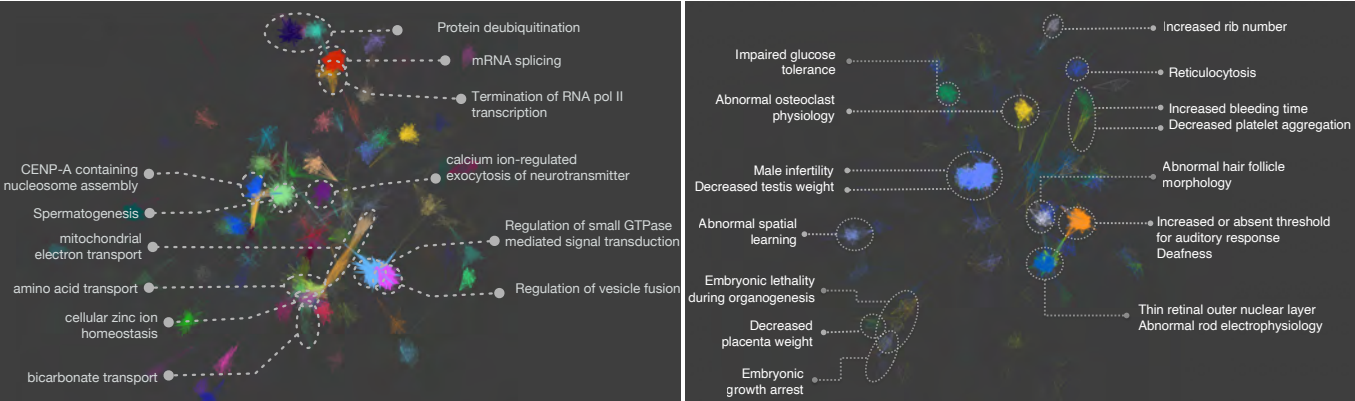
f Part of GOBP tree showing terms associated with TP53 (most studied gene) and TGFB1 (most annotated gene)



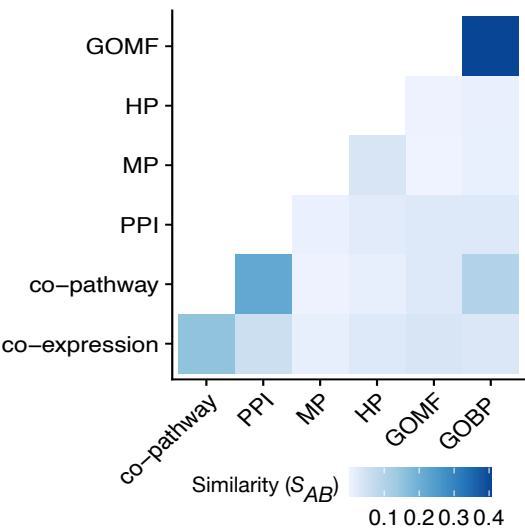
Supplementary Figure 2. Construction of the ontology-derived networks. **(a)** Pairwise comparison between different pairwise similarity measures (IC-based and frequency-based). The strong correlations show the robustness of ontology-based gene pair relationships. **(b)** Distribution of pairwise similarity of GO biological process annotations over all gene pairs. We filtered out all gene pairs with minimal similarity, i.e. those for which the only common ancestor is the root term of the ontology tree. **(c)** Overall pipeline for extracting ontology-based relationships. We measured IC-based semantic similarity of all gene pairs, then removed weak and redundant edges using the disparity filter. **(d)** Distribution of pairwise similarities of GO biological process annotations across all gene pairs before and after applying the disparity filter. The disparity filter corresponds to a dynamic cutoff. While gene pairs with low similarity scores are generally removed and pairs with high similarity scores remain virtually unaffected, gene pairs with medium similarity scores either remain or are discarded depending on the strength of the similarity scores with all involved neighbors. **(e)** Relationship between disparity filter *p*-value (see Methods for details) and similarity score according to GO biological process annotations for all gene pairs. Networks derived from semantic similarity measures favor gene pairs with similar annotation depths (e.g., from the same protein families) over highly, but diversely annotated gene pairs. Gene pairs with high similarity scores often belong to the same protein family such as the ER membrane protein complexes, olfactory receptors, and membrane transporters, and tend to share a large fraction of annotated GO terms. The barplot on the right shows the GO terms shared by four example gene pairs (green indicates shared terms, blue and red terms unique to the first and second gene, respectively). **(f)** Portion of the biological process branch of the GO highlighting terms annotated to TP53 (most studied gene) and TGFB1 (most annotated gene). While both genes are well characterized, with 87 and 176 annotated GO terms, respectively, only ten annotations are shared, indicating that they are involved in distinct biological processes. As a result, the computed similarity score and subsequent disparity *p*-value failed to reach the significance threshold, meaning that the two genes are not connected.

Supplementary Figure 3

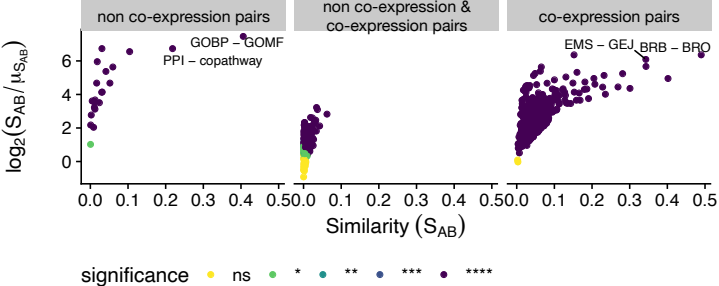
a Clusters identified by ontology-derived networks recapitulated original gene similarity and annotation



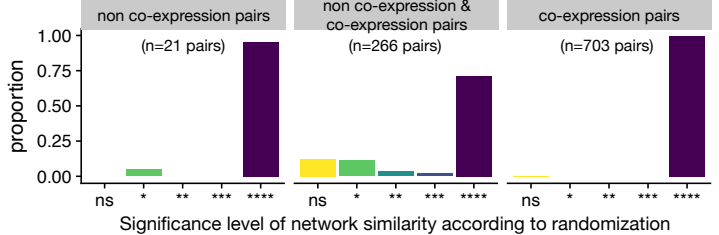
b Overlap similarity score across layers



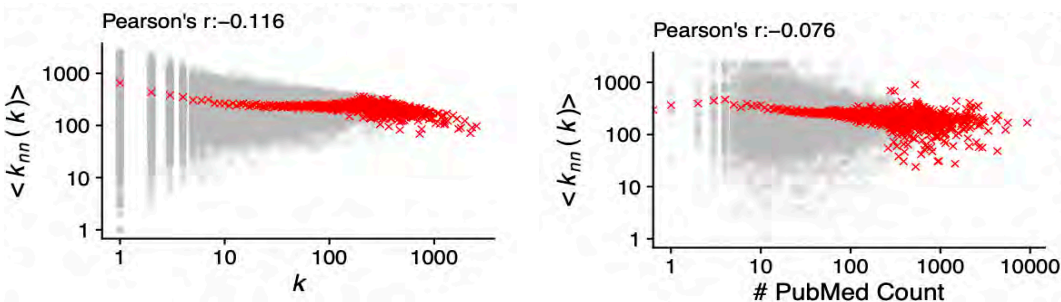
c Pairwise similarity of networks involved in the analyses



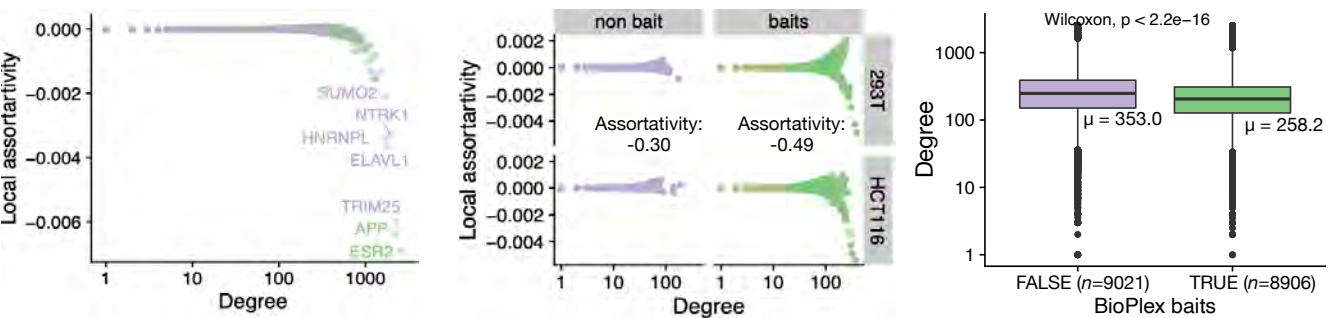
d Corresponding network similarity significance level



e Disassortativity measures: relationship between the average degrees of neighbors and node characteristics



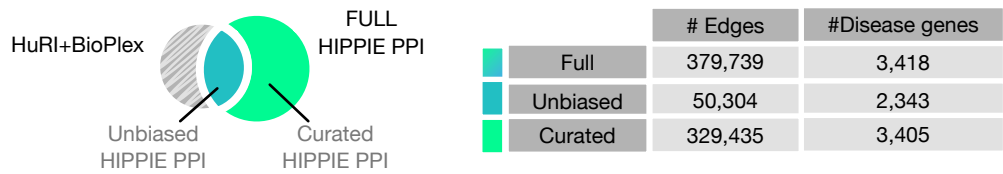
f Disassortativity in the PPI are partially derived by the experimental design on high-throughput studies



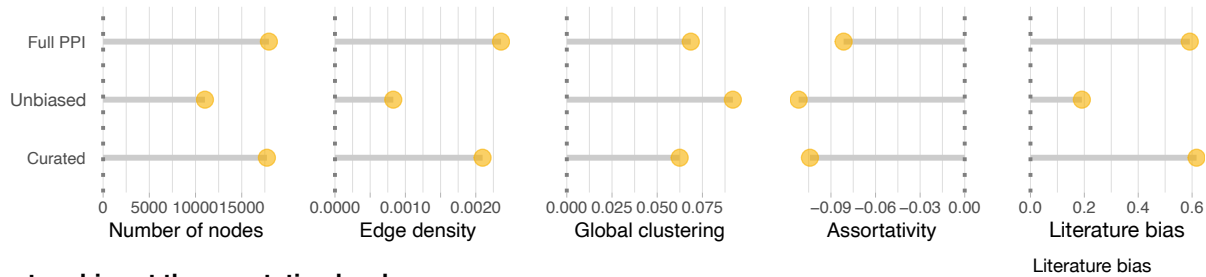
Supplementary Figure 3. Network properties and characterization. **(a)** Visualizations of the GO biological process (left) and mammalian phenotypic similarity (right) network layers with clusters highlighted that correspond to significantly enriched network communities (clusters with Benjamini-Hochberg corrected enrichment hypergeometric p-value $< 1e-20$ were labelled). **(b)** The pairwise edge overlap (S_{AB}) between the different layers is generally low, indicating that each layer provides different pieces of information and that the overall redundancy between the layers is low. **(c)** Pairwise similarity between all networks. The x-axis shows similarity scores (S_{AB}) of network pairs, the y-axis represents the effect size as measured by log fold-change of S_{AB} relative to the expected value from randomized network pairs of the same sizes. The panels show the values of non co-expression layers (n=21), non-coexpression and co-expression (n=266), and among co-expression network pairs (n=703), from left to right, respectively. **(d)** Bar charts summarizing the empirical significance of pairwise network similarity in (c), showing that the overlap is larger than expected by chance, in particular among the co-expression layers. This indicates that certain biological mechanisms are represented across different network layers. (Randomization p-values threshold: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****, one-sided Z-test, see Methods). **(e)** Average degree of neighbors of a gene vs. its degree (left) and PubMed count (right). Grey dots show the values for each gene, red dots the averages. Peripheral nodes (low k) tend to connect to nodes of higher degree, and vice versa, resulting in a weak overall correlation (Pearson's $r = -0.12$). Similarly, more studied genes tend to be connected to lower degree genes (Pearson's $r = -0.08$). **(f)** Local assortativity of genes within the BioPlex network for bait proteins (green) and prey proteins (purple). Left: The relationship between local assortativity score and network degree of the BioPlex network reveals disassortativity among hubs. Middle: Similar analysis, measuring interactions from the two cell lines used in the experiments separately, further revealed that bait proteins contribute to the overall disassortativity of the network. Right: Comparison of the average degree of neighbors between the bait and prey proteins: Bait proteins show lower average degree of neighbors than prey proteins (with degrees of 258 and 353; n=8906 and 9021 proteins, respectively; p-value $< 2.2e-16$, two-sided Wilcoxon test) and contribute more to the overall disassortativity of the network (degree assortativity = -0.49 and -0.30 for bait and prey proteins, respectively). Bounds of box represent 25th and 75th percentiles, center the median, whiskers 10th and 90th percentiles, respectively

Supplementary Figure 4

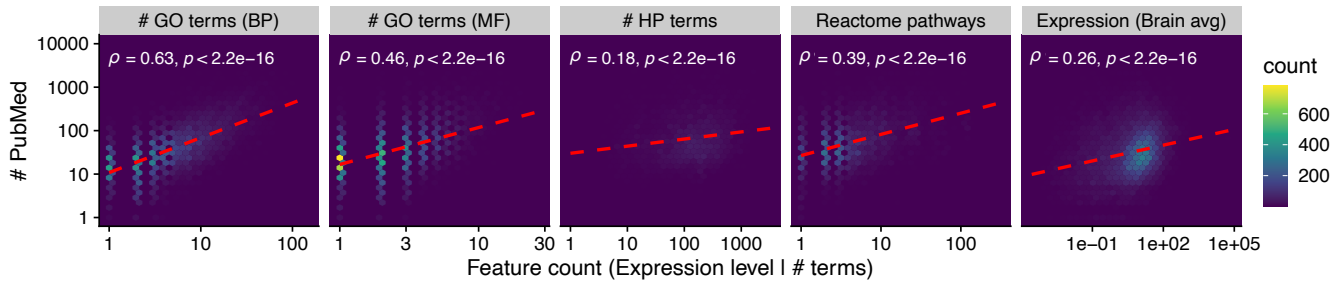
a Derivation of unbiased and curated subsets of the PPI network



