AIX-MARSEILLE UNIVERSITY

Ecole Doctorale des Sciences de la Vie et de la Sante

Thesis submitted for the degree of Doctor of Philosophy (Ph.D.)

in Biology of health: Bioinformatics and Genomics

Approaches to explore multiplex biological networks and application to study premature aging diseases

---

Alberto VALDEOLIVAS URBELZ

Defended on the 15th of March 2019 in front of the jury members:

Dr. Anaïs BAUDOT	Supervisor
Prof. Pierre CAU	Co-supervisor
Prof. Laura I. FURLONG	Reporter
Dr. Antonio RAUSELL	Reporter
Dr. Nicolas THIERRY-MIEG	Examinator
Prof. Jacques VAN HELDEN	President

AIX-MARSEILLE UNIVERSITÉ

Ecole Doctorale des Sciences de la Vie et de la Sante

Thèse présentée pour obtenir le grade universitaire de Docteur d'Université

en Biologie de la santé : Bioinformatique et Génomique

Approches pour explorer les réseaux biologiques multiplex et application aux maladies du vieillissement prématuré

Alberto VALDEOLIVAS URBELZ

Soutenue le 15/03/2019 devant le jury composé de :

Dr. Anaïs BAUDOT	Directrice de thèse
Prof. Pierre CAU	Co-directeur de thèse
Prof. Laura I. FURLONG	Rapportrice
Dr. Antonio RAUSELL	Rapporteur
Dr. Nicolas THIERRY-MIEG	Examinateur
Prof. Jacques VAN HELDEN	Président du jury

Abstract

Genes and proteins do not act isolated in cells but rather interact to perform their functions in signaling pathways, molecular complexes, or, more generally, biological processes. These interactions can be represented as large networks in which nodes are genes or proteins and edges represent their interactions. Various graph-theory based approaches have been developed to extract the functional knowledge contained in biological networks. Nevertheless, these methods have been mainly applied to individual networks, ignoring the diversity of biological interactions. We state here that these different types of interactions can be represented as multiplex networks, i.e. collections of networks sharing the same nodes, leading to a more accurate description of biological systems.

This thesis focuses on the extension from individual to multiplex networks of some of the state-of-the-art guilt-by-association methods in computational biology, and on their application to the study of human diseases. Concretely, I expanded the random walk with restart algorithm in order to explore multiplex and heterogeneous networks. We demonstrated that our method is able to take advantage of different interaction sources to predict disease-associated genes. I also participated to the *Disease Module Identification DREAM Challenge*, an international project aiming at testing different clustering algorithms in a bio-medical context, with a multiplex-tailored algorithm developed by our team. Overall, multiplex network find more biologically relevant modules than individual networks.

On the application side, we concentrate on premature aging diseases, also called progeroid syndromes, a group of rare genetic disorders that resemble some aspects of physiological aging at an early age. In this framework, I first applied random walks with restart on multiplex networks to unveil the deregulated biological processes in specific premature aging diseases. Then, we applied an extension of this algorithm to detect the modules associated to more than 70 disorders annotated with at least one premature aging related phenotype. The results revealed the landscape of perturbed molecular processes in premature aging diseases, which can be paralleled with the hallmarks of physiological aging to help identifying common and specific features.

Finally, the last chapter describes the analyses of *proteomics* and *phosphoproteomics* mass spectrometry data derived from prostate cancer cell lines. These cell lines represent different stages of the disease's progression. We identified some molecular events potentially responsible for the progression of prostate cancer to metastatic stages, which are highly aggressive and incurable.

Keywords: biological networks, multiplex networks, random walk with restart algorithm, clustering algorithms, premature aging diseases, proteomics, phos-

phoproteomics, prostate cancer.

Résumé

Les gènes et les protéines n'agissent pas de manière isolée dans les cellules, mais interagissent plutôt pour faire leurs fonctions dans les voies de signalisation, les complexes moléculaires ou, plus généralement, dans les processus biologiques. Ces interactions peuvent être représentées sous forme de grands réseaux dans lesquels les nœuds sont des gènes ou des protéines et les arêtes représentent leurs interactions. Diverses approches basées sur la théorie des graphes ont été développées pour extraire la connaissance fonctionnelle contenue dans les réseaux biologiques. Néanmoins, ces méthodes ont été principalement appliquées à des réseaux individuels, en ignorant la diversité des interactions biologiques. Nous déclarons ici que ces différents types d'interactions peuvent être représentés sous la forme de réseaux multiplexes, c'est-à-dire des ensembles de réseaux partageant les mêmes nœuds, ce qui permet une description plus précise des systèmes biologiques.

Cette thèse est focalisée sur le développement de nouveaux algorithmes étendant aux réseaux multiplexes certaines méthodes populaires de la théorie des graphes en biologie computationnelle, ainsi que sur leur application à l'étude des maladies humaines. Concrètement, j'ai étendu l'algorithme *random walk with restart* aux réseaux multiplex et hétérogènes. Nous avons démontré que notre méthode est capable de prendre avantage des différentes sources d'interaction pour prédire les gènes associés à une maladie. J'ai également participé à *DREAM Challenge*, un projet international visant à tester différents algorithmes de clustering dans un contexte biomédical, avec un algorithme adapté au multiplex développé par notre équipe. Globalement, le fait de considérer plusieurs réseaux améliore la détection des modules.

Du côté des applications, nous nous concentrons sur les maladies liées au vieillissement prématuré, également appelées syndromes progéroïdes, un groupe de maladies génétiques ressemblant à certains aspects du vieillissement physiologique à un âge précoce. Dans ce cadre, j'ai d'abord appliqué des *random walk with restart* sur des réseaux multiplex afin de dévoiler les processus biologiques dérégulés dans maladies particuliers liées au vieillissement prématuré. Ensuite, nous avons appliqué une extension de ces algorithmes pour détecter les modules associés à plus de 70 syndromes annotés avec au moins un phénotype lié au vieillissement prématuré. Les résultats ont révélé le paysage des processus moléculaires perturbés dans les maladies du vieillissement prématuré, qui peuvent être mis en parallèle avec les caractéristiques du vieillissement physiologique.

Enfin, le dernier chapitre décrit les analyses des données de spectrométrie de masse, *proteomics* et *phosphoproteomics*, dérivées de lignées cellulaires du cancer de la prostate. Ces lignées cellulaires représentent différentes étapes de

la progression de la maladie. Nous avons identifié certains événements moléculaires potentiellement responsables de la progression du cancer de la prostate vers les stades métastatiques, qui sont très agressifs et incurables.

Mots clés: Random Walk, réseaux biologiques, réseaux multiplex, clustering, maladies du vieillissement prématuré, protéomique, phosphoprotéomique, cancer de la prostate.

Acknowledgements

To be honest, I never thought about doing a PhD just after finishing my bachelor degree in Physics. However, sometimes life takes you through unexpected paths and now that my time in Marseille is coming to an end, I can say without hesitation that I really feel glad to have chosen this route. It allowed me to work on something that I really enjoy as well as meeting many interesting people from all over the world.

First and foremost, I would like to thank the two people who allowed me to get on board with them on this adventure: **Anaïs Baudot** and **Pierre Cau**, my thesis supervisors. I want to express my sincerest gratitude to **Anaïs** for her support and guidance during these last years. You showed me how doing research that we can proudly stand for. I appreciate very much the freedom you gave me to explore different avenues and to build my own research interests. I could not forget to thank you for your comprehensive and picky text corrections that helped me to improve my writing skills :-). Before moving on to the next person, I just want to say loudly THANKS FOR BEING SO NICE. **Pierre**, many thanks for trusting on our work and therefore funding my PhD. I have really appreciated our interesting scientific discussions where you shared with us your vast knowledge. Thanks for providing us with a complementary perspective that we were missing. I honestly think that you are an example to others concerning how to be passionate about scientific work.

I am also very grateful to the members of my thesis committee: **David Salgado**, **Pascal Hingamp** and **Laurence Calzone**. They encouraged me to keep working enthusiastically and gave me some precious advises. I would like to thank all the members of the jury, especially **Antonio Rausell** and **Laura Furlong** for willingly accepting the revision of this manuscript. My sincere thanks to my examinators: **Nicolas Thierry-Mieg** and **Jacques van Helden**. I hope that you can find the lecture of this manuscript enjoyable and not very tedious.

I feel proud to have been part of the *Institut de Mathématiques de Marseille* and the *Marseille Medical Genetics* unit. This gave me the opportunity to work with many brilliant colleagues that I would like to thank. A big thank you to **Elisabeth Rémy** and **Laurent Tichit** for their valuable scientific feedback and for being always ready to help me with any topic either personal or work-related. Thank you very much to those who are part of the team for a longer time: **Firas Hammami**, **Brigitte Mosse**, **Elva Novoa**, **Léonard Herault** and **Alain Guenoche**; and to those who have most recently joined us: **Léo Pio-López** and **Pooya Zakeri**. I really hope that we can keep working together in the future.

It was also a pleasure to be a member of the *ProGeLife* family, which made discovering the exciting complexity of aging. I would like to especially thank **Sophie Perrin** and **Claire Navarro** for trying to validate experimentally our pre-

dictions and for spending their time explaining us how they did it. I would also like to mention the other members of *ProGeLife*: **Nicolas Lévy**, **Eric Dessaud**, **Gaëlle Odelin**, **Pascale Klopp**, **Clovis Bondu** and **Héloïse Laroye**.

Occasionally, a project can face many obstacles and last longer than expected. However, it is also possible to meet great colleagues in those situations. I feel grateful for having the chance to work with **Luc Camion** and **Maria Katsogiannou**. I wish you all the best for your forthcoming projects.

Well, I think it is time to shift towards the more personal side of the acknowledgements. During these last years, I had the chance to live in Marseille, a city that produces mixed feelings among its inhabitants: sometimes you love it, sometimes you hate it. I write these lines short before leaving and I can assure you now that the moments of love for the city clearly exceed those of hate. I will certainly miss you Marseille. You were also my base camp for my expeditions to the amazing beaches and mountains of *La Provence*. Nevertheless, discovering these places would not have been such a rewarding experience for me without the friends who accompanied me.

First of all, I would like to thank all the flatmates that have passed through *La République* apartment, one of the most popular spots in Marseille's nightlife :-); okay okay, I am exaggerating. The original crew was composed of **Diogo** (Troncooooo), football and a beer is our perfect evening plan, **Annamaria**, always busy organizing hundred different plans, and **Guillaume**, trying to correct our broken French. Then, the Croatian team came into scene with two top players **Luksa** (Kuuuna) and **Ante** (Eeeepaña). Guys, I hope to visit you soon in Munich and Cagliari. Many thanks to **Alessandro** for his intensive Italian courses, *mi manca parlare italiano con te!* I would also like to thank **Marie** and **Yves**, the last signings of *La République*. Although one of those signings was a bit controversial, right **Marie**?

Participating in the *Café des Langues* gave me the chance to make very good friends from all over the world. I had the enormous pleasure of meeting Dr. Bachata (**Alejandro**), the music dictators (**Lamia**, **Lolita** and **Wi-Fi**), the Mexican team and associates (**Santiago**, **Jaime**, **Claire** and **Ariana**), any Italian in Marseille? Virtually none...(**Elena**, **Serena**, **Andrea**, **Paolo** and **Federico**), El danés errante (**Michael**), the Pointu crew (**Juan** and **Pavlo**) and some good friends that I have to mention but are not easy to categorize (**Marta** (Troncaaaaa), **Claudio**, **Jordi**, **Belén**, **Lan**, **Nico** and **Rasa**). Thanks guys for all the moments we shared. I will be waiting for your visit in my next destination!

Tomar la decisión de cambiar un empleo estable y bien remunerado en tu país de origen por un doctorado en un país donde no conoces la lengua no es fácil. **Neli** siempre me ayudó y me apoyó para que siguiera los dictados de mi corazón. No es fácil de expresar con palabras lo eternamente agradecido que te estoy. Por ello, simplemente quiero añadir lo siguiente: MUCHÍSIMAS GRACIAS.

Quiero enviar desde aquí un beso enorme para la mujer con más paciencia del mundo, **mi madre**. Además de querernos y mimarnos, La **señora Urbelz** ha sido capaz de guiar a buen puerto una nave cuya tripulación está compuesta por tres Valdeolivas varones. Creánme si les digo que esto es una empresa titánica. Un fuerte abrazo a la persona que me inculcó el amor por la ciencia, **mi padre**. Aún recuerdo las recompensas de 5 pesetas por cada especie de árbol que yo adivinaba durante nuestros paseos por el parque de Morales. **Señor Valdeolivas**, espero poder devolver el favor mostrándote las plantas de *Les Calanques*. Parece que ya estoy llegando al personaje de mi hermano, **Quique**. Como yo soy el hermano mayor, te agradezco por tus futuras acciones que todavía no han tenido lugar. Gracias **Doctor Valdeolivas** por cuidarme y darme consejos médicos cuando sea viejo (¡aún más viejo que ahora!)

Como el tamaño de mi familia es reducido, me voy a permitir el capricho de citarla al completo. Muchas gracias a mi **tía Pilar** por animar incansablemente las fiestas familiares. Además tiene distinto repertorio para calentar las fiestas, lo mismo te canta que te genera una discusión. Gracias a mi **prima Lorena** por los bellos recuerdo de infancia. Aunque en aquellos tiempos me hacías sufrir mucho con el chocolate, ahora esas imágenes mentales evocan muy felices momentos para mí. Pese a que le tengo un poco de miedo por ser más fuerte que yo, quiero también dar las gracias a mi **primo Óscar**, ¡Vaya mazada niño! Y por supuesto, gracias a **Lucía** por convertir cada visita a Santander en una nueva aventura. Es mejor que estés bien preparada para mi próxima vez, ¡Merengue vuelve más poderoso que nunca! Quiero también agradecer a la pequeñaja, **Alexia**. Esas fotos que recibo en mi móvil donde tú eres la protagonista siempre consiguen sacarme una sincera sonrisa. ¡Gracias a ti también, **Robe**!

Y esto se está empezando a acabar, ¡pero todavía no se ha acabado del todo! No podría despedirme sin antes nombrar a **Suzana**. No hace mucho tiempo que te cruzaste en mi vida, sin embargo has llamado con fuerza a las paredes de mi corazón (Yeah, yeah, colder than ice, hard as a rock!). Hvala ti puno! Faleminderit shumë!

Alberto Valdeolivas

Marseille, February 2019

Contents

Preface

I have always been fascinated by the fact that the number of cells in a human being is estimated to be greater than the number of galaxies in the whole universe. This huge number of entities must communicate and organize among themselves in a marvelous complex way to ensure the proper functioning of our body. Moreover, this complexity is also observed at a smaller scale given that around 10^{10} proteins work in a coordinated manner inside a medium-sized eukaryotic cell to keep us alive. The question that arises is: can we fully understand what is actually happening inside a cell? I would rather reply no if I had to bet on an answer. Nevertheless, I strongly believe that we are in an unprecedented moment to improve our comprehension on the subject. We can nowadays take advantage of the massive amount of biological data supplied by the new high-throughput technologies.

This accumulation of large-scale data called for the development of computational tools able to explore and mine the vast amount of biological knowledge they contain. This culminated in the birth and exponential growth of the field of bioinformatics, bringing together two of the most appealing research fields for me. I felt that, from my very modest contribution, I had to be part of this scientific community at such decisive moments in which, for example, the human genome has been completely decoded. Thus, I decided to quit my job in the banking sector in 2014, and started studying a master in bioinformatics that eventually drove me to Marseille, where the work detailed in the forthcoming lines took place.

This thesis manuscript presents the result of my research conducted under the supervision of Prof. Pierre Cau (*ProGeLife*) and Dr. Anaïs Baudot (*Marseille Medical Genetics, MMG*). This double supervision led to a close collaboration between *ProGeLife*, a biotechnology R&D company dedicated to the development of therapeutic solutions for rare genetic diseases characterized by premature aging, and the *Networks and Systems Biology for Diseases* group from the *MMG* unit. In addition, we intently work closely with our former group, *Mathématiques et Algorithmique pour la Biologie des Système (MaBios)*, from the *Institut de Mathématiques de Marseille, I2M*. In this interdisciplinary framework, I tried to contribute to the research carried out in these groups by providing the knowledge acquired during my education in Physics, Biophysics and Bioinformatics along with my working experience in the banking sector as a software developer.

In this manuscript, I describe first the research path I followed during the last four years from my own perspective (Personal summary). Then, in the first three chapters, I introduce the background state-of-the-art notions, experimental techniques and algorithms relevant for my work and results. To be more precise, the first chapter details some general notions of graph theory, which correspond to the mathematical scaffolds underlying my work. The second chapter describes

the different types of biological networks, how they are constructed and their properties. I next introduce the concept of network medicine in the third chapter, and discuss how biological networks can help us to better understand and treat diseases.

The later three chapters describe the results I obtained during my PhD studies. More specifically, chapter fourth focuses on the development of new algorithms to deal with combinations of different types of biological networks. We show how these methods outperform traditional approaches that take as their input single networks, within the context of network medicine. Then, I applied extended versions of these new approaches to study a group of diseases displaying a premature aging phenotype. These results are detailed in chapter fifth, and aim at improving our knowledge about the molecular factors involved in these disorders and in physiological aging. In chapter sixth, I describe the results yielded by my analyses on *proteomics* data derived from prostate cancer cell lines. We identified some molecular mechanisms likely to be implicated in the disease's progression to resistance. The end of the manuscript is dedicated to discussion of the overall results and personal conclusions.

Personal summary

In this section, I try to describe how the projects arose from a chronological point of view and how they evolved from my personal perspective. The main goal is to locate the reader in the general framework in which this thesis was accomplished. It is to note that citations are not included in this section for the sake of simplicity, but they are provided in the forthcoming sections.

Master Thesis: Proteomics and Phosphoproteomics for Prostate Cancer

I arrived in Marseille in 2015 to accomplish my Master Thesis under the supervision of Dr. Anaïs Baudot in the *Institut de Mathématiques de Marseille, I2M, UMR7373, Campus de Luminy*. The project involved a collaboration between different wet and dry laboratories, and concerned a disease of considerable impact on today's society, prostate cancer (PC). Indeed, PC is the second leading cause of male cancer-related death in industrialized countries. It is often treated by castration therapy, but eventually recur as Castration-Resistant Prostate Cancer (CRPC), which is highly aggressive and incurable. It is therefore essential to identify the molecular events responsible for the progression of PC to the castration-resistant stage in order to design specific therapies. Such molecular variations can be monitored by system-wide approaches such as *transcriptomics* and *proteomics*. In this line, our collaborators in the CRCM, *Centre de Recherche en Cancérologie de Marseille*, conducted a SILAC-based mass spectrometry approach to identify and quantify the proteomes and phosphoproteomes of four widely used prostate cell lines. These cell lines represent different stages of the disease progression, from prostate healthy tissue to CRPC.

During my master research stay and the beginning of my PhD thesis, I carried out rigorous statistical analysis procedures over these *proteomics* and *phosphoproteomics* data. My results allowed characterizing each cell line from a functional point of view. In addition, we pointed out potential biomarkers differentially expressed or phosphorylated in PC and CRPC compared to normal. We finally turned to our specialty and integrated *proteomics* and *phosphoproteomics* markers in a molecular network, permitting to highlight the functional mechanisms differentially expressed in the CRPC stage. More information about this project and an associated article detailing our results are available in the final chapter of the present manuscript, chapter 6.

Katsogiannou, M.†, Boyer, J.B.†, **Valdeolivas, A.†**, Remy, E., et al. (2018). *In-*

tegrative proteomic and phosphoproteomic profiling of prostate cell lines. Under Review. †Equal contribution.

PhD Project: Networks for Premature Aging diseases

After a few months working on the aforementioned project, I successfully defended my master thesis, and Dr. Anaïs Baudot and Prof. Pierre Cau gave me the opportunity to undertake a doctoral thesis under their guidance. *ProGeLife*, the biotechnology R&D company founded by Pierre Cau and Nicolas Levy, would fund my work through a *CIFRE* scholarship. *ProGeLife* is dedicated to the development of therapeutic solutions for rare genetic diseases characterized by Premature Aging (PA). Anaïs has extensive experience in the systems biology field, and more specifically in biological networks. Consequently, Pierre and Anaïs offer me to develop network-based methods, and apply them to study PA diseases as a thesis topic.

The general issue we aimed to address was to identify genes/proteins potentially related to PA diseases, in order to get insights about the processes deregulated in these syndromes, and predict new potential therapeutic targets. In addition, this information can provide clues about how physiological aging occurs. Our working framework is the one of network medicine, which states that the complex phenotype of a disease does not only arise from the isolated punctual mutation in its causative gene, but from the global perturbation induced by that mutation in biological networks.

Network-based algorithms

Genes and proteins do not act isolated, but rather interact with each other to perform their functions in cells. These interactions can be represented as large networks in which nodes are the genes or proteins, and edges represent their interactions. As interactions are the foundations of gene and protein cellular processes, functionally-related proteins tend to lie in the same network neighborhood within biological networks. These molecular arrangements are at the core of the guilt-by-association strategies, which state that interacting proteins are more likely to be functionally related than non-interacting proteins. Guilt-by-association procedures are based on two major graph-theory approaches:

1. Network propagation ranking methods prioritize network nodes according to their topological distances to a previously defined group of genes of interest, which are called the seeds. In particular, random walk with restart

captured my attention since they have the ability to explore the global topology of networks. It is among the most widely-used ranking techniques in computational biology, and are described on section 3.1.1.

2. Clustering algorithms allow identifying communities of tightly connected proteins, which are therefore likely to be involved in the same biological processes. An overview of the different clustering algorithms, with a special emphasis on those frequently used in network biology, can be found on section 3.1.2. Among the various approaches, we can highlight the algorithms aiming at optimize the modularity, a measure describing the strength of division of a network into communities.

Multiplex networks

The guilt-by-association algorithms mentioned above are usually applied on single (monoplex) networks, mainly representing protein-protein direct physical interactions. However, other type of physical interactions and functional relationships exist among genes and proteins. For instance, proteins can also be members of the same molecular complexes without having a direct physical contact, or can participate in the same pathways. Biological interactions are identified thanks to different experimental techniques, including yeast-two-hybrid or mass spectrometry. In addition, different functional relationships among genes or proteins can be inferred from *-omics* data, such as co-expression associations from *transcriptomics*) data.

Overall, we can build different networks of gene/protein relationships from these interactions sources. Each networks contain nodes corresponding to genes or proteins, and edges belonging to different interaction categories. The different networks display various topological features, and each has its own strengths and weaknesses. The richness of this information can be integrated into multiplex networks, i.e. collections of network layers (monoplex networks) sharing the same nodes, but in which edges represent interactions of different categories. Multiplex networks have been shown to be a good approach to represent complex systems, since they provide an effective way to integrate individual sources of information while keeping track of their individual features. We therefore decided to direct my research work towards the development of algorithms dealing with multiplex biological networks. To leverage a maximum of biological information, we also considered heterogeneous networks, i.e., networks composed of different type of nodes such as diseases linked by phenotypic relationships. The core of the work presented in this document lies in the innovative extension of network propagation ranking methods and clustering algorithms to the multiplex-heterogeneous framework. Undoubtedly, we always kept in mind their applications to PA diseases.

Random walks with restart on multiplex networks

Random walk with restart (RWR) is the state-of-the-art guilt-by-association ranking algorithm used in network computational biology. It was initially applied to prioritize candidate disease genes/proteins. The algorithm ranks all the network nodes according to their proximity to known disease-associated genes/proteins taken as seeds. I developed two extensions of the RWR algorithm, in order to explore multiplex networks and multiplex-heterogeneous networks. A multiplex network is composed of different layers, sharing the same nodes but different sets of interactions. A heterogeneous network is composed of networks having both different nodes and edges, linked through bipartite interactions. We demonstrated that applying random walk with restart on a multiplex (RWR-M) network or on a multiplex-heterogeneous (RWR-MH) network performs better than its application to monoplex isolated networks in the task of predicting disease genes. Moreover, I applied these new methods to explore the network vicinity of PA syndromes, in order to unveil their associated diseases and pathways. This work is described in detail on section 4.3, where I also added the following peer-reviewed published article:

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., et al. (2018). *Random walk with restart on multiplex and heterogeneous biological networks.* **Bioinformatics**, (August), 1–9.

This work also lead to the publication of an *R* package on *Bioconductor* (see Appendix A.2 for details). The application of this method on a PA disease led to the selection of a particular gene for further biological analyses. My *ProGeLife* colleagues conducted a set of experimental validations to determine if the selected gene is involved in aging. These results are presented in the discussion of this manuscript.

Clustering on multiplex networks

Biological networks usually present a community structure, meaning that groups of proteins are more densely connected between each other than with the rest of the network. These groups of tightly connected entities, usually called modules, contain nodes likely to be involved in the same biological functions. Therefore, the accurate extraction of these modules is an important guilt-by-association strategy. For instance, the identification of the modules where a mutated gene participates can reveal the cellular and molecular mechanisms underlying diseases.

A plethora of methods have been proposed for community detection based

on different theoretical bases and diverse assumptions to define biological modules. In this context, the Disease Module Identification (DMI) DREAM challenge has been set up to investigate the performance of different community detection algorithms in a bio-medical context, by providing biological networks and a controlled benchmark to evaluate the modules detected by the different methods. We participated to the challenge with an enhanced version of *MolTi*, a modularity-based software designed to detect communities in multiplex networks previously developed by our team. The overall results of the challenge are detailed in the consortium article to which I participated (Appendix B).

In addition, we performed an in-depth analysis of the communities detected by the the new version of *MolTi* on both simulated data and the DREAM challenge benchmark. Concisely, we are able to detect more relevant modules taking advantage of the multiplex approach. These results, along with a detailed discussion about the challenges of the detection of communities on real multiplex biological networks, are presented in section 4.4, where the following peer-reviewed published article is also included:

Didier, G., **Valdeolivas, A.**, & Baudot, A. (2018). *Identifying communities from multiplex biological networks by randomized optimization of modularity.* **F1000Research**, 7(0), 1042.

The landscape of premature aging diseases

Premature aging syndromes, also called progeroid syndromes, are a group of rare genetic disorders that phenotypically resemble some of the aspects of physiological aging at an early age. The aged appearance of PA disease patients can be the result of alterations in molecular mechanisms also occurring at an advancing age on healthy individuals. Consequently, PA diseases are the subject of a significant number of studies aiming at identifying their altered processes and transferring this knowledge to physiological aging. These syndromes are clinically and genetically heterogeneous, affecting different tissues and displaying dissimilar severity degrees. They can therefore potentially reveal different agents involved in the aging process.

My last project was to define a functional landscape for a set of diseases displaying phenotypes related to PA. To this goal, I developed a seed-expanding approach to generate modules around node(s) of interest in multiplex and multiplex-heterogeneous networks. The method is based on the recursive application of the previously mentioned RWR-M and RWR-MH algorithms. We applied the extended method to generate modules associated to different PA diseases. The

functional modules identified revealed the landscape of perturbed molecular processes in PA diseases, which can be paralleled with those altered in physiological aging. We also classified these PA diseases in different groups aiming at identifying their common and specific features. We are currently gathering these results and writing the article detailed below, which can be found on chapter 5.

Valdeolivas, A., authors to be defined & Baudot, A. (2018). *A Multiplex Seed-expanding Approach to Describe the Landscape of Premature Aging Diseases*.

Introduction

1 Graph theory

Graph theory is the area of mathematics studying graphs; it is usually called network science when it is applied to real-world systems. Indeed, a graph is a mathematical objects describing the pairwise relations between the entities of a system, i.e. a network. The use of graphs have strongly gained attention in the last decades since they may be applied to model and analyze problems into a plethora of fields. For instance, a wide variety of applications can be found in diverse areas such as computer science (Deo 2017), linguistics (Filippova 2010), physics (Estrada 2013), chemistry (Trinajstić 1983), social sciences (Borgatti et al. 2009), biology (Verwoerd 2007) and ecology (Bascompte 2007).

This chapter provides a formal introduction to graph theory by detailing some of the mathematical concepts sustaining my work. I tried to keep a biological perspective, and hence give examples of graph-theory applications in the framework of computational biology. It is to note that the biological networks considered in the results chapters of this manuscript are large-scale, finite and undirected networks, so I herein restrict definitions to such graphs.

1.1 History of graphs

The former city of Königsberg in Prussia, nowadays Kaliningrad in Russia, is at the mouth of the River Pregel. The river separates the city in four land masses connected by seven bridges, as represented in Figure 1.1

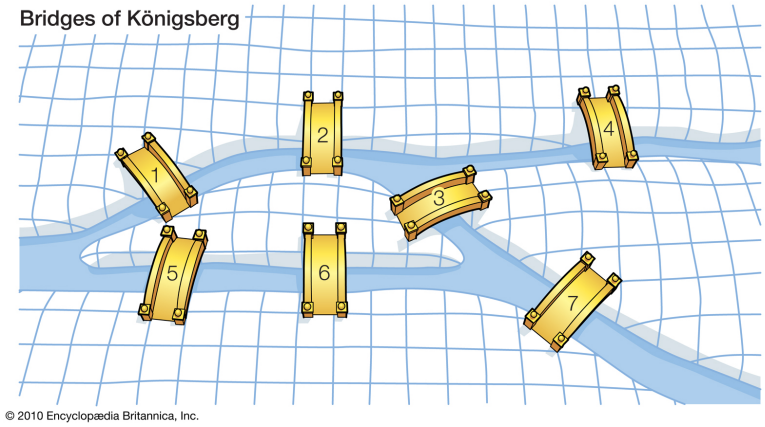


Figure 1.1 – Schematic representation of the city of Königsberg showing the seven bridges, the four land masses and the Pregel river. Figure adapted from (Carlson 2017)

The inhabitants of the city thought about the possibility of finding a path to

travel across the four land masses, crossing each of those bridges once and only once. In 1736, Leonhard Euler determined with mathematical rigor that such path does not exist (Euler 1736). Doing so, Euler demonstrated the first theorem in graph theory. The key idea is to notice that the walk inside each land mass is not relevant, and they can thereby be represented with a point (node or vertex). These points can be connected with other points according to the city bridges (edges) whose crossing sequence will uniquely determine the route of the walk. In this context, Euler eliminated all features of the city except the list of nodes and edges (Wilson et al. 1976; Shields 2012). This mathematical structure is a graph.

Some mathematicians thought that the answer to this problem was trivial, included Euler himself when he was first asked about it. However, the importance of Euler's resolution lies on the new mathematical formulation he used to demonstrate it, giving rise to graph theory.

1.2 Formal definition of graphs

A graph, G, is defined as a pair of sets $G = (V_G, E_G)$ where V_G is a finite collection of points, called vertices or nodes, and E_G is a finite collection of 2-element subsets of V_G, called edges (Wilson et al. 1976). The edges link the vertices of the graph establishing a given relation between them. G is an **undirected graph** if its edges are unordered pairs $\{u, v\}$, where u and v are vertices in V_G.

In order to illustrate the mathematical object defined by a graph, let us consider the graph G represented in Figure 1.2. This graph can be fully defined as follows:

$$G = (V_G, E_G) \quad \text{where} \quad V_G = \{1, 2, 3, 4, 5, 6\} \quad \text{and}$$
$$E_G = \{\{1, 2\}, \{2, 3\}, \{2, 6\}, \{3, 4\}, \{4, 5\}, \{5, 6\}\}$$

The graph presented in Figure 1.2 is defined as a **simple graph**, i.e., an undirected graph with neither multiple edges nor loops. Additional attributes such as weights, labels or colors, can be assigned to vertices and edges. However, these kind of graphs are not enough to model and represent some biological systems. For instance, two atoms in a molecule can be linked by more than one chemical bond (**multi-edges**); and some transcription factors can regulate their own expression (**self-loops**) (J. Gross 2018; Cargnin et al. 2005).

The biological networks detailed in chapters 4, 5 and 6 are simple graphs. Nevertheless, for the sake of comparison of different approaches, the reader should bear in mind that I sometimes consider aggregated networks. Those particular networks, being the union of different networks, can have multi-edges. For instance, a pair of proteins having a physical interaction between them as well as being co-expressed in the same tissues, will be linked by two edges in an aggregated network. I also usually considered unweighted networks in this manuscript.

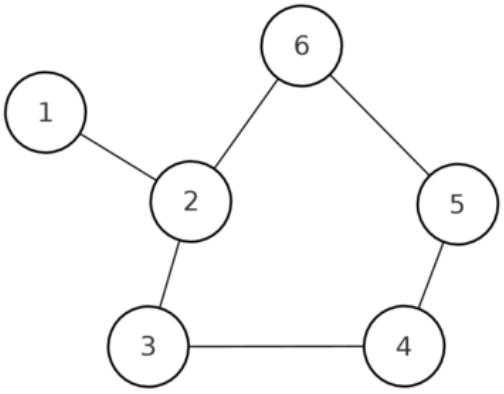


Figure 1.2 – A drawing of a simple graph containing 6 vertices and 6 edges.

Noteworthy, weighted networks are sometimes useful in computational biology, since weighted edges can account for the reliability of the interaction described.

1.3 Terminology for graphs

Let us consider an undirected graph $G = (V_G, E_G)$ with a set of vertices V_G and a set of edges E_G. This section introduces some of the most commonly used terms in graph theory:

— The number of vertices in G ($|V_G|$) is named the **order** of G.
— The number of edges in G ($|E_G|$)is called the **size** of G.
— Vertices linked by an edge are called their **endpoints**. In this context, vertices u and v are the endpoints of the edge $\{u, v\}$.
— Two vertices u and v are **neighbors** or **adjacent vertices** if they are linked by an edge (if $\{u, v\} \in E_G$). The **neighborhood** of a vertex v is the set of the neighbors of v.
— Edges are said to be **adjacent edges** if they have a common vertex.
— A **self-loop**, or simply a **loop**, is an edge whose endpoints are the same vertex. An edge of the structure (v, v) is a loop.
— An edge joining two distinct nodes is named a **proper edge**.
— Two distinct edges are called **parallel** if they have the same endpoints.
— A **simple graph** has no parallel edges and no loops.
— A **multigraph**, as opposed to a simple graph, can contain multiple edges (parallel edges) and loops.
— The **degree** or **valence** of the node v, usually denoted as $d(v)$, $deg(v)$ or k_v, is the number of edges with v as an endpoint. By convention, loops are counted twice and parallel edges contribute independently.

— A **complete graph** possess a set of edges containing every possible edge between all of the vertices. The complete graph on n vertices is usually denoted by K_n. An example of a complete graph is illustrated in Figure 1.3 A.
— A **regular graph** is a graph where all their nodes have the same degree. A regular graph whose vertices have degree k is called a **k-regular graph**. Figure 1.3 B shows an example of a regular graph.
— A **bipartite graph**, B, is a graph whose vertex set, V_B, can be divided into two disjoint and independent subsets V_{B1} and V_{B2}, such that every edge contains an endpoint in V_{B1} and another endpoint in V_{B2}.