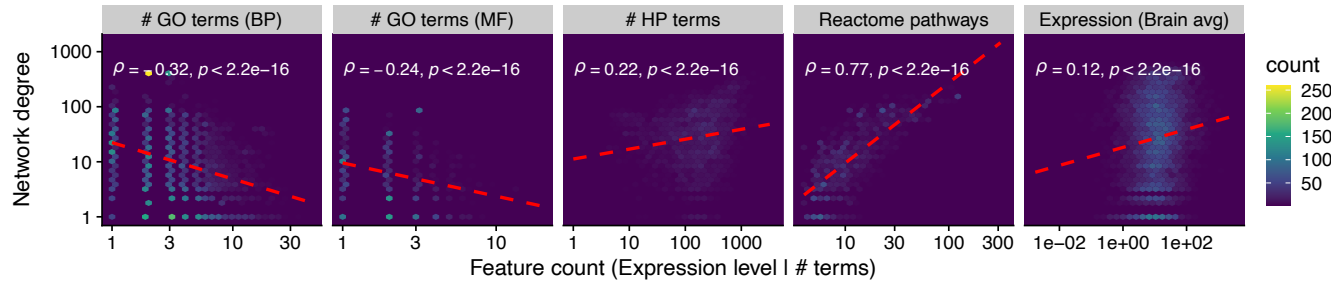
b Topological properties of the unbiased and curated PPI subsets compared to the full PPI



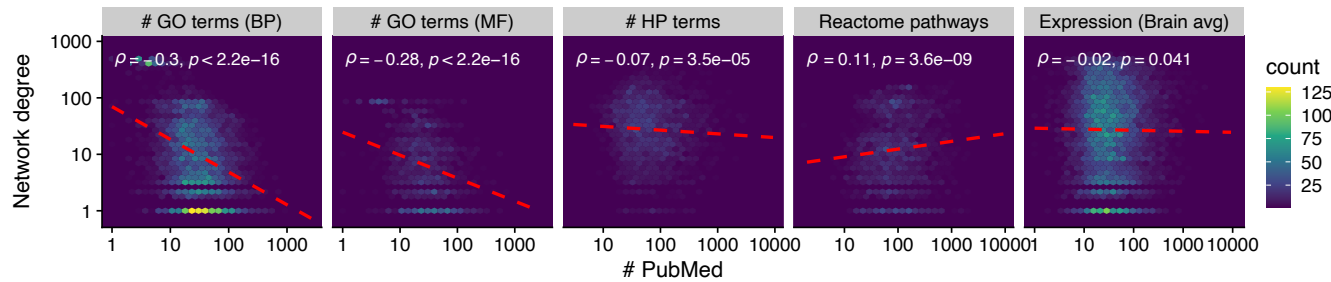
c Literature bias at the annotation level



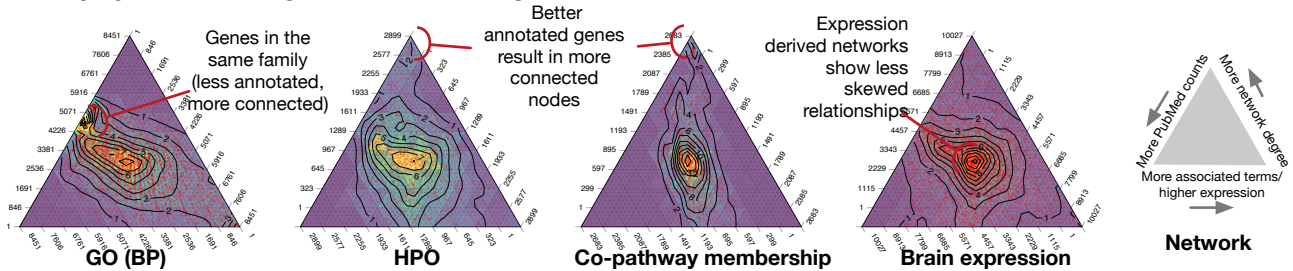
d Relationship between network degrees and the annotation/feature count



e Literature bias at the network level

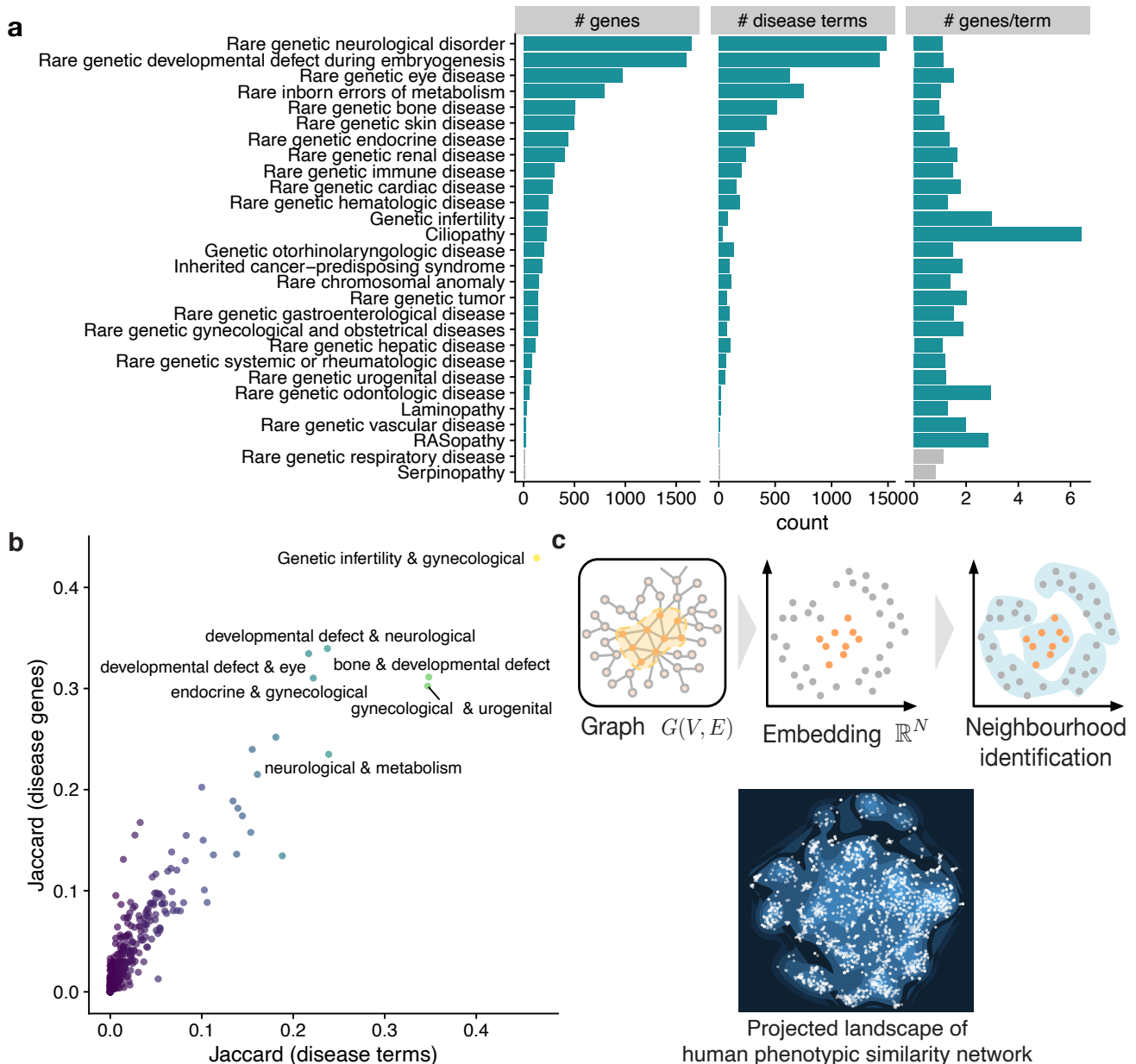


f Interplay between the gene feature, the degree, and the literature bias



Supplementary Figure 4. Interplay between data curation, literature bias and network characteristics. **(a)** Splitting the PPI network into two categories to investigate the impact of interactions curated from small-scale experiments on the prediction performance: the unbiased PPI with interactions from systematic high-throughput screens (Y2H-based HuRI and MS-based BioPlex), and the curated PPI for the rest. The unbiased and curated categories make up 13% and 87%, respectively, of the edges contained in the full PPI. **(b)** Topological properties of the PPI subsets compared to the full PPI. **(c)** Literature bias at the annotation level. The number of publications and gene features (expression level or annotation terms) are generally positively correlated, in particular for GO annotations. Data are represented as density. Red dashed line shows line of best fit, with Spearman correlation coefficient and the corresponding p-value (Fisher z-transformation, two-sided). **(d)** Relationship between network degree and number of annotations. **(e)** Literature bias at the network level. The correlation is reversed for the GO layer due to the similarity measurement (see Supplementary Fig. 2). **(f)** Ternary plots showing the interplay between number of annotations, PubMed count and network degree. The skew of the data from the center towards an edge or a corner represents anomalies (correlation) of these features. In the case of GO, hubs tend to emerge from groups of functionally similar, and in most cases, less diversely annotated genes. For example, the region with high density of genes towards the middle of the left edge represents a group of genes with high degree and average publication counts.

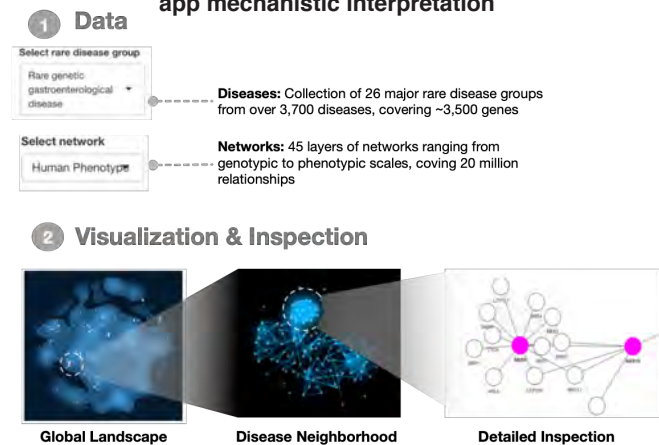
Supplementary Figure 5



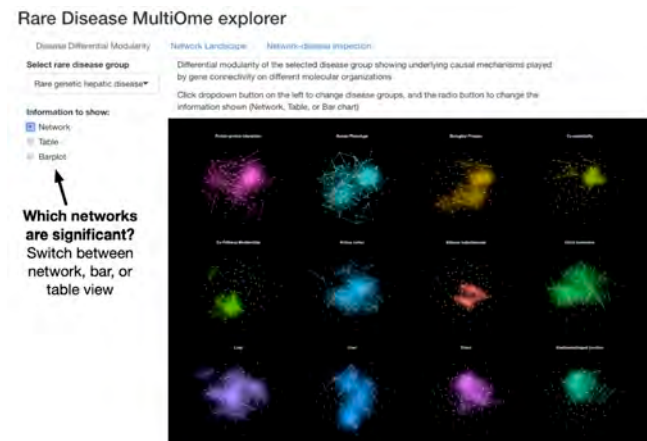
Supplementary Figure 5. Characterization of Orphanet disease groups. **(a)** Number of genes (left), individual disease term (middle) and their ratio (right) for the considered 26 disease groups. Grey bars represent terms with insufficient gene set size (i.e., less than 20) which were discarded from the analysis. Despite the wide range of the number of associated genes in different disease groups, the respective gene/term ratio remains consistent across all groups. **(b)** Pairwise similarity of disease groups as measured by the Jaccard index for overlapping descendant terms (x-axis) and overlapping annotated genes (y-axis). For example, rare genetic infertility is most closely related to rare genetic gynecological diseases. Overall, the disease groups are uniquely defined, with 90.5% of disease pairs having a Jaccard Index for shared genes < 0.1 . **(c)** Illustration of the construction of the network landscapes using the node2vec graph embedding algorithm followed by t-SNE projection onto 2d Euclidean space. White dots are disease genes, their positions reflect their connectivity on the corresponding network. We utilize this method to visually inspect large networks where their modularity can otherwise be difficult to observe.