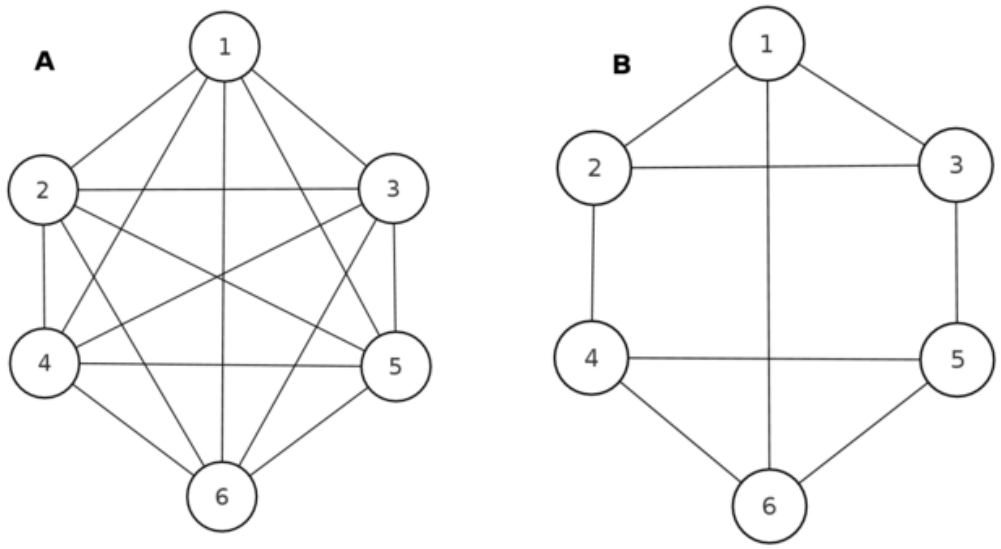


Figure 1.3 – Complete and regular graphs. A) the complete graph of degree 5, K_6. B) A 3-regular graph.

Regular and complete graphs are sometimes used in biology. For instance, let us consider expression levels of proteins in different tissues. We can compute the correlation among these levels for every pair of proteins. A complete graph can be constructed by taking as weighted edges the correlation values among all the protein pairs. A regular graph can be build by imposing that each protein should be linked to a given number of proteins.

In addition, bipartite graphs provide an interesting tool in computational biology. In section 4.3, I detail the use of a bipartite network, where nodes are both genes and diseases. Bipartite edges are the links between diseases and their causative genes.

For further terminology about graphs and more rigorous mathematical definitions, we address the interested reader to some of the following references (Chen 1997; J. Gross 2018; Shirinivas et al. 2010; Cvetković et al. 1995; Chartrand et

al. 2012).

1.4 Matrix representation

There are diverse methods of graph representation: in the previous sections of this chapter, for instance, graphs have been drawn as diagrams, or described as set of edges. Nonetheless, these representations are only reasonable when the size and order of graphs are small. Graphs are commonly represented in form of matrices allowing computers to efficiently manipulate them. In addition, matrix representation gives access to the power of linear algebra (J. Gross 2018). For instance, well known operations of matrices can account for the computation of different paths within a graph such as the random walks, which are widely used in computational biology (see section 3.1.1 and 4.3).

The **adjacency matrix** of an undirected graph $G = (V_G, E_G)$, denoted by A_G, is a square symmetrical matrix of size $|V_G| \times |V_G|$. In a simple graph, the matrix element $A_G[u, v]$ is one when there is an edge between vertex u and vertex v, and zero if that edge does not exist (Godsil et al. 2001):

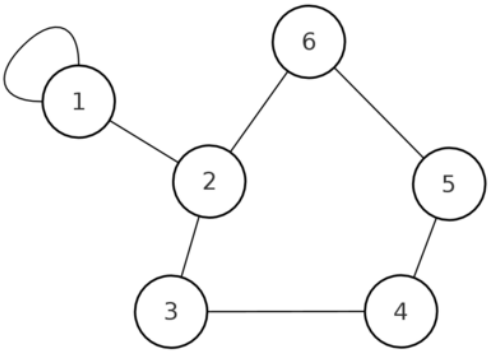
$$A_G[i, j] = \begin{cases} 1 & \text{if } (u, v) \in E_G \\ 0 & \text{otherwise.} \end{cases}$$

The diagonal of the matrix is composed of zeros, since there are no loops in simple graphs. The definition of adjacency matrix can be easily extended to multigraphs and graphs with loops. To do so, the corresponding matrix item should account for the total number of edges between two nodes. In addition, diagonal elements are permitted to be different from zero. For convention, loops are usually counted twice. Figure 1.4 shows the diagram representation of a graph with loops and its adjacency matrix. To construct an adjacency matrix of a weighted network, we have to replace each element with value one for its corresponding edge weight.

1.5 Distances and walks

A large number of methods in graph theory are related to the idea of exploring a graph. For instance, a graph may describe the subway network of a specific city. The vertices represent the stations and the edges account for the rails connecting pairs of stations. Many questions may emerge: what is the distance between two given stations?; What is the fastest way to travel between those stations?; Which areas of the city are more densely connected?

These applications call for the definition of the concept of walks in graphs. A **walk** W in a graph G, starting from vertex u and ending on vertex v, is an alternating sequence of vertices and edges such that consecutive vertices are



$$A_G = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 1.4 – Drawing of a graph along with its adjacency matrix.

adjacent:

$$W = \langle u = v_0, e_1, v_1, e_2 \dots v_{n_1}, e_n, v_n = v \rangle$$

where $n \geq 0$ and v_{i-1} and v_i are adjacent for $i = 1, 2, \dots, n$. The walk W is **closed** if it starts and finish at the same vertex, $u = v$, otherwise it is called an **open walk**. The **length** of a walk is the number of edges in the sequence walk. A walk with no repeated edges is called a **trail**. A **path** is a trail with no repeated vertices, excluding the starting and ending vertices.

A graph is **connected** if there is a walk between every pair of vertices. A classical **distance** defined in connected graphs G is the length of the shortest path. It is denoted by $d_G(u, v)$ or simply $d(u, v)$. The **diameter** of a connected graph G, denoted by $diam(G)$, is the greatest distance between any two vertices. The diameter of biological networks tend to be quite small due to some of their intrinsic properties (see section 2.8).

1.6 Subgraphs, modules and motifs

A **subgraph** of a graph G is another graph H constructed from a reduced set of the vertices and edges of G. It can be denoted as $H \subseteq G$, if $V_H \subseteq V_G$ and $E_H \subseteq E_G$. In this situation, G is also called a **supergraph** of H. An **induced subgraph** is comprised of all the edges of the parental supergraph whose endpoints belong to V_H. A **component** of a graph G is a maximal connected subgraph of G. There is only one component in a connected graph which is the whole graph (J. Gross 2018; Chartrand et al. 2012).

Partitioning an initial graph into a set of smaller graphs, usually induced subgraphs, is a constant challenge in network theory. For instance, clustering algorithms aim at finding **modules/communities** within graphs. Modules are groups of nodes more densely connected with each other than with the rest of the network. This is a very active field in computational biology since groups of tightly

connected proteins are usually involved in the same biological processes, as discussed in section 3.1.2.

Another useful application is the detection of **motifs**, which are defined as subgraphs more repeated than expected by chance in a specific network. Motifs present a particular topology or pattern of interactions between vertices, such as triangles. They may reflect functional properties of the network or reveal how it evolves. Motif repetition in biological networks has been applied to determine common evolutionary mechanisms. For example, the same motifs have been found in the transcription networks of bacteria, yeast and higher organism, suggesting common underlying adaptive mechanisms (R. Milo et al. 2002; Bonnici et al. 2013).

1.7 Topological measures and metrics in graphs

Several standard measures and metrics have been defined to capture, quantify and describe particular features and attributes of graphs. These properties are key factors that can help to understand the underlying nature of the real-world system described by a particular graph. In addition, some of these features are the base to compare different graphs, and hence different systems.

Let us consider an undirected graph $G = (V_G, E_G)$ with a set of vertices V_G and a set of edges E_G. The number of nodes, i.e. order of G, is defined by $|V_G| = n$ and the number of edges, i.e. size of G, is denoted by $|E_G| = m$. Let $A = (a_{v,u})$ be the adjacency matrix, i.e. $a_{v,u} = 1$ if vertex v is linked to vertex u, and $a_{v,u} = 0$ otherwise. The metrics used in this thesis manuscript are introduced in the next two sections.

1.7.1 General metrics

In this section, I define some of the most relevant metrics of graph-theory in the context of biological networks:

1. The **density** or **connectance**, ρ, of a graph is defined as the ratio between its number of edges and the maximum possible number of edges. The maximum possible number of edges in a simple graph is given by $\binom{n}{2}$:

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}$$

   A graph is said to be **dense** if ρ is close to one, whereas it is said to be **sparse** if ρ is close to zero. Complete graphs, like the one illustrated in Figure 1.3 A, have the maximum possible number of edges and their density is therefore $\rho = 1$.

2. The number of edges connected to a node i is its degree and it is denoted by k_i, as we defined previously. An important property of a graph is its **average degree** which is defined by the following equation:

$$\langle k \rangle = \frac{2m}{n}$$

The factor 2 emerges from the fact that each edge contributes to the degree of its two endpoints in undirected graphs.

3. The **degree distribution** p_k of a graph is defined as the probability that a randomly elected vertex has a given degree k:

$$P(k) = \sum_{k=1}^{n} p_k = 1; \qquad p_k = \frac{n_k}{n}$$

where n_k is the number of vertices of degree k. The degree distribution plays a key role in understanding how networks are organized and therefore their behaviour. For instance, under perturbations, a network containing nodes with roughly the same degree is not going to behave in the same manner than a network containing few high-degree nodes and many poorly connected nodes. In the context of a network of physical interactions between proteins, if one protein is mutated, the global impact on the network, and hence on the organisms, may depend on its degree. Overall, the more interaction partners has the mutated protein, the more it may affect the network (see section 2.8.1.3 for further details).

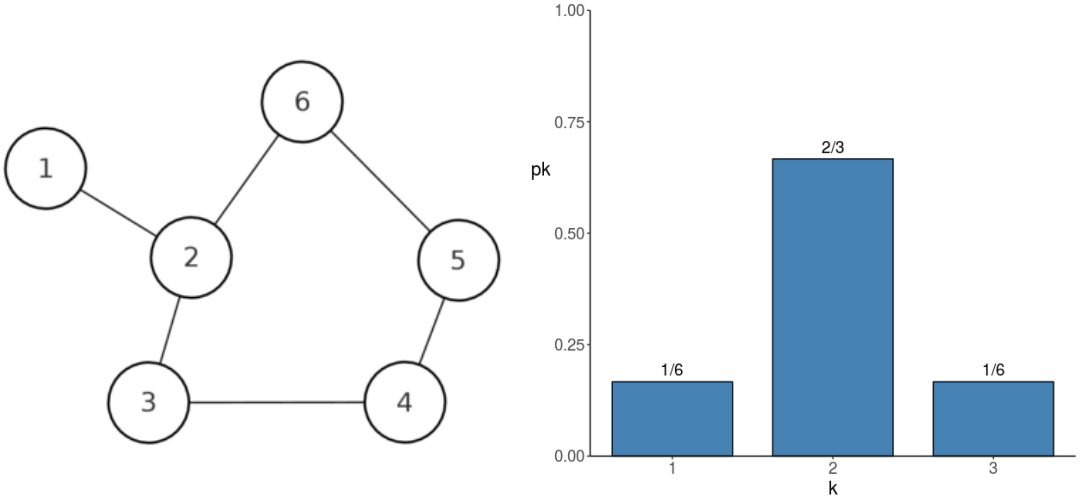


Figure 1.5 – Drawing of a simple graph along with its degree distribution.

4. The **clustering coefficient** estimates the extent to which nodes in a graph tend to group together. It can be described from the perspective of individual nodes, **local clustering coefficient**, or from a global view of the

network, **global clustering coefficient**. Formally, the local clustering coefficient of a vertex i with degree k_i is:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

where e_i is the number of connections between the ki neighbors of the node i.

The global clustering coefficient computes the average probability that two neighbors of a node are also connected, i.e. they are also neighbors themselves:

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i$$

$\langle C \rangle$ represents a probability and therefore its value is between 0 and 1. This measure is relevant in biological networks because it indicates the extent to which network nodes can be grouped into different modules or communities. As we commented before, proteins or genes that belong to the same module may be functionally related.

5. Another metric to measure the strength of division of a network into modules is the **Modularity**. It was defined by Newman and Girvan (Newman and Girvan 2004) as:

$$Q(G) = \frac{1}{2m} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta_{c_i, c_j}$$

where m is the total number of edges of G, A its adjacency matrix, k_i and k_j are the degree of the nodes i and j respectively, and δ_{c_i,c_j} is 1 if i and j belong to the same module and 0 otherwise. Biological networks usually exhibit a high degree of modularity and, consequently, the optimization of this value is one of the most popular approaches to detect modules within them (see sections 3.1.2 and 4.4). It is to note that I used a slightly different notation for the definition of modularity than for the remaining metrics. I tried to stay aligned with the equations presented in the section 4.4.

1.7.2 Centrality measures

The identification of the most relevant nodes in a network is another main question in graph theory. However, the definition of relevance is very subjective. It is usually defined by the concept of **centrality**, closely related to the degree of a node. But the importance of a node strongly depends on the problem to be addressed and, consequently, these measures should be carefully chosen. The most used centrality measures are presented below:

1. The most straightforward centrality measure is given by the degree of the vertices of a network. Even though **degree centrality** is a very simple centrality measure, it is very convenient in multiple contexts. For instance, a person with many connections within a social network might be one of the most influential in the group (Newman 2018). The same reasoning can be applied for proteins, where highly connected proteins (hubs) are likely to be more essential for organisms (see section 2.8.1.3).
2. The **eigencentrality** or **eigenvector centrality** is based on the same concept of the degree centrality. However, eigencentrality takes into account the importance of the vertices connected to each. Each node is scored proportionally to the sum of the degrees of its neighbors (Newman 2018). The eigenvector centrality score of a vertex v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{u \in M(v)} x_u = \frac{1}{\lambda} \sum_{u \in G} a_{v,u} x_u$$

where $M(v)$ is the set of neighbors of v and λ is a constant. This equation can be written as the eigenvector equation:

$$Ax = \lambda x$$

There will usually be many different eigenvalues λ for which a non-zero solution exists. However, the centrality measure is given by the unique largest eigenvalue results according to the Perron–Frobenius theorem (Newman 2010). A popular clustering algorithm is based on this measure (Newman 2006) (see sections 3.1.2 and 5.3 for further details).
3. Eigenvector centrality is well suited to be applied on dense graphs, however some problems can arise for sparse graphs. **Katz centrality** overcomes this limitation by taking into account the total number of paths between pairs of nodes (Hanneman et al. 2005). Connections made through remote vertices are penalized by an attenuation parameter $\alpha \in (0, 1)$. The connections between two nodes through intermediaries are represented by the powers of A, and therefore the Katz centrality score of a vertex v can be written as:

$$x_v = \sum_{k=1}^{\infty} \sum_{u=1}^{n} \alpha^k (A^k)_{vu}$$

For values of α close to zero the contribution of paths longer than one abruptly decreases. Long paths are devalued smoothly when the attenuation parameter increases, contributing more to the Katz score. It is suggested to choose $\alpha < \frac{1}{\lambda_{max}(A)}$, where $\lambda_{max}(A)$ is the largest eigenvalue of A. Noteworthy, Katz-centrality based approaches have been applied to

prioritize candidate genes following guilty-by-associations assumptions (J. Zhao et al. 2011) and section 3.1.

4. Let us illustrate a situation where a node with large Katz centrality is connected to many others. According to the equations presented in the previous point, the neighbors of that node will also get high centrality scores. However, there are many situations where this is not convenient. For instance, an important website with high centrality links to many other websites. Some of these sites are not supposed to play an important role in the network and, therefore, they do not deserve a high centrality score (Newman 2018). ***PageRank* centrality** elude this problem considering that the centrality transferred from network neighbors is proportional to their centrality divided by their degree (Brin et al. 1998). Therefore, the *PageRank* centrality score of a vertex v can be written as:

$$x_v = \alpha \sum_u a_{uv} \frac{x_u}{k_u} + \frac{1-\alpha}{n}$$

where k_u is the degree of the node u. *PageRank* is the main algorithm used by *Google Search* to rank websites. This measure is also one of the foundations of the random walk with restart algorithm, a guilt-by association method widely extended in computation biology (see sections 3.1.1 and 4.3).

5. **Closeness centrality** brings a totally different approach to measure centrality in a graph. The underlying idea relies on the fact that the more central nodes are those closest to the remaining nodes of the network. It is therefore computed as the reciprocal of average length of shortest paths between the vertex and all other vertices in the network (Sabidussi 1966):

$$x_v = \frac{n}{\sum_{\substack{u \in G \\ u \neq v}} d(u,v)}$$

where $d(u,v)$ is the distance between nodes u and v. Closeness centrality is widely used in several domains because describes node relevance in a very intuitive way. Nevertheless, the range of distances within biological networks is rather small and, consequently the closeness centrality scores tend to stack together (Newman 2018) (see section 2.8). However, some studies have also applied guilty-by-associations techniques in biological networks based on the grounds of this measure (Tran et al. 2014).

6. Yet another different approach to the concept of centrality is the so called **betweenness centrality**. Let us consider a flow of information being transferred between every pair of nodes of a graph along their connecting edges. Larger flows of information will go thought vertices connecting pairs of nodes in a shorter way. In other words, betweenness centrality evaluates

the amount of times a vertex operates as a link along the shortest path between two other vertices (Freeman 1977). It seems reasonable that nodes transferring a large amount of information are ranked as the most relevant. The betweenness centrality is defined as:

$$x_v = \sum_{s \neq v \neq u} \frac{\sigma_{su}(v)}{\sigma_{su}}$$

where σ_{su} is the total number of shortest paths from vertex s to vertex u and $\sigma_{su}(v)$ is the number of those paths passing through v. Betweenness centrality certainly contrasts with the previously described centrality measures. A specific vertex can actually have low scores for any of those metrics and still have high betweenness (Newman 2018). An example to illustrate this situation is presented in Figure 1.6. All the shortest paths between a node of group A and a node of group B should pass through node 1. Therefore, node 1 has a high betweeness, even thought its degree is two and it is not expected to have large centrality scores on the remaining metrics.

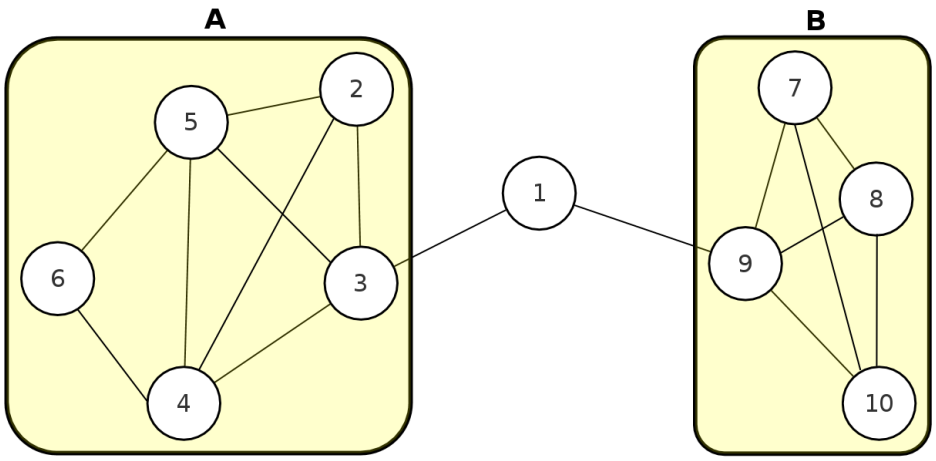


Figure 1.6 – Drawing of a simple graph. The drawing highlights an example of a vertex, node 1, with high betweeness centrality but lower scores for the remaining centrality metrics.

Betweenness centrality is a very important measure regarding failures or attacks on the network (see section 2.8.1.3). Let us suppose that the graph presented on Figure 1.6 models the transport system of a city. A failure on the node 1 totally removes the communication between the two areas of the city. On the other hand, failures on nodes 6, 8 or 10 entail a reduced local effect on the global transport of the city.

(Newman and Girvan 2004) suggested that nodes and edges with higher betweenness are most likely to lie between modules, rather than inside them. This situation is exploited to divide networks in modules (J. Yoon et al. 2006).

2 Building biological networks

The cells are the building blocks of all living organisms. Multicellular organisms contain cells specialized in different roles, which group together into different levels of organization to accomplish their biological functions. Biological organisms are therefore very complex and the identification and characterization of all their individual constitutive cells is not enough to understand how they function.

At a smaller scale, a parallel situation occurs with proteins, which are the functional units at the cellular level. Nowadays, we can identify and measure the expression levels of virtually all the proteins contained in a cell. We can obtain valuable information by analyzing this data with a **reductionist** approach, which attempts to explain the system behaviour in terms of their individual components. For instance, a reductionist procedure is usually applied to identify deregulated genes or proteins in diseases. However, this kind of approaches are not sufficient to understand global cell properties and functioning. In opposition to reductionism, **system biology** aims at understanding how a system operates by analyzing it as a whole, not just as a collection of parts. We should therefore look at multiple interacting components simultaneously to better discern the complex functioning of cells. A suitable way to do so, and hence to describe such biological systems is by using **networks**, as we described in chapter 1.

This chapter focuses on the description of large biological networks representing different types of interactions or functional relationships between genes and/or proteins. I first describe different types of interactions existing between genes/proteins. Then, I describe how to build biological networks from these interactions that are identified at large-scale thanks to different experimental techniques. In particular, I discuss about: *i)* networks of physical binary protein-protein interaction; *ii)* networks of proteins complexes; *iii)* networks of biological pathways; and *iv)* networks of co-expressed genes/proteins. These types of networks were my main working instruments to obtain the results exposed in chapters 4 and 5. Afterwards, I briefly comment about other types of biological networks that are relevant in computational biology. I end up the chapter describing the topological properties of biological networks.

It is to note that we aim at exploring and mining the vast amount of data contained in these large-scale networks. Nevertheless, large-scale networks are mainly static, and hence do not reflect the dynamics of the aforementioned biological interactions. A totally different set of approaches are developed in systems biology, to unveil the behaviour of a specific dynamic processes containing a limited number of interactions (e.g. a single pathway). These smaller networks can be modeled using Boolean dynamics, in which a node has two states (inactive or active), or using ordinary differential equations, in which a node has sev-

eral states representing different concentrations of genes/proteins. Up to date, these methods are not used on the same scale, and they are only applicable to networks of reduced size due to the huge amounts of parameters involved. I do not enter into details about network modelling because it is out of the scope of my thesis.

2.1 Biological interactions between genes and proteins

We can link the nodes of a network composed of genes and/or proteins through different types of edges. The edges may describe real physical interactions between the biological molecules, but also any kind of functional relationship. In this section, I briefly introduce the different types of interactions considered in the networks used during my PhD thesis. Noteworthy, I herein consider gene and protein nodes equally, unless otherwise specified.

Proteins interplay with each other and with other biological molecules through different types of interactions. These interactions are of varied nature and can be classified attending to several criteria. Let us consider two different proteins: protein A and protein B. We can first distinguish if the interaction between A and B entails a physical contact between them or not. The easiest interaction to picture is the direct **binary** contact between proteins A and B. However, A and B can also interact by being part of the same protein **complex**, or by participating to the same **pathway**. In a protein complex, A and B are not necessarily in direct binary contact, but they can interact through different proteins or other molecules, such as non-coding RNAs. It is said that A and B have a **co-complex** interaction in that case. In a pathway, like a set of metabolic reactions, the product of a reaction catalyzed by A can be the substrate of a subsequent reaction catalyzed by B. Therefore, A and B are not in direct physical contact in this situation neither.

We can further classify the interactions among proteins according to their chemical nature or their lifetime span:

1. The vast majority of interactions between proteins are established by **non-covalent bonds**. Non-covalent interactions are in most cases weak because they are based on hydrogen bonds, Van der Waals interactions, or hydrophobic bonds between residues of the interacting proteins. On the other hand, **Covalent bonds** mediate interactions with high affinity. Covalent interactions between proteins are uncommon, aside from some post-translational modifications as ubiquitination and SUMOylation. These post-translational modifications can activate binding sites or change its shape entailing new potential interactions (Westermarck et al. 2013; Akiva et al. 2015).
2. **Permanent** interactions involve proteins in contact for a long time, usu-

ally taking part of **stable** complexes that carry out well defined biological functions. On the contrary, proteins can shortly interact usually through non-covalent bonds. These ones are called **transient** interactions, and they are often occur in a reversible manner. The latter are the classical features of the interactions that occur in biochemical cascades, i.e. biological pathways (Nooren et al. 2003). These definitions can be extended to protein complexes: **permanent** or **stable complexes** and **transient complexes**. Halfway between transient and stable complexes, we can find the so called **fuzzy protein complexes**, which have more than one structural form. These complexes can adapt their structure and function depending on the cellular conditions (Fuxreiter 2012).

3. Proteins that are able to reach their stable folded structure on their own are called **non-obligate proteins**. These proteins interact with other non-obligate proteins to form **non-obligate protein complexes**. Other proteins get only to their functional folded structure when they interact with other proteins and molecules or when they are part of a complex. It means that these particular interactions stabilize the structure of the proteins involved. These proteins are called **obligate proteins** and they integrate molecular complexes that are known as **obligate protein complexes**. Generally, interactions between obligate proteins are prone to last longer in time, forming stable complexes, whereas interactions among non-obligate proteins are more likely to establish **transient** interactions (Nooren et al. 2003).
4. We can also classify pathways according to their ultimate biological objective. More details in this direction are provided in section 2.4.

We should be aware of this diversity of interactions because the different technologies dedicated to the identification of interactions are often prone to detect one class of interactions over others. Several different experimental techniques have been proposed to identify protein interactions at large-scale. The most popular ones are described in the sections 2.2 and 2.3.

It is to note that we are here mainly describing interactions between proteins, but these interactions often encompass other types of biological molecules. For instance, long non-coding RNA act as scaffolds for some protein complexes (Ribeiro et al. 2018). Another major class of interactions can be found in the genetic regulation where proteins, protein complexes, DNA and RNA molecules operate in a coordinated way. More precisely, the transcription factors bind to specif DNA sequences to activate or repress the expression of particular genes by promoting or blocking the recruitment of RNA polymerase, a multiprotein complex (Latchman 1997).

2.2 Networks of physical binary interactions

Historically, the most widely-used network in computational biology is the one composed of proteins which are linked by their physical contacts. They are usually referred as protein-protein interactions (PPIs) networks. The **interactome** is defined as the complete collection of all physical PPI that can take place within a cell (Cusick et al. 2005). The identification of the interactome is a crucial challenge in biology because molecular interactions between proteins are essential to almost every process in cells. The interactome can therefore help us to disclose protein and cell functioning.

2.2.1 Yeast two-hybrid screening

Yeast two-hybrid screening (Y2H) is an *in vivo* molecular biology technique used to discover **binary** PPIs in living yeast cells. It was first developed in 1989 to identify individual physical interaction between two proteins. Some years later, the Y2H method led to a revolution in the field since it can be easily automated for high-throughput studies of binary protein interactions at a genome-wide scale (Fields et al. 1989; Brückner et al. 2009). Y2H results are the major information source for some interaction databases (e.g. **IntAct** (Kerrien et al. 2012), **BioGRID** (Chatr-Aryamontri et al. 2017), see section 2.2.2), and have strongly contributed to the definition of the human interactome (Gandhi et al. 2006).

The Y2H screening systems are based on the modular properties of most eukaryotic transcription factors. A few years before the development of the Y2H technique, Keegan and coworkers identified two different functional domains of Gal4, a transcriptional activator in yeast: a DNA binding domain (DBD) and a transcriptional activation domain (AD). These domains preserve their function when Gal4 is split into two fragments (Keegan et al. 1986; Verschure et al. 2006). Fields and Song demonstrated that the transcription process can still be activated when the two domains are indirectly connected through the physical interaction of two other proteins. They could therefore assess the existence of an interaction between pairs of proteins by monitoring the expression of the transcripted gene (Fields et al. 1989). A scheme of the process is illustrated on Figure 2.1. The protein of interest X and the DBD of the transcription factor are fused into a construct called bait. On the other hand, a second component called prey is built by fusing the candidate interacting protein Y and the AD of the transcription factor. The interaction between bait and prey, if it takes place, results in a functional transcription factor, which is able to recruit RNA polymerase II. In this context, the reporter gene is transcripted and its expression reveals a physical interaction between prey and bait, confirming the interaction between X and Y proteins (Brückner et al. 2009).

Some of the overall key advantages of Y2H method are simplicity, time efficiency, low cost and high scalability (Struk et al. 2018). Indeed, Y2H can be

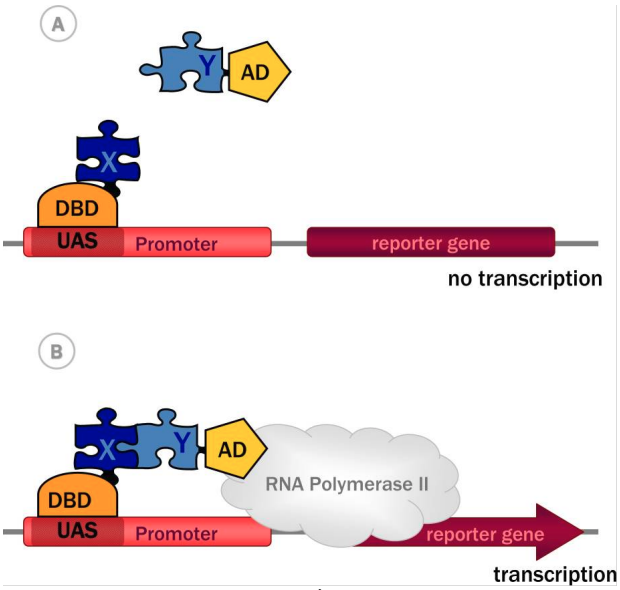


Figure 2.1 – A scheme of a Y2H assay. A) The bait is formed by the fusion of the protein X and the DBD of the transcription factor. It binds to the upstream activator sequence (UAS) of the promoter of the reporter gene. The prey is the result of merging the protein Y and the AD. B) The indirect connection between the two domains, mediated by the direct interaction between X and Y, activates the functional transcription factor giving rise to the recruitment of RNA polymerase II and the subsequent transcription of the reporter gene. Figure adapted from (Brückner et al. 2009).

used to screen a bait against a set of preys in a protein matrix, such as proteome-wide sets of full length open reading frames (ORFs) (Brückner et al. 2009). This technique is also sensitive to the detection of the challenging weak and transient interactions such as stress responses (Xiaoyun Liu et al. 2017), hormone signalling (Ryu et al. 2017) or immune signalling (Yang et al. 2018), to cite some recent works.

On the other hand, the main weaknesses of the Y2H assays is the potential generation of non-overlapping, non-reproducible results and thus an excess of false positives and false negatives (Auerbach et al. 2005; Deane et al. 2002). These limitations mainly emerge from the intrinsic setting of Y2H experiments. Indeed, even though Y2H is performed *in vivo* in the nucleus of yeast cells, many real contexts can not be reproduced and many interactions are systematically missed. For instance, some interactions occur after post-translational modifications in mammalian cells, which do not take place in yeast cells. In the same framework, Y2H assays significantly lack interactions involving membrane proteins, because once the AD of the transcription factor is fused to a membrane protein, it will be retained on the membrane avoiding its activation in the nucleus (Xia et al. 2006; Lalonde et al. 2008). In addition, Y2H can identify interactions that are possible from a biochemical point of view, but whose proteins are never co-expressed or co-localized (Koegl et al. 2007; Stynen et al. 2012). Despite these drawbacks, Y2H assays have successfully discovered thousands of protein interactions, fueling the creation of large interaction maps in diverse species. Moreover, many modifications and improvements of the initial Y2H technique have been implemented to reduce the impact of the aforementioned issues. For example, the membrane-based yeast two-hybrid (MbY2H) system allows detecting protein interactions directly at the membrane, based on the split-ubiquitin protein complementation assay (Johnsson et al. 1994; Lentze et al. 2008). A thoroughly review of the major methods applied for PPI identification can be found in the work of Rao and colleagues (Rao et al. 2014).

2.2.2 Databases containing PPIs

The massive amount of PPIs identified during the last years gave rise to the development of many databases collecting and storing this data in a structured way (Szklarczyk and L. J. Jensen 2015). The information contained in these PPI databases is usually diverse and they can be categorized in line with their sources:

1. *Primary dabases*: databases that collect primary experimental data such as **DIP** (Salwinski et al. 2004), **BIND** (G. D. Bader et al. 2001), **IntAct** (Kerrien et al. 2012), **BioGRID** (Chatr-Aryamontri et al. 2017), **HPRD** (Keshava Prasad et al. 2009) or **MINT** (Chatr-aryamontri et al. 2007) .
2. *Prediction dabases*: databases that besides experimental data include predicted interactions from *in silico* methods and text mining approaches (e.g.

STRING (Szklarczyk, Morris, et al. 2017), **GeneMANIA** (Warde-Farley et al. 2010) and **I2D** (Brown et al. 2005)) or predicted interactions alone, like **PIPs** (McDowall et al. 2009).

3. *Meta-databases* that result from the integration of the databases belonging to the previous categories. The most representative examples are **APID** (Alonso-López et al. 2016) and the **PINA** platform (M. J. Cowley et al. 2012).

We focus here in the primary databases because they are the ones we used in the forthcoming results sections 4 and 5. It is to note that the protein interactions contained in these databases not only originates from Y2H assays, but also from other techniques not detailed in this manuscript. In addition, these databases often contain other types of interactions between proteins besides binary interactions. Each database have its own philosophy, and tend to accumulate interactions detected with a given experimental or obtained under certain conditions. Some technologies are prone to detect one class of interactions over others. Overall, the information contained in these databases is highly valuable, but we should carefully extract and filter the data. We also have to be aware that PPI networks constructed from interactions stored in a specific database can be biased and generate misleading results (Lalonde et al. 2008).

The **IMEx** Consortium was created to alleviate this issue and ease the challenging integration of the data generated using different technologies and stored in different databases. It is a collaboration between public PPIs data providers who have agreed to improve the data quality and curation of molecular interaction. Some of the aforementioned databases, such as DIP (Salwinski et al. 2004), IntAct (Kerrien et al. 2012) and MINT (Chatr-aryamontri et al. 2007), are the core founders of IMEx. They have also proposed the Minimum Information about a Molecular Interaction Experiment (MIMIx) guidelines, defining a list of parameters to be supplied when describing molecular interactions in a publication (Orchard et al. 2007). The format has been broadly followed by molecular interaction databases, allowing the development of tools like **PSICQUIC**, which enables access to these resources through a standard Web Service and query language (del-Toro et al. 2013). Aiming to ensure reliable interactions, the PPI networks created in this work were built using PSICQUIC (see chapters 4 and 5).

2.2.3 Final remarks about PPI networks

Recent years have witnessed an exponential growth in the number of identified binary interactions between proteins due to the last advances in high-throughput technologies. We can therefore create large PPI maps for several species, including humans. These PPIs networks contain a vast amount of priceless information about biological functioning since proteins interact between them and with other molecules to perform virtually all the cellular processes. Consequently,

this kind of networks has been the most widely used in computational biology along the last last years. Nevertheless, we should bear in mind some additional considerations when using them.

Despite the large amount of PPIs accumulated in the aforementioned databases, the human interactome is far from complete. It is estimated that these technologies only cover around 20% of all potential pairwise protein interactions (Menche et al. 2015). It is also complicated to evaluate the extent to which any of the databases reveals the interactome. To give some numbers, the PPI used in chapters 4 and 5 has around $60\,000$ pairwise interactions among $12\,000$ proteins. Some studies have also suggested that the size of the interactome it is correlated with the biological complexity of the different species (Palmari et al. 2008).

In addition, we know that some technologies fail to detect interactions of some specific nature, like Y2H assays in the identification of interactions involving membrane proteins or depending on post-translational modifications mediated by enzymes (Lalonde et al. 2008). This lack of interactions of a certain type can also be reflected in the databases. Furthermore, databases are prone to present an over representation of interactions containing fashioned proteins, such as cancer related proteins, due to the unbalanced number of studies dealing with them during the last years.

It is therefore crucial to carefully inspect the different databases before building networks derived from them and extracting the information they contain. Indeed, none of databases is able to reflect the complexity of biological systems with different protein variants, modifications, and spatial and temporal dependencies (Szklarczyk and L. J. Jensen 2015). In our opinion, the integration of resources containing information of different nature can aid to reduce the incompleteness of individual sources and alleviate bias related issues. We therefore claim that other sources of information should be also consider to build biological networks. In this line, we next present networks of genes/proteins where the edges do not necessarily represent a direct physical contact between protein pairs (sections 2.3 and 2.4). We discuss about the integration of these networks in chapter 4.

2.3 Networks of proteins complexes

According to the data stored in the comprehensive yeast genome database, more than half of the yeast proteins participate in the formation of molecular complexes (Güldener et al. 2005). Some studies indicate that around 80% of the human proteins operate in complexes, and many proteins belong to several different complexes (Berggård et al. 2007). These complexes play critical roles in a large variety of biological processes (Giurgiu et al. 2018). Some well-known examples are the proteasome and the ribosome. Thus, not only the characterization of the interactome is essential, but also the description of the full set of

protein complexes functioning within cells, the (**complexome**).

2.3.1 Affinity purification coupled to mass spectrometry

In this section, we discuss about the leading technique to detect molecular complexes: Affinity Purification (AP) coupled to Mass Spectrometry (MS). Another popular approach combines biochemical co-fraction with quantitative mass spectrometry (Havugimana et al. 2012).

Affinity Purification coupled to Mass Spectrometry (AP-MS) is an *in vitro* approach to detect molecular complex and therefore to identify **co-complex** interactions among proteins (Gavin, Bösche, et al. 2002). Briefly, Affinity Purification is a popular purification technique which involves the fusion of a molecular marker, the tag (e.g. TAP, Strep-Tag, GFP, and c-Myc among others), to a protein under study. Mass spectrometry is a technique that ionizes chemical species and sorts the ions based on their mass-to-charge ratio.

In early 2000, tandem affinity purification coupled to mass spectrometry (TAP-MS) was the first approach of this type to purify and identify the interactions of an specific target protein. In the first successful approach, the TAP tag consisted of two IgG binding domains of *Staphylococcus aureus* protein A (ProtA) and a calmodulin binding peptide (CBP) separated by a TEV protease cleavage site. The construct, which we can call bait like in the Y2H technique, is produced under physiological conditions and introduced into the host cell or organism. A group of proteins (prey proteins) will attach to their natural interacting partner, the initial bait protein, to form a molecular complex. Afterwards, ProtA binds firmly to an IgG matrix and the whole complex is "fished out" (Puig et al. 2001). Subsequently, the complexes are purified and separated into its components according to their size. Then, the TAP process can be coupled with MS to detect the unknown prey proteins (Figure 2.2). In 2002, (Gavin, Bösche, et al. 2002) first developed this method in a high-throughput manner for systematic identification of protein complexes in *Saccharomyces cerevisiae*.

A high-throughput AP-MS assay requires a large number of purification experiments, each one of them aiming at identifying the prey proteins attached to the initial bait protein to form a molecular complex. The purification results should not be considered as a relevant molecular complex directly due to the presence of experimental noise, spurious interactions or the participation of the bait in different complexes. For this reason, several computational methods have been developed to score the protein interactions within the purification results and infer the real complexes (Krogan et al. 2006; Gavin, Aloy, et al. 2006; Friedel et al. 2009).

The main advantage of AP-MS methods is that identify interactions occurring in their native cell environment under near-physiological conditions (Van Leene et al. 2014). Therefore, unlike Y2H assays, TAP-MS allows to detect the interaction between proteins after their post-translational modifications (Cusick et al.

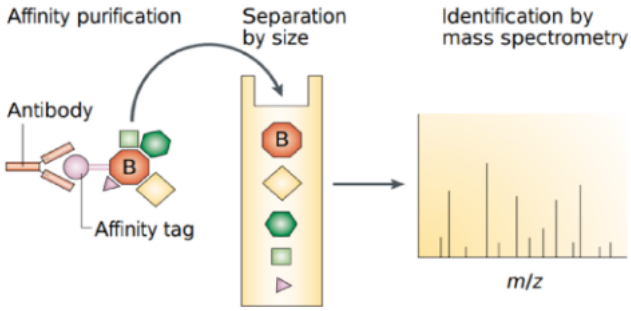


Figure 2.2 – A scheme of a AP-MS assay. The bait protein B is tagged and then introduced into a cell or an organism where their protein partners attach to it. The tag binds to a molecule (an antibody in this particular case) allowing to retrieve the whole complex, which is afterwards separated into its components. Then, the prey proteins are identified by a coupled MS technique. Figure adapted from (Grünenfelder et al. 2002).

2005). Moreover, some AP procedures are applicable to membrane proteins (Q. Li et al. 2004).

AP-MS also have some weaknesses. The tag of a protein can result in non-desirable interactions leading to its non-native folding. Furthermore, the over expression of the bait protein is usually required during the tagging, possibly influencing the stoichiometry of the complex (Struk et al. 2018). Then, the correct tag choices is a crucial step in the preparation of every AP-MS assay. Another drawback is the potential loss of interactions (or even gain of spurious ones) during the purification of the complexes (Rao et al. 2014). However, the last years have witnessed a considerable improvement in the affinity-based methods thanks to new purification techniques along with the increased sensitivity of the mass spectrometry techniques (Lackner et al. 2015; Qu et al. 2016).

I would like to note the fact that co-complex methods, due to their nature, are not able to discern between direct and indirect physical protein interactions. Consequently, there are different approaches to build a protein interaction network from this kind of large-scale data. Additional methods are presented in the review of Rao and co-workers (Rao et al. 2014).

2.3.2 Construction of protein networks from co-complex interactions

AP-MS assays identify molecular complexes containing several proteins. These type of proteins interactions are called co-complex because the proteins belong to the same molecular complex. In this context, co-complexes interactions include direct physical interactions (proteins in direct contact through a binding

interface) and indirect interactions (proteins belonging to the same complex, but without physical contact) (Schelhorn et al. 2011; X.-F. Zhang et al. 2015). Hence, unlike in the results of Y2H assays, a model is needed to translate the experimental group-based observations into pairwise protein interactions (de Las Rivas et al. 2010). The two major procedures are the **matrix model** and the **spoke-based model** (Gary D. Bader et al. 2002; Mering et al. 2002), which are schematically illustrated in Figure 2.3.

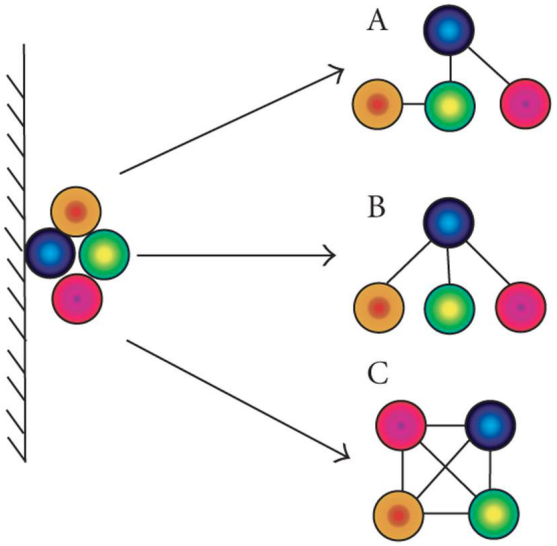


Figure 2.3 – Models to transform protein complexes to binary interactions. The balls represent polypeptide chains within a protein complex, the blue one being the bait protein. Lines are the pairwise interaction between the proteins. A) Unknown real topology of a protein complex. B) **Spoke model**, the bait interacts with each captured prey. (C) **Matrix model**, all possible interacting pairs within the complex are considered. Figure adapted from (Hakes, Robertson, Oliver, and Lovell 2007).

The choice of the model has an extraordinary impact on the topology of the network. However, the model selection is not evident and it should be carefully analyzed based on the available interaction data, and in accordance with the ultimate application of the network (Hakes, Robertson, and Oliver 2005). The network of complexes used in the results chapter 5.3 was generated applying the matrix model on data retrieved from the databases detailed in the next section 2.3.3. Our network of complexes contains $91\,502$ pairwise interactions between 8650 proteins.

2.3.3 Databases containing protein complexes

The **CORUM** database was created to collect the composition, cellular function, disease association and stoichiometry of hundreds of mammalian proteins complexes that had been analyzed in individual experiments (Ruepp et al. 2008). *CORUM* aims at providing a high-quality and carefully curated dataset, adding each complex to their records individually after a meticulous review of the associated literature. Moreover, high-throughput results are excluded since they can potentially contain a significant fraction of false-positives (Ruepp et al. 2008). The latest release of *CORUM* contains 4274 mammalian protein complexes, with a large predominant of human complexes (67%) (Giurgiu et al. 2018).

More recently, the Human Protein Complex Map (**hu.MAP**) was released (Drew et al. 2017). They constructed a comprehensive global map of human protein complexes by re-analyzing three independent large-scale human protein complex mass spectrometry experimental datasets (Huttlin et al. 2015; Hein et al. 2015; Wan et al. 2015). These three experimental setups explore the landscape of the human complexes by means of distinct methods (TAP-MS and co-fractionation-MS) in different samples (diverse cells and tissues) and applying different approaches for the affinity-tagged bait proteins. Consequently, the results of these three experiments are highly complementary in several facets and they do not show a significant overlapping panorama of the human complexome. In this context, *hu.MAP* developers built a protein complex discovery pipeline based on machine learning techniques, and trained with these complementary sets of protein complexes. The result was a unique exhaustive map of over 4600 human protein complexes (Drew et al. 2017).

2.3.4 Final remarks about networks of protein complexes

As in the case of binary physical interactions, we should bear in mind the different types of molecular complexes described on section 2.1. Some experimental techniques may display a preferential ability to detect some specific type of complexes. For instance, the detection of transient complexes is usually more challenging than the detection of stable ones in AP-MS assays (Perkins et al. 2010). In addition, we need to choose a model to transform the molecular complexes into a network of pairwise interactions between proteins. The properties of the resulting network will be quite different depending on this choice.

Noteworthy, we described in separated sections the networks constructed from binary protein interactions and the networks inferred from protein complexes. In the literature, binary interactions are sometimes treated along with co-complex interactions, and they are both together used to build protein-protein interaction networks. As we already described, these interactions are essentially of different nature, and they are detected at large-scale by very different technologies. Therefore, these two type of protein interactions are mainly not

overlapping and they provide complementary information. These are the main reasons that lead us to construct two different networks in the results presented in section 5.3. Accordingly, we used a matrix model to build our network of complexes using *CORUM* and *hu.MAP* data. Proceeding this way we aimed at reinforcing the idea of protein complex in opposition to binary protein interactions (see figure 2.3).

2.4 Networks of pathways

A biological pathway is a cascade of reactions leading to a certain resulting product or entailing a change/response in a cell. Pathways include biochemical reactions, transport processes and catalysis events (Cerami et al. 2011). Some of the most representative examples are the ordered series of chemical reactions transforming biological substances (metabolism), and the transmission of a signal to respond to an external situation (e.g. repairing the damage caused by a wound). The correct functioning of biological pathways is key for cell survival. They provide complementary information about protein functioning than the one disclosed by the interactions presented in the previous sections, because the nature and source of the interactions is different.

2.4.1 Types of biological pathways

The most common broad classification of biological pathways is detailed below:

1. *Metabolic pathways*: Metabolism is the set of life-sustaining chemical reactions that occur in all living cells of every organism. A metabolic pathway is any of the sequences of enzyme-mediated biochemical reactions that lead to biosynthesis (anabolism) or breakdown (catabolism) of biomolecules within a cell. These reactions are connected by their intermediates: the products of one reaction are usually the substrates for subsequent reactions, and so on (Schilling et al. 2000). Therefore, a metabolic pathway can be represented as a directed network of proteins (enzymes) which are not in direct physical contact. Some of the most relevant and best studied metabolic pathways are the glycolysis (Figure 2.4) and the Krebs cycle.
2. *Signal transduction pathways*: Signal transduction is the process of transferring a signal throughout an organism resulting in a cellular response. The sequence of these biochemical events, from the reception of a stimulus to the final cell reaction, constitutes a signaling pathway. The majority of signal transduction pathways involve the binding of signaling molecules, known as ligands, to membrane receptors that trigger events inside the cell. The nodes of a signaling network can be classified according to their function with regard of the initial stimulus. *First messengers* is the most

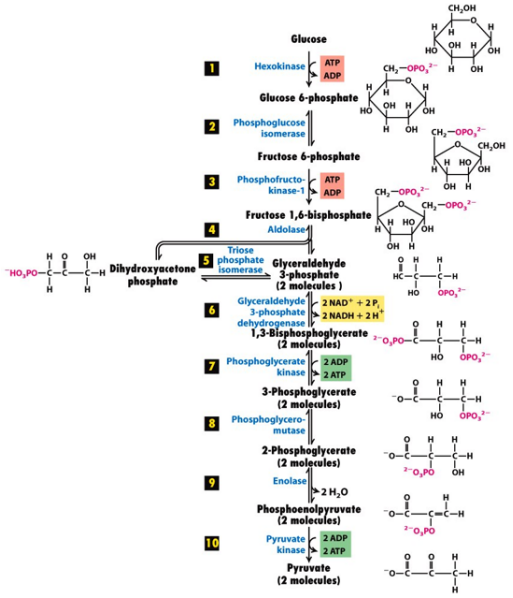


Figure 2.4 – The glycolysis pathway. Glucose is degraded to pyruvic acid following a sequence of enzyme-catalysed reactions. Several of these reactions are reversible (double arrows) and participate in the gluconeogenesis, the synthesis of glucose. Figure adapted from (Lodish 2016).

common way to name the ligands, whereas receptors are called *signal transducers*. They activate others membrane-bound proteins, the *primary effectors*, that provoke the release of molecules accomplishing intracellular signal transduction within the cytoplasm, *the second messengers*. The second messengers can trigger the final cellular response, such as gene activation or metabolism alterations. A classical example is the release of hormones from endocrine glands, such as the thyroid or pancreas. Some cells have specific receptors for the hormones, which trigger the activity of different pathways upon receiving a signal. For instance, the insulin can induce muscle cells to uptake and store glucose, whereas it will also force liver cells to cease producing it (Bradshaw et al. 2010; Lodish 2016; Yachie-Kinoshita et al. 2018).

These categories of pathways have in common the sequential and ordered nature of the involved reactions, thus they are usually described as directed networks. On the other hand, some of their topological features can be quite diverse. For instance, from an evolutionary point of view, metabolic pathways, especially for intermediary metabolism, are usually well conserved along the tree of life from mammals to bacteria whereas signaling pathways are quite divergent (Kanehisa and Goto 2000). This lead to significant differences in some of their properties, such as the network diameter and the clustering coefficient (see section 1.5 and 1.7). Collections of pathways of different types are stored in dedicated databases.

2.4.2 Databases containing collections of biological pathways

Pathway databases emerged with the purpose of providing an integrated view of well-established information about biological processes. Data stored in these databases is usually carefully curated and reviewed by experts from the relevant field on a regular basis (Joshi-Tope et al. 2005; Szklarczyk and L. J. Jensen 2015). There are two leading broadly-used resources, namely **KEGG** (Kanehisa, Furumichi, et al. 2017) and **Reactome** (Fabregat et al. 2018).

KEGG is the pathways database spanning the largest number of natural species with over 1500 different species. It contains a vast collection of manually drawn pathway maps representing well-established knowledge on the molecular interactions, reactions and circuits for cellular processes ranging from metabolism to pathways disturbed on human diseases. Each curated map aims at being applicable for many species. To this goal, every pathway have nodes grouping genes of different species This implementation makes KEGG database a very powerful tool for evolutionary comparative studies, because the maps reflect the areas of the pathway existing in a particular species, allowing a comparison between the generic and the specie specific pathways (Szklarczyk and L. J. Jensen 2015;

Kanehisa, Furumichi, et al. 2017).

Unlike KEGG, **The Reactome Knowledgebase** is mainly focused on a single organism, *Homo sapiens*. In Reactome, biological processes, such as signal transduction or protein synthesis, are systematically characterized in molecular detail to create an ordered network of biomolecular transformations. Each interaction is carefully annotated by experts, including directionality, type, localization, substrates, stoichiometry and known associated diseases. Moreover, human proteins are linked to their molecular functions to provide a global view of the maps. Reactome provides an in-depth description for each one of its 2244 annotated pathways involving more than $12\,000$ reactions among $10\,778$ different proteins (Szklarczyk and L. J. Jensen 2015; Fabregat et al. 2018).

There are other relevant pathways databases that are worthy to be, at least, briefly detailed:

1. **WikiPathways** is a community curated resource dedicated to biological pathways, and mainly focused in human and model organisms. Contributions are monitored by a group of supervisors (Kutmon et al. 2016).
2. **MetaCyc** is a curated database based on experimentally determined metabolic pathways. It contains 2666 pathways from 2960 different organisms.
3. **PANTHER** Pathways stores around 180 well-defined diagrams mainly focusing on signaling pathways (Mi et al. 2017).
4. **Netpath** is a database of curated human signaling pathways. It is centered on signaling cascades of the immune system (Kandasamy et al. 2010).
5. The Network Data Exchange (**NDEx**) is software platform that eases the sharing and publication of networks among reseraches. In addition, data from former relevant pathway databases, such as **NCI Pathway Interaction Database** (Schaefer et al. 2009), is now stored in NDEx (Pratt et al. 2015).
6. **Pathway Commons** is a collection of publicly available data from multiple organisms. It contains resources coming from both PPIs and pathways databases and it is focused on providing an easy-to-use web-based interface (Cerami et al. 2011). Pathway Commons data is also availabe in NDEx.
7. **ConsensusPathDB** integrates human interaction networks including binary and complex protein-protein, genetic, metabolic, signaling, gene regulatory as well as biochemical pathways. Data is retrieved from 32 public resources (Kamburov et al. 2013).

The information accumulated in these databases is rapidly growing. Nevertheless, the way these pathways are stored and displayed is very diverse calling for the development of integrative methods. In this framework, the Biological Pathway Exchange (**BioPAX**) project provides a standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data

from heterogeneous sources. BioPAX is defined in Web Ontology Language (OWL) and is represented in the RDF/XML format. Some of the aforementioned databases, such as Reactome and Netpath, offer BioPAX format (Demir et al. 2010).

2.4.3 Construction of networks from pathways sources

The databases previously described contain thousands of different pathways. This is a very valuable information about cell and protein functioning, complementary to the one provided by physical binary interactions and molecular complexes. It is therefore very interesting to fetch the knowledge stored in pathways databases and convert it into networks.

There are two main approaches to analyze these networks. The first one is based on retrieving all the pathways stored in one of the databases, or in several ones, and construct a large graph consisting on thousands of protein nodes linked by their reactions. The resulting graph contains a massive amount of biological information that can be highly valuable when mined using graph-theory based methods. This is the approach I followed during my thesis. To provide some details, we retrieved data from KEGG, Biocarta, NCI, Reactome and Panther databases. We built a network containing around $254\,766$ edges between $10\,534$ proteins (see chapters 4 and 5). Even though a pathway is a set of sequential reactions, we constructed an undirected network be aligned with the previously presented networks. The second approach is based in modeling a small set of reactions to reflect their dynamics, something that the large undirected networks are not able to do. We already further discussed about this point in the section 2.1.

2.5 Co-expression networks

Butte and Kohane constructed the first gene co-expression network from a public data set of 79 different RNA expression measurements for 2467 yeast genes. They computed the mutual information between RNA expression patterns for each pair of genes. Then, they set different thresholds to determine if two genes were co-expressed and hence linked by an edge in their network. Based on the hypothesis that genes showing similar expression patterns are related biologically (i.e. tightly connected in their network), they applied a clustering method on the co-expression network aiming at revealing groups of genes involved in similar functions (Butte et al. 2000).

Indeed, posteriors studies have demonstrated the interest of gene co-expression networks due to the fact that co-expressed genes tend to be controlled by the same transcriptional regulatory mechanism, functionally related, or members of the same pathway or protein complex (Wolfe et al. 2005; Weirauch 2011).

Recent improvements along with reduction of cost in transcriptomic technologies have favored a vast amount of studies dealing with gene expression levels under different conditions in numerous organism. Therefore, large co-expression networks can be built from this valuable source of information. Gene co-expression can supply complementary data or strengthen the information provided by the other networks described in this chapter, as described in section 2.1. In chapter 4, we show that the integration of a co-expression network and a PPI network performs better than these networks alone in the task of recovering gene-disease associations. .

2.5.1 Construction of gene co-expression networks

The overall strategy to construct gene co-expression networks have not substantially changed from the first one employed by Butte and Kohane (Butte et al. 2000). It usually involves two steps:

1. *Calculation of a similarity score:* the expression levels of genes under different conditions are usually log_2 transformed before calculating the similarity score in order to scale the values in the same range. Then, the score between gene pairs can be determined by applying different measures. The most straightforward method is computing the Euclidean distance, which measures the geometric distances between expression values. However, it has been shown that is not a very convenient approach under some conditions (Dehmer 2011). The most widely-used approaches are **Pearson correlation** and **Spearman's rank** that perform well in comparison with more computationally demanding methods like **mutual information** (L. Song et al. 2012; Ballouz et al. 2015). The Pearson correlation is the most popular method, although it assumes that the expression data follow a normal distribution and it can only detect linear relationships. Furthermore, it is sensitive to outliers, while Spearman's rank is more robust, but remarkably less powerful (Serin et al. 2016). The Pearson's correlation coefficient returns values between -1 and 1, allowing not only detect co-expressed genes, but also genes expressed in opposing directions.

   Other methods have been proposed as an alternative for Pearson's correlation, like "**bi-weight mid-correlation**" (bicor), whose authors claim to be more robust than Pearson correlation and to outperform mutual information in terms of elucidating gene pairwise relationships (L. Song et al. 2012)

2. *Significance threshold choice:* once the previous step has been accomplished, a cutoff is applied to generate the edges between genes in the co-expression network. There are several ways to choose this cutoff, the most basic one being the selection of an arbitrary cutoff and the generation of edges between genes with similarity scores above this cutoff value. Other methods compute a z-score for each similarity score based on the

number of samples. This z-score is then converted into a p-value. Different approaches calculate a p-value based on permutations of the expression values (Butte et al. 2000; Dehmer 2011). Afterwards, a classical p-value cutoff is applied (e.g. a p-value of 0.01 or 0.05), defining as significant some scores that become the network edges.

These methods are based on statistical routines, but they ignore the properties of the biological mechanisms controlling gene expression, gene regulatory pathways (see section 2.6.1). One of the prevalent protocols is to choose the cutoff based on Weighted Gene Coexpression Network Analysis (WGCNA). This method selects the threshold resulting in the network which best fits a power-law distribution, which is a key feature of the topology of biological networks (see section 2.8.1). According to their developers, the WGCNA method is able to capture more cohesive and biological relevant modules of genes than the purely statistical methods (Bin et al. 2005; Bassel et al. 2011).

Co-expression networks have demonstrated to provide a very useful source of biological information. However, we should keep in mind some considerations when working with co-expression networks. The major one is our goal of extracting some kind of reliable biological knowledge from the network, which it is often hindered when using large data sets inferring too noisy co-expression networks (Usadel et al. 2009). In addition, co-expression networks provide a static and an undirected description of their underlying process, gene regulatory pathways, that are highly dynamic and presents an oriented nature.

In the results presented in chapters 4 and 5, we used a co-expression network derived from RNA-Seq expression data publicly available on the Human Protein Atlas (Uhlen et al. 2015) (see section 2.5.2). We computed Spearman's correlation of gene expression data from many different tissues and cell lines, and selected the similarity scores above 0.7 to be included in the network. The result was a network of around $1\,000\,000$ edges between $10\,000$ genes. Publicly available resources containing gene expression data, as the Human Protein Atlas, are a valuable resource to build co-expression networks. They allow inspecting gene expression data for different tissues, different cell lines and under many different conditions (e.g. healthy versus diseased, control versus drug treatment). Moreover, any researcher can access these data and therefore create co-expression networks without conducting biological experiments.

2.5.2 Public gene expression data sources

In order to integrate the massive amount of transcriptomic data generated in different experimental designs, an international effort was made to define the minimal set of necessary information for microarray experiments to be properly interpreted and to be comparable with one another (Kellam 2001). This led to the

development of publicly available gene expression databases derived from microarray experiments, which are still frequently employed because they are relatively cheap and their analysis is highly standardized (Serin et al. 2016). Comprehensive microarray gene expression sets generated in roughly every published study are uploaded to public repositories, such as the Gene Expression Omnibus (GEO) (Edgar et al. 2002), Array Express (Brazma et al. 2003) or Genevestigator (Hruz et al. 2008).

More recently, the Human Protein Atlas project presented a publicly available map of the human proteome in cells, tissues and organs using various *-omics* technologies, including antibody-based imaging, mass spectrometry-based *proteomics* and quantitative *transcriptomics* (Uhlen et al. 2015). This project takes advantage of RNA-sequencing, using next-generation high-throughput sequencing technologies (RNA-Seq). The main advantage of RNA-Seq over the microarray platforms is its ability to cover the entire transcriptome, allowing therefore to unveil more complete representations of the underlying gene regulatory networks (Serin et al. 2016). RNA-Seq is superior in detecting low abundance transcripts, discriminating between different isoforms and allowing the identification of genetic variants. RNA-Seq also eludes the technical issues derived from the microarray pre-designed sequence detection probe (S. Zhao et al. 2014).

2.6 Other biological networks

The networks described above are the most common and widely used in computational biology, and are also the ones that I have handled the most during my thesis (see chapters 4 and 5). In those networks, the nodes are either genes or proteins, which are usually considered equally. From a biological point of view, the PPI network describes binary physical interactions between proteins. Nevertheless, their associated graph often comprises gene nodes because we lack information to decipher the specific gene product (i.e. protein isoform) that is actually interacting. Co-expression networks are built from *transcriptomics* data that is based on the measure of the genes mRNA transcripts, but again the most relevant information is the final protein expression levels.

I detail below other types of biological networks not described so far. For the sake of simplicity, I separated them into two sections: *i)* other biological networks where the nodes are also proteins or genes, like in the previously defined networks; and *ii)* other biological networks where the nodes are other biological entities.

2.6.1 Other networks of genes or proteins

A large amount of different types of networks can be build by linking the genes or proteins with edges of diverse nature. For instance, networks connecting

genes with shared Gene Ontology (GO) annotations reflect functional relations (Biological process: GO-BP) or cellular co-localization (cellular component: GO-CC) (Yongjin Li and J. Li 2012; Himmelstein et al. 2015). We can also infer functionally relevant networks by associating genes that are regulated by the same transcription factor, or are targets of the same drug (Himmelstein et al. 2015; Lin et al. 2017). Similarly, genes having a high degree of sequence similarity or genes sharing some kind of protein domain (i.e. protein domain co-occurrence networks) can also be functionally related (Atkinson et al. 2009; Z. Wang et al. 2011).

Conceptually more complex networks of genes and proteins can also be constructed. As an illustration, we can construct networks based on the premise that genes that co-evolve are likely to be functionally related, since those genes are often gained and lost together through molecular evolution (Y. Li et al. 2018). Another example can be provided by networks relating genes which produce some kind of response on cells (e.g. genes which that essential for the proliferation of different cancer cell lines) (G. S. Cowley et al. 2014; Marbach 2018).

2.6.2 Other networks containing different nodes

The nodes of all the networks I described so far are genes or proteins. However, we can build biological networks describing relations among other type of entities, i.e. networks where the nodes are not genes or proteins.

Networks of microRNA and Long non-coding RNA have gained a growing interest during the past few years, as more functions are disclosed for these transcripts. The edges connecting these molecules can arise from direct physical contacts among them or from functional relationships, such as the sequence similarity or interaction with the same protein or DNA regions. The construction of co-expression networks among these non coding RNA is also a fairly widespread procedure (Yuan et al. 2014; Meng et al. 2015; Chou et al. 2018; Do et al. 2018).

Another type of very appealing networks are those where the nodes represent different drugs. We can establish edges among these drugs based on different criteria. For instance, we can link drugs according to their structural similarity or concerning common properties of their target proteins. Also we can set drug–drug relations based on the influence, whether positive or negative, of one drug upon another (Brody 2018). Drug networks are widely used in the context of drug repurposing, network pharmacology or drugs synergies (Udrescu et al. 2016; Takeda et al. 2017; Lotfi Shahreza et al. 2018; H. Li et al. 2018) (see section 3.2.5). Furthermore, the co-occurrence of drugs can be applied to construct a network of similar adverse drug reactions (Xiaowen Chen et al. 2016).

There is also a considerable scientific literature about interaction networks among the amino acid residues that form the proteins. The analysis of the network topology can reveal the most relevant residues on particular proteins, which is of large interest in molecular evolution, protein docking and pharmacology

(Amitai et al. 2004; Rockah-Shmuel et al. 2015; Viswanathan et al. 2015) (Figure 2.5). However, this kind of approaches are usually restricted to a limited number of resolved protein structures. The use of this interaction networks can complement other existing methods in the prediction of protein structure and protein-protein binding interfaces (Viswanathan et al. 2015).

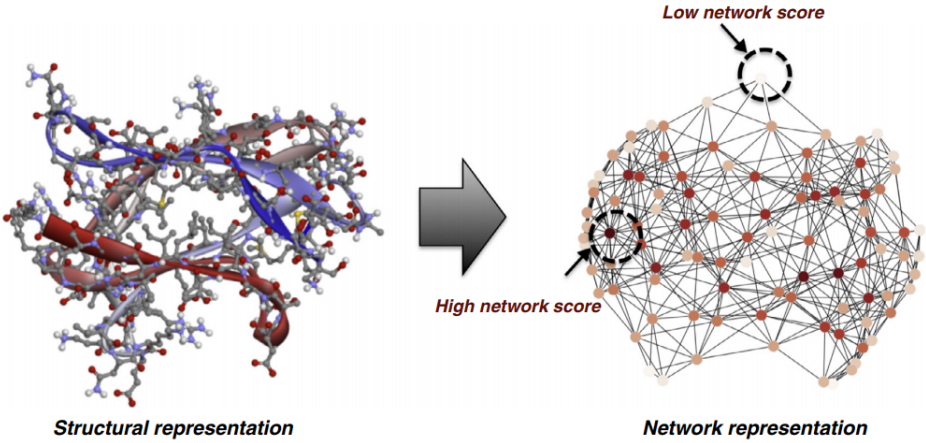


Figure 2.5 – Structural (left) and Network representation (right) of a domain of the dengue envelope protein. The intensity of the red color in the network denotes the relevance of each node within the network according to a defined topological-based score (for instance,, it can be computed based on different centrality measures). Figure adapted from (Viswanathan et al. 2015).

Several studies have generated networks connecting human diseases if they share some kind of molecular relationship or they have common clinical symptoms (Goh et al. 2007; X. Zhou et al. 2014). In this framework, shared phenotypic features among diseases may underlay a common altered biological process at the molecular level. In the literature, we can also find disease networks where the edges are derived from comorbidity associations (i.e. diseases that are more likely to co-occur than statistically expected). Comorbidity can arise from multiple factors, but again, a network analysis may reveal their underlying common molecular basis (Hidalgo et al. 2009; Khan et al. 2018). During my thesis, I constructed a disease-disease similarity network by connecting diseases that share relevant phenotypes (see section 4.3).

2.7 Integration of networks

I provide here some brief notes about network integration that is extended in chapter 4. The networks described in the present chapter can be combined into different types of networks of networks:

1. **Aggregated networks**: the different interactions of a group of networks with common nodes can be merged into an individual network. In this union of networks, a pair of nodes can be linked by two edges. Aggregated networks are hence multigraphs. This approach is not very suitable in some cases because the topology of the initial networks may be missed.
2. **Multiplex networks**: collections of networks sharing the same nodes, but in which the edges belong to different categories. For instance, I integrated the gene/protein networks described in sections 2.2, 2.3, 2.4 and 2.5 into a multiplex network (see chapters 4 and 5). Some studies have also built multiplex networks of diseases (Halu et al. 2017). In opposition to aggregated networks, multiplex networks keep track of the topology of the initial networks. We showed that multiplex networks outperform aggregated networks in the gene-disease association task (section 4.3).
3. **Heterogeneous networks**: they contain two networks with different types of nodes and edges, as well as a bipartite graph detailing associations between nodes of distinct nature of the two previous networks. To give some examples, a PPI network and a disease-disease network were integrated thanks to bipartite edges inferred from disease-causative genes, in order to predict new gene-disease associations (Yongjin Li and Patra 2010). Similar approaches can be found with drug-target networks (Xing Chen, M.-X. Liu, et al. 2012), or lncRNA-disease networks (Sun et al. 2014).

In chapter 4, I describe how I integrated a multiplex network of genes/proteins with a disease-disease similarity network to create a network both multiplex and heterogeneous. These multiplex-heterogeneous network was used to predict new gene-disease associations (see chapter 4) and to define the functional landscape of a group of premature aging diseases (see chapter 5).

2.8 Topological properties of biological networks

The previously described biological networks have different characteristics both inherent to the nature of the processes they describe, and emerging from the particular way in which they are constructed. Nevertheless, it has been demonstrated that the laws governing the formation and evolution of real-world networks, ranging from social to protein networks, are shared to a large extent (Strogatz 2001; N. et al. 2013). These results have boosted the utilization of network theory into the field of biology during the last years (A. L. Barabási and Oltvai 2004). In this section, we aim at describing these common topological properties in the context of networks whose nodes are genes or proteins.

2.8.1 Scale-free networks

Until early 90s, networks of real-world systems were usually described following the **random network model** stated by Paul Erdös and Alfréd Rényi. According to this model, all the graphs containing a specific number of nodes and edges are equally likely. In other words, networks of a given size and order are generated by placing random links among them with a previously defined probability (Erdös et al. 1960). The degree distribution (see section 1.7) of these networks follows a Poisson distribution, indicating that the majority of nodes have roughly the same number of edges (Figure 2.6 A).

Later works showed that the random network model fails to explain the topological properties of real-world networks. In 1999, Barabási and Albert proposed the **scale-free network model** in opposition to the random network model. They demonstrated that the degree distribution of networks describing many real-world systems follows a power law, instead of a Poisson distribution. According to this model, the probability that a given node has a degree k is $P(k) \sim k^{-\gamma}$, where γ is the degree exponent, whose value is $\gamma \in [2, 3]$ in most of the real networks (A.-L. Barabási and Albert 1999). Scale-free networks are non-uniform, meaning that some nodes count with a large number of edges, which are called **hubs**, while others are poorly connected (Figure 2.6 B) (A. L. Barabási and Oltvai 2004).

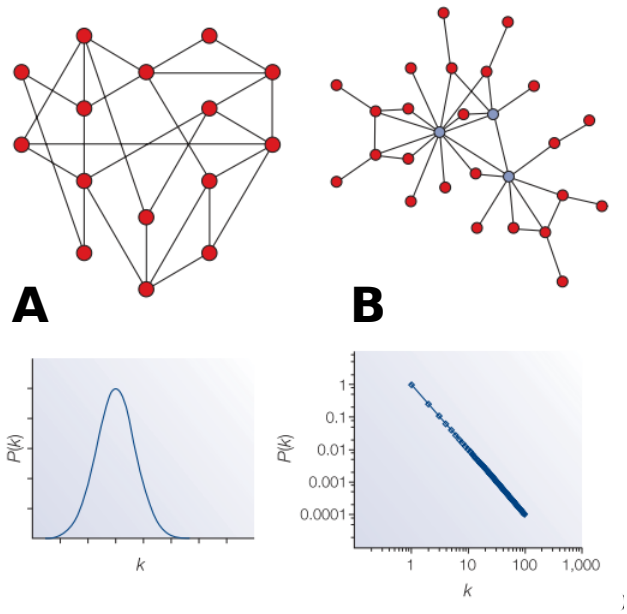


Figure 2.6 – Schematic examples of **(A)** a random network and **(B)** a scale-free network, along with their respective degree distributions. Figure adapted from (A. L. Barabási and Oltvai 2004).