Supplementary Figure 6

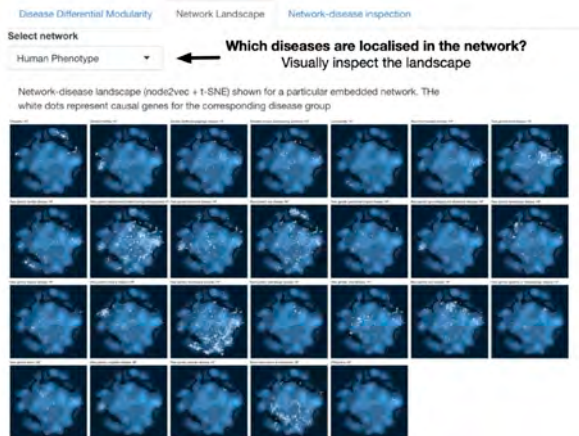
a An overview of the Explorer: complementary Shiny app mechanistic interpretation



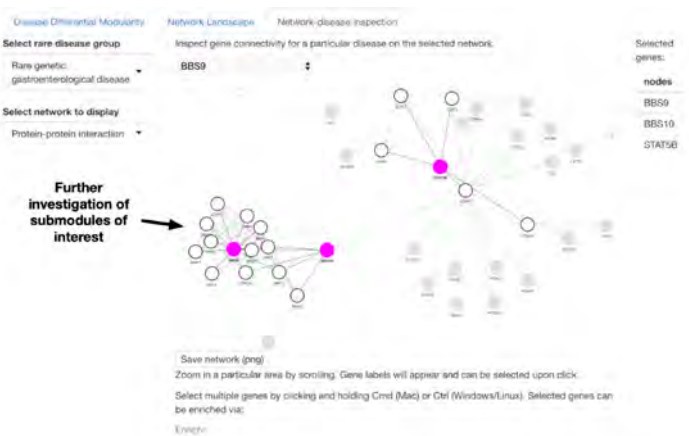
Tab 1: Differential modularity of disease groups on their relevant networks



Tab 2: Network landscape for all disease groups

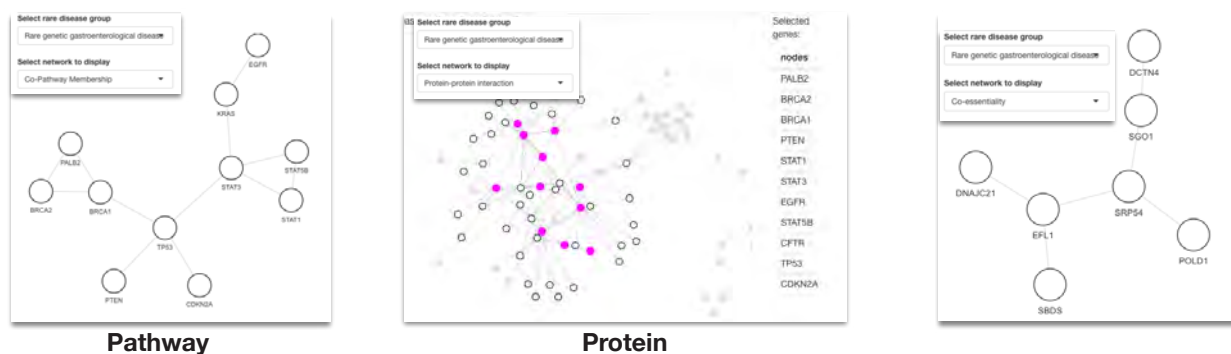


Tab 3: Detailed inspection of network submodules



b Mechanistic interpretation for different sub-modularities for rare gastroenterological diseases

1 Identify genes representing LCC on network layers of interest



2 Performing enrichment on respective databases to gain mechanistic insights

Index	Name	P-value	Adjusted p-value	Index	Name	P-value	Adjusted p-value
1	Downstream signal transduction Homo sapiens R-HSA-186763	1.011e-8	7.614e-7	1	ribosome assembly (GO:0042255)	0.0001603	0.009775
2	Signaling by SCF-KIT Homo sapiens R-HSA-1433557	7.585e-9	7.614e-7	2	nucleus localization (GO:0051647)	0.002098	0.03113
3	Signaling by FGFR1 in disease Homo sapiens R-HSA-5655302	3.620e-9	7.614e-7	3	centromeric sister chromatid cohesion (GO:0070601)	0.002098	0.03113
4	Signaling by PDGF Homo sapiens R-HSA-186797	1.492e-8	8.430e-7	4	SRP-dependent cotranslational protein targeting to membrane, translocation (GO:0006616)	0.002098	0.03113
5	Signaling by FGFR in disease Homo sapiens R-HSA-1226099	2.901e-8	0.000001311	5	nuclear migration (GO:0007097)	0.003146	0.03113
6	Signaling by cytosolic FGFR1 fusion mutants Homo sapiens R-HSA-1839117	1.005e-7	0.000003787	6	fatty acid homeostasis (GO:0055089)	0.003146	0.03113
7	Diseases of signal transduction Homo sapiens R-HSA-5663202	2.574e-7	0.000007271	7	intracellular protein transmembrane transport (GO:0065002)	0.003844	0.03113
8	Growth hormone receptor signaling Homo sapiens R-HSA-982772	2.489e-7	0.000007271	8	establishment of protein localization to endoplasmic reticulum (GO:0072599)	0.005239	0.03113
9	FGFR1 mutant receptor activation Homo sapiens R-HSA-1839124	5.516e-7	0.00001385	9	mitotic sister chromatid cohesion (GO:0007064)	0.005239	0.03113
10	Signaling by Interleukins Homo sapiens R-HSA-449147	0.000001182	0.00002672	10	guanosine-containing compound metabolic process (GO:1901068)	0.005587	0.03113

Reactome

Name	P-value	Adjusted p-value
BAX complex	2.355e-14	8.429e-12
Survivin complex	4.683e-14	8.429e-12
Fatty acid synthase complex	1.074e-12	1.288e-10
BCL-2 complex	2.101e-12	1.890e-10
Bcl-2 family protein complex	7.958e-12	5.730e-10
PTEN phosphatase complex	3.000e-11	1.800e-9
B cell receptor complex	5.997e-11	3.084e-9
Phosphatidylinositol phosphate phosphatase complex	7.516e-11	3.382e-9
DNA polymerase processivity factor complex	8.080e-10	2.909e-8
PCNA complex	8.080e-10	2.909e-8

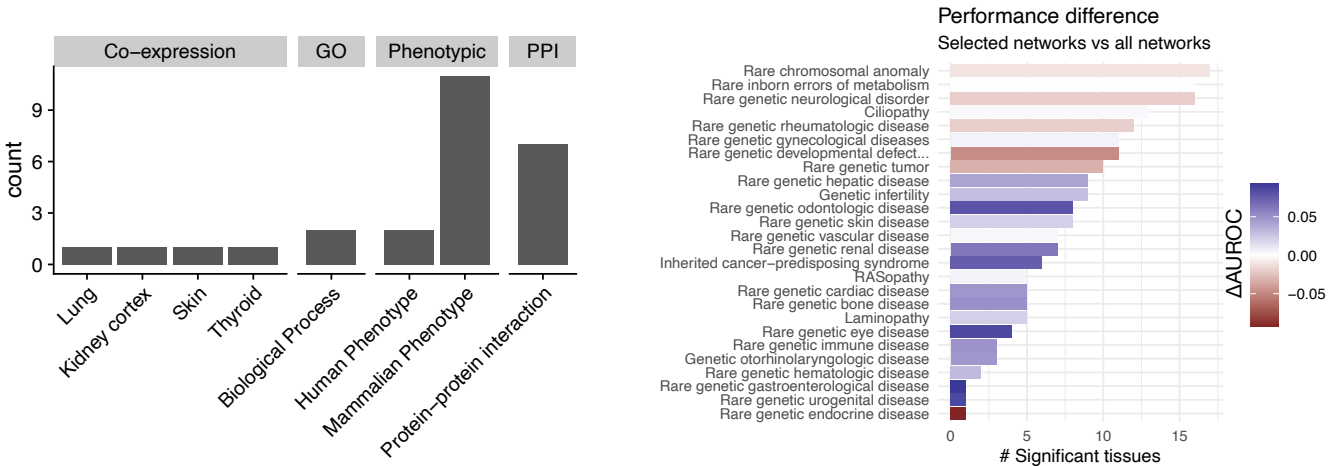
Jensen Compartments

Go Biological Process

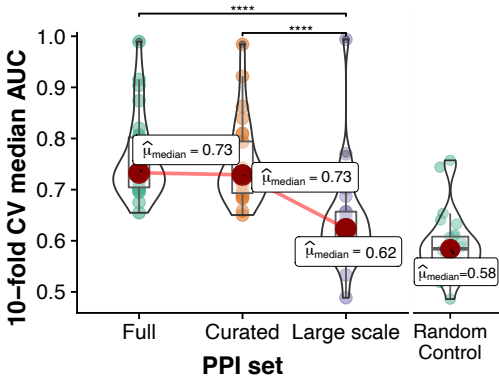
Supplementary Figure 6. Overview of the MultiOmeExplorer web app (www.menchelab.com/MultiOmeExplorer). **(a)** The interactive Shiny app provides visualizations of all network-disease relationships that were considered in this study and allows for detailed inspection of individual disease modules of interest. The first tab gives an overview of the connection patterns of a particular disease on all network layers that were identified as significant. The second tab shows the landscape of all rare disease groups for a selected network layer. The third tab allows for a detailed inspection of submodules of interest and extract genes for downstream enrichment analysis. **(b)** Example usage for interpreting significant connection patterns among rare gastroenterological diseases. On pathway and protein levels, genes connected within the respective LCCs correspond to genes causing pancreatic carcinoma. These genes are enriched for apoptotic processes such as Bax and Survivin complexes, and major cancer pathways such as SCF-KIT, FGFR1, and PDGF. On the co-essentiality level, the LCC represents causal genes of Shwachman-Diamond syndromes, where they are co-dependent for ribosome biogenesis.

Supplementary Figure 7

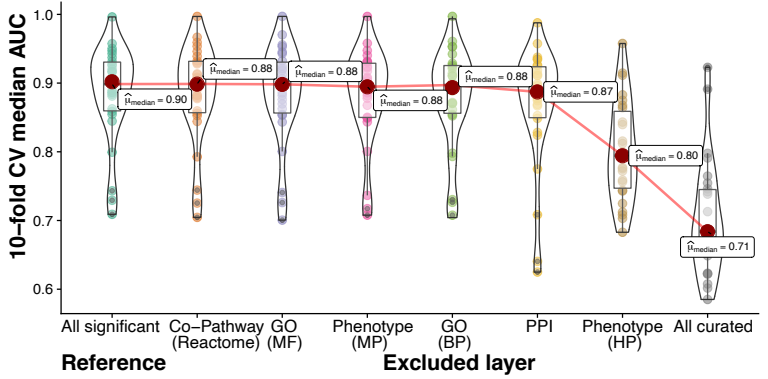
a Most relevant single layer for each disease group **b** Heterogeneous disease groups perform better using all layers



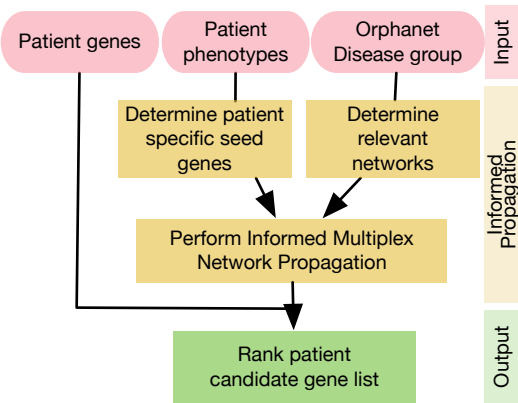
c The performance of different PPI sets



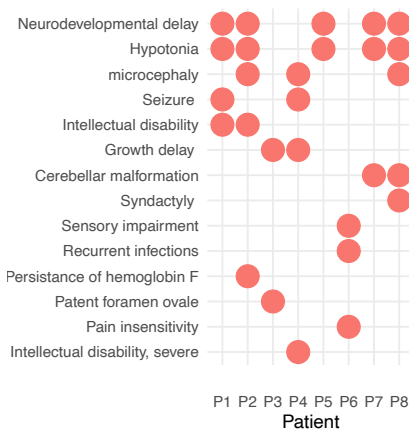
d The performance upon network layer removal



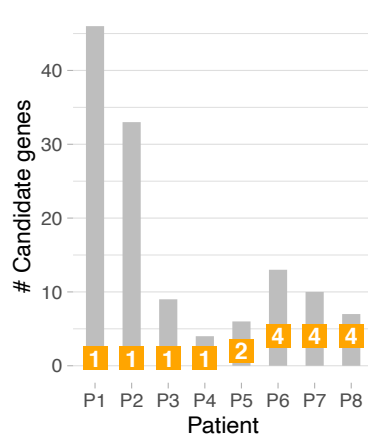
e Workflow: patient-specific informed propagation



f Local cohort: phenotypes



g Local cohort: Causal gene ranking

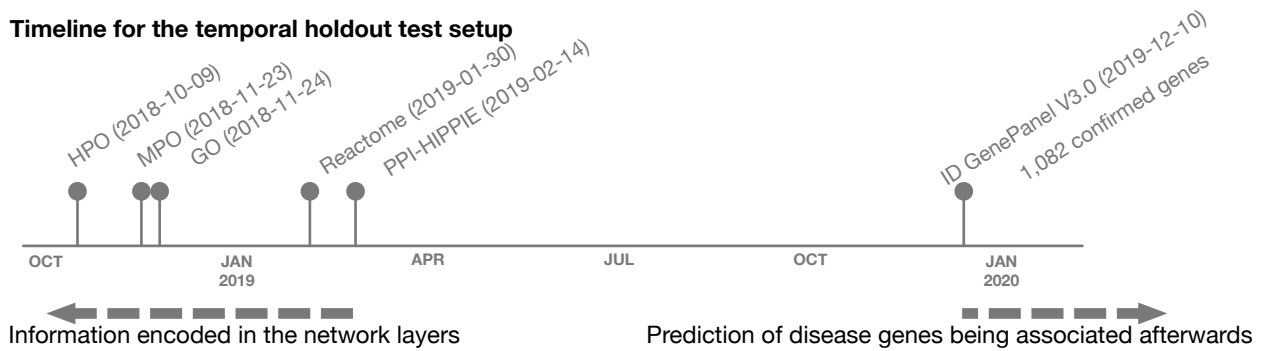


Supplementary Figure 7. Disease gene prediction and causal gene prioritization in patients.