A large number of mechanisms within cells seem to be governed by scale-free networks. For instance, the presence of very promiscuous proteins interacting with a significantly greater number of partners than the average degree in PPI networks, suggests the existence of protein hubs (Vallabhajosyula et al. 2009; Vidal et al. 2011). This situation is even accentuated in metabolic pathways, where a scarce number of metabolic substrates participate in many reactions and function as metabolic hubs, such as co-enzyme A which is used by around 4% of cellular enzymes and is present in every living organism (Daugherty et al. 2002). On the other hand, most metabolic substrates participate in a reduced number of biochemical reactions. The gene co-expression networks of organisms spanning from bacteria to human display scale-free distributions (E. Gross 2012).

However, the scale-free model does not apply for all biological networks. The topology of the networks associated with genetic regulatory pathways is better described by a mixed model. A reduced number of TFs regulate hundreds of genes, whereas the number of genes simultaneously regulated by a many TFs is almost negligible. It suggests that, in this case, the outgoing degree distribution is scale-free, but the incoming degree distribution fits better into an exponential distribution (A. L. Barabási and Oltvai 2004; Deplancke et al. 2006; Vidal et al. 2011).

I discuss in the following points about further properties of biological networks emerging from the scale-free topology. Then, I talk about some studies aiming at finding an explanation to this set of topological properties from the optics of evolutionary biology. Finally, I comment on a recent publication that brought an intense debate about the scale-free nature of real-world networks.

2.8.1.1 Small-world networks

The **small-world** property can be easily illustrated with the six degree of separation concept (Milgram 1967). According to this idea, a maximum of six steps is needed to reach every node in a network where all the human beings in the world are connected by friendship relations (knowing someone "personally") (Milgram 1967; Guare 1990). More formally, the expected distance between two randomly selected nodes increases proportionally to the logarithm of the number of nodes in small-world networks. Consequently, they tend to have a significant smaller mean-shortest path length (see section 1.5) than random networks. Small-world networks have higher clustering coefficients (see section 1.7) and an over-abundance of hubs (Watts et al. 1998).

The mean-shortest path length of scale-free networks is even shorter than the one predicted by the small-world effect. Therefore, we usually describe scale-free networks as ultra small-world (Cohen et al. 2003). This effect can be clearly seen in the metabolic networks, where a few reactions can link all the metabolites (Wagner et al. 2001).

2.8.1.2 Modules and motifs

Scale-free networks show a large clustering coefficient leading to the community structure usually present in biological networks.It means that groups of nodes are more densely connected internally than with the rest of the network (see section 1.6). These groups of tightly connected entities, usually called **modules**, contain proteins likely to be involved in the same biological functions (Hartwell et al. 1999). The accurate extraction of these modules is decisive in a bunch of fields. For instance, it can reveal modules associated with a given disease, and therefore provide clues about the deregulated biological processes on that disease (Marbach 2018). The metabolic network provides again a remarkable evidence of the high clustering coefficient displayed on biological networks. In this case, the high clustering coefficient is the same for metabolic networks of the different species, and therefore independent of the network size (Ravasz et al. 2002).

As defined in section 1.6, network **motifs** are characteristic subgraphs that appear significantly more frequently than on randomized versions of the same network (networks with the same set of nodes and an equal degree distribution, but where the edges are placed aleatory) (Milo et al. 2002). Real-world networks can be characterized by its set of over-represented motifs. For instance, some studies have revealed the evolutionary conservation of proteins involved in motifs in PPIs. For instance in the yeast protein interaction network, a study has shown that the conservation is particularly significant for those proteins participating in motifs that define a fully connected subgraph of four and five nodes (Wuchty et al. 2003). These findings suggest that motifs may represent evolutionary conserved topological units of biological networks (Wuchty et al. 2003; Shoval et al. 2010; Vidal et al. 2011).

2.8.1.3 Network robustness and hubs

Network robustness is an essential attribute of real-world networks; it can be defined as the capacity of resisting against failures and perturbations. We can easily notice the robustness of biological networks. For example, metabolic pathways are able to adapt to nutrient scarcity while maintaining their key activities for cell survival. The analysis of robustness in networks is performed by measuring the response of the network to the removal of nodes or edges. Without going into technical details, we can perceive robustness in an intuitive way thanks to this sentence: "the larger is the amount of nodes we should remove to disconnect a connected network, the more robust the network is" (Stauffer et al. 2014).

Scale-free networks are extraordinarily robust against accidental failures due to their intrinsic topological architecture. A high percentage of their nodes are poorly connected, thus the withdrawal of a randomly selected node does not highly affect the rest of the network. Even when many of their individual nodes are disabled at the same time, scale-free networks keep a large degree of in-

tegrity, i.e. a great number of nodes still form a large connected component (see section 1.5). On the other hand, scale-free networks are vulnerable to the so-called targeted attacks. The removal of a few key hub nodes disintegrates the network into a large number of small isolated components (A. L. Barabási and Oltvai 2004).

In accordance with the robustness properties of scale-free networks, it is expected that molecular hubs are crucial for cell survival. Indeed, it has been shown in model organisms that hubs correspond to essential genes, are evolutionary more conserved, and have a tendency to present higher expression levels (Jeong et al. 2001; Fraser et al. 2002; Ivanic et al. 2009). Other studies in yeast has shown that protein connectivity also correlates with genetic pleiotropy (Yu et al. 2009). The evolutionary selective pressure over mutations can somehow explain the conservation of hubs. Mutations that take place in hubs have a large phenotypic impact, making complicate for the host to have offspring, and therefore preventing their transmission to the population (Vidal et al. 2011).

2.8.1.4 Scale-free networks and gene duplication

Some studies have suggested that the properties of biological scale-free networks can be explained through gene duplication (Vázquez et al. 2002; Pastor-Satorras et al. 2003). Gene duplication is one of the major mechanisms to generate new genetic material during evolution (J. Zhang 2003). Just after a duplication episode, the protein generated by the new gene is exactly the same as the original one. In this context, both protein copies will have the same structure and will therefore interact with the same set of partners. Consequently, each of the proteins partners that interacted with the ancestor gains a interaction with the new protein (Vidal et al. 2011). Afterwards, the duplicated gene will evolve in a different way than the original one, potentially resulting in a new protein product. Assuming that all genes have the same probability of undergoing a duplication process, hubs are more likely to interact with genes who has undergone a duplication than poorly connected proteins. This situation is known as preferential attachment and it is a well documented process on the Internet network topology (A. L. Barabási and Oltvai 2004).

2.8.1.5 Are scale-free networks so frequent in nature?

A recent published article has intensely shaken the network-science community. In this study, Anna Broido and Aaron Clauset evaluated a large corpus of more than 900 real-world networks from different scientific domains, and concluded that pure scale-free networks are infrequent. They systematically applied state-of-the-art statistical tests on the networks under study. More precisely, for each network, they estimated the best-fitting power-law model, test its statistical validity, and then compared it via a likelihood ratio test to alternative non-scale-free distributions. Moreover, they defined quantitative criteria in order to bet-

ter classify networks into five categories accounting for differing strengths and evidences from the scale-free structure. According to their results, more than 40% of the networks under study showed no direct or indirect evidence of scale-free structure. On the other hand, less than 5% of the networks displayed the strongest level of direct evidence for scale-free structure, what we can define as genuine scale-free networks (Broido et al. 2018). Broido and Clauset's work seems technically sound and it has unleashed a vivid debate and certain controversy. In particular, Albert-László Barabási, the main reference regarding the scale-free theory, has argued that real networks have predictable deviations from a pure power law. He claims that these deviations arise from the fact that preferential attachment is not the solely mechanism leading to the scale-free topology. From his point of view, these results do not weaken the idea that scale-freeness underlies many or most complex networks (Klarreich 2018). Moreover, Barabási rejected the statistical tests performed by Broido and Clauset. He discussed that even the exact model of scale-free networks, following a pure power law, fails their test (A. L. Barabási 2018).

This debate has already appeared in the context of biological networks and, in particular, for co-expression networks. Even though many studies state the scale-freeness of the biological networks, some statistical analyses have also refuted the power law distribution (Khanin et al. 2006; Lima-Mendez et al. 2009). This is a very important point since, as we have presented, many properties of biological networks are presented as consequence of their scale-free topology.

3 Analysis of biological networks in a bio-medical context

Correct cellular activity relies on the accomplishment of an intricate set of interactions among proteins and other biomolecules, such as metabolites. Environmental factors and genetic abnormalities may alter the structure and properties of biological molecules, and hence have an impact on their interactions with other molecules. These perturbations can propagate through networks leading to disease states that are difficult to explain considering punctual protein failures. Therefore, the clinical manifestations of a disease, i.e. the disease phenotype, can be regarded as the result of numerous altered biological process at the molecular level (Zhong et al. 2009; Schadt 2009). In this context, **network medicine**, the utilization of network theory to study human diseases, sounds very convenient towards a better understanding of the genotype-phenotype relationship in diseases. Network medicine is an excellent example of a systems biology approach. It addresses the study of diseases from the pluralism of causes and their effects in biological networks, instead of relying on a single disease symptom or on an individual causative effect. I start this chapter by introducing the most widely extended graph-theory methods applied to network medicine. Then, after a section describing the differences and similarities between common and rare diseases, I comment their diverse application by providing clinical examples. As in previous chapters, the content is focused on static networks rather than network modelling.

3.1 Graph-theory methods on network medicine

Functionally-related proteins tend to lie in the same neighborhood within biological networks as a consequence of their ensemble of interactions to carry out cellular functions. Similar or closely related diseases are caused by alterations of the same biological processes, and therefore their causative genes/proteins are also expected to be nearby located in biological networks. This situation allows predicting protein cellular functions and identifying potential gene-disease associations using guilt-by-association strategies, which are based on two major graph-theory approaches: Network propagation ranking methods and clustering algorithms. Under these premises, the forthcoming sections are centered on the identification of gene-disease associations, disease biomarker discovery, the inference of genotype-phenotype relations and pharmacology in the context of network medicine. However, such guilt-by-association strategies can also be applied to other biological questions, such as the functional characterization of poorly studied genes/proteins. Furthermore, they are global methods that can

be applied to any kind of network describing a real-world system.

3.1.1 Network propagation ranking methods

The identification of genes associated with diseases is one of the main goals of network medicine. A few years ago, genes associated with diseases were discovered by linking genomic intervals to a particular phenotype. More recently, genome-wide association studies (GWAS) were developed aiming at identifying single nucleotide polymorphisms (SNPs) potentially related with diseases. The result of both techniques is a large number of disease-gene candidates, where it is not straightforward to discover the genetic variants involved in the disease (Köhler et al. 2008; A. L. Barabási, Gulbahce, et al. 2011). Therefore, network-based computational methods were developed to prioritize among these disease-gene candidates.

The first approaches were parsing the direct interactors of disease protein candidates in a PPI network (Oti et al. 2006). Then, more elaborated algorithms computing the shortest paths between candidates and known disease proteins were developed (Franke et al. 2006; George et al. 2006). Afterwards, network propagation ranking algorithms, which are able to exploit the global topology of the PPI network, were shown to largely outperform initial methods for the identification of disease genes (Köhler et al. 2008; Vanunu et al. 2010). Network propagation is a family of closely related methods and, under certain conditions, mathematically equivalent techniques. They include **random walks** on a graph, **diffusion processes** on a graph and **current computations in electric networks** (Cowen et al. 2017). Since the methods are roughly equivalent, I solely focus on Random Walks with Restart (RWR) without getting into the technical details, which are provided in section 4.3.

Briefly, RWR starts from a node or a set of nodes - called seed(s) - that have some kind of relevant information for us (e.g. genes associated with a specific diseases of interest). Then, an imaginary particle explores the network following randomly its edges. The particle have also a previously defined probability of jumping back to the seed(s) after each step. After a large number of repetitions, the number of times the particle visited each particular network node can be considered as a measure of its distance to the seeds (Figure 3.1). Several extensions of the RWR algorithm have been proposed to further improve the prediction of disease candidate genes. They are mainly based on the integration of the PPI network with other biological networks or on slightly modifications of the original RWR algorithm (Yongjin Li and Patra 2010; Yongjin Li and J. Li 2012; Xie et al. 2015; Z. Q. Zhao et al. 2015).

In addition to the application of RWR to identify genes associated with diseases, similar apporaches have been proposed to predict drug-target interactions (Xing Chen, M.-X. Liu, et al. 2012; H. Liu et al. 2016), adverse drug reactions (Xiaowen Chen et al. 2016), lncRNA-disease association (Xing Chen, You,

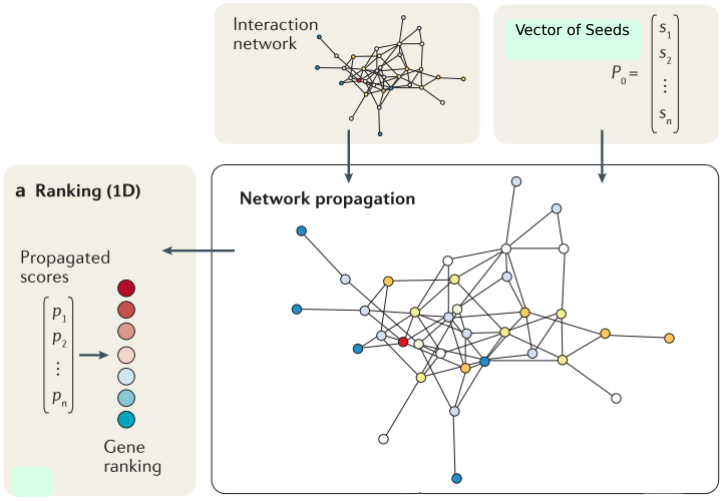


Figure 3.1 – RWR takes an input a network and a vector indicating our prior knowledge about the network (Seed genes). An imaginary particle explores the neighborhood of the seeds and it provides a score that is used to rank the remaining genes of the network based on their distance to the seeds. Figure adapted from (Cowen et al. 2017).

et al. 2016; Gu et al. 2017) and to characterize gene functions (Blatti et al. 2016), among others. Smedley and colleagues developed Exomiser, where RWR is applied to prioritize genes and variants in the context of whole-exome sequencing (Smedley, Köhler, et al. 2014; Smedley, Jacobsen, et al. 2015).

The studies commented in this section are focused on the development of the computational tools rather than in providing clinical results, which are described in forthcoming sections of the current chapter.

3.1.2 Clustering algorithms

As we already mentioned in section 2.8, biological networks present a modular structure as a consequence of the scale-free topology and the small-world property. Therefore, we can find within the network groups of densely connected proteins likely to be implicated in the same biological functions or associated to the same phenotype (Hartwell et al. 1999). The accurate extraction of these modules is decisive because it allows hypothesizing about the function of poorly characterized proteins. Moreover, we can identify modules associated with different diseases, and hence investigate about potential therapeutic actions on the processes altered on these diseases.

Clustering is the process of grouping objects in such a manner that the objects placed in a group (also called cluster, module or community) are to some extent more similar than to those placed in other groups. Clustering methods are based

on different similarity measures among the objects to be grouped. This measure is not data-type specific, allowing to apply the same clustering algorithm to different kind of objects (e.g. genes/proteins in biological networks or individuals in social networks) (Sharma et al. 2017). In addition, in contrast with network propagation ranking methods, clustering algorithms do not usually require previous knowledge about the input data. For example, a clustering algorithm applied on a protein network computes the resemblance between nodes based on the topological features of the network, but it ignores the functions of the genes of their associations with diseases. Another feature that we should bear in mind is that cluster analysis is not a fully automatic process, but an iterative technique of knowledge discovery. In other words, clustering algorithms are usually very sensitive to changes in their initial parameters and some kind of data preprocessing may be performed to obtain satisfactory results (e.g. clusters containing a number of genes/proteins that is relevant from a biological point of view) (Table 3.1). These steps, as well as the selection of the appropriate clustering method, strongly depend on the input data and the goal of the analysis.

A plethora of very diverse clustering methods have been proposed for module extraction from biological networks. This vast amount of different algorithms emerges from the lack of a precise definition of module. Moreover, the different approaches select different strategies to extract the modules in a efficient way from a computational point of view. We can classify clustering methods into five categories namely, **Partitioning**, **Hierarchical**, **Density based**, **Model-based** and **Graph-based** (Andreopoulos et al. 2009; Wiwie et al. 2015) (Table 3.1):

1. *Partitioning methods:* the user provides a defined number of initial clusters k, and the process iteratively assigns objects to the closest cluster based on a metric, such as the Euclidean distance. The most representative examples are the ***k*-Means** (MacQueen 1967) and the Partitioning Around Medoids or ***k*-Medoids** (Kaufmann et al. 1987) algorithms. Their main drawback is that the resulting number of cluster is provided as an input parameter, which is not very advisable when dealing with large networks.
2. *Hierarchical methods:* they build a hierarchy of clusters based on a distance measure (e.g. Euclidean distance, Manhattan distance) and a linkage criteria that determines the distance between groups as a function of the pairwise distances between their individual objects (Langfelder et al. 2008). The result is a tree or dendrogram that is a very useful tool in some areas of boinformatics since the clusters can be explored at different levels. They are usually slow and therefore not recommended for large biological datasets. In addition, sometimes very large modules are merged, neglecting potentially interesting smaller local clusters (Andreopoulos et al. 2009). The popular **Spectral Clustering** (Ng et al. 2001) can be placed within this category (Wiwie et al. 2015).
3. *Density-based methods:* clusters are defined as areas containing a higher

density of objects than the remainder areas of the data set. Objects in these sparse areas are usually considered to be noise or points to separate the other clusters. The most popular density-based cluster is **DBSCAN**, which is based in the number of points that are reachable within a previously defined distance from a given point. The points not reachable from any other point are considered to be outliers and are not assigned to any group (Ester et al. 1996).

4. *Model-based methods:* they assume that the objects to be clustered match a model, which is usually a statistical distribution. Then, the method try to find the modules resulting in the best fit of the distribution (Andreopoulos et al. 2009). The main issue of these approaches emerges from the fact that models are an oversimplification of the real nature of the data. Another drawback is that they are normally computationally quite expensive (Wiwie et al. 2015). **Self-Organizing Maps** (Kohonen 1982) are one of the most representative algorithms falling into this category.
5. *Graph-based methods:* they are widely used when dealing with biological networks since they consider their input as a graph. They aim at finding the modules based Based on theoretical problems of graph theory like simulating random walks or clique discovery. Therefore, these methods extract modules from the networks according to their topological properties. They take advantage of the strength of graph theory, but they are sensitive to user-defined parameters and often computationally expensive (Andreopoulos et al. 2009; Wiwie et al. 2015) **Markov clustering** (Marinus van Dongen 2000) and **Affinity Propagation** (Frey et al. 2007) are some of the best known examples of graph-based clustering methods. Other methods try to optimize the modularity (see section 1.7). Identifying the set of clusters maximizing the modularity is NP-complete (Brandes et al. 2008). Therefore, many different metaheuristic approaches have been proposed when dealing with large groups, as the **Louvain** algorithm (Lefebvre et al. 2008).

It is to note that one of the categories described above is specifically called graph-based methods because it is based on graph theory well known problems. Nevertheless, the methods of the remaining categories are also applicable to networks since they can take as input the network adjacency matrix.

The development of clustering methods is a very active research field, and new methods are published frequently. For instance, the graph-based **Leiden** algorithm has been very recently published (Traag et al. 2018). The Leiden algorithm tries to overcome some of the issues related to the Louvain method, in particular the arbitrarily badly connected communities over and above the resolution limit. To this goal, the new method incorporates a combination of different moves along the network nodes: a smart local move, a fast local move and a random neighbour move (Traag et al. 2018).

Table 3.1 – Technical details about the most popular clustering methods of the different categories.

Method Name	Complexity	Parameters	Category
***k*-Means**	$O(n * K * I * d)$	K	P
***k*-Medoids**	$O(K(n-K)^2 I)$	K	P
Hierarchical	$O(n^2)$ or $O(n^2 \log n)$	K	H
Spectral	$O(n^3)$	K	H
Louvain Modularity	$O(n \log n)$	γ	D
DBSCAN	$O(n^2)$	ϵ and $Dens$	D
Self-Organizing Maps	$O(n^2)$	$Grid$	M
Markov clustering	$O(Nk^2)$	Exp and Inf	G
Affinity Propagation	$O(N^2 I)$	$Damp$	G

Categories: P, partioning methods; H, hierarchical methods; D, density-based methods; M, model-based methods; G, graph-based methods. Complexity: number of objects, n; number of clusters, K; number of iterations, I; number of attributes, d; threshold for number of resources allocated per node, k. Parameters: number of clusters, K; resolution parameter to tune the size of modules, γ; density threshold , $Dens$; size or radio of the module neighborhood, ϵ; size of the grid to tune the size of modules, $Grid$; expansion, Exp and inflation Inf to control the size of the clusters; damping factor to control the size of the modules $Damp$.

As we mentioned before, all these methods can be applied on biological networks, but which method is the most suitable to address a specific biological question? The problem is complicated since there is not a gold benchmark to evaluate the relevance of the extracted modules in a specific biological context. Some alternatives are the generation of simulated networks with a known community structure or the evaluation of the biological information contained in the genes/proteins belonging to each module. Both methods have limitations and they cannot evaluate whether the extracted modules can reflect novel and relevant biological knowledge. With this in mind, the **Disease Module Identification (DMI) DREAM challenge** aimed at investigating the performance of different community detection algorithms in a biomedical context by providing a controlled benchmark for all the participants (see Appendix B for the article detailing the results of the DREAM challenge). We participated to this challenge with an algorithm based on the Louvain algorithm for modularity optimization (see section 4.4).

It is important to note that clustering algorithms are less effective than network propagation ranking methods at associating proteins with their functional roles or with diseases through a guilt-by-association strategy (Sharan et al. 2007; J. Song et al. 2009). Once again, clinical results are detailed in the forthcoming sections of the current chapter.

3.2 Clinical applications in network medicine

The multifactorial origin of diseases as well as their diverse clinical manifestations are the key motivations for the use of network theory in medicine. Indeed, network medicine has already provided very valuable clinical results and its use

is expected to widely span in the coming years.

In this section, I first describe the differences between common and rare diseases and state why network medicine is a suitable approach in both cases. Then, I present relevant results of network medicine in different application frameworks.

3.2.1 Common and rare diseases

The name of **common diseases** arises from the large number of patients affected by any of these diseases. For instance, it is estimated that prostate cancer affects one in six men during their lifetime in the United States (*Cancer Screening Guidelines | Detecting Cancer Early* 2018) (see chapter 6). Common human diseases are multifactorial or "complex" diseases, since they emerge from the interplay of several aspects. Some of these aspects can be deemed as intrinsic to the patient (e.g. DNA modifications, age or gender), whereas external environmental factors, such as diet or exposure to toxins, play also a key role in the development of these diseases. The joint action of these factors determines the onset, severity and progression of the diseases. A large number of diseases fall into this category, including several types of cancer and many neurodegenerative disorders. Some examples include Alzheimer's disease, osteoporosis, diabetes, Parkinson's disease and coronary heart disease (Motulsky 2006). In all such diseases, we can identify many of the so-called age-related diseases whose prevalence dramatically increase with the age of individuals. A clear example is provided by Alzheimer's disease, affecting an estimated of one in 14 people over the age of 65 and one in every 6 people over the age of 80, according to the National Health Service from the UK. Therefore, in the context of an increasingly aging society, it is fundamental to explore new research avenues like the one provided by network science. Moreover, the complex interplay of many factors leading to common diseases seems to fit with the inherent nature of networks.

In opposition to common diseases, **rare diseases** have a low frequency of occurrence in the world population. According to the World Health Organization (WHO), a disease is classified as rare if the percentage of affected individuals ranges from $0.65 - 1‰$ of the total population. It is complicated to state the exact number of rare diseases identified up to date, due to their low prevalence and the blurred boundaries among some of them. However, we can assert that there are between 6000 and 8000 known rare diseases. Altogether, rare diseases may affect around 30 million European Union citizens and a similar number of US individuals. Most rare disease are genetic and thus present throughout the patient's whole life. In addition, a significant percentage of them are life-threatening (*About Rare Diseases* 2014; Stoller 2018). Despite this extensive social impact, over 95% of rare diseases lack a FDA-approved therapeutic agent (Pushpakom et al. 2018). This low drug coverage arises from the large number of different rare

diseases, combined with the reduced number of patients for each one of them. Indeed, private companies and public administrations have conceived research on rare diseases as expensive and unprofitable so far.

Rare diseases are usually caused by mutations in a single gene in opposition to what occurs in common diseases. Nonetheless, we claim that rare diseases are also highly complex. First and foremost, these diseases are often associated to high allelic heterogeneity. An striking example is provided by the laminopathies, which are a collection of heterogeneous diseases caused by more than 300 distinct mutations in the *LMNA* gene. Laminopathies exhibit more than 15 phenotypically diverse disorders, including both tissue-specific and systemic disorders. Among laminopathies, dilated cardiomyopathy 1A (MIM code: 115200) mainly affects muscles, whereas Dunnigan-type familiar partial lipodystrophy (MIM code:151660) impacts on adipose tissue, and Hutchinson-Gilford Progeria Syndrome (HGPS; MIM code:176670) affects multiple tissues (Rankin et al. 2006; Szeverenyi et al. 2007; Dittmer et al. 2014). Furthermore, phenotypic severity can vary considerably across individuals carrying the same punctual mutation, even within the same family (Brodsky et al. 2000). Taking everything together, we can not deny the complexity associated with human rare diseases.

We can also indicate some potential molecular relations between common and rare diseases. It has been proposed that genetic variants related to rare diseases could contribute to the risk of developing some common diseases, highlighting comorbidity associations between rare disorders and multifactorial diseases (Blair et al. 2013). For instance, given the common phenotypes displayed between Xeroderma pigmentosum (OMIM: 278700) and aggressive skin cancers, a study investigated and confirmed a comorbidity relation between them. This fact points towards the alteration of common molecular mechanisms in both diseases, concretely to the failure of DNA repair in this case, resulting in common phenotypes (Cleaver 2005) Some potentially inverse comorbidities between rare and common diseases have also being studied, such as the possible protection against cancer of HGPS patients (La Rosa et al. 2013; Fernandez et al. 2014). Moreover, a decrease in the expression levels of the proteins encoded by the gene causing HPGS, *LMNA*, seems to be correlated with the severity of some cancers such as lung adenocarcinoma or prostate adenocarcinoma (Guinde et al. 2018; L. Zuo et al. 2018).

Network medicine is mostly focused on the study of common diseases given their multifactorial nature and the large number of individuals affected. However, the genetic background of rare disease patients should also be considered due to the potential presence of *modifier genes* that may affect the severity of the phenotype. Therefore, the complexity is also a hallmark of rare monogenic diseases. We can look upon rare diseases phenotypes as the result of the propagation within biological networks of the altered interactions suffered by mutated proteins. As a matter of fact, the causative gene of rare monogenic disorders is occasionally known and well documented, providing a valuable departure point

for network theory approaches. Therefore, what we can learn from common diseases is often applicable to rare syndromes and vice versa.

Another reason for the reduced number of network medicine works on rare diseases is the challenging acquisition of an enough number of patient samples. Therefore, the data that we can use as input for the different methods is very limited. This issue is common to many methods besides network-approaches, and to diverse applications. As a matter of fact, I personally believe that the integration of data into networks can somehow overcome the statistical limitations associated to the reduced amount of available data.

3.2.2 Network medicine for the identification of disease-implicated genes

Network medicine methods led to the discovery of several new genes involved in diseases. Goehler and colleagues achieved one of the first remarkable results in this field by building a PPI network around *Huntingtin* gene, whose mutation causes Huntington's disease (MIM code: 143100). They revealed new potential interactions for *Huntingtin*, which eased the discovery of the *GIT1* gene. *GIT1* enhances the aggregation of *Huntingtin* and its function is altered during disease pathogenesis, turning it into a potential target for therapeutic intervention (Goehler et al. 2004).

A paradigmatic example of a guilt-by-association approach in network medicine is provided by the work of Lim and co-workers (Lim et al. 2006). They performed Y2H assays to detect interaction partners of 23 known ataxia-causing genes. The gene *PLEKHG4*, a common binder of many of the initially known ataxia genes, was shown to cause ataxia-like phenotypes in mice after its deletion (Amino et al. 2007; A. L. Barabási, Gulbahce, et al. 2011).

Procedures developed upon the same basics have been used to identify genes involved in breast cancer (Pujana et al. 2007), schizophrenia (Camargo et al. 2006) or Alzheimer's disease (Forabosco et al. 2013), among others.

Another interesting tool for the identification of disease-causative genes is the *Exomiser*. *Exomiser* prioritizes genes and variants from exome sequencing data by using random walk analysis of protein interaction networks (Smedley, Jacobsen, et al. 2015).

3.2.3 Network medicine for the identification of disease biomarkers

Molecular diagnostics are the different techniques used to identify and analyse biological markers in order to diagnose and monitor diseases, detect potential risk of developing diseases and choose the therapies that suit the patients best.

In the framework of network medicine, molecular diagnostics is usually accomplished based on the concept of differential networks, which refers to a com-

parative analysis between networks representing distinct states. Overall, gene expression data from patients and healthy controls is used to build co-expression networks or it is mapped into a PPI network or a signaling network. Clustering-based methods allow to identify modules of genes whose collective expression significantly differs between diseased and healthy individuals. These network-based approaches can overcome the limitations of classical methods and detect genes which are not individually deregulated, but participating into functional modules which are essential for the disease pathogenesis. Consequently, the concept of disease biomarker gains a broader context, potentially improving the molecular diagnosis and prognosis process (R. Liu et al. 2014).

Grounded on those ideas, several studies have applied network medicine methods to improve molecular diagnosis of different diseases. We can cite some articles analyzing diverse types of cancer, such as lung (Y.-C. Wang et al. 2011), breast (Bertoli et al. 2015) and pancreatic cancer (M. Zhou et al. 2016), or neurogenerative diseases, like Alzheimer's disease (Padmanabhan et al. 2017) and Parkinson's disease (Santiago et al. 2015).

Similar methodologies have also been employed for early diagnosis of diseases. It is crucial to determine the molecular factors behind the transition from a pre-diseased to a diseased state. The goal is twofold: to point out to individuals susceptible to develop a disease, and to improve our knowledge about the functional aspects triggering pathogenesis. Some examples of this line of research can be found in articles investigating diabetes (Xiaoping Liu et al. 2013; M. Li et al. 2013) and respiratory diseases (X. Wu et al. 2014).

Another key point in the clinics is the prognosis, which comprises the prediction of how a disease will evolve, the appearance of further complications, the impact on the patient's quality of life and the potential influence on his/her life expectancy. A suitable clinical prognosis permits to adapt the treatment to the patient, avoiding for instance, adverse side effects related to excessive treatment of low-risk individuals. We can cite some methods that incorporated both network-based approaches and machine learning techniques to accurately estimate survival rate in breast, ovarian or colorectal cancer patients (G. Wu et al. 2012; Shi et al. 2012). More recently, a method systematically evaluated the contribution of individual genes to patient survival in four cancer types by integrating *omics* data of diverse nature in a PPI network (F. Zhang et al. 2016).

3.2.4 Networks analysis of diseases comorbidities

In medicine, comorbidity is usually defined as the co-occurrence of diseases in an individual. We can therefore establish a direct comorbidity relationship between two diseases when they account for a higher-than-expected joint occurrence in individuals (Hidalgo et al. 2009). For instance, the incidence of celiac disease is more frequent in type 1 diabetes patients than in the general population (Smyth et al. 2008). Moreover, there are some diseases that seem to have

a "protective" effect against the development of other syndromes. This lower-than-expected probability of the occurrence of a concrete disease in individuals diagnosed with another medical condition is known as inverse comorbidity (Tabarés-Seisdedos et al. 2011). A well documented example is the decreased incidence of some cancers in patients with Alzheimer's disease (Ou et al. 2013).

Comorbidities are a major health issue, as they increase patient mortality and entangle the election of suitable treatment plans. Moreover, the co-occurrence of diseases increases with age, challenging health systems in the context of an elderly population (Divo et al. 2014; Hu et al. 2016). Various factors could be at the foundation of comorbidities, such as drug-side effects, lifestyle or genetics aspects. From a network medicine perspective, the hypothesis is that common molecular mechanisms could underlay comorbidities relationships. We therefore aim at identifying this altered shared mechanisms between different diseases (Figure 3.2).

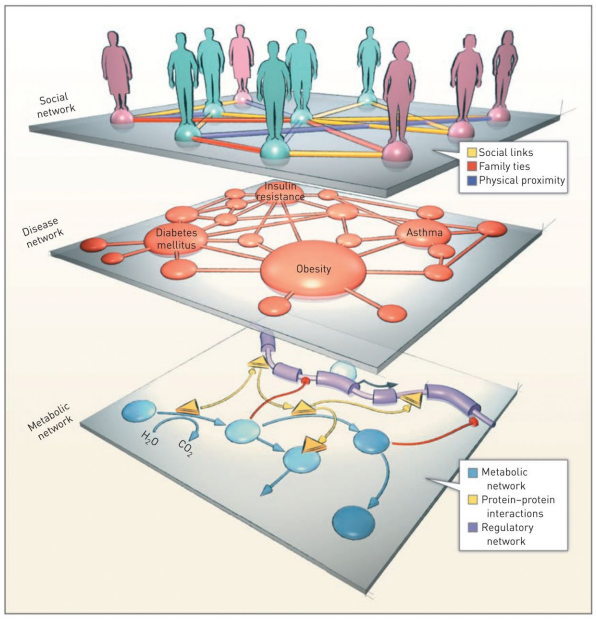


Figure 3.2 – Graphical representation of some of the key actors inter-playing in the comorbidities scenario. A disease network (middle layer), where two diseases are linked if they share symptoms, they have a direct comorbidity relation or they have common genetic background. A social networks (upper layer), that may have a large influence upon the disease network (spread of an infectious disease or lifestyle). A cellular network, such as metabolism (lower layer), which are perturbed under environmental and genetic factors leading to disease states. Figure adapted from (A.-L. Barabási 2007; Faner et al. 2014).

Some examples of studies addressing comorbidities from a network medicine perspective are the following:

1. Lee and colleagues demonstrated that diseases caused by enzymes catalyzing adjacent metabolic reactions have comorbidity patterns in patients. They also showed that diseases with higher degree in their metabolic network were associated with a larger mortality rate. They discussed that patients diagnosed with hub diseases are more likely to develop closely related diseases, increasing together the mortality rate of the original condition (Lee et al. 2008).
2. Infections are associated with a higher probability of developing other diseases. In this context, a study investigated the comorbidity association between 2 infectious diseases, SARS and HIV, with other 7 common diseases (eg. breast cancer and kidney disorder). They pointed out to genes that seem to be deregulated in both types of diseases, suggesting common underlying mechanisms.

3.2.5 Network pharmacology

Conventional drug discovery approaches have generally been based on the design of highly specific ligands targeting single proteins that cause a specific disease. Nevertheless, The number of new drugs effectively implemented in the clinic has been significantly reduced during the last decades. The main reasons for this drug development decline are the lack of efficacy due to drug resistance and patients heterogeneity along with clinical toxicology (Poornima et al. 2016). The main goal of network pharmacology is addressing the true complexity of diseases by detecting the perturbations they induce in biological networks. Then, based on chemical properties, it attempts identifying molecules capable of targeting the set of the most critically affected proteins. Network pharmacology represents a paradigm shift from "one target - one drug" to "network-targeted - multi-component therapeutics" (Hopkins 2008; Poornima et al. 2016).

Network pharmacology frequently relies on *omics* techniques measuring molecules at cellular level in response to a concrete pathophysiology and/or drug treatment. In this context, the applications of network pharmacology span from modelling synergistic combinations of several drugs to the identification of new potential adverse drugs effects reactions (Azmi et al. 2010; Panossian et al. 2013; Xiaowen Chen et al. 2016). Recently, network pharmacology has also been proposed as a promising approach to understand the pleiotropic effects of natural products. A natural product can target several pathways at the same time and alleviate the symptoms of different diseases. Therefore, the identification of the interactions between the chemical compounds of the natural product and their targeted molecules can lead to new strategies in drug development. For instance, the natural active compounds of herbal medicines represent a very valuable resource for drug discovery as a result of their multiplicity in structure, bioactivity and reduced toxicity. (Xu 2006; Poornima et al. 2016; H. Zuo et al. 2018)

Another branch of network pharmacology is the so-called network-based drug

repositioning. Drug repositioning aims at finding new uses for approved existing drugs, which have already demonstrated to be safe for human treatment. The main advantage of this procedure is the drastic reduction of costs and time when compared to the development of new drugs (H. Chen et al. 2015). Consequently, it can be a very suitable approach for treating rare diseases since their low profit scope has prevented large economic investments in their research. Lotfi Shahreza and co-workers recently published a comprehensive literature review of these methods and some specific clinical applications. Moreover, they compared the different methods by evaluations based on their performance (Lotfi Shahreza et al. 2018).

Results

4 Multiplex networks in biology

In this chapter, I first define multiplex networks from a general perspective. Then, I briefly discuss some of the articles using multiplex networks within the framework of computational biology. I finally present some the results of my thesis illustrated in two peer-reviewed published articles. The results described in this chapter are focused on the computational development rather than on the clinical application. My clinically-oriented results are presented on chapters 5 and 6.

4.1 Definition and relevance of multiplex networks

Real-world systems are composed of sets of components connected to each other in with various types of relationships. Moreover, the features of the different types of interactions can be totally different and these relationships are often time-dependant. A straightforward example is provided by social networks, where relations of different categories (eg., family, work, friendship ...) exist between individuals. Individual networks, also called monoplex networks, are not enough to consider this variety of information and do not provide a truthful scheme to investigate complex systems (Kivelä et al. 2014).

The field of multilayer networks is currently under formal development to address this issue, and hence still lack standardized naming conventions. In this manuscript, we consider the definition stated in this paragraph. A multilayer network can be non-formally defined as a network composed of different network layers, each of which representing a specific type of interactions or the same type of interaction occurring at different time points. A multiplex network is a particular case of a multilayer network where the layers share the same set of nodes (Kivelä et al. 2014). This new way of representing and describing complex systems is capturing the attention of research groups working in very diverse scientific areas. Indeed, describing real-world systems using multiplex networks have yielded valuable results in fields such as social network analysis (Battiston et al. 2016), ecology (Stella et al. 2017), transport (Cardillo et al. 2013) or computational neuroscience (De Domenico et al. 2016). We present in the following section some of the most relevant results obtained in computational biology by means of the use of multiplex network.

4.2 Multiplex networks in computational biology

We state that the complexity of biological systems can be modeled in a more precise way using multiplex networks. This is due to the fact that biological

macromolecules such as proteins, interact with each other through multiple types of different relationships. In addition, as discussed in chapter 2, not only physical binary interactions are nowadays screened using high-throughput technologies. We can indeed represent many types of *-omics* data as networks. Overall, we are nowadays able to build multiplex biological networks containing a great amount of biologically relevant information.

Nevertheless, few studies have used multiplex networks in computational biology so far. On the analysis side, some clustering algorithms have been developed to identify biological communities from multiplex networks. Overall, the communities of genes/proteins identified by these methods are more relevant than those detected in individual monoplex networks (Bennett et al. 2015; Didier et al. 2015; Cantini et al. 2015). For instance, (Cantini et al. 2015) constructed a 4-layer multiplex network, and applied popular clustering methods to detect gene communities within each layer. Then, they applied a consensus clustering method across the four layers to identify the definitive communities (Lancichinetti et al. 2012). Thanks to this procedure, they pointed out to some genes potentially driving different cancer types. It is to note that their approach relies on a consensus clustering method rather than a dedicated multiplex-network clustering method. A multiplex-tailored community detection method was applied to find modules in a multiplex network where the nodes represent human diseases. The identified communities suggest potential comorbidity relationships between diseases (Halu et al. 2017). Our team also developed a clustering method, called *Molti*, suited to multiplex networks (Didier et al. 2015). We participated to the disease module identification DREAM challenge with an updated version of *Molti*. This challenge aims at testing the ability of different clustering algorithms to extract relevant modules from biological monoplex and multiplex networks in a biomedical context. Further details about *Molti* and the disease Module Identification DREAM Challenge are provided in section 4.4.

Large-scale PPIs are usually lacking tissue-specific contexts. Multiplex networks can represent a valuable resource to define tissue-specific features. For instance, (Zitnik et al. 2017) built a multiplex network in which each layer accounts for PPIs in different human tissues. They afterwards applied a network-embedding approach, i.e. methods to represent network nodes as numeric vectors in such a way that similar vectors account for nearby network nodes (Duran-Frigola et al. 2018), to predict tissue-specific protein functions. Moreover, they gave different importance to the different network layers by including hierarchy relations among tissues, which allowed to improve their predictions (Zitnik et al. 2017).

Another promising line of research uses multiplex networks to investigate the interaction patterns between cells. Indeed, describing biological processes at cellular scale may be very appealing. Let us suppose a group of cells of the same type, i.e. a cell population. The cells of this cell population interchange information between them, interact with cells of different types, and its number

can evolve with time. This diversity of interactions can be properly described with multiplex networks. An interesting example is the one presented in the work of Gosak et al. 2015. In order to describe the information flow between beta cells in islets of Langerhans, they constructed a weighted multiplex network. Each node represents a cell; intralayer interactions stand in the first layer for the propagation of the depolarization of membrane potential, and in the other layer for the propagation of Ca^{2+} waves; interlayer connections among the same cells reflect the time lag of the Ca^{2+} signal with respect to the depolarization. This multiplex approach revealed that high-degree nodes exhibit a larger time lag between the membrane potential and the Ca^{2+} signal than less connected nodes. The authors associated these results with a higher activity of endoplasmic reticulum calcium pumps in the most connected cells (Gosak et al. 2015).

4.3 Random walks with restart on multiplex networks

Functionally or phenotypically-related proteins have a tendency to be located in the same area in biological networks. Thus, topological distances between proteins in networks are a key resource to infer some of their cellular functions, as well as to predict and study disease-associated proteins. Random Walk with Restart (RWR) is one of the state-of-the-art network propagation ranking methods. It computes the distance between a set of initial node(s), called seed(s), and all the remaining nodes of the network. RWR measures the distances between network nodes in a more effective way than methods relying on direct topological measures, such as shortest-paths based approaches, since it takes into account the global topology of the network (i.e. it explores all the possible paths between pairs of nodes).

The initial goal of my thesis was trying to predict new genes/proteins related to aging using network-based approaches. The state-of-the-art guilt-by-association RWR method appeared as a suitable strategy to achieve this purpose. RWR has been successfully applied to predict gene-disease associations and to characterize protein functions, as I previously described on section 3.1.1. Even though RWR has been applied to several types of networks (eg., monoplex, heterogeneous and aggregated networks), it had never been applied to multiplex networks. We expected that the ability to explore simultaneously various layers of gene/protein interactions, each one of them associated with its own topological features, weaknesses and strengths, can improve the results obtained with previous methods. Hence, during my thesis work, I formally expanded the RWR method to consider multiplex networks. I demonstrated that RWR on multiplex networks significantly outperforms RWR on monoplex or aggregated networks, when applied to recover gene-disease associations.

Biological networks whose nodes are not genes or proteins can also be built.

In this work, we constructed a disease-disease similarity network where two diseases are connected if they share significant phenotypes. We created a heterogeneous network linking the network of diseases to the gene/protein networks based on bipartite known gene-diseases associations. We consequently extended the RWR algorithm in order to make it suitable to explore networks both multiplex and heterogeneous. We showed that the performance is remarkably better than the one achieved by RWR on multiplex networks and RWR on heterogeneous networks alone. I finally applied the new method to gain insights about the molecular mechanisms underlying two rare diseases. These results are described in the following publication:

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., et al. (2018). *Random walk with restart on multiplex and heterogeneous biological networks.* **Bioinformatics**, (August), 1–9.

Further results related to this article are available on Appendix A and in the general discussion. In particular: *i)* Appendix A.1: contains the supplementary material of the article; *ii)* Appendix A.2: includes the vignette explaining the use of the R package I developed associated with this article; *iii)* Discussion: describes the application of the algorithm to a a premature aging disease aiming at identifying new age-related genes. Based on this results, I build my general discussion about my thesis work.

Bioinformatics, 2018, 1–9
doi: 10.1093/bioinformatics/bty637
Advance Access Publication Date: 18 July 2018
Original Paper

Systems biology

Random walk with restart on multiplex and heterogeneous biological networks

Alberto Valdeolivas[1,2,*], Laurent Tichit[1], Claire Navarro[2,3], Sophie Perrin[2,3], Gaëlle Odelin[2,3], Nicolas Levy[3], Pierre Cau[2,3], Elisabeth Remy[1] and Anaïs Baudot[1,*]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, 13009, Marseille, France, [2]ProGeLife, 13001, Marseille and [3]Aix Marseille Univ, INSERM, MMG, 13005, Marseille, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 20, 2017; revised on June 13, 2018; editorial decision on July 11, 2018; accepted on July 16, 2018

Abstract

Motivation: Recent years have witnessed an exponential growth in the number of identified interactions between biological molecules. These interactions are usually represented as large and complex networks, calling for the development of appropriated tools to exploit the functional information they contain. Random walk with restart (RWR) is the state-of-the-art guilt-by-association approach. It explores the network vicinity of gene/protein seeds to study their functions, based on the premise that nodes related to similar functions tend to lie close to each other in the networks.

Results: In this study, we extended the RWR algorithm to multiplex and heterogeneous networks. The walk can now explore different layers of physical and functional interactions between genes and proteins, such as protein–protein interactions and co-expression associations. In addition, the walk can also jump to a network containing different sets of edges and nodes, such as phenotype similarities between diseases. We devised a leave-one-out cross-validation strategy to evaluate the algorithms abilities to predict disease-associated genes. We demonstrate the increased performances of the multiplex-heterogeneous RWR as compared to several random walks on monoplex or heterogeneous networks. Overall, our framework is able to leverage the different interaction sources to outperform current approaches. Finally, we applied the algorithm to predict candidate genes for the Wiedemann–Rautenstrauch syndrome, and to explore the network vicinity of the SHORT syndrome.

Availability and implementation: The source code is available on GitHub at: https://github.com/alberto-valdeolivas/RWR-MH. In addition, an R package is freely available through Bioconductor at: http://bioconductor.org/packages/RandomWalkRestartMH/.

Contact: alberto.valdeolivas@etu.univ-amu.fr or anais.baudot@univ-amu.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent years have witnessed the accumulation of physical and functional interactions between biological macromolecules. For instance, protein–protein interactions (PPI) are nowadays screened at the proteome scale for many organisms, revealing thousands of physical interactions between proteins. Interaction data are commonly represented as networks, in which the nodes correspond to genes or proteins, and the edges to their interactions. The availability of large-scale PPI networks led to the application of graph-theory based approaches for their exploration, with the ultimate goal of extracting the knowledge they contain about cellular functioning. These methods exploit the tendency of functionally-related proteins

to lie in the same network neighborhood. For instance, clustering algorithms allow identifying communities of proteins involved in the same biological processes (Arroyo *et al.*, 2015; Brohée and van Helden, 2006; Chapple *et al.*, 2015; Katsogiannou *et al.*, 2014), and guilt-by-association strategies explore topological relationships to predict protein cellular functions (Schwikowski *et al.*, 2000).

Network-based guilt-by-association strategies, in particular, have been widely used to identify new disease-associated genes. The first approaches were parsing the direct interactors of disease proteins in a PPI network (Oti *et al.*, 2006). Then, more elaborated algorithms computing the shortest paths between candidates and known disease proteins were developed (Franke *et al.*, 2006; George *et al.*, 2006). But algorithms able to exploit the global topology, such as network propagation or random walk, were finally shown to largely outperform initial methods in the identification of disease genes (Köhler *et al.*, 2008; Vanunu *et al.*, 2010).

Random walks were first developed to explore the global topology of networks, by simulating a particle that iteratively moves from a node to a randomly selected neighboring node (Lovász, 1993). The idea of restart, which led to the random walk with restart (RWR) algorithm, was first introduced for Internet search engines. It intent to simulate the behavior of an internet user. The user surfs randomly from a web page to another thanks to the hyper-links, but he can also restart the navigation in a new arbitrary web page. Thereby, depending on the topological structure of the pages and hyper-links, some pages will be visited more frequently than others. The number of visits is considered as a proxy measure of each web page relevance (Brin and Page, 1998). Moreover, if one forces the particle to always restart in the same node or set of nodes—called seed(s)—RWR can be used to measure a proximity between the seed(s) and all the other nodes in the network (Pan *et al.*, 2004).

RWR became the state-of-the-art guilt-by-association algorithm in network computational biology. It was initially applied, as commented above, to prioritize candidate disease genes. All the network nodes are ranked by the RWR algorithm according to their proximity to known disease-associated nodes taken as seeds (Köhler *et al.*, 2008). Several extensions of the RWR algorithm further improved the prediction of disease candidate genes, mainly by considering also phenotype data (Li and Li, 2012; Li and Patra, 2010; Xie *et al.*, 2015; Zhao *et al.*, 2015). For instance, Li and Patra (2010) described a RWR on a heterogeneous network. A heterogeneous network is composed of two networks, each having its own nodes and edges, which belong to different categories, and which are linked through bipartite interactions. Li and Patra (2010) connected a PPI network with a disease–disease similarity network using known bipartite gene-disease associations.

However, a common feature and limitation of these approaches is that they perform the walks in a single network of interactions between genes and proteins. Doing so, they ignore a rich variety of information on physical and functional relationships between biological macromolecules. Indeed, not only PPI are nowadays described on a large-scale: immuno-precipitation experiments followed by mass-spectrometry can inform on the *in vivo* molecular complexes (Ruepp *et al.*, 2010), pathways interaction data are cured and stored in dedicated databases such as Reactome (Fabregat *et al.*, 2016) and Kegg (Kanehisa *et al.*, 2008). In addition, other functional interactions can be derived, for instance from transcriptomics expression data by constructing a co-expression network, or from gene ontology (GO) annotations (Ashburner *et al.*, 2000) by constructing a co-annotation network.

Each interaction source has its own meaning, relevance and bias: some networks contain links of high relevance (e.g. curated signaling pathways), while others contain thousands or even millions of interactions prone to noise (e.g. co-expression networks) (Didier *et al.*, 2015). The combination of the different sources is expected to provide a complementary view on gene and protein cellular functioning (Menche *et al.*, 2015). But networks can be combined in different ways. Generally, the different networks are merged into an aggregated network. For instance, Li and Li (2012) adapted the RWR algorithm to a network in which PPI and co-annotation interactions were aggregated. However, aggregating interaction sources as a single network dismisses the individual features and topologies of each network. In this context, the multiplex framework offers an interesting alternative. Collections of networks sharing the same nodes, but in which the edges belong to different categories or represent interactions of a different nature are called multiplex (alt. multi-slice, multi-layer) networks (Battiston *et al.*, 2014). In a biological multiplex network, each layer contains a different category of physical and functional interactions between genes or proteins.

We present here two extensions of the RWR algorithm to explore multiplex networks (RWR-M) and multiplex-heterogeneous networks (RWR-MH). We constructed a multiplex network composed of three layers of physical and functional interactions between genes and proteins, and a disease–disease network based on phenotype similarities. We applied a leave-one-out cross-validation (LOOCV) strategy to compare the RWR-M and RWR-MH algorithms to alternatives, including RWR on monoplex networks, aggregated networks and heterogeneous-only networks. We showed that considering many interaction sources through a multiplex-heterogeneous network framework enhances remarkably the performances of disease-gene prioritization. Finally, we applied the RWR-MH algorithm to predict candidate genes for being implicated in the Wiedemann–Rautenstrauch syndrome (WRS), whose responsible gene(s) remain unknown. We also explored the network vicinity of the SHORT syndrome (SS) and its associated gene, *PIK3R1*, and unveiled associated syndromes and pathways.

2 Materials and methods

2.1 Random walk on graphs

Let us consider an undirected graph, $G = (V, E)$ with adjacency matrix A. An imaginary particle starts a random walk at an initial node $v_0 \in V$. Considering the time is discrete, $t \in \mathbb{N}$, at the t-th step the particle is at node v_t. Then, it walks from v_t to v_{t+1}, a randomly selected neighbor of v_t following matrix M (Lovász, 1993). Therefore, we can write: $\forall x, y \in V, \forall t \in \mathbb{N}$

$$\mathbb{P}(v_{t+1} = y \,|\, v_t = x) = \begin{cases} \dfrac{1}{d(x)} & \text{if } (x,y) \in E \\ 0 & \text{otherwise,} \end{cases}$$

where $d(x)$ is the degree of x in the graph G. Defining $p_t(v)$ as the probability for the random walk to be at node v at time t, we can describe the evolution of the probability distribution, $\mathbf{p}_t = (p_t(v))_{v \in V}$, with the equation:

$$\mathbf{p}_{t+1}^T = M\mathbf{p}_t^T \tag{1}$$

where M denotes a transition matrix that is the column normalization of A. The stationary distribution, solution of the equation $\mathbf{p}_*^T = M\mathbf{p}_*^T$, represents—if it exists—the probability for the particle to be located at a specific node for an infinite amount of time.

In the RWR version, at each iteration, the particle can also restart by jumping to any randomly selected node in the graph, with a defined restart probability, $r \in (0, 1)$. This avoids the walk to be trapped in a dead end, and assures the existence of the stationary distribution (Langville and Meyer, 2004). Moreover, we can restrict the restart of the particle to specific node(s), called seed(s) (Pan *et al.*, 2004). Doing so, the particle will explore the graph focusing on the neighborhood of the seed(s), and the stationary distribution can be considered as a measure of the proximity between the seed(s) and all the other nodes in the graph.

Formally, based on Equation (1), RWR equation can be defined as:

$$\mathbf{p}_{t+1}^T = (1-r)M\mathbf{p}_t^T + r\mathbf{p}_0^T. \tag{2}$$

The vector $\mathbf{p}_0$ is the initial probability distribution. Therefore, in $\mathbf{p}_0$, only the seed(s) have values different from zero. After several iterations, the difference between the vectors $\mathbf{p}_{t+1}$ and $\mathbf{p}_t$ becomes negligible, the stationary probability distribution is reached, and the elements in these vectors represent a proximity measure from every graph node to the seed(s). In this work, iterations are repeated until the difference between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ falls below 10^{-10}, as in previous studies (Li and Patra, 2010; Erten *et al.*, 2011; Zhao *et al.*, 2015).

We set the global restart parameter to $r = 0.7$, as in previous studies (Köhler *et al.*, 2008; Li and Li, 2012; Li and Patra, 2010; Smedley *et al.*, 2014; Zhao *et al.*, 2015), for all versions of the RWR algorithm. For the sake of simplicity, we have considered unweighted graphs. However, the extension of the algorithms to weighted graphs is straightforward, and can be achieved by replacing the adjacency matrices by matrices of the weighted edges.

2.2 Random walk with restart on multiplex graphs

2.2.1 Definition

A multiplex graph is a collection of L undirected graphs, considered as layers, sharing the same set of n nodes (De Domenico *et al.*, 2014; Kivelä *et al.*, 2014). Each layer $\alpha = 1, \ldots, L$, is defined by its $n \times n$ adjacency matrix $A^{[\alpha]} = (A^{[\alpha]}(i,j))_{i,j=1,\ldots,n}$, where $A^{[\alpha]}(i,j) = 1$ if node i and node j are connected on layer α, and 0 otherwise (Battiston *et al.*, 2014). We do not consider auto-interactions $(A^{[\alpha]}(i,i) = 0 \, \forall \, i = 1, \ldots, n)$, and v_i^α stands for the node i in layer α. A multiplex graph is characterized by its adjacency matrix:

$$\mathbf{A} = A^{[1]}, \ldots, A^{[L]} \tag{3}$$

and is defined as $G_M = (V_M, E_M)$, where:

$$V_M = \{v_i^\alpha, \, i = 1, \ldots, n, \, \alpha = 1, \ldots, L\},$$

$$E_M = \{(v_i^\alpha, v_j^\alpha), \, i,j = 1, \ldots, n, \, \alpha = 1, \ldots, L, \, A^{[\alpha]}(i,j) \neq 0\} \bigcup$$

$$\{(v_i^\alpha, v_i^\beta), \, i = 1, \ldots, n, \, \alpha \neq \beta\}.$$

2.2.2 RWR-M: extension of RWR to multiplex graphs

The particle can walk from its current node v_i^α to any of its neighbors within a layer, or jump to any node v_i^β with $\beta \neq \alpha$ (De Domenico *et al.*, 2013), and thereby change from one to another layer, as schematically displayed in Figure 1A.

We can thus extend the classical RWR algorithm to a multiplex graph (RWR-M) by building a $nL \times nL$ matrix, A. The matrix A contains the different types of transitions that the simulated particle can follow at each step, and is defined as:

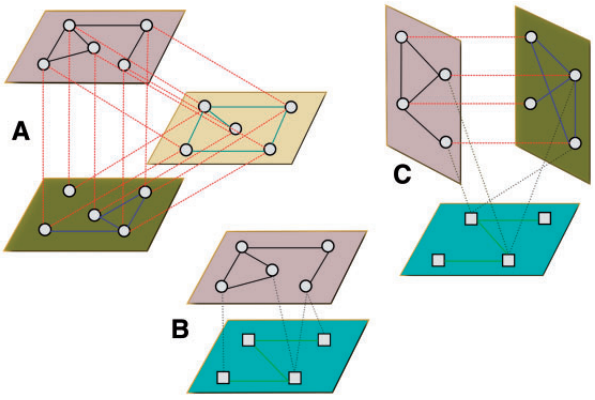


Fig. 1. Multiplex, heterogeneous and multiplex-heterogeneous graphs. (**A**) A multiplex graph composed of three layers. The particle can navigate within each layer or jump to the same node in another layers. (**B**) A heterogeneous graph composed of two graphs. The particle can navigate within each graph or jump to the other graph according to bipartite associations between the two different types of nodes. (**C**) A multiplex-heterogeneous graph

$$A = \begin{pmatrix} (1-\delta)A^{[1]} & \frac{\delta}{(L-1)}\mathbf{I} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \frac{\delta}{(L-1)}\mathbf{I} & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}\mathbf{I} & \frac{\delta}{(L-1)}\mathbf{I} & \cdots & (1-\delta)A^{[L]} \end{pmatrix} \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $A^{[\alpha]}$ is the adjacency matrix of the layer α, as described in (3). The elements in the diagonal represent the potential intra-layer walks, whereas the off-diagonal elements account for the possible jumps between different layers. The parameter $\delta \in [0, 1]$ quantifies the probability of staying in a layer or jumping between the layers: if $\delta = 0$ the particle will always stay in the same layer after a non-restart step.

Let us denote the transition matrix M obtained by a column normalization of A. Equation (2) in the multiplex case becomes:

$$\bar{\mathbf{p}}_{t+1}^T = (1-r)M\bar{\mathbf{p}}_t^T + r\bar{\mathbf{p}}_{RS}^T \tag{5}$$

where $\bar{\mathbf{p}}_t = [\mathbf{p}_t^1, \ldots, \mathbf{p}_t^L]$ and $\bar{\mathbf{p}}_{t+1} = [\mathbf{p}_{t+1}^1, \ldots, \mathbf{p}_{t+1}^L]$, $t \in \mathbb{N}$, are $n \times L$ vectors representing the probability distribution of the particle in the multiplex graph. These vectors are composed of the probability distributions in every layer. The restart vector, $\bar{\mathbf{p}}_{RS}$, represents the initial probability distribution. We define it as $\bar{\mathbf{p}}_{RS} = \tau \cdot \bar{\mathbf{p}}_0$, where the vector parameter $\tau = [\tau_1, \ldots, \tau_L]$ measures the probability of restarting in the seed(s) of each layer in the multiplex graph. It is to note that it is possible to tune the importance of each layer by modifying the parameter τ.

We established an equal restart probability in all the layers, $\tau = (1/L, 1/L, \ldots, 1/L)$, and we also considered an equal probability for staying in a layer or jumping between the layers, $\delta = 0.5$.

When the stationary probability distribution is reached, every node is associated to L proximity measures, one for each layer of the multiplex graph. We compute the global score for every node as the geometric mean of its L proximity measures. The geometric mean

penalizes nodes with a good score in one layer, but low scores in the remaining layers.

2.3 Random walk with restart on heterogeneous graphs

2.3.1 Definition

A heterogeneous graph contains two graphs with different types of nodes and edges, as well as a bipartite graph containing bipartite associations between them (Lee *et al.*, 2013). Let us consider the graphs $G_V = (V, E_V)$ with $V = \{v_1, \ldots, v_n\}$, $G_U = (U, E_U)$ with $U = \{u_1, \ldots, u_m\}$, and the bipartite graph $G_B = (V \cup U, E_B)$ with $E_B \subseteq V \times U$. The edges of the bipartite graph only connect pairs of nodes from the different sets of nodes, V and U. We can now define a heterogeneous graph, $G_H = (V_H, E_H)$, as:

$$V_H = \{V \cup U\}$$

$$E_H = \{E_V \cup E_U \cup E_B\}.$$

2.3.2 RWR-H: extension of RWR to heterogeneous graphs

Li and Patra (2010) proposed a RWR on a heterogeneous graph. This heterogeneous graph was composed of a PPI network, a disease–disease similarity network, and a bipartite graph containing protein-disease associations. The particle walks on the PPI network, on the disease–disease similarity network, and can also jump between the two networks following the bipartite associations (Fig. 1B). Equations and technical details of the approach proposed by Li and Patra (2010) are described in the Supplementary Methods.

2.4 Random walk with restart on multiplex-heterogeneous graphs

2.4.1 Definition

Let us consider a L-layers multiplex graph, $G_M = (V_M, E_M)$, with $n \times L$ nodes, $V_M = \{v_i^\alpha, i = 1, \ldots, n, \alpha = 1, \ldots, L\}$. Let $G_U = (U, E_U)$ be a graph with m nodes, $U = \{u_1, \ldots, u_m\}$. In order to build a heterogeneous graph composed of G_M and G_U, we need to link the nodes in every layer of the multiplex graph G_M to their associated nodes in the graph G_U, according to their bipartite associations, E_B. Since the same nodes are present in every layer of the multiplex graph, it is necessary to have L identical bipartite graphs, $G_B^{[\alpha]} = (V_M \cup U, E_B^{[\alpha]})$ to define the multiplex-heterogeneous graph. We can then describe a multiplex-heterogeneous graph, $G_{MH} = (V_{MH}, E_{MH})$, as:

$$V_{MH} = \{V_M \cup U\}$$

$$E_{MH} = \{\cup_{\alpha=1,\ldots,L} E_B^{[\alpha]} \cup E_M \cup E_U\}.$$

2.4.2 RWR-MH: extension of RWR to multiplex-heterogeneous graph

We finally extended the RWR algorithm to multiplex-heterogeneous networks (RWR-MH). At a given step, let the particle be at a specific node within a layer of the multiplex graph. At the next non-restart step, the particle can either (i) walk within the same layer or (ii) jump to the same node in a different layer or (iii) jump to the other graph if a bipartite association exists (Fig. 1C).

Let consider a multiplex graph composed of n gene/protein nodes and L-layers, with an adjacency matrix $A_{M(nL \times nL)}$, like the one described in Equation (4). Let also consider a disease–disease similarity graph characterized by its adjacency matrix, $A_{D(m \times m)}$, where m is the total number of diseases. The bipartite graphs with adjacency matrices $B_{(n \times m)}^{1,\ldots,L}$ associate the gene/protein nodes in each layer of the multiplex graph to diseases. These bipartite graphs are identical, we define them as $B_{(n \times m)}$, and construct the bipartite adjacency matrix of the multiplex-heterogeneous graph by sticking $B_{(n \times m)}$ L times.

$$B_{MH} = \begin{pmatrix} B_{(n \times m)} \\ B_{(n \times m)} \\ \vdots \\ B_{(n \times m)} \end{pmatrix}. \quad (6)$$

Then, we can define the global adjacency matrix of the multiplex-heterogeneous graph as $A = \begin{bmatrix} A_M & B_{MH} \\ B_{MH}^T & A_D \end{bmatrix}$, where B_{MH}^T represents the transpose of B_{MH}. From this point, we can proceed in an analogous way to the one describing the RWR on heterogeneous graphs (Supplementary Methods). We define a global transition matrix for the multiplex-heterogeneous network and calculate its components using the same equations. We just have to replace the adjacency matrix of the PPI network, $A_{P(n \times n)}$, by the adjacency matrix of the multiplex network$A_{M(nL \times nL)}$, and the bipartite adjacency matrix, $B_{(n \times m)}$, by the adjacency matrix of the bipartite graph of the multiplex-heterogeneous graph, $B_{MH(nL \times m)}$.

In order to apply the Equation S5 (Supplementary Methods), we have to consider that the vectors $\tilde{\mathbf{p}}_{t+1}$, $\tilde{\mathbf{p}}_t$ and $\tilde{\mathbf{p}}_{RS}$ are now of dimension $((n \times L) + m)$, since the RWR-MH algorithm is ranking n proteins in L different layers and m diseases at the same time. It is to note that it is possible to tune the importance of each network by defining $\tilde{\mathbf{p}}_{RS} = \begin{bmatrix} (1-\eta)\mathbf{u}_0 \\ \eta \mathbf{v}_0 \end{bmatrix}$, where $\mathbf{u}_0$ defines the initial probability distribution of the multiplex graph, as described for the RWR on heterogeneous graphs (Supplementary Methods), and $\mathbf{v}_0$ the initial probability distribution of the disease-disease similarity network.

2.5 Network sources

Network details can be found in Supplementary Methods: sizes and densities (Supplementary Table S1), degree distributions (Supplementary Fig. S1A) and overlaps between nodes and edges (Supplementary Fig. S1B and C). Network figures are represented using Cytoscape (Shannon *et al.*, 2003).

2.5.1 Biological networks

We constructed three biological networks containing genes or proteins as nodes (genes and proteins are here considered equally): a PPI network, a network connecting proteins according to pathway interaction data, and a network in which the links correspond to co-expressed genes (Supplementary Methods). The networks were generated from downloads on November 23 and 24, 2016, and from the source codes available on GitHub. The PPI network contains 12 621 no and 66 971 edges. The Pathway network contains 10 534 nodes and 254 766 edges, and the Co-expression network is composed of 10 534 nodes connected by 1 337 347 edges.

2.5.2 Disease–disease similarity network

Diseases and their associated phenotypes were obtained from the Human Phenotype Ontology Project (HPO) (Köhler *et al.*, 2014), and we constructed a disease-disease similarity network-based on phenotype similarities between every pair of diseases. The similarity value is computed according to the relevance of the shared phenotypes. We estimated the relevance of each phenotype from the

information content (IC) given by its frequency in the HPO database, as proposed by Westbury *et al.* (2015) (Supplementary Methods).

2.5.3 Gene–disease bipartite associations

We connected the nodes in each layer of the multiplex network with the disease–disease similarity network thanks to bipartite gene–diseases associations extracted from OMIM (Hamosh *et al.*, 2005), using biomaRt (Durinck *et al.*, 2009) (downloads December, 2016). We obtain 4496 associations between genes/proteins and diseases.

2.6 Leave-one-out cross-validation

The performances of the different RWR algorithms were evaluated using a LOOCV strategy. Known disease-gene associations from OMIM (Hamosh *et al.*, 2005) and DisGeNET v4.0 (Piñero *et al.*, 2016) were used as a benchmark: for each disease-associated to at least two genes, each associated gene is removed one-by-one, and considered as the *left-out gene*. The remaining genes are used as seed(s) in the RWR algorithms. All the network nodes are then scored and ranked according to their proximity to the seed(s), and the rank of the left-out disease-gene is recorded (Supplementary Methods).

3 Results

Our main goal was to design a RWR algorithm able to exploit multiple biological interaction sources. We first constructed three biological networks: a PPI network, a pathway-derived network and a co-expression network (Materials and methods). These networks can be considered independently as monoplex networks. They can also be merged as an aggregated network, with nodes and edges corresponding to the union of the monoplex networks. The aggregated network is composed of 17 559 nodes and 1 659 084 edges (Supplementary Table S1). Finally, we also studied the three networks as a multiplex network. A multiplex network is a collection of networks considered as layers, sharing the same set of nodes, but in which the edges belong to different interaction categories. In our multiplex network, the layers share the same set of 17 559 nodes, also corresponding to the union of all network nodes. The genes/proteins absent in a layer are added as isolated nodes in this layer.

We also constructed a disease–disease similarity network, in which the nodes correspond to diseases, and the edges to the most significant phenotype similarities between the diseases (Materials and methods). Finally, in order to construct a multiplex-heterogeneous network, we linked the disease–disease similarity network to the multiplex network thanks to bipartite gene–disease associations.

We next devised different RWR algorithms, which each leverage the different networks and combinations thereof, and we compared their efficiencies.

3.1 Random walk with restart on multiplex networks are more efficient than on monoplex networks

The classical RWR algorithm takes as input a monoplex network. Here, we first adapted the RWR algorithm to navigate a multiplex network (RWR-M). Basically, at each step, the particle can walk from one node to another in the same layer, as in a monoplex network, but it can also move to the same node in another layer of the multiplex network (Materials and methods). We compared the performances of the classical RWR and multiplex RWR-M algorithms in retrieving disease-associated genes, thanks to a LOOCV strategy (Materials and methods). For that, we created a test set composed of diseases associated to at least two genes in the set of 4529 protein nodes common to the three networks. This test set contains 273 diseases and 1312 gene–disease associations. For every disease, each of its associated genes is iteratively left-out, and the remaining gene(s) are considered as seed(s) to run the algorithms. We then compared the ability of the different RWR algorithms to retrieve the left-out gene (Fig. 2).

Focusing first on monoplex networks, the worst performance is observed for the classical RWR algorithm applied to the co-expression network. It seems difficult to retrieve known disease-associated genes from a network built from correlations of mRNA expression data alone. The Pathway-derived network achieves the best performance among the monoplex networks, probably because pathways databases are usually built on established biological knowledge and curated. Noticeably, the RWR algorithm is not able to predict disease-associated genes from randomized versions of these biological networks (Supplementary Results).

The RWR-M algorithm, exploiting more than one interaction source in a multiplex framework, performs better than the classical RWR. In particular, despite the low ranking capacities of the co-expression network alone, its integration as a layer in a multiplex framework of two or three layers enhances the performance of the algorithm. Overall, the best results are obtained with the integration of the three network layers (Fig. 2).

3.2 Random walk with restart on multiplex networks are more efficient than on aggregated networks

In a second step, we compared the performances of the RWR on multiplex network (RWR-M) with the classical RWR run on the three networks aggregated as a single monoplex network. In the aggregated network, two proteins can be linked by up to three edges (corresponding to the three network sources), and the particle can choose between these different edges to move from a node to one of its neighbors, as in Li and Li (2012). The ranking ability of RWR-M and classical RWR on the aggregated network are again tested by LOOCV. In this case, we created the test set with diseases associated

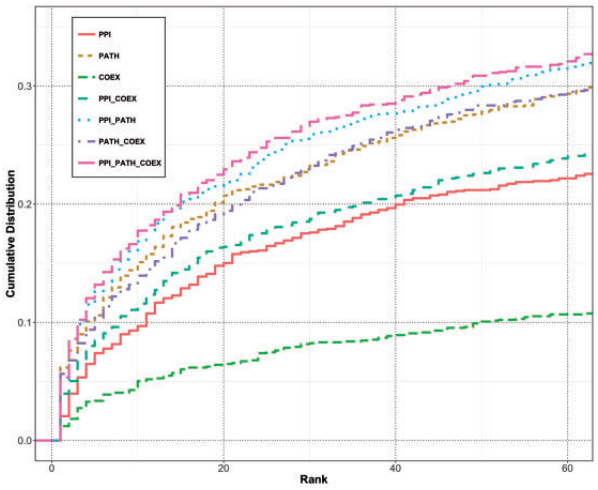


Fig. 2. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied to the protein–protein (PPI), Pathway (PATH) and co-expression (COEX) monoplex networks. RWR-M algorithm is applied to combinations of two or three of these networks, considered as layers of a multiplex network

to at least two nodes in the total of 17 559 nodes corresponding to the union of the nodes of the three networks. The test set contains 537 diseases and 2892 gene–disease associations.