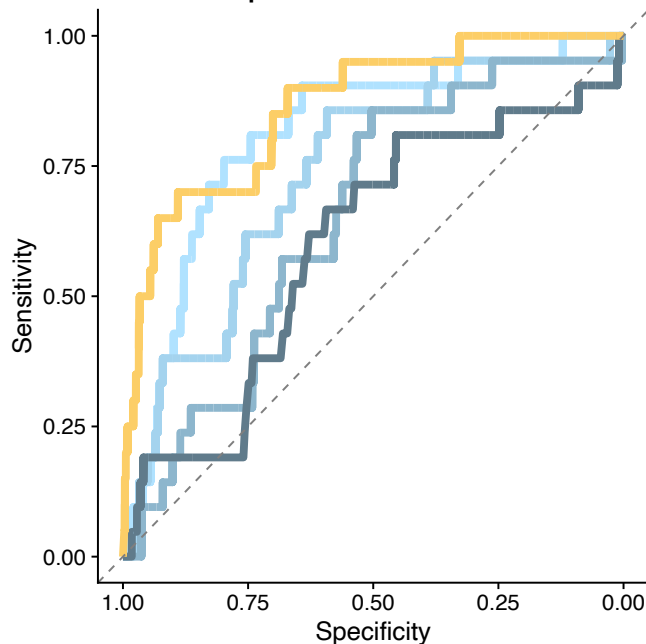
(a) Number of times that a particular network layer was found to be the most relevant layer (i.e., with the most significant modularity) for a particular disease. **(b)** Performance difference (Δ AUROC) between using only relevant layers and all layers. Red: All layers perform better, blue: relevant layers perform better. Inclusion of all network layers yields better retrieval performances compared to selected relevant networks in syndromic disease groups (i.e., diseases that manifest in multiple physiological systems). **(c)** 10-fold cross validation retrieval performance of disease genes for different PPI subsets ($n = 26$ disease groups). The curated PPI performs equally well as the full PPI (median AUROC = 0.73), whereas the unbiased PPI sees a significant performance drop (median AUROC = 0.62, p -value = $1.76e-9$, FDR-corrected Durbin-Conover non-parametric test). Random subnetworks of curated PPI that have the same number of edges as the curated PPI show a comparable performance drop (with a median AUROC of 0.58, even slightly reduced). Threshold for p -values: $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***, $p < 0.0001$:****. Bounds of box represent 25th and 75th percentiles, center the median, whiskers 10th and 90th percentiles, respectively. **(d)** Prediction performance of the multiplex network upon removing network layers derived from curated databases ($n = 26$ disease groups). For most layers, the 10-fold cross validation AUROC performance is only slightly decreased after their removal (median AUROC between 0.87 and 0.88 for pathway, GO, MPO and PPI layers), only the removal of the HPO layer had a stronger impact (AUROC = 0.80, p -value = 0.0003, FDR-corrected Durbin-Conover non-parametric test). Removal of all layers that involve curated data (Reactome, GO, HP, MP, and PPI) resulted in the lowest performance (AUROC = 0.71, p -value = $1.17e-11$). Elements of boxplots are as described in (c). **(e)** Schematic of the patient-specific informed propagation framework for prioritizing patients' causal genes. **(f)** Phenotypic terms (HPO) associated with different patients in the local cohort. **(g)** Causal gene ranking for the local cohort (8 patients). Grey bars represent the number of candidate genes in each patient (mean = 16), yellow boxes indicate the ranking of the actual causal genes provided by our framework.

Supplementary Figure 8

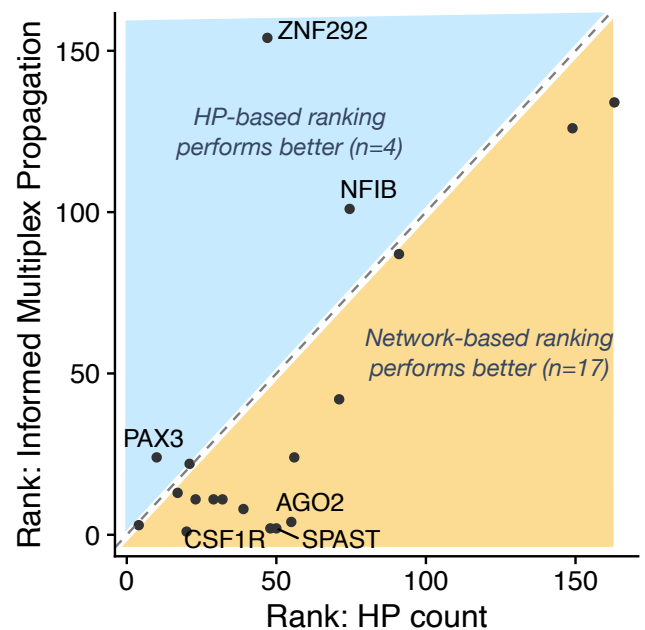
a Timeline for the temporal holdout test setup



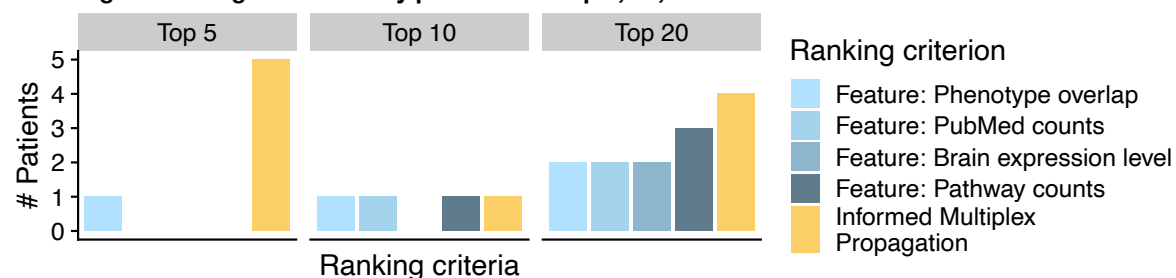
b Performance of temporal holdout test



c Comparison of the two best performing methods



d Ranking of causal genes correctly predicted at top 5, 10, and 20



Supplementary Figure 8. Temporal-holdout analysis of causal genes discovered after network construction. (a) Timeline for identifying patients with causal genes that were discovered after the curated databases (GO, MPO, HPO, and the PPI) were retrieved. (b) ROC curve of our approach (yellow, AUROC = 0.95) and various gene level based benchmarks for the temporal-holdout set of 21 patients with causal genes discovered after data curation. The overall performance as measured by the AUROC remained high for all tested prediction methods, slight reductions (e.g. from 0.90 to 0.86 for the informed multiplex propagation) were within the 10-fold interquartile range in most cases (c) Comparison between the predicted ranks of true genes according to the two best performing methods, the informed multiplex propagation, and the phenotypic overlap. The former yields better ranking in 17 cases (81%). (d) Number of patients for which the true causal gene was prioritized among the top five, 10, and 20 for all considered methods.

4

Discussion

Network-based methodologies and visualization have become established practices within data science and have been embraced across all disciplines. Rare disease analytics and diagnostics is among the disciplines where methodologies developed in network biology have been utilized, leading to accurate means for disease gene discovery and interpretation of causality. The work presented in this doctoral thesis further demonstrated that heterogenous biological datasets can be leveraged to address several practical and conceptual challenges in rare disease research, particularly related to data scarcity. This chapter aims to provide a detailed discussion on key findings in the publication. The points of discussion cover (i) various aspects of network construction, including the definition of biological scales, the comparison of structural characteristics and the complementarity across networks layers, (ii) the quantification of modularity and performance in disease gene retrieval and (iii) clinical applications. Alternative methods are also discussed where applicable. The chapter ends with future prospects and an outlook of the challenges described in this thesis as well as in network medicine in general.

4.1 Scale representation, network construction and characterization

A motivation of this work is based on observations in previous studies that despite leaps in PPI mapping technologies, the uncharted territories of the human interactome remain vast. With the majority of the interactions being collected based on small-scale experiments, our study showed that in addition to being incomplete, the human interactome is largely impacted by social bias, an observation that is in line with previous studies (Luck et al., 2020; Huttlin et al., 2021; Menche et al., 2015). On this issue, a particularly noticeable network property that

reflects the interplay between the interaction discovery process that led to the social biases is the disassortativity observed in the PPI layer. This is an intriguing process and to our knowledge not completely understood yet. Our conclusion was that the observed disassortativity of the PPI may arise both from hypothesis-driven, small-scale experiments, as well as from systematic, large-scale efforts: Over time, well-known proteins with important functions aggregate more and more interactors from a large number of small-scale experiments, thus resulting in the observed strong correlation between PubMed count and PPI degree. If we assume that the interactors, on average, do not receive the same level of attention as the hub protein itself, we would observe disassortative connection patterns even though the complete underlying network may be uncorrelated. Similar arguments can also be made for large-scale experiments: While they can be designed to avoid social biases by selecting a wide range of bait proteins, the interactors of the selected bait proteins may still be mapped out more thoroughly compared to the non-baits. In summary, we conclude that as long as the search space of the PPI has not been mapped at uniform depth and full proteome scale, connectivity biases are likely to remain prevalent. The true nature of the correlation structure of the PPI network thus remains an open question.

To overcome the limitation of the PPI as well as to gain better perspectives of the whole system, we incorporate additional data to complement the PPI. These data are represented as ‘scales’, the hidden interconnectivity between genotypes to phenotypes. From our perspectives, each of these network scales provide a snapshot of the system and how the effect of a perturbation at the bottom, the genetic scale, ripples through intermediary layers and subsequently emerges at the surface level, the phenotypes. In practice, it is not entirely how these intermediary scales are defined, or which available data can best represent them. Some studies even dismiss the entire network notion of biological organization and chose to envision them as hierarchical system (Ma et al., 2018; Yu et al., 2016). When representing biological system as multiscale networks, a major challenges is distinctive from other systems such as transport or social networks, where layers in the systems were clearly determined and readily represented as network data. As introduced in Section 1.3 and illustrated in Figure 1.2, the scales in biology can be observed at spatial, temporal, as well as functional sense. Our scale representation was originally inspired by the idea of the central dogma of molecular biology (Crick, 1958) that genetic information transfer starts at genes before being transcribed into transcripts and translated into proteins. In our work, we wanted to observe such processes at the interaction level, and therefore representing them as networks at the genomic, transcriptomic and proteomic scales. Combined with three additional layers based on functional annotations derived from pathways and ontologies data, we were able to extract insights at the functional level, bridging the gap between proteins and phenotypes. Our framework has demonstrated that data of different types, from structured formats such as ontologies to continuous numeric data such as gene expression, can be converted into a unified framework of multiplex networks. Nevertheless, the computational cost of converting diverse sets of data into such frameworks can be prohibitive. This primarily resulted from the resource-

intensive computation of semantic similarity for every gene pair across different ontologies, which has to be performed in a regular basis when newer versions of each dataset are generated.

Another challenge is the extraction of data used to represent these biological scales. Several factors can affect the resulting networks and therefore all interpretation based thereon: This includes (i) biological factors: the context such as cell lines or tissues in which the data was sampled, (ii) experimental factors: the ‘readout’ of the data which provides the accuracy and resolution of the measurements, and (iii) technical factors: this includes algorithms and processing steps towards capturing the relationships among entities that were measured. The first two factors are subject to the availability of data and technologies that best represent biological scales. The influence of the last factor has to a certain extent been demonstrated in Chapter 3.

For example, at the genetic level, representing genetic interactions has proven to be challenging at the experimental, biological and technical levels. Genetic interactions are defined as the deviation of observed phenotypes upon digenic mutations in comparison to the effects observed in individual mutations (Mani et al., 2008), and have been systematically characterized in large-scale especially in yeast (*Saccharomyces cerevisiae*) (Costanzo et al., 2019; van Leeuwen et al., 2016; Kuzmin et al., 2018). This effort is more difficult to be translated for human genes because the size of the yeast genome is less than a third of that of human, and such interactions in yeast are usually quantified by simple readouts such as cell growth. Even with latest technologies such as CRISPR, studies in human genes were only able to perform double gene knockouts at only a small fraction of all possible gene pairs (Han et al., 2017; Thompson et al., 2021). With the limitation of scalability of digenic perturbation, there were efforts of using indirect measurements such as correlated fitness profiles upon gene knockouts across cell lines (Kim et al., 2019). Networks resulting from such experimental setup may only represent a distant proxy to the underlying genetic interactions, but they represent general and genome-level interactions at a scale that is currently prohibitive in targeted experiments.

The transcriptome scale, in contrast, has benefited from the efforts by the GTEx consortium to map out transcriptional profiles across 53 tissues (GTEx Consortium, 2015; eGTEx Project, 2017). This enabled the construction of tissue-specific co-expression networks. Samples from tissues that exhibit similar expression profiles were merged to increase sensitivity, resulting in 38 unique tissues (Paulson et al., 2017). Indeed, we found that tissue specificity and consequently their unique co-expression profiles being a key in investigating disease mechanisms. However, the ‘common core’, i.e., the co-expression between gene pair observed across multiple or all tissues, which we have identified to be enriched by essential and house-keeping genes, are also found to play significant roles. Our results show that they contribute to multiple rare diseases groups, especially for those affecting multiple physiological systems. Our approach of extracting the common core as its own layer aimed to reduce the issue that the propagation across the multiplex network may likely favour these group of genes due to their high occurrences across tissues, and could potentially mask tissue-specific signals. In total, the transcriptome scale

was mediated by 39 network layers, each representing tissue characteristics, and constitute the majority (84%) of all network layers in consideration.