The ranks of the left-out disease genes are better with the RWR-M than with the classical RWR on the aggregated network (Fig. 3). The aggregated and multiplex networks use the same biological data and interaction network sources, but the multiplex framework further keeps tracks of the individual topological structures in each network layer.

3.3 Random walk with restart on multiplex-heterogeneous networks are more efficient than on multiplex or heterogeneous networks alone

We previously compared the performances of RWR algorithms on different combinations of networks containing the same nodes but edges belonging to different interaction categories. We now wish to extend these comparisons to heterogeneous networks, i.e. networks containing different sets of nodes, such as genes/proteins and diseases.

We first coded the heterogeneous RWR-H algorithm as proposed by Li and Patra (2010) (Materials and methods). The RWR-H algorithm takes as input a heterogeneous network composed of a PPI network and a disease–disease similarity network. We constructed the disease–disease similarity network by computing the phenotype similarity between a pair of diseases (Materials and methods). The PPI and the disease–disease similarity networks are connected by bipartite gene-disease associations. In the RWR-H algorithm, the particle can move from the PPI network to the disease–disease similarity network thanks to these bipartite associations.

We here compared the ranking capacities of RWR-M and RWR-H by LOOCV. In this case, we created a test set of diseases associated to at least two genes in the set of 12 621 nodes present in the PPI network. The test set contains 242 diseases and 880 gene-disease associations. We can observe first that RWR-M and RWR-H perform better than the classical RWR on the monoplex PPI network (Fig. 4). This stands for other types of heterogeneous networks, built by combining pathway and disease–disease similarity networks, or co-expression and disease–disease similarity networks (Supplementary Fig. S3).

In this context, an algorithm able to execute a RWR on both multiplex-heterogeneous networks is expected to have better performances. Therefore, we extended our RWR-M approach to heterogeneous networks, defining a RWR on multiplex-heterogeneous networks, RWR-MH (Materials and methods). The RWR-MH displays a remarkable amelioration of performances in the prioritization task, since over 45% of the left-out genes are ranked within the top 20 (Fig. 4).

Finally, we further checked the influence of the different parameters involved in the RWR-MH algorithm using the LOOCV strategy. Overall, the RWR-MH is a very robust algorithm since variations in the parameters do not lead to large variations in the ranking performances (Supplementary Results, Supplementary Figs S4 and S5).

3.4 Candidate genes for the undiagnosed Wiedemann–Rautenstrauch syndrome

The Wiedemann–Rautenstrauch neonatal progeroid syndrome (MIM code: 264 090) is characterized by intrauterine growth retardation with subsequent failure to thrive and short stature (Toriello, 1990). Patients also display a progeroid appearance, decreased subcutaneous fat, hypotrichosis and macrocephaly (Kiraz *et al.*, 2012). Only a few published cases have been documented, and to our knowledge, no gene has been described as causative of the syndrome yet.

To illustrate the application of our approach for disease-associated gene prediction, we applied the RWR-MH algorithm using as seed only the WRS disease node. We then considered the top 25 ranked genes as putative candidates for playing a role in WRS (Fig. 5). Many of these top predicted candidate genes, such as *FIG4*, *RNF113A* or *LMNA*, are implicated in diseases directly connected to WRS from phenotype similarities. Mutations in *LMNA* are responsible for the Hutchinson–Gilford progeria syndrome (MIM code: 176 670) and other premature aging syndromes such as

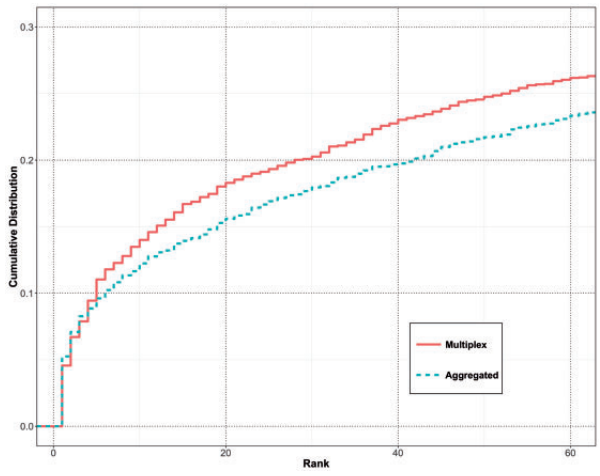


Fig. 3. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied on the three networks aggregated as a single monoplex network, and RWR-M algorithm is applied to combinations of the three networks as layers of a multiplex network

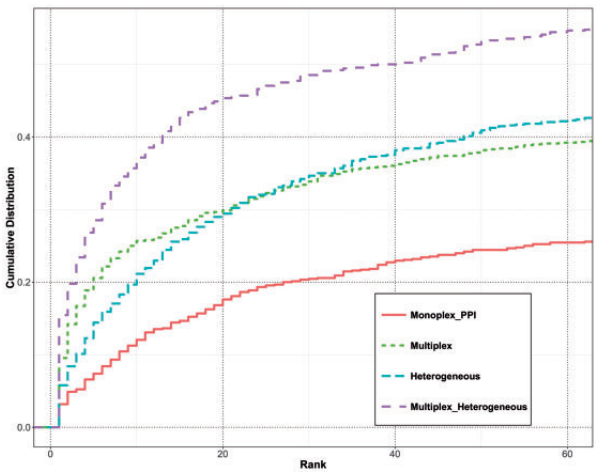


Fig. 4. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied to the monoplex PPI network, RWR-M is applied to the combinations of the three monoplex networks as layers of a multiplex network, RWR-H algorithm is applied to the heterogeneous network composed of the PPI network and the disease-disease similarity network, and RWR-MH algorithm is applied the multiplex-heterogeneous network composed of the three-layers multiplex network and the disease-disease similarity network

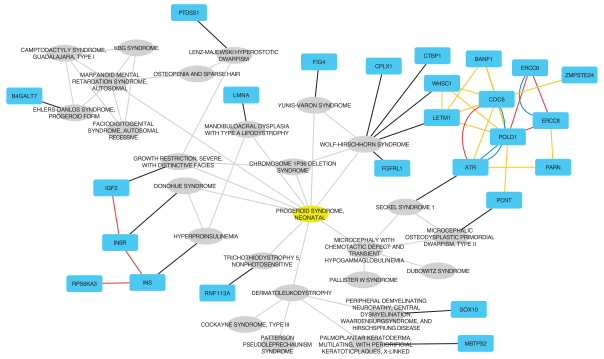


Fig. 5. Network representation of the top 25 ranked genes and diseases when the RWR-MH algorithm is executed using WRS as seed (yellow node). Gray elliptical nodes are diseases; turquoise rectangles are genes/proteins. Black edges are bipartite gene-disease associations from OMIM (Hamosh *et al.*, 2005); grey edges are the similarity links in the disease–disease network; blue edges are PPI interactions; yellow edges are co-expression relationships; red edges are pathway interactions. It is to note that results are represented as an aggregated network only for visualization purposes

Mandibuloacral Dysplasia with type A lipodystrophy (MIM code: 248 370). However, the targeted sequencing of *LMNA* in few WRS patients did not identify mutations (Hou, 2009; Kiraz *et al.*, 2012). The RWR-MH algorithm also top ranked *ZMPSTE24*, which is known to cause severe progeroid syndromes such as Restrictive Dermopathy (MIM code: 275 210) (Navarro *et al.*, 2006). But here also, no mutations were found for this gene in five WRS patients (Hou, 2009).

Another set of interesting candidates is given by the subnetwork composed of the four genes *IGF2*, *INS*, *INSR* and *RPS6KA3*. All these genes participate in the insulin pathway, and are associated to diseases sharing phenotypes with WRS [i.e. Donohue Syndrome (MIM code: 147 670), hyperproinsulinemia (MIM code: 176 730) and severe growth restriction (MIM code: 147 470)]. The insulin pathway is suspected to play a role in WRS (Arboleda *et al.*, 2007). Similarly, a cluster of proteins related to the cell cycle and DNA repair is connected to WRS through the Wolf–Hirschhorn syndrome (MIM code: 194 190), and DNA repair defects are also suspected to be involved in WRS (Hou, 2009).

3.5 Exploring vicinity of *PIK3R1* and SHORT syndrome

SS (MIM code: 269 880) is a rare disease with clinical features defined by its acronym: short stature, hyperextensibility of joints and/or inguinal hernia, ocular depression, Rieger abnormality and teething delay (Gorlin, 1975). However, these phenotypes do not describe the full range of SS phenotypes, and other clinical features include, for instance, partial lipodystrophy and insulin resistance (Avila *et al.*, 2016). Mutations in the *PIK3R1* gene are described as the main cause of SS (Chudasama *et al.*, 2013; Dyment *et al.*, 2013; Thauvin-Robinet *et al.*, 2013).

We applied the RWR-MH algorithm using the *PIK3R1* gene and the SS disease as seeds, and explored the top 25 ranked diseases and genes, along with their interactions and associations (Supplementary Fig. S6). Many of the top ranked diseases recapitulate phenotypes associated to SS. For instance, permanent neonatal diabetes mellitus (MIM code: 606176) accounts for SS phenotypes associated to insulin resistance. Mandibuloacral dysplasia with type B lipodystrophy (MIM code: 608 612) and other diseases associated to lipodystrophy are also top ranked, as well as the growth hormone insensitivity syndrome (MIM code: 262 500) that share with SS the phenotypes related to short stature, among others.

Some of the identified subnetworks are very appealing. For instance, we can observe a loop linking the SS, its associated gene, *PIK3R1*, the Lowe oculocerebrorenal syndrome (MIM code: 309 000) and its associated gene *OCRL*. These two diseases share a noticeable amount of phenotypes, including growth retardation and glucose intolerance. The *PIK3R1* and *OCRL* genes are coding proteins involved in the same pathway: synthesis of phosphatidylinositol phosphates at the plasma membrane (reactome code: R-HSA-1 660 499). Therefore, we can hypothesize a common deregulation of this pathway in the two diseases, leading to shared phenotypes.

Similarly, we can point to the subnetwork containing the *ELN* gene, implicated in the Williams–Beuren syndrome (MIM code: 194 050). Many phenotypes associated to this syndrome are similar to SS and Lowe oculocerebrorenal syndrome. In this case, the *ELN* gene is linked to the *PDGFRB* gene by a co-expression relationship. *PDGFRB* is highly connected to many nodes in the subnetwork, including to *PIK3R1*, by pathway interactions. The co-expression interaction between *PDGFRB* and *ELN* is intriguing because the two genes are, to our knowledge, not described to be involved in the same pathway or process. However, they seem to be regulated by the same microRNA-29 family (Cushing *et al.*, 2015; Zhang *et al.*, 2012). Overall, these results could also allow pointing to other candidate genes predicted to be involved in the SS. This is interesting as, for instance, Dyment *et al.* (2013) did not find any mutation in the *PIK3R1* gene in one of the seven tested patients.

4 Discussion

Physical and functional relationships between genes and proteins are diverse. They are identified or derived from various approaches, each having its own features, strengths and weaknesses. In this context, the integration of different sources of interaction, exploiting data pluralism, is expected to outperform approaches dealing with single networks. Indeed, the combination of different large-scale interaction datasets increases the available biological information, and potentially reduce the bias and incompleteness of isolated sources (Menche *et al.*, 2015).

We and others also hypothesized that the multiplex framework, which retains information on the topology of the individual networks, would perform better as compared to the aggregation of the different interaction sources (Battiston *et al.*, 2014; Didier *et al.*, 2015; Kivelä *et al.*, 2014; Kurant and Thiran, 2006). We have shown previously, for instance, that the multiplex framework is more efficient than network aggregations to extract communities from biological networks (Didier *et al.*, 2015). We extended here the RWR algorithm by designing the RWR-M algorithm able to leverage multiplex networks. The performances of the RWR-M algorithm are clearly improved as compared to previous algorithms navigating monoplex networks, such as RWR on PPI networks (Köhler *et al.*, 2008) or RWR on aggregated networks (Li and Li, 2012). It is particularly interesting to note that even if a monoplex network, such as the co-expression network, displays poor ranking performances isolated, its integration as a layer of a multiplex network leads to an increase of the performance, thereby demonstrating the potential of the RWR-M strategy.

Moreover, we extended our algorithm to deal with multiplex-heterogeneous networks. To this goal, we first built a disease–disease similarity network-based on the IC of the shared phenotypes between every pair of diseases. Previous approaches building

disease–disease networks, such as the ones proposed by (Li and Patra, 2010; Li and Li, 2012), were based on MimMiner (van Driel *et al.*, 2006). MimMiner mines OMIM full-text and clinical synopsis to compute similarity between diseases. Contrarily, our approach is based on the controlled classification of phenotypes in an ontology, and considers both the ontological structure and the frequencies of phenotypes.

Thanks to the LOOCV, we demonstrated that when the new RWR-MH algorithm is applied on this complex multiplex-heterogeneous network, the prioritization results are far better than those of all other versions of the algorithm. We have also demonstrated that the RWR-MH algorithm displays a robust behavior upon variations of the different parameters. This was previously observed for variations in the parameters of a RWR-H algorithm (Li and Patra, 2010; Zhao *et al.*, 2015). The particle keeps exploring the different network layers thanks to the jumps, and still leverage the complementary biological information. This stability is however observed for the average ranking of left-out genes in the LOOCV, but a focused analysis and network representation of the top 25 ranked genes and diseases in real-case applications would reveal variations.

We focused our applications on a multiplex network composed of a PPI, a pathway and a co-expression network. Other biological networks could be collected or constructed from—omics data, and integrated into our multiplex-heterogeneous framework. Functional interactions can be derived, for instance, by connecting genes annotated for the same GOterms (Ashburner *et al.*, 2000). It would also be valuable to include networks with transcription factors— targets genes, non-coding RNAs as well as drug and therapeutic targets.

Funding

A.V. is the recipient of a CIFRE grant 2015/0982 from the French 'Agence Nationale de la Recherche et de la Technologie'. The project leading to this publication has received funding from the Excellence Initiative of Aix-Marseille University - A*Midex, a French 'Investissements d'Avenir' program.

Conflict of Interest: none declared.

References

Arboleda,G. *et al.* (2007) The neonatal progeroid syndrome (Wiedemann-Rautenstrauch): a model for the study of human aging? *Exp. Gerontol.*, **42**, 939–943.

Arroyo,R. *et al.* (2015) Systematic identification of molecular links between core and candidate genes in breast cancer. *J. Mol. Biol.*, **427**, 1436–1450.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Avila,M. *et al.* (2016) Clinical reappraisal of SHORT syndrome with PIK3R1 mutations: toward recommendation for molecular testing and management. *Clin. Genet.*, **89**, 501–506.

Battiston,F. *et al.* (2014) Structural measures for multiplex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **89**, 1–16.

Brin,S. and Page,L. (1998) The anatomy of a large scale hypertextual Web search engine. *Comput. Networks ISDN*, **30**, 107–117.

Brohée,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Chapple,C.E. *et al.* (2015) Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.*, **6**, 7412.

Chudasama,K.K. *et al.* (2013) SHORT syndrome with partial lipodystrophy due to impaired phosphatidylinositol 3 kinase signaling. *Am. J. Hum. Genet.*, **93**, 150–157.

Cushing,L. *et al.* (2015) Disruption of miR-29 Leads to Aberrant Differentiation of Smooth Muscle Cells Selectively Associated with Distal Lung Vasculature. *PLoS Genet.*,**11**, e1005238–e1005227.

De Domenico,M. *et al.* (2013) Mathematical formulation of multilayer networks. *Phys. Rev. X*, **3**, 1–15.

De Domenico,M. *et al.* (2014) Navigability of interconnected networks under random failures. *Proc. Natl. Acad. Sci. USA*, **111**, 8351–8356.

Didier,G. *et al.* (2015) Identifying communities from multiplex biological networks. *PeerJ.*, **3**, e1525.

Durinck,S. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1134.

Dyment,D.A. *et al.* (2013) Mutations in PIK3R1 cause SHORT syndrome. *Am. J. Hum. Genet.*, **93**, 158–166.

Erten,S. *et al.* (2011) DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Min.*, **4**, 19.

Fabregat,A. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

George,R.A. *et al.* (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.

Gorlin,R.J. *et al.* (1975) A selected miscellany. *Birth Defects Orig. Artic. Ser.*, **11**, 39–50.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Hou,J.W. (2009) Natural course of neonatal progeroid syndrome. *Pediatr. Neonatol.*, **50**, 102–109.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Katsogiannou,M. *et al.* (2014) The functional landscape of Hsp27 reveals new cellular processes such as DNA repair and alternative splicing and proposes novel anticancer targets. *Mol. Cell. Proteomics*, **13**, 3585–3601.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Kiraz,A. *et al.* (2012) Wiedemann-Rautenstrauch syndrome: report of a variant case. *Am. J. Med. Genet. A*, **158A**, 1434–1436.

Kivelä,M. *et al.* (2014) Multilayer networks. *J. Complex Netw.*, **2**, 203–271.

Köhler,S. *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.

Kurant,M. and Thiran,P. (2006) Layered complex networks. *Phys. Rev. Lett.*, **96**, 4.

Langville,A. and Meyer,C. (2004) Deeper Inside PageRank. *Internet Math.*, **1**, 335–380.

Lee,S. *et al.* (2013) PathRank: ranking nodes on a heterogeneous graph for flexible hybrid recommender systems. *Expert Syst. Appl.*, **40**, 684–697.

Li,Y. and Li,J. (2012) Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics*, **13** (Suppl. 7), S27.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Lovász,L. (1993) Random walks on graphs: a survey. In: *Combinatorics, Paul ErdÖs Is Eighty*, Vol. 2. Keszthely, Hungary, pp. 1–46.

Menche,J. *et al.* (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.

Navarro,C.L. *et al.* (2006) Molecular bases of progeroid syndromes. *Hum. Mol. Genet.*, **15**, R151–R161.

Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.

Pan,J-Y. *et al*. (2004) Automatic multimedia cross-modal correlation discovery. In: *KDD '04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 653–658.

Piñero,J. *et al*. (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Ruepp,A. *et al*. (2010) CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res*., **38**, D497–D501.

Schwikowski,B. *et al*. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnol*., **18**, 1257–1261.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Smedley,D. *et al*. (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**, 3215–3222.

Thauvin-Robinet,C. *et al*. (2013) PIK3R1 mutations cause syndromic insulin resistance with lipoatrophy. *Am. J. Hum. Genet*., **93**, 141–149.

Toriello,H.V. (1990) Syndrome of the month: Wiedemann-Rautenstrauch syndrome. *J. Med. Genet*., **27**, 256–257.

van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.

Vanunu,O. *et al*. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*., **6**, e1000641.

Westbury,S.K. *et al*. (2015) Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med*., **7**, 36.

Xie,M. *et al*. (2015) Network-based phenome-genome association prediction by bi-random walk. *PLoS One*, **10**, e0125138–e0125118.

Zhang,P. *et al*. (2012) Inhibition of MicroRNA-29 enhances elastin levels in cells haploinsufficient for elastin and in bioengineered vessels-brief report. *Arterioscler. Thromb. Vasc. Biol*., **32**, 756–759.

Zhao,Z.Q. *et al*. (2015) Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem*., **57**, 21–28.

4.4 A clustering algorithm for multiplex networks

My colleagues extended the definition of modularity to multiplex networks. This allowed applying the Louvain algorithm, which optimizes the modularity, to detect communities from multiplex networks. They demonstrated that the multiplex clustering outperforms network aggregation and consensus approaches in the identification of modules derived from simulated networks. Then, they constructed a multiplex network containing four layers of physical or functional interactions between genes and proteins, namely *i)* a PPI, *ii)* a pathway-derived network, *iii)* a co-expression network and *iv)* a network of molecular complexes. Their method, which was called *MolTi*, allowed recovering a larger number of biologically annotated communities on the multiplex network than on their aggregated counterpart networks. They concluded that considering the multiplex nature of protein networks leads to better-defined functional modules (Didier et al. 2015). At the time of the publication of the *MolTi* article, there were hardly any other methods tailored to detect modules in multiplex networks. It was therefore no possible to evaluate the performance of *MolTi* against similar methods.

The DMI DREAM challenge was launched during summer 2016, aiming at investigating the performance of different community detection algorithms in a biomedical context. The challenge organizers provided participants with six anonymous networks, as well as a controlled benchmark (more details are provided on Appendix B, where the article describing the DMI DREAM challenge results is available). The challenge was divided in two sub-challenges to identify communities from the six networks independently (sub-challenge 1) or jointly from the six networks (sub-challenge 2). We participated to the sub-challenge 2 with our multiplex network approach *MolTi*. During the test phase of the challenge, we implemented various extensions of *MolTi* that allowed us to recover larger number of disease-associated modules than its original version. To be more precise, the new version of *MolTi* runs a randomized version of the Louvain algorithm, consider edge and layer weights, and performs iterative clustering. The new version is called *MolTi-DREAM*, and its results within the framework of the DMI DREAM challenge are described in the following article:

Didier, G., **Valdeolivas, A.**, & Baudot, A. (2018). *Identifying communities from multiplex biological networks by randomized optimization of modularity.* **F1000Research**, 7(0), 1042.

The article describing the global results of DMI DREAM challenge is available on the Appendix B.

METHOD ARTICLE

REVISED Identifying communities from multiplex biological networks by randomized optimization of modularity [version 2; referees: 3 approved, 1 approved with reservations]

Gilles Didier [1], Alberto Valdeolivas[1-3], Anaïs Baudot [1,3]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
[2]ProGeLife, Marseille, France
[3]Aix Marseille Univ, Inserm, MMG, Marseille, France

v2 **First published:** 10 Jul 2018, **7**:1042 (https://doi.org/10.12688/f1000research.15486.1)
Latest published: 22 Nov 2018, **7**:1042 (https://doi.org/10.12688/f1000research.15486.2)

Abstract

The identification of communities, or modules, is a common operation in the analysis of large biological networks. The *Disease Module Identification DREAM challenge* established a framework to evaluate clustering approaches in a biomedical context, by testing the association of communities with GWAS-derived common trait and disease genes. We implemented here several extensions of the MolTi software that detects communities by optimizing multiplex (and monoplex) network modularity. In particular, MolTi now runs a randomized version of the Louvain algorithm, can consider edge and layer weights, and performs recursive clustering.

On simulated networks, the randomization procedure clearly improves the detection of communities. On the *DREAM challenge* benchmark, the results strongly depend on the selected GWAS dataset and enrichment ***p***-value threshold. However, the randomization procedure, as well as the consideration of weighted edges and layers generally increases the number of trait and disease community detected.

The new version of MolTi and the scripts used for the DMI DREAM challenge are available at: https://github.com/gilles-didier/MolTi-DREAM.

Keywords

Biological Networks, Multiplex, Multi-layer, Community identification, Clustering, DREAM challenge

This article is included in the DREAM Challenges gateway.

Open Peer Review

Referee Status: ✓ ? ✓ ✓

	Invited Referees 1	2	3	4
REVISED **version 2** published 22 Nov 2018	✓ report			✓ report
version 1 published 10 Jul 2018	? report	? report	✓ report	? report

1 **Emre Guney**, Pompeu Fabra University, Spain

2 **Lenore J. Cowen**, Tufts University, USA

3 **Yasir Suhail**, Yale University, USA

4 **Arda Halu**, Harvard Medical School, USA

Discuss this article

Comments (0)

Corresponding authors: Gilles Didier (gilles.didier@univ-amu.fr), Anaïs Baudot (anais.baudot@univ-amu.fr)

Author roles: Didier G: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Valdeolivas A**: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Baudot A**: Conceptualization, Funding Acquisition, Investigation, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The project leading to this publication has received funding from the Centre National de la Recherche Scientifique (PEPS BMI IMFMG), the French ``Plan Cancer 2009–2013'', and the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" programme.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Didier G, Valdeolivas A and Baudot A. **Identifying communities from multiplex biological networks by randomized optimization of modularity [version 2; referees: 3 approved, 1 approved with reservations]** *F1000Research* 2018, **7**:1042 (https://doi.org/10.12688/f1000research.15486.2)

First published: 10 Jul 2018, **7**:1042 (https://doi.org/10.12688/f1000research.15486.1)

REVISED **Amendments from Version 1**

In the revised version, we further described the dream challenge data and evaluation procedure. Concerning our approach, we detailed the description of the SBM protocol to generate random networks with a community structure, and added the resolution parameter to the modularity formula.

Several additional experiments were performed, and additional data, displayed as figures and tables, are provided in the response to reviewers, for size constraint reasons. In particular, we now provide the results of the Dream challenge on communities obtained by varying the modularity parameter (as compared to the recursion procedure), simulations including standard deviation, Rand index comparisons of the monoplex networks communities, as well as many discussions around the points suggested by the referees.

We finally want to emphasize the fact that, as a companion paper of the Dream challenge channel, our manuscript is included in a set of papers strongly linked to the main consortium paper, available in BioRxiv. Some of the results and experiments can be fully understood only in light of the consortium paper, since we are required not to duplicate the information. This concerns in particular the data and evaluation protocol of the Dream challenge, which are fully described in the consortium paper.

See referee reports

Introduction

Biological macromolecules do not act isolated in cells, but interact with each other to perform their functions, in signaling or metabolic pathways, molecular complexes, or, more generally, biological processes. Thanks to the development of experimental techniques and to the extraction of knowledge accumulated in the literature, biological networks are nowadays assembled on a large scale. A common feature of biological networks is their modularity, i.e., their organization around communities - or functional modules - of tightly connected genes/proteins implicated in the same biological processes[1,2].

The Disease Module Identification (DMI) DREAM challenge aims at developing a benchmark to investigate different algorithms dedicated to the identification of communities from biological networks[3]. The challenge has been divided into two sub-challenges, to identify communities either i) from six biological networks independently, or ii) from all these networks jointly. The second sub-challenge, in particular, intend to test if some approaches can leverage complementary information from multiple networks jointly to define integrated communities. The clustering approaches proposed by the participants are assessed regarding their capacity to reveal *disease communities*, defined as communities significantly associated with genes implicated in diseases in GWAS studies[3,4].

The challengers proposed various strategies and clustering approaches, including kernel clustering, random walks or modularity optimization. We competed with an enhanced version of MolTi, a modularity-based software that we recently developed[5]. We focused on the subchallenge dedicated to the identification of communities from multiple networks as MolTi was initially developed to cluster multiplex networks, i.e., networks composed of different layers of interactions. Molti extended the modularity measure to multiplex networks and adapted the Louvain algorithm to optimize this multiplex-modularity. We have demonstrated that this multiplex approach better identifies the communities than approaches merging the networks, or performing consensus clusterings, both on simulated and real biological datasets[5].

Grounded on these initial results, we here extended and tested our MolTi software, both on simulated data and on the DMI challenge framework. We improved MolTi with the implementation of a randomization procedure, the consideration of edge and layer weights, and a recursive clustering of the classes larger than a given size.

With simulated data, we observed that considering more than one network layer improves the detection of communities, as already noted in Didier *et al.*, 2015[5], but also that communities are better detected with the randomization procedure. With the DMI benchmark, we pointed to a great dependence on the GWAS dataset used for the evaluation and on the FDR threshold defined, but, overall, randomizations and edge and layer weights increase the number of disease communities detected.

Methods

MolTi-DREAM: communities from multiplex networks

We detected communities with an extended version of MolTi[5], a modularity-based software. Although MolTi was specifically designed for multiplex networks, (i.e., networks composed of different layers of interactions), it deals with monoplex networks (i.e. single-layer network) by considering them as multiplex networks composed of a single layer. All the networks are here considered undirected. The new version of MolTi, MolTi-DREAM, and the scripts used for the DMI DREAM challenge are available at https://github.com/gilles-didier/MolTi-DREAM.

Modularity. Network modularity was initially designed to measure the quality of a partition into communities[6], and subsequently used to find such communities. Since finding the partition optimizing the modularity is NP-complete, we applied the meta-heuristic Louvain algorithm[7]. This algorithm starts from the community structure that separates all vertices. Next, it tries to move each vertex from its community to another, picks the move that increases modularity the most, and iterates until no change increases the modularity any more. It then replaces the vertices by the detected communities and performs the same operations on the newly obtained graph, until the modularity cannot be increased any more. In order to handle multiplex networks, we use a multiplex-adapted modularity and an adaptation of the Louvain algorithm for optimizing this multiplex-modularity.

Edge and layer weights Modularity approaches can deal with weighted networks[8], and we modified MolTi to handle weighted networks. We also added the possibility to weight each layer of the multiplex network: the contribution of each layer

in Equation (1) is multiplied by its weight when computing the multiplex modularity.

Multiplex modularity The modularity measure to detect communities in a multiplex network $(X^{(g)})_g$ can be written as

$$\sum_g \frac{w^{(g)}}{2m^{(g)}} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j}^{(g)} - \gamma \frac{S_i^{(g)} S_j^{(g)}}{2m^{(g)}} \right) \delta_{c_i,c_j}, \quad (1)$$

where $X^{(g)}$ denotes the (monoplex) network of the layer g, $w^{(g)}$ is the user-defined weight associated to the network g, $m^{(g)}$ is the sum of the weights of all the edges of $X^{(g)}$, $X_{i,j}^{(g)}$ is the weight of the edge $\{i, j\}$ in $X^{(g)}$, $S_i^{(g)}$ is the sum of the weights of all the edges involving vertex i in $X^{(g)}$, δ_{c_i,c_j} is equal to 1 if i and j belong to a same community and to 0 otherwise, and γ is the resolution parameter modulating the size of the communities detected.

Randomization. We implemented a randomized version of the Louvain algorithm, similar to the one in GenLouvain[9]. Rather than updating the current partition by picking the move leading to the greatest increase of the modularity, we randomly pick a move among those leading to an increase of the modularity. Different runs of the randomized Louvain generally return different partitions, even if the results are often close. MolTi-DREAM runs the randomized Louvain algorithm a user-defined number of times, and returns the partition with the highest modularity.

Simulations of Multiplex Networks with a known community structure

We simulated random multiplex networks with a fixed known community structure and various topological properties by using Stochastic Block Models (SBMs) as in Didier *et al.*, 2015[5]. SBMs model networks with a given community structure under the key assumption that all edges are drawn independently conditionally on the communities to which their nodes belong. In our simulations, we considered multiplex networks with 1,000 vertices split into 20 balanced communities. Each individual network of these multiplex networks is then simulated by independently drawing edges with fixed intra and inter community edge probabilities: 0.1 and 0.01 for sparse networks and 0.5 and 0.2 for dense ones. Dense (resp. sparse) multiplex networks contain only dense (resp. sparse) networks, while mixed networks contain both sparse and dense networks. Multiplex networks with missing data are obtained by randomly removing half of the vertices (and the edges involving them) of the multiplex networks simulated from SBMs.

The relevance of a community structure is assessed by computing the adjusted Rand index[10] between the detected communities and the ones used to simulate the multiplex networks.

The Disease Module Identification challenge benchmark

Biological Networks. The DMI challenge provided six human biological networks: two protein-protein interactions, one literature-curated signaling, one co-expression, one network linking genes essential for the same cancer types, and one network connecting evolutionary-related genes. These six networks have various sizes and edge densities (Table 1). All networks have weighted edges, and all networks but the signaling network are undirected. However, we considered the signaling network as undirected.

Evaluations with GWAS data. The communities identified by the different challengers were evaluated according to the associations of their member genes with GWAS data, using the PASCAL tool described in Lamparter *et al.*, 2016[4]. The procedure leverages the SNP-based *p*-value statistics obtained from 180 GWAS datasets, covering common diseases and traits. The communities are associated with p-values, then corrected for multiple testing, and an FDR threshold is used to determine the number of significant disease communities in a given partition[3,4]. We used three datasets: the "Leaderboard" (76 GWASs) and "Final" (104 GWASs), which were used during the challenge, and their union in a "Total" dataset (180 GWASs).

Obtaining modules in a given size range. The DMI challenge set up two constraints on the submitted communities: no overlap and a size ranging from 3 to 100 nodes. We here post-filtered all partitions to keep only classes containing from 7 to 100 nodes.

Resolution parameter Modularity-based clustering approaches are often associated to a resolution parameter γ to tune the size of the obtained communities. We tested different values of this parameters ($\gamma = 1$, $\gamma = 5$, $\gamma = 10$, $\gamma = 100$), but the leaderboard tests showed clearly better results for the recursive approach. We chose to keep the default $\gamma = 1$ and focused on this recursive procedure.

Recursion procedure We re-clustered all the communities above a certain size (here 100 vertices) by extracting the corresponding subgraphs from the networks and applying recursively the MolTi algorithm. We iterated the process until obtaining only communities with less than 100 vertices, if possible (some communities with more than 100 vertices cannot be split by considering modularity).

Table 1. Number of vertices, of (non-zero-weighted) edges and density of the biological networks used in the DMI challenge.

Network	Number of nodes	Number of edges	Density
1-ppi	17,397	2,232,405	1.48×10^{-2}
2-ppi	12,420	397,309	5.15×10^{-3}
3-signal	5,254	21,826	1.34×10^{-3}
4-coexpr	12,588	1,000,000	1.26×10^{-2}
5-cancer	14,679	1,000,000	9.28×10^{-2}
6-homology	10,405	4,223,606	7.80×10^{-2}

Results

Randomization improves community detection on simulated multiplex networks

To evaluate the accuracy of the community structures detected from the initial MolTi and its improved version that includes the randomization procedure, we simulated random multiplex networks with a fixed, known community structure, and various features (Methods). We observed that considering a greater number of layers always improves the inference of communities, as already observed[5] (Figure 1). In addition, communities are better detected from sparse multiplex networks than from dense ones. We also observed that the randomizations improve the accuracy of the detected communities, in particular for dense multiplex networks, with or without missing data. Increasing the number of randomizations improves the results up to four randomization runs.

Finding disease modules with MolTi

We applied the improved MolTi to the networks provided by the DMI challenge (Methods). We focused on the sub-challenge 2, which was dedicated to the identification of communities from multiple networks. We considered the six DMI biological networks as layers of a multiplex network, and applied the recursion procedure to obtain communities in the required size range. The significant disease communities were selected regarding their enrichments in GWAS-associated genes (Methods). We observed first that the number of detected disease communities is strongly dependent on the GWAS dataset and FDR threshold used (Figure 2). For the FDR threshold used during the challenge, i.e., FDR lower than 0.05, the number of significant disease modules detected slightly increases after randomization (Figure 2).

Multiplex versus monoplex. We next evaluated the added value of the multiplex approach as compared to the identification of modules from the individual networks. When analyzing the significant disease modules obtained for an FDR threshold of 0.1, we observed that combining biological networks in a multiplex generally increases the number of significant modules (Figure 3). However, this does not stand for the cancer and/or homology networks, which lower the number of significant modules retrieved when added as layers of the multiplex. We hypothesize that the community structures of these networks (if they exist) are so unrelated that it is pointless to seek for a common structure by integrating them.

These observations are consistent with the DMI challenge observations, in which the top-scoring team in the sub-challenge 2 handled only the two protein-protein interaction networks. Our algorithm also performs well with the two protein-protein interaction networks, but the highest number of disease modules is retrieved by considering network combinations that exclude the cancer and homology network layers (Figure 3).