With the lack of comparable tissue-level information for protein-protein interactions, several studies have chosen to impose tissue specificity on the expression level directly onto the interactome (Luck et al., 2020; Basha et al., 2020). This assumes that the absence of transcripts in a tissue rules out that their protein products interact. We chose not to pursue such approaches for three main reasons: First, the extent to which protein levels can be predicted from mRNA expression is still under debate (Fortelny et al., 2017; Wilhelm et al., 2014). Filtering protein interactions based on co-expression data may thus introduce additional inaccuracies. Second, in the context of disease aetiology, protein interactions across different tissues may play an important role, for example between signaling proteins that are secreted from one tissue and their corresponding receptors in another. The absence of co-expression of such gene pairs in the same tissue does not exclude the presence of the interactions of their protein products in such crosstalks, which could potentially be relevant in a disease context. Third, as co-expression and PPI networks provide complementary insights at different biological levels, we deemed it useful to also keep them distinct from a methodological perspective. In our context, a co-expressed gene pair represents the co-variability of their expression levels across all samples in a given tissue. The detection of such interactions can be used to inform higher-level regulatory mechanisms, for example related to tissue identity.

In summary, we carefully constructed networks from diverse types of data, with the goal of obtaining networks that are comparable in size and complementary in utility. A set of measures on both topology and similarity among the networks was applied. Our results reveal not only a wide structural diversity which reflects intrinsic orchestration of individual biological scales, but also subtle technical and historical details resulting from data curation processes. There remain challenges which primarily derived from the consistency of networks constructed from vast amount of resources of heterogeneous types, and the sustainability in keeping the constructed networks up to date as new knowledge and data are being generated. In Section 4.4, alternative methodologies for storing such large datasets while maintaining general utilities are discussed.

4.2 Mapping rare disease genes and modularity quantification

A primary motivation for our work was to show that network methodologies originally developed for complex diseases can also be applied to rare diseases. A major challenge, however, is the fact that there are as many rare diseases as there are rare disease genes, according to the Orphadata rare disease-gene association (<http://www.orphadata.org/>). This means that the vast majority of rare diseases are associated with only one or a few genes. The quantification of disease modularity using similar approaches as for complex diseases would not be possible

due to the small gene sets. To this end, we first showed that after aggregating rare diseases into groups, those with similar phenotypes exhibit the same network characteristics as complex diseases, namely that they aggregate in disease specific network neighborhoods. This finding allowed us to overcome the scarcity of information that is particular to individual rare diseases and represents not only a major roadblock for diagnosis and treatment, but also for applying network approaches that rely on the existence of a well defined seed neighborhood. With this seed neighborhood in place, we can now leverage the large body of existing literature on network methods used for studying complex diseases. Specifically, disease gene prediction methods, including diffusion-based methods as the one applied in this thesis, as well as multilayer network approaches have been shown to be effective in several diseases. The aggregation of rare diseases as group, however, relies on the quality of the data and classification of diseases provided by curators and developers, in this case the Orphanet Rare Disease Ontology. We have observed that top-level disease groups were primarily organized based on physiological presentation, but also mechanistic description such as chromosomal anomalies or broad phenotypic terms such as infertility. Additionally, depending on the activity of specific communities and consortia, the level of sub-classification, curation and maintenance of annotation for different disease groups may vary. This results in vastly different disease terms and their corresponding number of associated genes. This may partly explain why disease modules in certain rare disease groups are less well defined, as well as why the predictive power of gene prioritization for those disease groups are lower.

Beyond the quantification of modularity, we wanted to visualize disease modules so that they can be inspected and compared against one another, which could facilitate further understanding of disease relationships. However, performing such task in up to 46 genome-scale networks can be difficult, as projecting large networks onto limited space can easily result in intertwined hairballs. To solve this issue, we mapped the networks onto the ‘disease landscape’ - the embedding of network information onto two-dimensional Euclidean space via node embedding algorithm node2vec (Grover & Leskovec, 2016). We found that the projected landscape aided in providing additional insights on potentially affected tissues or likely molecular mechanisms at a glance. However, parameterization on both node embedding (here, node2vec) and the dimensionality reduction (here, t-SNE) can influence the final coordinates of the nodes at the low Euclidean dimension. As a result, we chose to restrict the projected landscapes for visualizing purposes rather than for prediction and interpretation.

Finally, our choice of using the largest connected component (LCC) as the measurement for network-disease significance (π) was based on previous observation (Menche et al., 2015) that disease genes may form a more significantly connected cluster compared to random gene sets of the same size, but not necessarily more *densely* connected to apply measures such as clustering coefficients. Alternative methods in disease module identification involving various techniques including similarity matrices, optimization algorithms, or ensemble methods have been compared

(Choobdar et al., 2019).

4.3 Disease gene retrieval and prediction performance

We investigated the extent of signals being picked up in network-based prediction methods are due to the circular confirmatory bias, i.e. whether the performance reflects disease knowledge already embedded in the networks rather than the intrinsic utility of the networks beyond the known information. Indeed, while the ability to pick up subtle relationships hidden in large databases is a particular strength of integrative network methods, this also poses challenges for disentangling the extent to which they actually predict previously unseen relationships. For large-scale studies integrating a broad range of data sources and considering thousands of rare disease genes such as the one presented in this thesis, it is unfeasible to retrace every piece of information contained in a particular knowledge base whether it ultimately stems from a rare disease gene discovery. To address potential circular confirmatory biases, we carefully performed three sets of analyses and investigated the predictive power of these following scenarios: (i) on the bias of the interactome: we performed analyses on various PPI subsets containing different levels of curation, as well as of (ii) the bias of networks derived from curated data: we compared the results under removal of different network layers derived from curated information. Finally, to minimize the information embedded in the networks that might affect the prediction, we implemented (iii) a temporal holdout setting for prioritization tasks using only disease genes that were discovered later than all data used to build the networks.

To assess the impact of PPI subsets, we analyzed the influence of interactions curated from small-scale experiments on the prediction performance of the PPI network as follows: We first collected systematic high-throughput PPI from two largest studies: the Human Reference Interactome (HuRI) (Luck et al., 2020) and the BioPlex interactome (Huttlin et al., 2021). The two resources provide technically complementary information as the former are based on yeast two-hybrid screening (Y2H) while the latter was constructed from affinity-purification mass spectrometry (AP-MS) profiling. We can therefore split the PPI network in two groups: (i) the ‘unbiased PPI’ for interactions that are confirmed in at least one of these two resources and (ii) the ‘curated PPI’ for the remaining interactions which have been curated from small-scale studies without being validated in the two large-scale resources. We observed that the unbiased and curated categories make up 13% and 87%, respectively, of the edges contained in the full PPI. The unbiased PPI is thus considerably smaller, and three times sparser than the full PPI as measured by edge density, more disassortative and, as expected, less affected by literature bias compared to the full PPI. To assess the performance of the two PPI subsets in retrieving known disease genes, the accuracy via 10-fold cross validation for all rare disease groups using the single layer random walk with restart on all subsets of the PPI were compared. The curated PPI performs equally well as the full PPI (median AUROC = 0.73), which was expected since the former contains

87% of the edges of the latter. The unbiased PPI, however, sees a significant drop in retrieval performance compared to the full PPI (median AUROC = 0.62). This again is expected due to the fraction of information it contains. We further assessed whether the performance drop in the unbiased PPI is due to the significantly reduced network size. In doing so, the analysis was repeated on bootstrapped samples from the curated PPI that contain similar number of nodes. We observed that the retrieval AUROC of these random curated PPI subsets also dropped to a comparable level to that of the unbiased PPI. This led to our conclusion that the performance drop of the unbiased PPI subsets is mostly due to its reduced network size, and that the presence of curated interaction in our PPI network play only minor roles in confirmatory bias.

In addition to the PPI, we also further assessed whether the retrieval performance of rare disease genes is influenced by other network layers that were constructed based on curated databases. This includes layers based on phenotypic similarity (HPO and MPO), pathway co-membership (Reactome), and GO term similarity (BP and MF branches). Initially, individual layers were removed and the retrieval performance via 10-fold cross validation AUROC was computed. We observed that the performance stay robust against removal of a network layer for most parts (median AUROC between 0.87 and 0.88). However, our results showed that the removal of HPO-based phenotypic similarity network may have a stronger impact (AUROC drops to 0.80 upon removal). This is not beyond expectation as the phenotype is one of the closest proxies to disease classification. In addition to individual layer removal, the predictive performance upon removing all layers based on curated databases, i.e. Reactome, GO, HP, MP, and PPI, was assessed. In other words, only co-expression in relevant tissues and the co-essentiality network remained for the retrieval process. Our result showed that the predictive power carried by these high-throughput data was significantly lower compared to when all relevant network layers were incorporated.

Another point of discussion is whether the performance in disease gene retrieval is primarily driven by particular layers. Our results show that the respective most relevant network alone (as measured by the disease modularity) is able to reach satisfactory predictive power. Interestingly, it was clear that adding more layers, and ultimately all network layers, decrease the predictive performance in almost all disease groups. Furthermore, we found that the performance is rather robust against removing any individual layer, regardless of which one, and only drops considerably when removing several layers. We can therefore assume that the predictive capability of multiplex network is a collective power of all layers that are relevant to the disease and not being primarily driven by specific layers. In this sense, the layers that consistently provide useful information are those that show significant network modularity for most diseases, i.e., the PPI, the phenotypic similarities (HP, MP), and GOBP. All layers considered in this study were relevant to two or more diseases. Note, however, that also the node coverage should be considered when assessing the contribution of a particular layer. For instance, while the phenotypic layers are highly significant across many diseases, they only contain less than 5,000

genes (3,342 and 4,365 nodes in the HPO and MPO layers, respectively). This means that on their own, they cannot offer any information on the remaining 75% of the genome, again advocating for a view that assesses the network layers collectively, rather than on an individual basis. We show in our results that the pairwise overlap between the different layers is generally low. This indicates that each layer provides different pieces of information and that the overall redundancy between the layers is low. Furthermore, our randomization analysis revealed that the overlap is larger than expected by chance, in particular among the co-expression layers. This indicates that certain biological mechanisms are represented across different networks, which may contribute to the robustness of the multiplex network towards the removal of individual network layers.

Lastly, we aimed to directly assess potential confirmatory biases for the specific application of patient gene prioritization by using a patient subset with minimal possible entanglement between test data and data used to generate the networks. To achieve this, a list of disease genes that were not yet known at the time of network data collection were extracted. This minimizes the likelihood of information related to disease association being incorporated in the network construction process. As shown in our results for intellectual disability (ID) gene prioritization, we collected a reference list of genes associated to the disease before March 2019 (the time of network construction). Of 131 patients with confirmed causal genes in the cohort, 21 patients had causal genes outside of the reference list. These patients can therefore be regarded as patients with novel causal genes that were unknown to the scientific community at the time, and such disease-gene association had not yet been embedded in the networks. We found that the overall performance of predicting causal genes (measured by AUROC) based on this subset of patients remained high throughout various methods applied. In addition, the proportion of true causal genes ranked in top five for this patient subset is at a comparable level to that of all patients. Taken together, these complementary analyses suggest that the impact of confirmatory biases in the prediction performance is very limited.

4.4 Future Prospects

As techniques in molecular biology have become more affordable and scalable, the volume of biological data has experienced rapid growth. Network theory has become one among various computational and mathematical methods that have enabled management, analysis, interpretation and visualization of such large datasets. Despite impressive milestones, the increasing volume and the multidimensional nature of biological data has posed additional challenges that urgently need to be addressed.

Early conceptualization of graph-based representation of knowledge was first introduced around 2000s with the introduction of frameworks such as the Semantic Web, Resource Description

Framework (RDF)¹ and Web Ontology Language (OWL)². The goals are to disambiguate Internet data by representing them in structured format. A way to do so is to store the knowledge as entities along with their relationships, creating a web of knowledge that are interconnected - the concept that was later popularized by Google as the ‘Knowledge Graph’ (Singhal, 2012). The Knowledge Graph was used to show further results that are *relevant* to the original queries in a panel alongside the web queries. Graph databases were implemented to optimize data storage and retrieval, and various graph-based query languages were developed (Angles et al., 2017). Graph-native data storage has enabled tasks such as traversal or multi-hop queries to be completed with simpler syntaxes and faster execution time compared to relational databases. In the context of biological data integration where entities such as genes, proteins, variants, drugs and phenotypes are linked, employing the concept of knowledge graph can be intuitive for representing the current knowledge base as well as facilitating discovery of novel relationships. Recent studies have developed biomedical knowledge graphs (Santos et al., 2020), incentivized the community to employ public infrastructure of WikiData (Waagmeester et al., 2020), or incorporated modern graph databases that allow web-based user interface queries and visualization (Himmelstein et al., 2017). Major biological databases have adopted such architectures including Reactome (Fabregat et al., 2018), the Monarch Initiative (Shefchek et al., 2020) and Open Targets (Kafkas et al., 2017). As the interest of knowledge graphs as both data warehouses and powerful tools for various downstream applications continues to grow across all domains, unified and efficient frameworks on biological data can hopefully be implemented and community-powered graph-based analytics will become a household tool in computational biology tasks.

The development of methods in modern network science and machine learning were parallel yet often not interoperable. This is because the former operates on network-based data while the latter often on Euclidean space. Recent development of graph embedding algorithms has closed this gap by enabling the transformation of characteristics of nodes in a network into coordinates in a Euclidean space, often as dense and continuous vectors (Chen et al., 2020; Goyal & Ferrara, 2018). This allows fast growing machine learning algorithms to be embraced on network data, and the novel disciplines of graph and geometric machine learning were developed. As discussed above, current implementations of graph embedding algorithms which preserve only certain node characteristics may not be fully representative of the underlying network structure. Furthermore, their sensitivity to parameterization can affect embedded results and consequently the downstream machine learning tasks (Gu et al., 2021). The growing interests in the community, however, will lead to the maturation of the field as well as the optimal frameworks for their respective applications. Extension of the approaches to more advanced graph architectures such as the embedding of knowledge graph consisting of multiple entities has also been explored (Ren* et al., 2019; Chang et al., 2020). As analyzing and interpreting big data

¹<https://www.w3.org/TR/rdf11-concepts/>

²<https://www.w3.org/TR/owl2-overview/>

relies increasingly on more efficient and more sophisticated machine learning algorithms, their newfound interoperability with network data can also potentially fuel the future developments of big network analytics.

Beyond the context of data representation and integration, network-based methodologies will also contribute to more precise identification of rare disease causality at variant resolution. Current standards of variant interpretation normally involve estimating the deleteriousness of a variant to determine the likelihood of pathogenicity. However, this approach may have missed non-deleterious pathogenic variants that disrupt PPI interfaces. Three-dimensional protein structures can further provide more insightful information whether the mutation is on functional domain or interaction interface of a protein. Indeed, experimentally validated three-dimensional structure of proteins is currently limited, let alone their interaction information that are often acquired via protein co-crystallization. However, recent databases (Porrás Millán et al., 2017; Meyer et al., 2018) have computationally predicted variants affecting the PPI. More recent advances in deep learning models that are able to predict protein structures close to experimentally validated resolution (Jumper et al., 2021; Baek et al., 2021) will also play major roles in fulfilling the missing PPI information as well as network biology at variant level. Such knowledge has already been applied in the context of ‘edgotype’, an idea that has been built around network biology concepts. It aims to describe how certain phenotypes are exhibited when a mutation disrupts PPI interface, enabling interpretability of disease heterogeneity in genetic diseases and cancers (Yi et al., 2017; Sahni et al., 2015). With the more complete PPI information powered by recent computational methods, network medicine is poised to play more major roles from systems analyses to clinical diagnostics on the patient levels.

Bibliography

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.
- Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, Chapter 7, Unit7.20.
- Aguirre-Plans, J., Piñero, J., Sanz, F., Furlong, L. I., Fernandez-Fuentes, N., Oliva, B., & Guney, E. (2019). GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets. *J. Mol. Biol.*, 431(13), 2477–2484.
- Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the World-Wide web. *Nature*, 401(6749), 130–131.
- Aleta, A., & Moreno, Y. (2019). Multilayer networks in a nutshell. *Annu. Rev. Condens. Matter Phys.*, 10(1), 45–62.
- Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoč, D. (2017). Foundations of modern query languages for graph databases. *ACM Comput. Surv.*, 50(5), 1–40.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1), 25–29.
- Auyang, S. Y. (1998). *Foundations of Complex-system Theories: In Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., & Baker, D.

- (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876.
- Barabási, A.-L. (2009). Scale-free networks: A Decade and beyond. <https://barabasi.com/f/303.pdf>. Accessed: 2021-12-6.
- Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, 12(1), 56–68.
- Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.*, 5(2), 101–113.
- Basha, O., Argov, C. M., Artzy, R., Zoabi, Y., Hekselman, I., Alfandari, L., Chalifa-Caspi, V., & Yeger-Lotem, E. (2020). Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes. *Bioinformatics*, 36(9), 2821–2828.
- Beadle, G. W., & Tatum, E. L. (1941). Genetic control of biochemical reactions in neurospora. *Proc. Natl. Acad. Sci. U. S. A.*, 27(11), 499–506.
- Birgmeier, J., Haeussler, M., Deisseroth, C. A., Steinberg, E. H., Jagadeesh, K. A., Ratner, A. J., Gudur, H., Wenger, A. M., Diekhans, M. E., Stenson, P. D., Cooper, D. N., Ré, C., Beggs, A. H., Bernstein, J. A., & Bejerano, G. (2020). AMELIE speeds mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.*, 12(544).
- Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, 33 Suppl, 228–237.
- Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7), 1177–1186.
- Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W. P., Links, A. E., Washington, N. L., Haendel, M. A., Robinson, P. N., Boerkoel, C. F., Adams, D., Gahl, W. A., Boycott, K. M., & Brudno, M. (2015). PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum. Mutat.*, 36(10), 931–940.
- Caldera, M., Müller, F., Kaltenbrunner, I., Licciardello, M. P., Lardeau, C.-H., Kubicek, S., & Menche, J. (2019). Mapping the perturbome network of cellular perturbations. *Nat. Commun.*, 10(1), 5140.
- Cava, C., Bertoli, G., Colaprico, A., Olsen, C., Bontempi, G., & Castiglioni, I. (2018). Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis. *BMC Genomics*, 19(1), 25.
- Cerrone, M., Remme, C. A., Tadros, R., Bezzina, C. R., & Delmar, M. (2019). Beyond the one Gene-One disease paradigm: Complex genetics and pleiotropy in inheritable cardiac disorders. *Circulation*, 140(7), 595–610.

- Chang, D., Balazevic, I., Allen, C., Chawla, D., Brandt, C., & Taylor, R. A. (2020). Benchmark and best practices for biomedical knowledge graph embeddings.
- Chen, F., Wang, Y.-C., Wang, B., & Kuo, C.-C. J. (2020). Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9.
- Cho, H., Berger, B., & Peng, J. (2016). Compact integration of Multi-Network topology for functional analysis of genes. *Cell Syst*, 3(6), 540–548.e5.
- Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., Natoli, T., Narayan, R., DREAM Module Identification Challenge Consortium, Subramanian, A., Zhang, J. D., Stolovitzky, G., Kutalik, Z., Lage, K., Slonim, D. K., Saez-Rodriguez, J., Cowen, L. J., Bergmann, S., & Marbach, D. (2019). Assessment of network module identification across complex diseases. *Nat. Methods*, 16(9), 843–852.
- Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., & Andrews, B. (2019). Global genetic networks and the Genotype-to-Phenotype relationship. *Cell*, 177(1), 85–100.
- Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., San Luis, B.-J., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A.-C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., & Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306).
- Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, 12, 138–163.
- De Domenico, M., Granell, C., Porter, M. A., & Arenas, A. (2016). The physics of spreading processes in multilayer networks. *Nat. Phys.*, 12(10), 901–906.
- de Ligt, J., Willemssen, M. H., van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B. B. A., Brunner, H. G., Veltman, J. A., & Vissers, L. E. L. M. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.*, 367(20), 1921–1929.
- Dickinson, J. A. (2016). Lesser-spotted zebras: Their care and feeding. *Can. Fam. Physician*, 62(8), 620–621.
- Drew, K., Wallingford, J. B., & Marcotte, E. M. (2021). hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.*, 17(5), e10016.
- eGTEx Project (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.*
- Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., Wu, G.,

- Stein, L., D'Eustachio, P., & Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. *PLoS Comput. Biol.*, *14*(1), e1005968.
- Feiglin, A., Allen, B. K., Kohane, I. S., & Kong, S. W. (2017). Comprehensive analysis of tissue-wide gene expression and phenotype data reveals tissues affected in rare genetic disorders. *Cell Syst*, *5*(2), 140–148.e2.
- Ford, C. E., & Hamerton, J. L. (1956). The chromosomes of man. *Nature*, *178*(4541), 1020–1023.
- Fortelny, N., Overall, C. M., Pavlidis, P., & Freue, G. V. C. (2017). Can we predict protein from mRNA levels? *Nature*, *547*(7664), E19–E20.
- Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., Bonner, D., Kernohan, K. D., Marwaha, S., Zappala, Z., Balliu, B., Davis, J. R., Liu, B., Prybol, C. J., Kohler, J. N., Zastrow, D. B., Reuter, C. M., Fisk, D. G., Grove, M. E., Davidson, J. M., Hartley, T., Joshi, R., Strober, B. J., Utiramerur, S., Undiagnosed Diseases Network, Care4Rare Canada Consortium, Lind, L., Ingelsson, E., Battle, A., Bejerano, G., Bernstein, J. A., Ashley, E. A., Boycott, K. M., Merker, J. D., Wheeler, M. T., & Montgomery, S. B. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.*, *25*(6), 911–919.
- Gallagher, R., Appenzeller, T., Normile, D., & Service, Robert F (1999). Beyond reductionism.
- Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(21), 8685–8690.
- Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94.
- Graessner, H., Zurek, B., Hoischen, A., & Beltran, S. (2021). Solving the unsolved rare diseases in europe. *Eur. J. Hum. Genet.*, *29*(9), 1319–1320.
- Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *KDD*, *2016*, 855–864.
- GTEx Consortium (2015). Human genomics. the Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.
- Gu, W., Tandon, A., Ahn, Y.-Y., & Radicchi, F. (2021). Principled approach to the selection of the embedding dimension of networks. *Nat. Commun.*, *12*(1), 3772.
- Guney, E., Menche, J., Vidal, M., & Barabási, A.-L. (2016). Network-based in silico drug efficacy screening. *Nat. Commun.*, *7*, 10331.
- Haendel, M. A., Vasilevsky, N., Brush, M., Hochheiser, H. S., Jacobsen, J., Oellrich, A., Mungall, C. J., Washington, N., Köhler, S., Lewis, S. E., Robinson, P. N., & Smedley, D. (2015a). Disease insights through cross-species phenotype comparisons. *Mamm. Genome*, *26*(9-10), 548–555.
- Haendel, M. A., Vasilevsky, N., Brush, M., Hochheiser, H. S., Jacobsen, J., Oellrich, A., Mungall, C. J., Washington, N., Köhler, S., Lewis, S. E., Robinson, P. N., & Smedley, D. (2015b). Disease insights through cross-species phenotype comparisons. *Mamm. Genome*, *26*(9-10), 548–555.

- Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., & Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.*, *35*(5), 463–474.
- Hartley, T., Lemire, G., Kernohan, K. D., Howley, H. E., Adams, D. R., & Boycott, K. M. (2020). New diagnostic approaches for undiagnosed rare genetic diseases. *Annu. Rev. Genomics Hum. Genet.*, *21*, 351–372.
- Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, *6*.
- Hu, J. X., Thomas, C. E., & Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, *17*(10), 615–629.
- Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., & Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst*, *6*(4), 484–495.e5.
- Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., Zhang, T., Rad, R., Pan, J., Nusinow, D. P., Paulo, J. A., Schweppe, D. K., Vaites, L. P., Harper, J. W., & Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, *184*(11), 3022–3040.e28.
- Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaite, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., Obar, R. A., Guruharsha, K. G., Li, K., Artavanis-Tsakonas, S., Gygi, S. P., & Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*.
- Jacobs, P. A., Baikie, A. G., Court Brown, W. M., & Strong, J. A. (1959). The somatic chromosomes in mongolism. *Lancet*, *1*(7075), 710.
- Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., & Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.*, *5*, 4022.
- Julkowska, D., Austin, C. P., Cuttillo, C. M., Gancberg, D., Hager, C., Halftermeyer, J., Jonker, A. H., Lau, L. P. L., Norstedt, I., Rath, A., Schuster, R., Simelyte, E., & van Weely, S. (2017). The importance of international collaboration for rare diseases research: a european perspective. *Gene Ther.*, *24*(9), 562–571.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohli, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

- Kafkas, Ş., Dunham, I., & McEntyre, J. (2017). Literature evidence in open targets - a target validation platform. *J. Biomed. Semantics*, 8(1), 20.
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Genome Aggregation Database Consortium, Neale, B. M., Daly, M. J., & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443.
- Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A. B., Silverstein, M. C., Lachmann, A., Kuleshov, M. V., Ma'ayan, A., Stathias, V., Terryn, R., Cooper, D., Forlin, M., Koleti, A., Vidovic, D., Chung, C., Schürer, S. C., Vasilias, J., Pilarczyk, M., Shamsaei, B., Fazel, M., Ren, Y., Niu, W., Clark, N. A., White, S., Mahi, N., Zhang, L., Kouril, M., Reichard, J. F., Sivaganesan, S., Medvedovic, M., Meller, J., Koch, R. J., Birtwistle, M. R., Iyengar, R., Sobie, E. A., Azeloglu, E. U., Kaye, J., Osterloh, J., Haston, K., Kalra, J., Finkbiener, S., Li, J., Milani, P., Adam, M., Escalante-Chong, R., Sachs, K., Lenail, A., Ramamoorthy, D., Fraenkel, E., Daigle, G., Hussain, U., Coye, A., Rothstein, J., Sareen, D., Ornelas, L., Banuelos, M., Mandefro, B., Ho, R., Svendsen, C. N., Lim, R. G., Stocksdales, J., Casale, M. S., Thompson, T. G., Wu, J., Thompson, L. M., Dardov, V., Venkatraman, V., Matlock, A., Van Eyk, J. E., Jaffe, J. D., Papanastasiou, M., Subramanian, A., Golub, T. R., Erickson, S. D., Fallahi-Sichani, M., Hafner, M., Gray, N. S., Lin, J.-R., Mills, C. E., Muhlich, J. L., Niepel, M., Shamu, C. E., Williams, E. H., Wrobel, D., Sorger, P. K., Heiser, L. M., Gray, J. W., Korkola, J. E., Mills, G. B., LaBarge, M., Feiler, H. S., Dane, M. A., Bucher, E., Nederlof, M., Sudar, D., Gross, S., Kilburn, D. F., Smith, R., Devlin, K., Margolis, R., Derr, L., Lee, A., & Pillai, A. (2017). The library of integrated Network-Based cellular signatures NIH program: System-Level cataloging of human cells response to perturbations. *Cell Syst.*
- Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., Parkinson, H., & Schriml, L. M. (2015). Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, 43(Database issue), D1071–8.
- Kim, E., Dede, M., Lenoir, W. F., Wang, G., Srinivasan, S., Colic, M., & Hart, T. (2019). A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Sci Alliance*, 2(2).
- Kitsak, M., Sharma, A., Menche, J., Guney, E., Ghiassian, S. D., Loscalzo, J., & Barabási, A.-L. (2016). Tissue specificity of human disease module. *Sci. Rep.*, 6, 35241.
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2013). Multilayer networks.
- Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the interactome for priori-