Evaluation of the edge and layer weighting. All the six biological networks used in the DMI challenge have weighted edges. We compared the number of disease modules obtained by considering or not considering these weights in the MolTi partitioning, for different FDR thresholds (Table 2). We observed that intra-layer edge weights only has a slight effect on the number of significant disease modules identified, except for the very low significance threshold of 0.01, where it seems pertinent to use these weights.

MolTi-DREAM allows assigning weights to each layer of the multiplex network, for instance to emphasize the layers known to contain more relevant biological information. Given the results obtained on individual networks, we decided to test a combination of weights that would lower the importance of the 5-cancer and 6-homology network layers. We observed that this led to detecting more disease modules (Figure 4). Conversely, less disease modules are detected when higher weights are given to these networks (Figure 4).

Discussion and conclusion

We applied here the MolTi software and various extensions to identify disease-associated communities following the DMI challenge benchmark. The new version of MolTi, MolTi-DREAM, runs a randomization procedure, takes into account edge and layer weights, and performs a recursive clustering of the classes that are larger than a given size. We finished tied for second in the challenge. However, even if we obtained higher scores than monoplex approaches, the difference was not significant and the organizers of the DREAM challenge declared the sub-challenge 2 vacant.

In the simulations, all the networks are randomly generated from the same community structure. These networks can thereby be seen as different and partial views of the same underlying community structure. Combining their information in a suitable way is thereby expected to recover the original structure more accurately. In contrast, combining networks with unrelated community structures (or no structure at all) is rather likely to blur the signal carried by each network. The DMI biological networks are constructed from different biological sources that might correspond to unrelated community structures. This may explain the results of the sub-challenge 2, in which the top-performer used only the two protein-protein interaction networks. With MolTi, we tried to leverage information from the 6 networks together. However, we do not obtain the highest number of disease modules from a multiplex containing all the six networks, but rather from a subset of these networks.

From a biological perspective, the protein-protein networks and the pathway networks are expected to contain mainly physical or signaling interactions between proteins. It has been shown that interacting proteins tend to be co-expressed[11], which could explain why the co-expression network also provides complementary information. In contrast, the cancer network is determined from processes operating at a very different level. Overall, these results show that the sources of biological information that are added as layers of a multiplex need to be evaluated thoroughly.

Evaluating the relevance of the community structure detected from real-life datasets is a very complicated problem since the

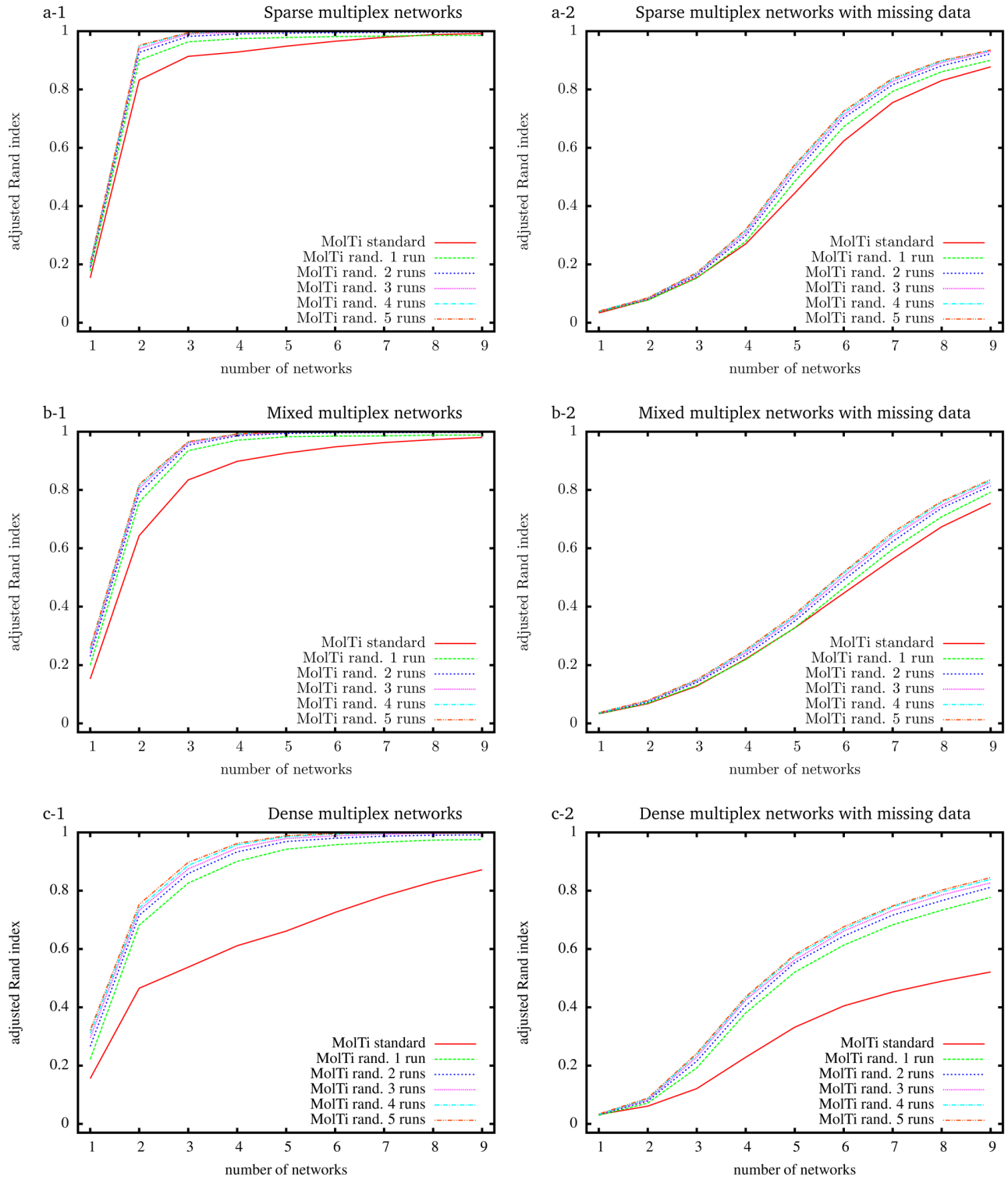


Figure 1. Adjusted Rand indexes between the reference community structure used to generate the random multiplex networks, and the communities detected by standard and randomized MolTi with 1 to 5 randomization runs. Multiplex networks contain from 1 to 9 graph layers. The indexes are averaged over 2,000 random multiplex networks of 1,000 vertices and 20 balanced communities. Each layer of sparse (resp. dense) multiplex networks is simulated with 0.1/0.01 (resp. 0.5/0.2) internal/external edge probabilities. Mixed multiplex networks are simulated by uniformly sampling each layer among these two pairs of edge probabilities. Multiplex networks with missing data (right column) are generated by removing vertices from each layer with probability 0.5.

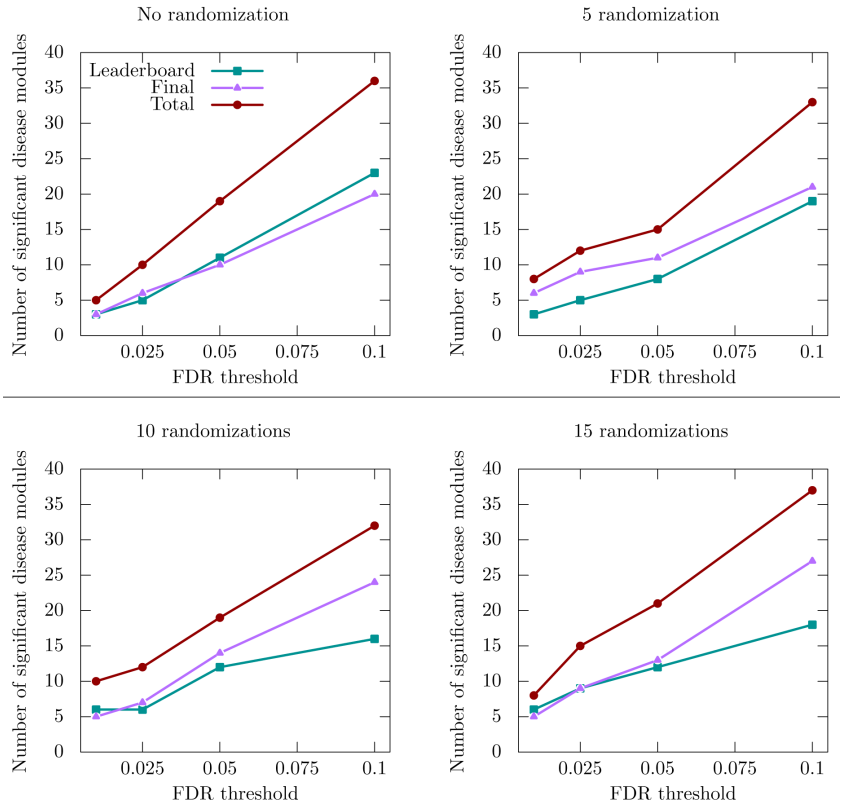
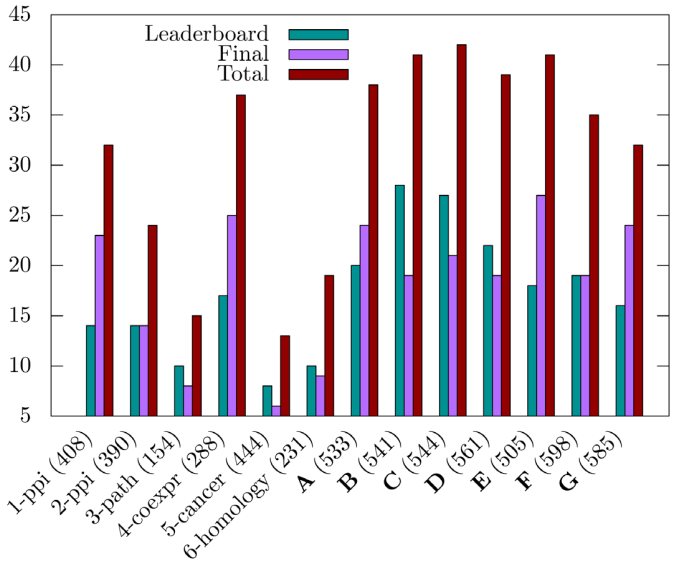


Figure 2. Number of significant disease modules identified from the multiplex network for different GWAS datasets and FDR thresholds. "Leaderboard" and "Final" datasets were used during the training and final evaluation of the challenge, respectively, whereas the "Total" dataset is the union of the two previous ones. The total number of considered communities is 605 in the absence of randomization, 584 for 5 randomizations, 585 for 10 randomizations and 582 for 15 randomizations.



A: 1-ppi, 2-ppi; **B**: 1-ppi, 4-coexpr; **D**: 1-ppi, 2-ppi, 3-path; **E**: 1-ppi, 2-ppi, 3-path, 4-coexpr; **F**: 1-ppi, 2-ppi, 3-path, 4-coexpr, 5-cancer; **G**: 1-ppi, 2-ppi, 3-path, 4-coexpr, 5-cancer, 6-homology.

Figure 3. Number of significant disease modules identified for different combinations of multiplex network layers. Ten randomizations have been applied, and the FDR threshold is set to 0.1. The total number of considered communities for each multiplex network is displayed in parenthesis.

Table 2. Number of significant disease modules detected. over 615 and 585 considered modules in the unweighted and weighted contexts, respectively

FDR	Unweighted	Weighted
0.01	5	10
0.025	13	12
0.05	20	19
0.1	30	32

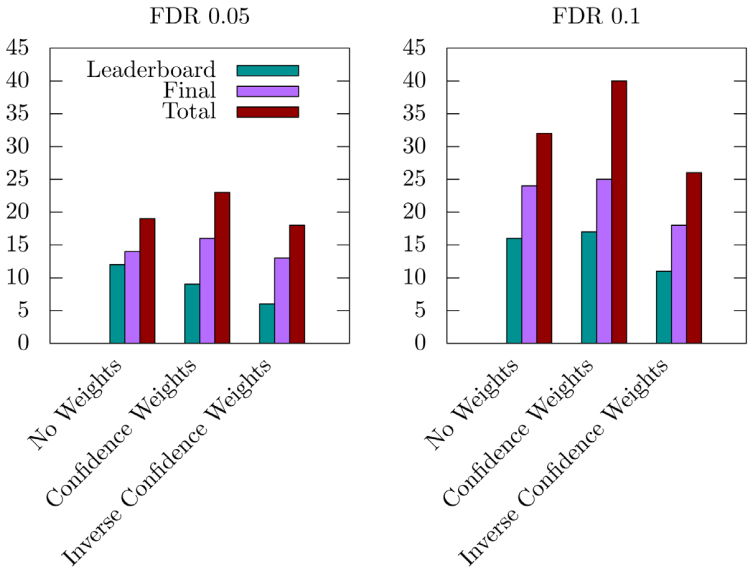


Figure 4. Number of significant disease modules identified with FDR thresholds 0.05 and 0.1, and from three different inter-layer weightings: *No Weights*, i.e., equal weights for all layers (585 modules in total), *Confidence Weights*, i.e., weights proportional to the expected biological relevance: 1-ppi=1, 2-ppi=1, 3-path=1, 4-coexpr=0.5, 5-cancer=0.1, 6-homology=0.1 (555 modules in total), and *Inverse Confidence Weights*, i.e., weights inversely proportional to the expected biological relevance: 1-ppi=0.1, 2-ppi=0.1, 3-path=0.1, 4-coexpr=0.5, 5-cancer=1, 6-homology=1 (648 modules in total).

actual structure is hidden and generally unknown. In this context, the only possibility for assessing the detected communities is to consider indirect evidence provided by some independent biological information. Different teams are thereby developing proxies to evaluate the communities, mainly based on testing the enrichment of genes contained in each community in Pathways or Gene Ontology annotations. The approach followed by the DMI DREAM challenge is based on GWAS data. This GWAS-based evaluation is specific in the sense that it considers *p*-value-weighted annotations rather than usual binary ones, i.e., "annotated/not annotated". This probably contributed to the volatility of the results observed with the DMI DREAM challenge framework.

Data availability

MolTi-DREAM and the scripts used for the DMI DREAM challenge: https://github.com/gilles-didier/MolTi-DREAM

Archived scripts and source code for MolTi-DREAM as at time of publication: http://doi.org/10.5281/zenodo.1468950[12]

License for MolTi-DREAM: GNU 3

Author information

GD designed MolTi and its extensions, AB and AV applied MolTi during and after the challenge. All authors participated in

the design of the study, the interpretation of the results and the writing of the manuscript.

Grant information

The project leading to this publication has received funding from the Centre National de la Recherche Scientifique (PEPS BMI IMFMG), the French "Plan Cancer 2009–2013", and the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" programme.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Hartwell LH, Hopfield JJ, Leibler S, *et al.*: **From molecular to modular cell biology.** *Nature.* 1999; **402**(6761 Suppl): C47–52.
PubMed Abstract | Publisher Full Text
2. Mitra K, Carvunis AR, Ramesh SK, *et al.*: **Integrative approaches for finding modular structure in biological networks.** *Nat Rev Genet.* 2013; **14**(10): 719–32.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Choobdar S, Ahsen ME, Crawford J, *et al.*: **Open community challenge reveals molecular network modules with key roles in diseases.** *bioRxiv.* 2018.
Publisher Full Text
4. Lamparter D, Marbach D, Rueedi R, *et al.*: **Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics.** *PLoS Comput Biol.* 2016; **12**(1): e1004714.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Didier G, Brun C, Baudot A: **Identifying communities from multiplex biological networks.** *PeerJ.* 2015; **3**: e1525.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Newman ME, Girvan M: **Finding and evaluating community structure in networks.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2004; **69**(2 Pt 2): 026113.
PubMed Abstract | Publisher Full Text
7. Blondel VD, Guillaume JL, Lambiotte R, *et al.*: **Fast unfolding of communities in large networks.** *J Stat Mech-Theory E: Theory and Experiment.* 2008; **2008**(10): P10008.
Publisher Full Text
8. Newman ME: **Analysis of weighted networks.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2004; **70**(5 Pt 2): 056131.
PubMed Abstract | Publisher Full Text
9. Mucha PJ, Richardson T, Macon K, *et al.*: **Community structure in time-dependent, multiscale, and multiplex networks.** *Science.* 2010; **328**(5980): 876–8.
PubMed Abstract | Publisher Full Text
10. Santos JM, Embrechts M: **On the use of the adjusted rand index as a metric for evaluating supervised classification.** In C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, *Artificial Neural Networks – ICANN 2009*, 175–184, Berlin, Heidelberg, Springer Berlin Heidelberg, 2009.
Publisher Full Text
11. Rual JF, Venkatesan K, Hao T, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature.* 2005; **437**(7062): 1173–1178.
PubMed Abstract | Publisher Full Text
12. Didier G: **gilles-didier/MolTi-DREAM: v1.1 (Version v1.1).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1468950

5 Landscape of premature aging Diseases

From the point of view of evolutionary biology, **human aging** can be defined as a gradual decline of natural physiological functions, leading to an increase in the mortality rate and a decrease in the reproductive rate (Flatt 2012). Nevertheless, it is challenging to state a global suitable definition with which the scientific community would agree. In contrast, there is no doubt about the fact that aging affects all the human beings, causing a vast impact in our economic and health systems. Indeed, aging is the major risk factor to develop some of the diseases causing most deaths worldwide, such cardiovascular diseases and neurodegenerative disorders (Folgueras et al. 2018).

In this context, it is essential to improve our understanding about the molecular mechanisms driving aging. Several studies have been made in this direction in the last decades, mainly by analyzing the effects of different actions, such as gene silencing or dietary restriction, on the life spans of model organisms (Friedman et al. 1988; Clancy et al. 2001; Wilkinson et al. 2012; Anderson et al. 2009). These works showed that the same altered processes (e.g. nutrient sensing pathways) affected life span in very different organisms, suggesting that they may also be associated with human aging. In addition, some rare diseases present clinical symptoms resembling certain features of physiological aging at an early age. We can hypothesize that the molecular mechanisms deregulated in these human diseases are also involved in physiological aging (Carrero et al. 2016). The interest of studying these premature aging (PA) diseases is therefore twofold: improve the clinical conditions of the patients, and their use as models to investigate aging-associated molecular deregulations.

In this chapter, I first comment about the biological processes that have been described as implicated in human aging so far. Then, I describe some of the PA diseases, i.e. diseases that present some phenotype similar to the physiological aging-phenotype. Finally, I include a draft of an article that we are currently writing. In this work, we aim at defining from a molecular point of view the functional landscape of a group of PA disease

5.1 The hallmarks of aging

It is widely accepted that the accumulation of cellular damage through time is the general cause of aging. However, aging is a very complex process, and this damage accumulates as the result of many different processes occurring in parallel (Gems et al. 2013). López-Otín et al. 2013 gathered and recapitulated the main aging-related mechanisms and proposed nine cellular and molecular

hallmarks of the aging process. These mechanisms, which are briefly describe in the forthcoming lines, seem to be related and interplay among them, determining together the aging phenotype. In section 5.3, we try to associate them with the deregulated biological processes in PA diseases. It is to note that the information presented below summarizes what is exposed in the original article (López-Otín et al. 2013). The nine hallmarks were further grouped into three categories (see Figure 5.1): *(i)* primary hallmarks; *(ii)* antagonistic hallmarks; and *(iii)* integrative hallmarks.

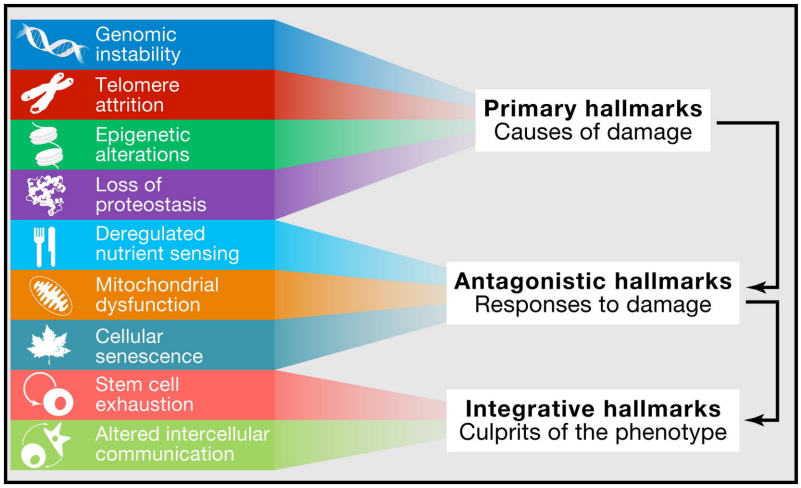


Figure 5.1 – The hallmarks of aging can be grouped into three categories: *i)* The primary hallmarks are the main responsible for the cellular damage caused during aging; *ii)* The antagonistic hallmarks are processes whose primary function is to compensate the damage. However, they can cause harmful effects if their activity increases above some levels or becomes chronic ; *iii)* the integrative hallmarks are responsible of the physiological aging-phenotype as the results of the effects of the previous hallmarks. Figure adapted from (López-Otín et al. 2013)

The primary hallmarks have invariably a negative effect on cellular damage:

1. **Genomic instability:** genetic damage increases throughout life (Moskalev et al. 2013). Various exogenous and endogenous agents can compromise DNA integrity, potentially resulting in different types of genetic abnormalities such as point mutations, translocations or telomere shortening, among others. The organisms possess a large variety of DNA repair mechanisms intended to diminish these damages.
2. **Telomere attrition:** shortening of telomeres is observed during physiological aging in humans (Blasco 2007). Telomerase is the enzyme responsible for the replication of the telomeric DNA. Nevertheless, telomerase is not express in most mammalian somatic cells leading to an intensified accumulation of DNA damage in the telomeres.

3. **Epigenetic alterations:** several epigenetic alterations, including histone modifications, DNA methylation and chromatin remodeling, come together with physiological aging. However, the relationship between these epigenetic alterations and aging is not clear to date. For instance, some works reported a global age-associated hypomethylation, whereas other studies revealed hypermethylation of some loci with age (Maegawa et al. 2010). There are multiple enzymatic systems ensuring the maintenance of epigenetic patterns
4. **Loss of proteostasis:** cells have several mechanisms controlling the biogenesis, correct folding, trafficking and degradation of proteins. These systems are coordinated in such a way that new functional proteins are produced to replace degraded ones. The activity of these processes declines with aging. For instance, it is well documented the accumulation of misfolded proteins with age as a consequence of a decreased activity in the proteasome and lysosome (Rubinsztein et al. 2011; Tomaru et al. 2012).

The antagonistic hallmarks of aging are involved in processes that try to maintain the correct cellular functioning by alleviating the damage caused by other mechanisms. Therefore, they have a positive impact in the organisms in normal conditions. Nevertheless, if their activity increases above some levels or becomes chronic, they can cause a harmful effect. These are the antagonistic hallmarks:

1. **Deregulated nutrient-sensing:** dietary restriction have shown to increase lifespan, or to result in healthier conditions in primates at advanced age (Mattison et al. 2012). In this context, the main actor is the insulin and IGF-1 signaling pathway. Its antagonistic role in aging can be explained as follows: decreased activity on this pathway results in lower rates of cell growth and metabolism, leading to lower rates of cellular damage. However, very low levels of of signaling in the insulin and IGF-1 pathway are lethal (Renner et al. 2009).
2. **Mitochondrial dysfunction:** ATP generation decreases with age due to a reduction in the efficacy of the respiratory chain. Historically, mitochondrial dysfunction has been associated to the increased production of reactive oxygen species (ROS) (Sanz et al. 2008). On the contrary, more recent studies have reported that an increased amount of ROS can extend lifespan (Ristow et al. 2011). To combine these visions, it has been proposed that ROS promotes proliferation and survival in response to stress conditions. Cellular damage and stress increase with age resulting in an larger production of ROS to compensate the increased damage. ROS levels above a particular threshold may become deleterious.
3. **Cellular senescence:** the arrest of the cell cycle is triggered by cellular damage like telemore attrition or other DNA lesions. Therefore, the purpose of senescence is positive: avoid propagation of damaged cells. In

addition, it activates the immune response to remove those cells. On the other hand, senescent cells should be replaced by functional ones, but the number of arrested cells increases with age along with the accumulation of DNA damage. At some point, the regeneration systems may become inefficient, turning cellular senescence into a negative mediator of aging (Collado et al. 2007).

We finally enumerate the integrative hallmarks of aging, which are the results of the seven previously described processes. They directly affect tissue homeostasis, causing thereby the physiological aging-phenotype:

1. **Stem cell exhaustion:** stem cells regenerate cells damaged by the previously described processes. As the global damage increases with age, stem cells become exhausted and lose their regenerative ability, leading to some of the age-related symptoms. For example, the production of immune cells decrease with age as a consequence of hematopoiesis exhaustion. Lower number of immunce cells translate into a higher incidence of some diseases such as anemia (Shaw et al. 2010).
2. **Altered intercellular communication:** Molecular damage not only affects the interior of the cell. Deregulations in the composition and organization of the extracellular matrix may increase with age, hindering neurohormonal signaling such as renin-angiotensin system and insulin-IGF1 signaling (H. E. Yoon et al. 2014; Anisimov 2003).

In the next chapter, I describe diseases displaying premature aging phenotypes at an early age. Some altered processes in these diseases are shared with those of physiological aging. A better understanding of these diseases can shed light about the interplay of the mechanisms described above.

5.2 Premature aging diseases

Hutchinson–Gilford Progeria Syndrome (HGPS; MIM code: 176670) was first described in 1886 by Jonathan Hutchinson (Hutchinson 1886). In 1897, it was also described in an independent study by Hastings Gilford (Rischbieth 1913). HGPS is an extremely rare autosomal dominant genetic disorder clinically characterized by postnatal growth retardation, midface hypoplasia, micrognathia, premature atherosclerosis, absence of subcutaneous fat, alopecia, and generalized osteodysplasia with osteolysis and pathologic fractures. HGPS patients typically die in their mid-teens, usually due to coronary artery disease (De Sandre-Giovannoli et al. 2003). HGPS phenotypically recapitulates some of the aspects of physiological aging at an early age. Indeed, the word progeria comes from Greek words roughly meaning "premature old age". After the first clinical characterization of HGPS, other syndromes displaying aging-related phenotypes at

a premature age were also described. These group of diseases were consequently named **progeroid syndromes** or **premature aging (PA) diseases**.

Progeroid syndromes share some clinical features, in particular the aforementioned PA appearance. Nevertheless, the molecular mechanisms underlying those common features can differ depending on the mutated gene(s) and the biological processes that are hence altered. In this context, progeroid syndromes are usually classified in two main categories (Navarro, Cau, et al. 2006; Carrero et al. 2016):

1. **Syndromes caused by mutations in components of the nuclear envelope and the nuclear matrix**. The nuclear lamina is a dense fibrillar barrier that separates the nucleus from the cytoplasm in eukaryotic cells. Besides providing mechanical support, the nuclear lamina regulates major cellular processes like DNA replication, cell division and chromatin organization. It is composed of lamins, which are encoded by *LMNA* and *LMNB1* genes, and nuclear lamin-associated membrane proteins. A specific mutation in *LMNA* causes HPGS, whereas several different mutations on it cause atypical progeria syndromes whose clinical features are similar to those of HGPS (Barthélémy et al. 2015). Restrictive dermopathy (MIM code: 275210) is a perinatal lethal PA disease caused by mutations in the *ZMPSTE24* gene, which is involved in a post-translational modification of the protein encoded by *LMNA* gene (Navarro, Cadiñanos, et al. 2005). A mutation in the *BANF* gene, whose encoded protein binds to the nuclear lamina, causes the Néstor-Guillermo progeria syndrome (MIM code: 614008). This disease shares many features with HGPS, excepting cardiovascular defects (Cabanillas et al. 2011).
2. **Syndromes caused by mutations in genes encoding DNA repair proteins**. As described in the hallmarks of aging section, several external and internal factors can affect DNA integrity. To counter this deterioration, the organisms have several DNA repair mechanisms. Then, mutations in genes involved in DNA repair may result in diseases displaying increased levels of DNA damage at an early age. These disease can therefore present clinical symptoms resembling the ones happening during physiological aging. We can further classify these syndromes according to the different DNA repair processes compromised by their associated mutations. Werner syndrome (MIM code: 277700), Bloom syndrome (MIM code: 210900) and Rothmund-Thomson syndrome (MIM code: 268400) are caused by mutations in different RecQ helicases (*WRN*, *BLM* and *RECQL4*, respectively) that participate in the repair of double-strand breaks during DNA replication. In addition to a PA appearance, these syndromes also show an increased susceptibility to cancer (Carrero et al. 2016). Another group of progeroid syndromes is related to mutations in genes encoding proteins implicated in the nucleotide excision repair pathway. Cockayne syndrome (mutations

in *ERCC6* or *ERCC8*), xeroderma pigmentosum (mutations in *XPA*, *XPB*, *XPC*, *XPG*, *ERCC4*, *ERCC6*, *DDB2* or *POLH*) and trichothiodystrophy (mutations in *XPB*, *XPD* or *TFB5*) are the most representative diseases of this group and are characterized by developmental defects and neurodegeneration (Navarro, Cau, et al. 2006).

The diseases presented above are considered as the canonical PA syndromes. Nevertheless, there are more diseases for which PA features have been described. Some of them are related to mutations in proteins implicated in well-known aging-related processes. This is the case of dyskeratosis congenita and Hoyeraal-Hreidarsson syndrome, linked to mutations in components of the telomerase complex and, therefore leading to an exacerbated telomere shortening (Carrero et al. 2016). In other diseases hand, the molecular mechanisms suggesting a link with physiological aging are not clear. For instance, Keppen-Lubinsky syndrome is a very rare disorder characterized, among others, by an aged appearance (Masotti et al. 2015). Its causative gene, *KCNJ6*, encodes a potassium channel located in the plasma membrane. Finally, the causative gene of some PA diseases, such as the Storm syndrome, remain unknown.

5.3 Mining multiplex-heterogeneous networks to study premature aging diseases

The study of PA disorders is important to improve the clinical conditions of patients. Moreover, the knowledge of the processes deregulated in these diseases can potentially be transferred to gain a better understanding of physiological aging at molecular level. With this in mind, we are currently working in the article presented below.

Valdeolivas, A., authors to be defined & Baudot, A. (2018). *A Multiplex Seed-expanding Approach to Describe the Landscape of Premature Aging Diseases*.

The supplementary material of this article containing additional tables and figures is available on Appendix C.

A seed-expanding method to define the landscape of Premature Aging Diseases from a multiplex-heterogeneous network

Alberto Valdeolivas, authors to be defined, and Anaïs Baudot

January 2019

Abstract

Motivation: premature aging syndromes are a group of rare genetic disorders that phenotypically recapitulate some of the aspects of physiological aging at an early age. They are however clinically and genetically heterogeneous. Premature aging diseases are the subject of a significant number of studies aiming at identifying their altered processes and transferring this knowledge to physiological aging.
Results: In this work, we present a seed-expanding method that generates modules starting from a given node(s). It is based on a extended version of our previous work, which described a Random Walk with Restart on multiplex and multiplex-heterogeneous networks. To our knowledge, this is the first module generation seed-expanding algorithm in multiplex-heterogeneous networks. We applied the new method to extract the communities associated to 77 disorders annotated with at least one PA phenotype. A first global analysis of the module nodes unveils the genes, processes, and diseases most frequently associated with PA diseases. These include components of the extracellular matrix, but also signaling receptors and transcription factors, for instance. We then hypothesized that the shared PA phenotype observed in all these diseases could be explained by the disruption of common underlying molecular mechanisms. We measured the node overlaps between all the disease communities, and detected clusters of PA diseases. The clusters are annotated for different biological processes, such as DNA repair, extracellular matrix organization, melanogenesis, or mTOR signaling pathways. Overall, they reveal the landscape of perturbed molecular processes in PA diseases, which can be paralleled with the hallmark of physiological aging to help identifying common and specific features.
Availability: https://github.com/alberto-valdeolivas/xxxxxxxxx
Contact: alvaldeolivas@gmail.com, anais.baudot@univ-amu.fr

1 Introduction

Should biological aging be labeled as a disease? At the present time this topic is subject of a vivid debate (Bulterijs et al., 2015; Gavrilov, L. A. and Gavrilova, 2017). The objective fact is that aging affects all the human beings worldwide, causing a vast impact in our economic and health systems. Noticeably, aging is the major risk factor for the development of cancer, neurodegenerative disorders and cardiovascular diseases (Folgueras et al., 2018). However, the molecular mechanisms underlying aging are poorly described. López-Otín et al. (2013) gathered and recapitulated some of the main known processes and proposed nine cellular and molecular hallmarks considered to be responsible of the aging process and leading to the aging phenotype. In addition, aging-related genes are carefully stored and curated in the *GenAge* database. These genes have been identified thanks to studies in model organisms such as *Caenorhabditis elegans*, but also from human disorders caused by mutations leading to aging-related phenotypes (Tacutu et al., 2018).

Premature aging (PA) syndromes, also called Progeroid syndromes, are a group of rare genetic disorders that phenotypically recapitulate some of the aspects of physiological aging at an early age. These syndromes are clinically and genetically heterogeneous (Puzianowska-Kuznicka and Kuznicki, 2005; Navarro et al., 2006). They are usually monogenic, i.e., caused by mutation in singles genes, but can affect few or many tissues, different locus can lead to similar diseases, and, contrarily, the phenotypes and severity can vary considerably across individuals carrying the same mutations. Overall, to our knowledge, a precise description of the molecular mechanisms underlying PA diseases, as well as a comprehensive classification of these diseases do not exist.

In addition, it has been proposed that PA disorders could be used as models to better understand aging-associated molecular deregulations (Carrero et al., 2016). For instance, Hutchinson-Gilford Progeria Syndrome (HGPS) and Werner Syndrome (WS) (*ORPHANET* codes: 740 and 902) mimic many of the features of human aging (Dreesen and Stewart, 2011). Consequently, a significant number of cellular and animal models have been developed based on the mutations of their causative genes: *LMNA* and *WRN*, respectively. These studies allowed identifying some of the altered molecular mechanism in these disorders, such as telomere attrition, defects in DNA repair, or epigenetic alterations (Chang et al., 2004; McCord et al., 2013).

Genes and proteins do not act isolated in cells but rather interact with each other to perform their functions in molecular complexes, pathways, and other biological processes. Mutations in genes and proteins will thereby affect their interaction patterns, and consequently the biological processes in which they are involved (Zhong et al., 2009). Diseases hence arise from network perturbations, and studying the complex biological networks in which genes and proteins participate is a first step towards better understanding the genotype to phenotype relationships in diseases (Schadt, 2009).

Biological interaction data are accumulating since the development of experimental techniques allowing their identification on a large-scale. These interac-

tions are usually represented as large networks in which the nodes correspond to the genes or proteins, and the edges represent their physical or functional interactions. Biological networks are usually organized in communities, i.e. structured around groups of nodes more densely connected with each other than with the rest of the network. These groups of tightly connected nodes, usually called modules, contain genes/proteins likely to be involved in the same cellular functions or processes in cells (Hartwell et al., 1999). The accurate extraction of these modules is promising in biomedicine because studying the modules in which the mutated genes/proteins are involved can reveals the cellular and molecular mechanisms underlying diseases (Furlong, 2013). We hypothesized here that a systematic identification of the modules in which PA-associated genes participate would reveal i) the biological processes perturbed in these diseases, but also ii) define a comprehensive landscape of biological processes perturbed in PA disorders. A plethora of clustering methods have been proposed to extract modules from networks in general, and biological networks in particular. These approaches are based on various assumptions and algorithms, such as modularity optimization or kernel approaches, for instance (Choobdar et al., 2018). Random Walk based methods, such as Walktrap (Pons and Latapy, 2006) and Infomap (Rosvall et al., 2009), are among the most popular approaches.