- tization of candidate disease genes. *Am. J. Hum. Genet.*, 82(4), 949–958.
- Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C. M., Brown, D. L., Brudno, M., Campbell, J., FitzPatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S.-M., Riggs, E. R., Scott, R. H., Sisodiya, S., Van Vooren, S., Wapner, R. J., Wilkie, A. O. M., Wright, C. F., Vulto-van Silfhout, A. T., de Leeuw, N., de Vries, B. B. A., Washington, N. L., Smith, C. L., Westerfield, M., Schofield, P., Ruef, B. J., Gkoutos, G. V., Haendel, M., Smedley, D., Lewis, S. E., & Robinson, P. N. (2014). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, 42(Database issue), D966–74.
- Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurphy, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., Brudno, M., Buske, O. J., Chinnery, P. F., Cipriani, V., Connell, L. E., Dawkins, H. J. S., DeMare, L. E., Devereau, A. D., de Vries, B. B. A., Firth, H. V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jähn, J. A., James, R., Krause, R., F Laulederkind, S. J., Lochmüller, H., Lyon, G. J., Ogishima, S., Olry, A., Ouwehand, W. H., Pontikos, N., Rath, A., Schaefer, F., Scott, R. H., Segal, M., Sergouniotis, P. I., Sever, R., Smith, C. L., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, M. W. M., Vulliamy, T., Yu, J., von Ziegenweidt, J., Zankl, A., Züchner, S., Zemojtel, T., Jacobsen, J. O. B., Groza, T., Smedley, D., Mungall, C. J., Haendel, M., & Robinson, P. N. (2017). The human phenotype ontology in 2017. *Nucleic Acids Res.*, 45(D1), D865–D876.
- Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Usaj, M. M., van Leeuwen, J., Koch, E. N., Pons, C., Dagilis, A. J., Pryszyk, M., Wang, J. Z. Y., Hanchard, J., Riggi, M., Xu, K., Heydari, H., Luis, B.-J. S., Shuteriqi, E., Zhu, H., Van Dyk, N., Sharifpoor, S., Costanzo, M., Loewith, R., Caudy, A., Bolnick, D., Brown, G. W., Andrews, B. J., Boone, C., & Myers, C. L. (2018). Systematic analysis of complex genetic interactions. *Science*, 360(6386), eaao1729.
- Lalonde, E., Rentas, S., Lin, F., Dulik, M. C., Skraban, C. M., & Spinner, N. B. (2020). Genomic diagnosis for pediatric disorders: Revolution and evolution. *Front Pediatr*, 8, 373.
- Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., Ding, L., & Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, 47(2), 106–114.
- Li, J., Zhang, W., Yang, H., Howrigan, D. P., Wilkinson, B., Souaiaia, T., Evgrafov, O. V., Genovese, G., Clementel, V. A., Tudor, J. C., Abel, T., Knowles, J. A., Neale, B. M., Wang, K., Sun, F., & Coba, M. P. (2017). Spatiotemporal profile of postsynaptic interactomes integrates components of complex brain disorders. *Nat. Neurosci.*, 20(8), 1150–1161.
- Liu, X., Maiorino, E., Halu, A., Glass, K., Prasad, R. B., Loscalzo, J., Gao, J., & Sharma, A. (2020). Robustness and lethality in multilayer biological molecular networks. *Nat. Commun.*, 11(1), 6043.
- Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charlotiaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S.,

- Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., Knapp, J. J., Kovács, I. A., Lemmens, I., Mee, M. W., Mellor, J. C., Pollis, C., Pons, C., Richardson, A. D., Schlabach, S., Teeking, B., Yadav, A., Babor, M., Balcha, D., Basha, O., Bowman-Colin, C., Chin, S.-F., Choi, S. G., Colabella, C., Coppin, G., D'Amata, C., De Ridder, D., De Rouck, S., Duran-Frigola, M., Ennajdaoui, H., Goebels, F., Goehring, L., Gopal, A., Haddad, G., Hatchi, E., Helmy, M., Jacob, Y., Kassa, Y., Landini, S., Li, R., van Lieshout, N., MacWilliams, A., Markey, D., Paulson, J. N., Rangarajan, S., Rasla, J., Rayhan, A., Rolland, T., San-Miguel, A., Shen, Y., Sheykhkarimli, D., Sheynkman, G. M., Simonovsky, E., Taşan, M., Tejeda, A., Tropepe, V., Twizere, J.-C., Wang, Y., Weatheritt, R. J., Weile, J., Xia, Y., Yang, X., Yeger-Lotem, E., Zhong, Q., Aloy, P., Bader, G. D., De Las Rivas, J., Gaudet, S., Hao, T., Rak, J., Tavernier, J., Hill, D. E., Vidal, M., Roth, F. P., & Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, (pp. 1–7).
- Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*.
- Mani, R., St Onge, R. P., Hartman, J. L., 4th, Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proc. Natl. Acad. Sci. U. S. A.*, 105(9), 3461–3466.
- Matalonga, L., Hernandez-Ferrer, C., Piscia, D., Schüle, R., Synofzik, M., Töpf, A., Vissers, L. E. L., de Voer, R., Tonda, R., Laurie, S., Fernandez-Callejo, M., Picó, D., Garcia-Linares, C., Papakonstantinou, A., Corvò, A., Joshi, R., Diez, H., Gut, I., Hoischen, A., Graessner, H., & Beltran, S. (2021). Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur. J. Hum. Genet.*, (pp. 1–9).
- McCarthy, M., & Birney, E. (2021). Personalized profiles for disease risk must capture all facets of health. *Nature*, 597(7875), 175–177.
- Meehan, T. F., Conte, N., West, D. B., Jacobsen, J. O., Mason, J., Warren, J., Chen, C.-K., Tudose, I., Relac, M., Matthews, P., Karp, N., Santos, L., Fiegel, T., Ring, N., Westerberg, H., Greenaway, S., Sneddon, D., Morgan, H., Codner, G. F., Stewart, M. E., Brown, J., Horner, N., International Mouse Phenotyping Consortium, Haendel, M., Washington, N., Mungall, C. J., Reynolds, C. L., Gallegos, J., Gailus-Durner, V., Sorg, T., Pavlovic, G., Bower, L. R., Moore, M., Morse, I., Gao, X., Tocchini-Valentini, G. P., Obata, Y., Cho, S. Y., Seong, J. K., Seavitt, J., Beaudet, A. L., Dickinson, M. E., Herault, Y., Wurst, W., de Angelis, M. H., Lloyd, K. C. K., Flenniken, A. M., Nutter, L. M. J., Newbigging, S., McKerlie, C., Justice, M. J., Murray, S. A., Svenson, K. L., Braun, R. E., White, J. K., Bradley, A., Flicek, P., Wells, S., Skarnes, W. C., Adams, D. J., Parkinson, H., Mallon, A.-M., Brown, S. D. M., & Smedley, D. (2017). Disease model discovery from 3,328 gene knockouts by the international mouse phenotyping consortium. *Nat. Genet.*, 49(8), 1231–1238.
- Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 1257601.
- Meyer, M. J., Beltrán, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., & Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods*.
- Mungall, C. J., McMurtry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J. P., Jacobsen, J. O. B., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan,

- Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., & Haendel, M. A. (2017). The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, *45*(D1), D712–D722.
- Nathans, D., & Smith, H. O. (1975). Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu. Rev. Biochem.*, *44*, 273–293.
- Newman, M. E. J. (2011). Complex systems: A survey.
- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, *31*(13), 3812–3814.
- Nicosia, V., & Latora, V. (2015). Measuring and modeling correlations in multiplex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *92*(3), 032805.
- Paulson, J. N., Chen, C.-Y., Lopes-Ramos, C. M., Kuijjer, M. L., Platig, J., Sonawane, A. R., Fagny, M., Glass, K., & Quackenbush, J. (2017). Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics*, *18*(1), 437.
- Perry, S. (2017). Mouse phenotyping sheds light on rare disease. *Nat. Biotechnol.*, *35*(9), 831.
- Pesquita, C. (2017). Semantic similarity in the gene ontology. In C. Dessimoz, & N. Škunca (Eds.) *The Gene Ontology Handbook*, (pp. 161–173). New York, NY: Springer New York.
- Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., & Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, *9 Suppl 5*, S4.
- Pierson, E., GTEx Consortium, Koller, D., Battle, A., Mostafavi, S., Ardlie, K. G., Getz, G., Wright, F. A., Kellis, M., Volpi, S., & Dermitzakis, E. T. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.*, *11*(5), e1004220.
- Porras Millán, P., Duesbury, M., Koch, M., & Orchard, S. (2017). The MINTAct archive for mutations influencing molecular interactions. *Genomics Comput Biol*, (p. 100053).
- Ren*, H., Hu*, W., & Leskovec, J. (2019). Query2box: Reasoning over knowledge graphs in vector space using box embeddings.
- Rentsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, *47*(D1), D886–D894.
- Rieckmann, J. C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S. S., Lanzavecchia, A., Mann, M., & Meissner, F. (2017). Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.*
- Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., & Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, *24*(2), 340–348.
- Rode, J. (2005). Rare diseases: understanding this public health priority. https://stars.eurordis.org/sites/default/files/publications/princeps_document-EN.pdf. Accessed: 2018-2-20.

- Rolland, T., Taşan, M., Charlotiaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A.-L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., & Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5), 1212–1226.
- Ruiz, C., Zitnik, M., & Leskovec, J. (2021). Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.*, 12(1), 1796.
- Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., GTEx Consortium, Engelhardt, B. E., & Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.*, 27(11), 1843–1858.
- Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y.-Y., Yachie, N., Zhong, Q., Shen, Y., Palagi, A., San-Miguel, A., Fan, C., Balcha, D., Dricot, A., Jordan, D. M., Walsh, J. M., Shah, A. A., Yang, X., Stoyanova, A. K., Leighton, A., Calderwood, M. A., Jacob, Y., Cusick, M. E., Salehi-Ashtiani, K., Whitesell, L. J., Sunyaev, S., Berger, B., Barabási, A.-L., Charlotiaux, B., Hill, D. E., Hao, T., Roth, F. P., Xia, Y., Walhout, A. J. M., Lindquist, S., & Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3), 647–660.
- Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Geyer, P. E., Coscia, F., Wewer Albrechtsen, N. J., Mundt, F., Jensen, L. J., & Mann, M. (2020). Clinical knowledge graph integrates proteomics data into clinical Decision-Making.
- Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, 11(4), 361–362.
- Seyfried, N. T., Dammer, E. B., Swarup, V., Nandakumar, D., Duong, D. M., Yin, L., Deng, Q., Nguyen, T., Hales, C. M., Wingo, T., Glass, J., Gearing, M., Thambisetty, M., Troncoso, J. C., Geschwind, D. H., Lah, J. J., & Levey, A. I. (2017). A multi-network approach identifies Protein-Specific co-expression in asymptomatic and symptomatic alzheimer's disease. *Cell Syst*, 4(1), 60–72.e4.
- Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., Sahni, N., Thibault, D., Voung, L., Guo, F., Ghiassian, S. D., Gulbahce, N., Baribaud, F., Tocker, J., Dobrin, R., Barnathan, E., Liu, H., Panettieri, R. A., Jr, Tantisira, K. G., Qiu, W., Raby, B. A., Silverman, E. K., Vidal, M., Weiss, S. T., & Barabási, A.-L. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.*, 24(11), 3005–3020.
- Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglou, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X. A., Balhoff, J. P., Babb, L., Bello, S. M., Blau,

- H., Bradford, Y., Carbon, S., Carmody, L., Chan, L. E., Cipriani, V., Cuzick, A., Rocca, M. D., Dunn, N., Essaid, S., Fey, P., Grove, C., Gouridine, J.-P., Hamosh, A., Harris, M., Helbig, I., Hoatlin, M., Joachimiak, M., Jupp, S., Lett, K. B., Lewis, S. E., McNamara, C., Pendlington, Z. M., Pilgrim, C., Putman, T., Ravanmehr, V., Reese, J., Riggs, E., Robb, S., Roncaglia, P., Seager, J., Segerdell, E., Similuk, M., Storm, A. L., Thaxon, C., Thessen, A., Jacobsen, J. O. B., McMurry, J. A., Groza, T., Köhler, S., Smedley, D., Robinson, P. N., Mungall, C. J., Haendel, M. A., Munoz-Torres, M. C., & Osumi-Sutherland, D. (2020). The monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, 48(D1), D704–D715.
- Sin, C., & Menche, J. (2021). The network of networks involved in human disease. In *Networks of Networks in Biology: Concepts, Tools and Applications*, (pp. 147–171). Cambridge University Press.
- Singhal, A. (2012). Introducing the knowledge graph: things, not strings. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>. Accessed: 2021-12-3.
- Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., & Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat. Protoc.*, 10(12), 2004–2015.
- Smedley, D., & Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human mendelian disease genes. *Genome Med.*, 7(1), 81.
- Smith, C. L., & Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 1(3), 390–399.
- Stam, C. J. (2014). Modern network science of neurological disorders. *Nat. Rev. Neurosci.*, 15(10), 683–695.
- Strauss, M. J., Niederkrotenthaler, T., Thurner, S., Kautzky-Willer, A., & Klimek, P. (2021). Data-driven identification of complex disease phenotypes. *J. R. Soc. Interface*, 18(180), 20201040.
- Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Johnson, S. A., Lyons, N. J., Berger, A. H., Shamji, A. F., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D. Y., Hu, R., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Greenside, P., Gray, N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W.-N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., & Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6), 1437–1452.e17.
- Tamayo, T., Rosenbauer, J., Wild, S. H., Spijkerman, A. M. W., Baan, C., Forouhi, N. G., Herder, C., & Rathmann, W. (2014). Diabetes in europe: an update. *Diabetes Res. Clin. Pract.*, 103(2), 206–217.

- Thompson, N. A., Ranzani, M., van der Weyden, L., Iyer, V., Offord, V., Droop, A., Behan, F., Gonçalves, E., Speak, A., Iorio, F., Hewinson, J., Harle, V., Robertson, H., Anderson, E., Fu, B., Yang, F., Zagnoli-Vieira, G., Chapman, P., Del Castillo Velasco-Herrera, M., Garnett, M. J., Jackson, S. P., & Adams, D. J. (2021). Combinatorial CRISPR screen identifies fitness effects of gene paralogues. *Nat. Commun.*, *12*(1), 1302.
- Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Bérout, C., Gut, I. G., Hansson, M. G., 't Hoen, P.-B. A., Patrinos, G. P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J. E., Posada, M., Robinson, P. N., Bushby, K., & Lochmüller, H. (2014). RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J. Gen. Intern. Med.*, *29 Suppl 3*, S780–7.
- Turner, S., Klimek, P., & Hanel, R. (2018). Introduction to the theory of complex systems. In *Introduction to the Theory of Complex Systems*. Oxford University Press, 1 ed.
- Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., & Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, *35*(3), 497–505.
- van Leeuwen, J., Boone, C., & Andrews, B. J. (2017). Mapping a diversity of genetic interactions in yeast. *Current Opinion in Systems Biology*, *6*(Supplement C), 14–21.
- van Leeuwen, J., Pons, C., Mellor, J. C., Yamaguchi, T. N., Friesen, H., Koschwanez, J., Ušaj, M. M., Pechlaner, M., Takar, M., Ušaj, M., VanderSluis, B., Andrusiak, K., Bansal, P., Baryshnikova, A., Boone, C. E., Cao, J., Cote, A., Gebbia, M., Horecka, G., Horecka, I., Kuzmin, E., Legro, N., Liang, W., van Lieshout, N., McNee, M., San Luis, B.-J., Shaeri, F., Shuteriqi, E., Sun, S., Yang, L., Youn, J.-Y., Yuen, M., Costanzo, M., Gingras, A.-C., Aloy, P., Oostenbrink, C., Murray, A., Graham, T. R., Myers, C. L., Andrews, B. J., Roth, F. P., & Boone, C. (2016). Exploring genetic suppression interactions on a global scale. *Science*, *354*(6312).
- van Riel, R., & Van Gulick, R. (2019). *Scientific Reduction*. Metaphysics Research Lab, Stanford University, spring 2019 ed.
- Vespignani, A. (2018). Twenty years of network science. *Nature*, *558*(7711), 528–529.
- Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, *144*(6), 986–998.
- Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.*, *90*(1), 7–24.
- Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., Keating, S. M., Manske, M., Mayers, M., Mitchen, D., Mitraka, E., Pico, A. R., Putman, T., Riutta, A., Queralt-Rosinach, N., Schriml, L. M., Shafee, T., Slenter, D., Stephan, R., Thornton, K., Tsueng, G., Tu, R., Ul-Hasan, S., Willighagen, E., Wu, C., & Su, A. I. (2020). Wikidata as a knowledge graph for the life sciences. *Elife*, *9*.
- Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.*, *270*(1514), 457–466.

- Wang, Y., Sahni, N., & Vidal, M. (2015). Global edgetic rewiring in cancer networks. *Cell Syst*, 1(4), 251–253.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.
- Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F., & Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), 582–587.
- Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L., & Troyanskaya, O. G. (2021). Decoding disease: from genomes to networks to phenotypes. *Nat. Rev. Genet.*, (pp. 1–17).
- Wright, C. F., McRae, J. F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T. W., Jones, P., Prigmore, E., Rajan, D., Lord, J., Sifrim, A., Kelsell, R., Parker, M. J., Barrett, J. C., Hurles, M. E., FitzPatrick, D. R., & Firth, H. V. (2018). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.*
- Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M. R., Leduc, M. S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S. E., Lupski, J. R., Beaudet, A. L., Gibbs, R. A., & Eng, C. M. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, 369(16), 1502–1511.
- Yi, S., Lin, S., Li, Y., Zhao, W., Mills, G. B., & Sahni, N. (2017). Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.*, 18(7), 395–410.
- Yu, G. (2020). Gene ontology semantic similarity analysis using GOSemSim. In B. L. Kidder (Ed.) *Stem Cell Transcriptional Networks: Methods and Protocols*, (pp. 207–215). New York, NY: Springer US.
- Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J., Ng, C. T., Krogan, N., Sharan, R., & Ideker, T. (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Syst*, 2(2), 77–88.
- Žitnik, M., Janjić, V., Larminie, C., Zupan, B., & Pržulj, N. (2013). Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.*, 3, 3202.
- Zurek, B., Ellwanger, K., Vissers, L. E. L. M., Schüle, R., Synofzik, M., Töpf, A., de Voer, R. M., Laurie, S., Matalonga, L., Gilissen, C., Ossowski, S., 't Hoen, P. A. C., Vitobello, A., Schulze-Hentrich, J. M., Riess, O., Brunner, H. G., Brookes, A. J., Rath, A., Bonne, G., Gumus, G., Verloes, A., Hoogerbrugge, N., Evangelista, T., Harmuth, T., Swertz, M., Spalding, D., Hoischen, A., Beltran, S., Graessner, H., & Solve-RD consortium (2021). Solve-RD: systematic pan-european data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet.*

Appendix

Reporting summary for the main publication

Statistics

[Editorial Policy Checklist.](#)

<input type="checkbox"/>	<input type="checkbox"/>	group/condition,
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	Only common tests should be described solely by name; describe more complex techniques in the Methods section.
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	Give P values as exact values whenever suitable.
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Policy information about [availability of computer code](#)

blob/main/report/report_session.md.

<https://github.com/menchelab/MultiOme/>

<https://github.com/menchelab/MultiOme>

www.menchelab.com/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data generated in this study are provided in the Supplementary Information/Source Data file. The RDconnect Genome-Phenome Analysis Platform (GPAP) data are available under restricted access, which can be obtained by validated users via the platform at <https://platform.rd-connect.eu/>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For statistical analyses of network and disease characterization, sample sizes are provided in figure legends and in the Methods. For patient data, we include all patients with known causal genes for both cohorts in consideration.
Data exclusions	No data were excluded in the study.
Replication	Study does not involve experimental results that require replication. Since the study uses biological samples of patients with unique clinical phenotype and genetic profile, biological replication is not possible and not relevant.
Randomization	Randomization is not relevant to the study as there was no group allocation involved.
Blinding	Blinding is not relevant to the study as there was no group allocation involved. Genome analyses are ascertained to be blind.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

PISANU BUPHAMALAI

🌐 izepb.github.io

🐦 [izebuphamalai](#)

✉ pisanu.buphamalai@univie.ac.at

CONTACT Campus Vienna Biocenter, VBC5, Dr.Bohr-Gasse 9 1.104, A-1030 Vienna, AUSTRIA

EDUCATION **Medizinische Universität Wien**, Vienna, AT **2016 – Present**
PhD Study - Medical Informatics, Biostatistics & Complex Systems

KTH Royal Institute of Technology, Stockholm, SE & **2014 – 2016**
Aalto University, Helsinki, FI
Erasmus Mundus Master's Degree in Computational Systems Biology (euSYSBIO)

Mahidol University, Bangkok, TH **2009 – 2013**
B.Sc. in Physics, First Class Honours, Distinction

RESEARCH EXPERIENCES **CeMM, Austrian Academy of Sciences**, Vienna, AT **Sep 2016 – Present**
Max Perutz Labs, University of Vienna, Vienna, AT **Oct 2020 – Present**
Predoctoral Fellow
Project: Network-based approaches to rare diseases
Supervisor: Univ-Prof. Dr. Jörg Menche

Joint Research Center for Computational Biomedicine, **Feb – Jul 2016**
RWTH-Aachen University and Aachen University Hospital, Aachen, DE
Masters research student
Project: Statistical analysis of membrane transporters towards drug sensitivity and resistance
Supervisors: Prof. Julio Saez-Rodriguez and Dr. Marc Brehme

Probabilistic Machine Learning Group, **May - Jul 2015**
Department of Computer Science, Aalto University, Helsinki, FI
Summer intern
Project: Group Factor Analysis (GFA) hyperparameter tuning for capturing latent variables from a mixture of continuous and discrete datasets
Supervisors: Prof. Samuel Kaski and Dr. Pekka Marttinen

Computational Regulatory Genomics Group, **Jun – Nov 2012**
Faculty of Medicine, Imperial College London, UK
Undergraduate research student
Project: Genome-wide analysis of chromatin interaction – insights into long-range gene regulation
Supervisors: Dr. Boris Lenhard and Dr. Nathan Harmston

Department of Biochemistry, Faculty of Science, Mahidol University **2011–2012**
Undergraduate research trainee
Project: Statistical analysis of gene regulation in *Burkholderia pseudomallei* using microarray data
Supervisors: Dr. Varodom Charoensawan and Prof. Sumalee Tungpradubkul

HONOURS AND SCHOLARSHIPS **Winning team, CytoData2019**, DKFZ, DE **2019**
Won the competition for drug target prediction based on cellular morphology

The Lindau Nobel Laureate Meetings, Lindau, DE **2019**
Selected to participate the prestigious annual meeting

Outstanding Poster Presentation Award, European Molecular Biology Organization **2018**
Awarded at the EMBO conference on network biology, EMBL, DE Heidelberg

Erasmus Mundus Master's Scholarship, European Commission **2014**
Awarded full scholarship for master's studies

PUBLICATIONS	<p>Buphamalai P, Kokotovic T, Nagy V, and Menche J. Network Analysis Reveals Rare Disease Signatures across Multiple Levels of Biological Organization. <i>Nature Communications</i> 12 (1): 115, (2021).</p> <p>Nagy, V, Hollstein, R, Pai, TP, Herde, MK, Buphamalai P, Moeseneder, P,..., Penninger, J. HACE1 deficiency leads to structural and functional neurodevelopmental defects. <i>Neurology Genetics</i>, 5(3), (2019) e330.</p> <p>Sdelci, S, Rendeiro, AF, ... Buphamalai P ... Kubucek, S. MTHFD1 interaction with BRD4 links folate metabolism to transcriptional regulation. <i>Nature genetics</i>, 51(6), (2019):990-998.</p> <p>Caldera M*, Buphamalai P*, Müller F, Menche J. Interactome-Based Approaches to Human Disease. <i>Current Opinion in Systems Biology</i> 3 (June 2017): 8894.</p> <p>Chutoam P, Charoensawan V, Wongtrakoongate C, Kum-Arth A, Buphamalai P, and Tungpradabkul S. RpoS and oxidative stress conditions regulate succinyl-CoA: 3-ketoacid-coenzyme A transferase (SCOT) expression in <i>Burkholderia pseudomallei</i>. <i>Microbiology and immunology</i>, 57(9):605–615, (2013).</p>	
BOOK CHAPTERS	<p>Buphamalai P*, Caldera M*, Müller F, Menche J. Analyzing Network Data in Biology and Medicine, Chapter 10: Network Medicine. <i>Cambridge University Press</i>, 2019.</p>	
CONFERENCES / PRESENTATIONS	<p>NetSci2020: Conference on network sciences, Rome, IT (virtual) Sep 2020</p> <p>Genomics of rare diseases, Cambridge, UK, (virtual) Mar 2020</p> <p>Cold Spring Harbor Laboratory Meeting on Network Biology, NY, USA Mar 2019</p> <p>German Bioinformatics Conference (GBC), Vienna, AT Sep 2018</p> <p>EMBO Workshop, Heidelberg, DE Apr 2018</p> <p>Integrating Systems Biology: From Networks to Mechanisms to Models</p> <p>ISMB 2017, Prague, CZ Jul 2017</p> <p>The International Conference on Intelligent Systems for Molecular Biology</p> <p>Lake Como Summer School of Advanced Studies, Como, IT May 2017</p> <p>Complex Networks, Theories and Applications</p>	
OTHER COURSES/ QUALIFICATIONS	<p>Coursera</p> <ul style="list-style-type: none"> • <i>Computing for Data Analysis</i> by John Hopkins University (Accomplished with distinction) 	
OTHER WORKING EXPERIENCES	<p>CeMM & Institute Curie Joint Young Scientists Retreat, Bratislava, SK Sep 2018</p> <p><i>Organiser</i></p> <p>Prompt International Institute, Bangkok, TH Jan 2013 – Jun 2014</p> <p><i>Part-time tutor for GCSE, A-Level, and IB diploma in physics and mathematics</i></p> <p>The 42th International Physics Olympiad (42th IPhO), Bangkok, TH Jul 2011</p> <p><i>Liaison officer for the Israeli national candidates</i></p> <p>The Faculty of Science Students Association, Mahidol University, TH 2009 – 2012</p> <p><i>Assistant Vice President of External Affair (2011 – 2012)</i></p>	
FORMAL LANGUAGES	<p>R ● ● ● ● ●</p> <p>Python ● ● ● ● ○</p> <p>Typesettings: L^AT_EX, Markdown</p> <p>MATLAB ● ● ● ○ ○</p> <p>Bash ● ● ● ○ ○</p>	
NATURAL LANGUAGES	<p>Thai ● ● ● ● ● (native)</p> <p>English ● ● ● ● ○ (proficient)</p> <p>German ● ● ● ○ ○ (intermediate)</p> <p>Swedish ● ○ ○ ○ ○ (basic)</p>	
OTHER ACTIVITIES	<p>Hiking, Skiing, Cycling, Climbing, Swimming, Cooking, Stargazing</p>	