It is to note that the majority of module-identification algorithms take as input single isolated networks (aka monoplex networks). However, genes and proteins interact with each other in intricate configurations depicting different types of relationships among them. The most commonly studied interactions between proteins are their direct physical interactions, called protein-protein interactions. Nevertheless, proteins can also be members of molecular complexes without a direct physical contact, or participate in the same pathways. In addition, some functional relationships can be inferred from *-omics* data, such as co-expression associations from transcriptomics data. This kind of systems can be better represented by a multiplex networks, i.e. collection of network layers sharing the same nodes, but in which edges belong to different categories (Battiston et al., 2014; Kivelä et al., 2014). Multiplex networks have been shown to provide a more accurate description of complex systems in several fields, such as social network analysis (Battiston et al., 2016), ecology (Stella et al., 2017), computational neuroscience (De Domenico et al., 2016), or biological networks (Didier et al., 2015, 2018). We previously demonstrated the added value of random walks on multiplex and multiplex-heterogeneous networks to identify disease-related genes (Valdeolivas et al., 2018). A heterogeneous network is composed of networks having both different nodes and edges, linked through bipartite interactions. We are using a multiplex network composed of four layers of physical and functional relationships between genes and proteins (protein-protein, pathways, molecular complexes, and co-expression interactions), linked to a disease-disease network thanks to bipartite gene-disease associations.

We developed here a seed-expanding approach to generate modules from multiplex and multiplex-heterogeneous biological networks. To this goal, we extended our random walk with restart algorithms, which is able to explore

multiplex and multiplex-heterogeneous networks around seed node(s) of interest (Valdeolivas et al., 2018). The algorithms are now able to defined modules associated to these seed node(s). We applied the approach to extract modules for more than 70 diseases associated to a PA phenotype. A first global analysis of the modules reveals biological processes, genes and diseases frequently associated with PA diseases. Further classification of all PA diseases revealed a map of the biological processes perturbed in PA. We functionally characterized each cluster and alluded to their link with the hallmarks of physiological aging and age-related diseases. We finally focused on the description of the modules associated to two particular PA diseases, Werner syndrome (WS) and Keppen-Lubinsky syndrome (KLS). We pointed out specific molecular processes likely perturbed in these diseases, as well as some interactions that may account for their shared PA phenotypic feature.

2 Materials and methods

The source code is available on *GitHub* at:
https://github.com/alberto-valdeolivas/xxxxxxxxx

2.1 Disease and gene datasets

We carefully inspected the *Human Phenotype Ontology* (HPO) (Köhler et al., 2017) to select phenotype terms associated with premature aging. We selected the HPO term: *"HP:0007495: prematurely aged appearance"* along with all its descendant's in the HPO ontology graph (Figure 1A and Supplementary Table S-1). We then fetched the 77 diseases from *ORPHANET* (Rath et al., 2012) associated to at least one of these HPO terms. We also extracted from *ORPHANET* their 94 causative genes (Figure 1B and Supplementary Table S-2).

2.2 Interaction Network datasets

Networks sizes and densities are detailed in the Supplementary Table S-3. Network figures were generated using Cytoscape (Shannon et al., 2003).

2.2.1 Gene/Protein networks

We constructed a 4-layers multiplex network composed of a protein-protein interaction (PPI) network, a network connecting proteins according to pathway interaction data extracted from pathway databases, a network in which the links correspond to co-expressed genes constructed from transcriptomics data, and a network of molecular complexes. In a nutshell, the PPI network was built by merging interactions from *PSICQUIC* (Del-Toro et al., 2013) and the *CCSB* Interactome database (et al. Rolland T. Tasan M., 2014). The network of pathways was generated using the R package graphite (Sales et al., 2012). The co-expression network was derived from RNA-Seq data publicly available

on the Human Protein Atlas (Uhlen et al., 2015). We computed Spearman correlations of TPM expression data from 37 tissues and 64 cell lines, and selected the absolute values above 0.7 to be included in the network. The network of complexes was constructed using a matrix model on data retrieved from the *CORUM* database (Giurgiu et al., 2018) and the *Human Protein Complex Map* (Drew et al., 2017). These networks were generated on December 18, 2017.

2.2.2 Disease-disease similarity network

We constructed a disease-disease similarity network following our previous procedure (Valdeolivas et al., 2018). Breifly, all *ORPHANET* (Rath et al., 2012) diseases and their associated phenotypes are fetched from HPO (Köhler et al., 2017). A similarity score is then computed for every pair of diseases based on the relevance of their shared phenotypes. The relevance of each particular phenotype is given by its frequency of appearance in the *HPO* database, as proposed by Westbury et al. (2015). Then, a network is constructed by linking every disease to its 5 most similar diseases.

2.2.3 Gene-disease bipartite associations

The multiplex network and the disease-disease similarity network are integrated into a multiplex-heterogeneous networks. We connected the nodes in each layer of the multiplex network with the disease–disease similarity network thanks to bipartite gene–diseases associations extracted from *ORPHANET* (Rath et al., 2012). We obtained 3 216 associations between genes/proteins and diseases.

2.3 Seed-expanding algorithm for module identification

Based on the global clustering approach proposed by Macropol et al. (2009) on monoplex networks, we developed an algorithm to generate modules from multiplex and multiplex-heterogeneous networks. We used a modified versions of our previously developed algorithms, random walk with restart on multiplex (RWR-M) and random walk with restart on multiplex-heterogeneous networks (RWR-MH) (Valdeolivas et al., 2018). RWR-M and RWR-MH start from a initial set of seed node(s) (gene and/or disease node(s)) and compute a score describing the proximity of all the network nodes *wrt* to the seed(s). In order to generate modules, we applied RWR-M recursively: in each iteration, the top scored node is added to the set of seeds until the modules reached a previously defined size k. The pseudo-code for the repRWR-M is described on Figure 1C. In the RWR-MH version, the top scored disease at each step is also added to the set of seeds, in addition to the top-scoring gene node. We called these algorithms repeated RWR-M (repRWR-M) and repeated RWR-MH (repRWR-MH). RWR-M and RWR-MH were executed with their default parameters as described in Valdeolivas et al. (2018). For sake of comparison, we also applied the repRWR on the individual monoplex networks.

2.4 Clustering coefficient

The clustering coefficient measures the probability that the neighbours of a node are also connected among them, but its definition is not straightforward for multiplex or heterogeneous networks (Kivelä et al., 2014). To measure the clustering coefficient of our multiplex and multiplex-heterogeneous networks, we merged all the individual interaction sources to create an aggregated monoplex network.

2.5 Disease module networks

We computed the overlap between the gene nodes for every pair of disease modules. We constructed a matrix where rows and columns represent each PA disease, and filled it with this number of common genes for every pair of diseases modules. The matrix was used to induced a weighted graph on which we applied different algorithms to identify clusters of disease modules (Figure 1E). The clustering algorithms used are: Fast Greedy (Clauset et al., 2004), Infomap (Rosvall et al., 2009), Label Propagation (Raghavan et al., 2007), leading eigenvector (Newman, 2006), Louvain (Blondel et al., 2008) and Walktrap (Pons and Latapy, 2006). The adjusted Rand index (Rand, 1971; Santos and Embrechts, 2009), which ranges between 0 and 1, was used to compare the similarity among the different community partitions. The closer two community partitions, the greater their adjusted Rand index.

2.6 Enrichment Analyses

Functional enrichment tests were carried out using *g:Profiler* (Reimand et al., 2011). We selected the strong hierarchical filtering, and a p-value threshold of 0.01, after FDR multiple testing correction (Benjamini and Hochberg, 1995). To perform the enrichment tests, *g:Profiler* (Reimand et al., 2011) can access several annotation databases, and we focused here on the results retrieved from Gene Ontology (GO) (Gerald M. Rubin, 2000), Reactome (Fabregat et al., 2016) and KEGG (Kanehisa et al., 2008).

Fisher's exact tests was applied to assess over-representation in aging-related genes in modules (Fisher, 1922). To do so, we downloaded a total of 307 aging-related genes from the *GenAge* database (Build 19) (Tacutu et al., 2018). We also created a subset of 108 aging-related genes by excluding those without direct evidence of association with aging, i.e. associated to the following criteria: *"upstream"*, *"downstream"*, *"functional"* and *"putative"* (Tacutu et al., 2018). To avoid over-fitting, we removed the aging-genes that are also modules seeds when performing the enrichment tests.

3 Results

The goal of this work is two-fold: better understand the molecular mechanism underlying PA diseases, and transfer this knowledge to physiological aging. We

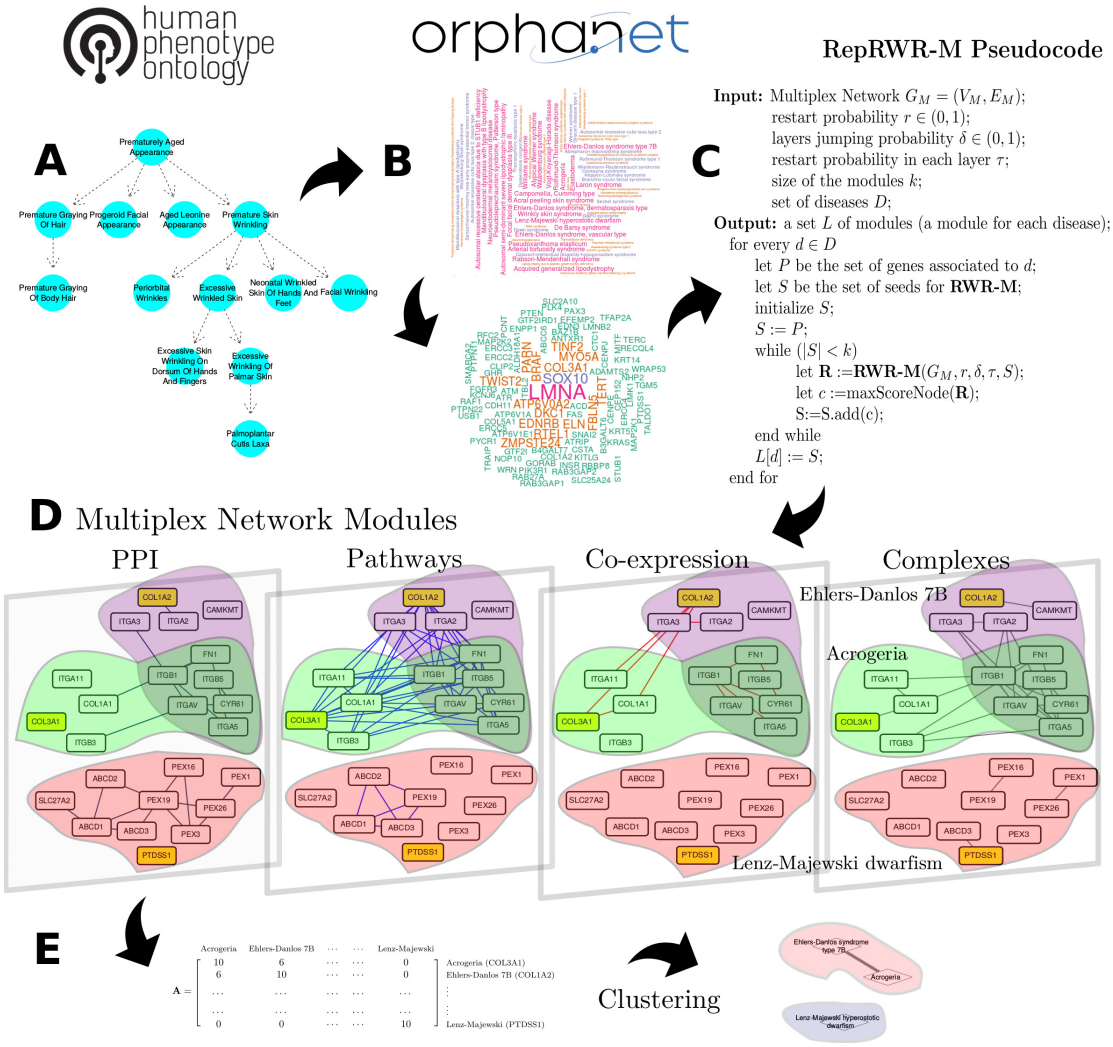


Figure 1: Workflow overview (A) Premature aging phenotypes terms selected from HPO (Köhler et al., 2017). **(B)** Set of PA diseases selected from *ORPHANET* (Rath et al., 2012), and associated causative genes. **(C)** These diseases and genes are the input seeds of the repRWR-M and repRWR-MH algorithms. The pseudocode of the repRWR-M algorithm is detailed. **(D)** Modules of size $k = 10$ obtained with repRWR-M are represented for Acrogeria (ORPHA:2500; *COL3A1*), Ehlers-Danlos syndrome type 7B (ORPHA:99876; *COL1A2*) and Lenz-Majewski hyperostotic dwarfism (ORPHA:2658; *PTDSS1*). **(E)** The number of overlapping genes between every pair of disease modules is used to build a matrix on which we apply a clustering algorithm to obtain a partition of the disease modules.

first identified PA disease modules thanks to new seed-expansion algorithms extracting communities from multiplex and multiplex-heterogeneous networks. Then, we analyzed the identified modules from a global point of view aiming at identifying processes, genes and diseases frequently associated to PA diseases. Then, we took advantage of the global set of modules to define the landscape of PA diseases and better classify them in functional clusters. We also compared these functional process to the hallmarks of physiological aging. Finally, we detailed the modules associated to WS and KLS.

3.1 Identification of Premature Aging Disease Modules with seed-expansion algorithms

We identified in *ORPHANET* 77 diseases associated to at least one premature aging phenotype according to the *Human Phenotype Ontology* (Material and Methods, Supplementary Table S-2). It is to note that some of the selected diseases, such as HGPS, are unambiguously classified as progeroid syndromes in the literature, while others are more debatable. For instance, the premature aging phenotype present in Griscelli syndrome patients is premature graying of hair. Therefore, our PA disease dataset encompass a wide diversity of diseases. These diseases are linked to a total of 94 causative genes (Supplementary Table), ranging from 0 to 11 associated genes per disease. For instance, Böök syndrome (*ORPHANET* code: 1262) has no associated causative gene, whereas Dyskeratosis congenita (*ORPHANET* code: 1775) is associated to 11 causative genes.

We constructed modules starting from the causative gene(s) of the PA diseases as seed node(s) for the repRWR-M algorithm, and starting from both the causative gene(s) and the PA disease as seed node(s) for the repRWR-MH algorithm (Material and Methods). The repRWR-M and repRWR-MH methods are able to generate modules of different sizes k. In order to identify the most suitable size for the PA disease modules, we assessed different topological and functional features of modules containing ($k = 10, 15, 20, 25, 30, 50, 100$) gene nodes.

3.1.1 Module clustering coefficient

The clustering coefficient measures the probability that the neighbours of a node are also connected among them, and therefore the cohesiveness of local groups (A. Barrat, M. Barthelemy, R. Pastor-Satorras, 2004). We computed the clustering coefficients of the modules obtained applying repRWR-M (Materials and Methods), and observed that the clustering coefficient is larger in average for the modules containing 25 and 30 genes (Figure S-1).

We then computed the clustering coefficient of the modules generated with the repRWR-MH algorithm on multiplex-heterogeneous networks. The clustering coefficients of these disease modules are remarkably lower than the ones obtained with the repRWR-M approach on the multiplex network. In addition,

the clustering coefficient of these modules decreases with their size (supplementary Figure S-2). These findings are expected since the multiplex-heterogeneous network contains the bipartite associations connecting genes/proteins with diseases, and these bipartite interactions are very sparse.

3.1.2 Module enrichments in aging genes

We next tested the over-representation of human aging-related genes (taken from *GenAge* database, Tacutu et al. (2018)) in the PA disease modules. The goal here was to evaluate the most relevant module size, but also to compare the modules obtained from monoplex networks, multiplex networks and multiplex-heterogeneous networks. The number of aging-related genes is significantly larger than expected by chance for the PA disease modules derived from all the networks but the co-expression network, with p-values ranging from $(10^{-8}, 10^{-39})$ (Figure S-3 and methods). Strikingly, we also observed that the PA disease modules built from the PPI monoplex network contains the most significant number of aging-related genes (Figure S-3). This could be due to genes added in the *GenAge* database because of interactions with aging-related genes rather than direct evidence of involvement in aging ((Tacutu et al., 2018) and Material and Methods).

We then repeated the over-representation tests only considering the genes with a direct evidence connecting them to aging (Figure S-4). All the p-values are again significant (¡0.05), except for PA disease modules obtained from the co-expression network. The modules discovered by the repRWR-MH are the most enriched in aging genes. The enrichments of the modules detected using the PPI network decrease noticeably, but are still comparable to the ones obtained using repRWR-M. We hypothesize that the co-expression layer might be hindering the performance of the repRWR-M and repRWR-MH. Concerning the module size, the most significant over-representation in aging-related genes, when applying repRWR-M and repRWR-MH, is obtained for modules containing 50 and 25 genes, respectively.

3.1.3 Module enrichments in biological annotations

Next, we explored the biological information contained in the disease modules by testing their enrichments in Gene Ontology (GO) Biological Process (BP) terms Gerald M. Rubin (2000) (Material and Methods). For every PA disease module, we retrieved the number of significant GO terms (FDR¡0.01, Material and Methods), and computed the average number of enriched GO terms in the ensemble of modules of a given size. The best performance is observed for the repRWR-MH (Figure 2). The performances of the repRWR on the multiplex network are comparable to those observed on the monoplex PPI and pathways networks (Figure 2). This could be due to the fact that some GO-BP annotations are inferred from physical interactions, and that here also, the co-expression layer could hamper the efficiency of the repRWR-M and repRWR-MH. The results are similar for the Gene Ontology annotations on molecular

function (MF) and cellular component (CC) (Figures S-5 and S-6). Once again, a good balance between the biological relevance of the modules and their size is reached for $k = 25$, 30 and 50.

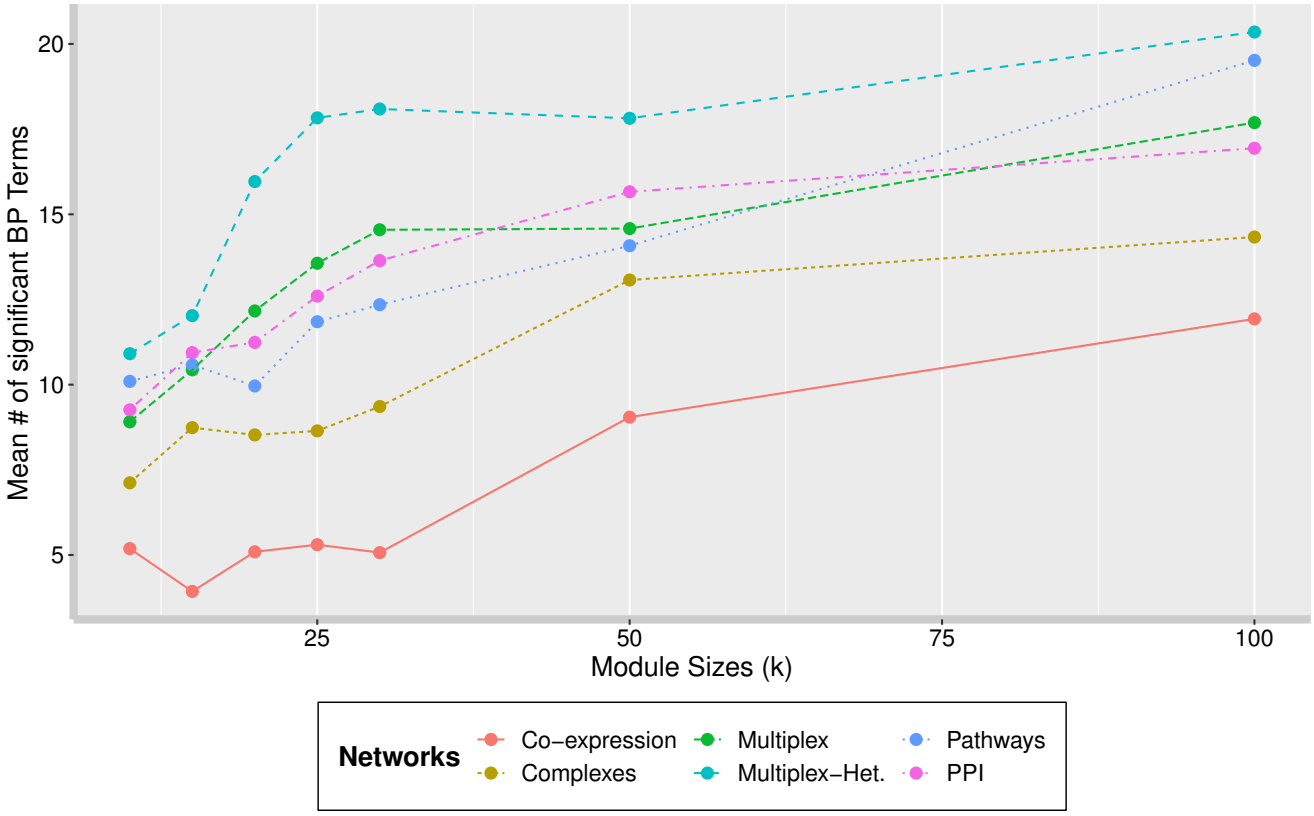


Figure 2: Mean number of significant Gene Ontology Biological Process annotations retrieved in the ensemble of PA disease modules of different sizes when the repRWR algorithms are applied on the different networks under study.

Following the results obtained with the study of the clustering coefficient and the module enrichments in aging and GO annotations, we decided to select a module size of $k = 30$ for the forthcoming analyses. In addition, in spite of the low ability of the co-expression network to capture annotated biological information, we decided to keep it as a layer of the multiplex network because it can provide complementary information for other poorly-annotated genes/proteins.

3.2 Global analysis of PA disease modules

For each one of the 77 PA diseases retrieved from *ORPHANET*, we constructed a module composed of 30 genes, i.e., modules of size $k = 30$, by applying repRWR-MH, a seed-expanding algorithm operating on multiplex-heterogeneous networks (Material and Methods). Gathering together all the nodes contained in the obtained 77 PA disease modules, we identified 704 different genes and 727 different diseases after removing the seeds. We first explored the common features of these modules by identifying the most frequent biological processes, genes and diseases.

3.2.1 Most frequent genes

Considering the 77 PA disease modules of size 30 and after removing the seeds used to create the PA disease modules, we identified a total number of 704 different genes. About 45% of them, i.e., 317 genes, are implicated in more than one module (see (Table 1 for the top-8 most frequent).

Gene name	Number of PA disease module memberships
FBN1	23
COL1A1	19
COL5A2	17
MLPH	14
PAX6	14
COL2A1	12
POLH	12
TGFBR2	12

Table 1: Top most frequent genes according to their memberships in PA disease modules. The modules were built using repRWR-MH on the multiplex-heterogeneous network

Noteworthy, none of top-8 genes shown on Table 1 are present in the *GenAge* database (Tacutu et al., 2018). The most frequent genes described in *GenAge* are *TP53*, *BDNF*, *CDKN2A* and *RET* appearing in from 7 to 10 different PA disease modules. The top-8 most frequent genes have not been described as directly associated to human aging. It has been shown that *FBN1* genetic deficiency accelerates some aspects of vascular aging in mice. Three collagen genes are included in the top-8 most frequent genes associated to PA disease modules. Mice with a mutation in the *COL1A1* gene display a shortened lifespan, when compared with wild-type littermates, and develop some features of premature aging, such as weight loss, decreased bone mineral density, and hypertension (Vafaie et al., 2014).

Another study demonstrated, again in mice, that the deregulation in the expression of the *PAX6* transcription factor is directly associated with aging-

related neuronal dystrophy (Tripathi and Mishra, 2012). Other frequent genes include *MLPH* involved in melanosome transport, the DNA polymerase *POLH* and the TGF-beta receptor *TGFBR2*.

3.2.2 Most frequent diseases

A total of 727 diseases are gathered in the 77 PA disease modules of size 30, among which 294 participate in more than one module. The most frequent disease is Familial melanoma (*ORPHANET* code: 618), associated to many different susceptibility factors, including for instance the *TERT* telomerase reverse transcriptase gene (Table 2). Other most frequent diseases are often associated to the most frequent genes described in the previous section. This is the case for the Ehlers-Danlos syndromes (*ORPHANET* code: 287, 90309, 90318) caused by mutations in collagen. It is to note that many Ehlers-Danlos syndromes are described in *ORPHANET*. Some of them being associated to PA phenotypes according to the HPO, and as such included in the initial list of PA diseases. This is also the case of the Familial thoracic aortic aneurysm and aortic dissection (*ORPHANET* code: 91387) and Neonatal Marfan syndrome (*ORPHANET* code: 284979), which are caused by mutations in *FBN1*. Finally, Griscelli disease type 3 (*ORPHANET* code: 79478), a cutaneous disease caused by mutations in *MLPH* or *MYO5A*.

Disease	**Number of PA disease module memberships**
Familial melanoma	18
Ehlers-Danlos syndrome, classic type	17
Ehlers-Danlos syndrome type 1	17
Ehlers-Danlos syndrome type 2	17
Familial thoracic aortic aneurysm and aortic dissection	17
Localized lipodystrophy	16
Neonatal Marfan syndrome	15
Griscelli disease type 3	14

Table 2: Diseases belonging to a larger number of PA disease modules. The modules were built using RepRWR-MH on the multiplex-heterogeneous network

3.2.3 Most frequent biological processes

We tested for each one of the 77 PA disease modules the enrichments in biological processes (Material and Methods), and checked the most frequent processes among all the modules. Focusing on pathways identified from the *KEGG* database (Kanehisa et al., 2008), we observed that almost half of the 53 *KEGG* pathways significant in more than 10 modules are related to cancer (Data not Shown). The increase risk of cancer in many premature aging diseases is well

documented (Carrero et al., 2016). For instance, Xeroderma pigmentosum (*ORPHA* code: 910) patients develop aggressive skin cancers, because of failures in DNA repair (Cleaver, 2005). And physiological aging is also the major risk of developing cancers (Aunan et al., 2017). Other pathways not directly annotated as cancer pathways are frequently significant in PA disease modules (Table 3). Most interesting pathways include the *PI3K-Akt* signaling pathway, and its closely related *FoxO* and *mTOR* signaling pathways. The most frequent pathways, Relaxin, Melanogenesis and AGE-RAGE are significant because they share reactions with the *MAPK*, *WNT* or *PI3K* canonical pathways. Other encouraging results are related to the cellular senescence and stem cells pluripotency pathways. These processes have been described as hallmarks of physiological aging (López-Otín et al., 2013).

Term ID	**Term Name**	**Number of PA disease module memberships**
KEGG:04926	Relaxin signaling pathway	28
KEGG:04916	Melanogenesis	22
KEGG:04933	AGE-RAGE signaling pathway in diabetic complications	20
KEGG:04151	PI3K-Akt signaling pathway	19
KEGG:04974	Protein digestion and absorption	18
KEGG:04150	mTOR signaling pathway	17
KEGG:04218	Cellular senescence	17
KEGG:04510	Focal adhesion	17
KEGG:01522	Endocrine resistance	16
KEGG:04068	FoxO signaling pathway	16
KEGG:04550	Signaling pathways regulating pluripotency of stem cells	15

Table 3: Top-11 most frequent non-cancer *KEGG* pathways associated to PA disease modules. The modules were built using repRWR-MH

3.3 Classification of PA diseases

The entire set of 77 premature aging diseases are annotated with at least one phenotype associated to the PA ontology term. However, apart from this shared phenotype, each disease is also associated to many other phenotypes, some of them not being related to aging. In addition, the degree of severity and affected tissues can be very heterogeneous.

We here intend to leverage the 77 PA disease modules to obtain a landscape of processes perturbed in PA diseases. To this goal, we built a network where each node represent a PA disease module, and the weighted edges between

modules depict their overlaps, i.e., the number of shared genes between every pair of PA disease modules (Material and Methods). Some disease modules are tightly linked, revealing that their underlying modules of genes are very similar (Figure 3). We hypothesize that these diseases are caused by perturbations of the same cellular processes. We then classified the different PA disease modules by applying several clustering algorithms (Material and Methods). The adjusted Rand index (Rand, 1971; Santos and Embrechts, 2009) was used to compare the similarity among the different partition obtained (Material and Methods). We selected the clusters obtained by the Walktrap algorithm (Pons and Latapy, 2006) for further analysis because it identifies the largest number of clusters, i.e., 10, and these clusters are on average more similar to the ones obtained by the other clustering algorithms (Figure S-6 and materials and methods). These 10 clusters contain from one to 17 diseases.

Cluster 1 contains the Werner Syndrome and four others disorders characterized by abnormal or degenerative conditions of the body's adipose tissue. The main enriched GO terms among the union of the genes contained in these five PA disease modules are related to cholesterol homeostasis and proteasome complex. We can easily relate the latter one with loss of proteostasis, one of the primary hallmarks of aging (López-Otín et al., 2013). Failures on proteostasis lead to protein misfolding, aggregation and other common components of aging-related neurodegenerative disorders like Alzheimer's disease and Parkinson's disease (Tanaka and Matsuda, 2014).

Cholesterol metabolism is also enriched in cluster 2. However, in this cluster, the most significant deregulated pathways are response to radiation and DNA repair, as it can be expected regarding some of of the diseases present in the cluster, such as Xeroderma pigmentosum (*ORPHANET* code: 910) and Rothmund-Thomson syndrome (*ORPHANET* code: 2909). Radiation leads to genetic instability, another primary hallmark of aging, resulting into an increased DNA repair activity. DNA repair initially alleviates the damage, but if its activity becomes excessive may also become harmful itself. In addition, some telomeric related functions appeared enriched in this cluster. Telomere attrition, is described as another hallmark of aging (López-Otín et al., 2013).

Cluster 3 and 4 are related to developmental process and some pathways that are well-known to be implicated in aging, such as *PI3K-Akt* and *EGFR* signaling pathways. Alterations in these pathways have been mentioned in the hallmark of aging (López-Otín et al., 2013). Cluster 3 has also a significant enrichment in the *RAP1* signaling pathway, whereas cluster 4 in the signaling by *RAS* mutants pathway. Both pathways are closely related to cancer.

The main enriched terms for clusters 5 and 6 are epidermis development and keratinization. Epidermal dysfunction contributes to an age-associated increase in the levels of inflammatory cytokines (Hu et al., 2017). Cluster 6 also presents several genes implicated in cell-cell adherens junction, whose impaired activity promotes vascular aging (Chang et al., 2017).

Cluster 7 contains, among others, HGPS the paradigmatic example of progeroid diseases. In this cluster, we can observe an enrichment in components of the extracellular matrix and focal adhesion pathway. Aging-associated changes in

the extracellular environment can alter intercellular communications, one of the hallmarks of aging (López-Otín et al., 2013).

Skin pigmentation and the MAPK1/MAPK3 signaling are the main enriched terms in cluster 8. Pigmentation changes with aging and it is mainly due to sun exposure. Moreover, darkly pigmented individuals suffer less radiation-related DNA damage than lightly pigmented subjects (GJ et al., 2002). Some studies suggest and age-associated impairment in the MAPK signaling pathway (Zhen et al., 1999).

The most enriched terms in the cluster 9 are proton-exporting ATPase activity, reactive oxygen species production in phagocytes and the *mTOR* signaling pathway. These processes can be easily linked to some of the hallmarks of aging, such as mitochondrial dysfunction and deregulated nutrient sensing (López-Otín et al., 2013).

Finally, we can dig into the cluster 10, which is solely integrated by Transaldolase deficiency (*ORPHANET* code: 101028). Its most relevant enriched term is cardiac chamber morphogenesis.

In this section, we classified our set of PA disorders according to the number of shared genes within their associated modules. This approach allowed us to describe the functional landscape of this group of disorders. We indicated the altered mechanisms in the different groups of diseases and linked the clusters to the hallmarks of physiological aging. We hope that these results can shed some light on the complex interplay of processes driving aging, and on how aging-related diseases can arise.

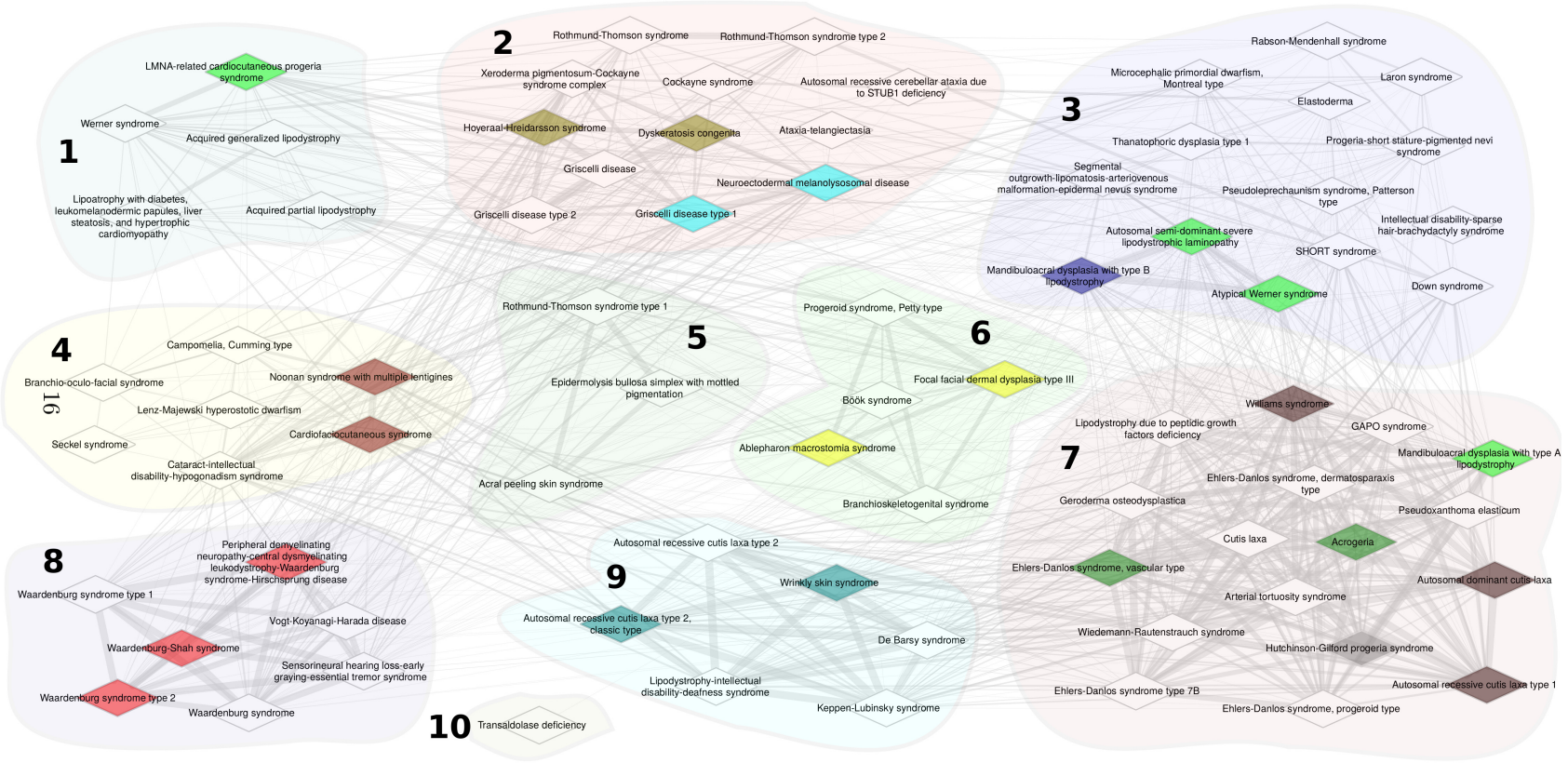


Figure 3: Network of PA disease modules. Each node represents a PA disease and the weighted edges depict the number of shared genes between modules of every pair of diseases. The communities detected by the Walktrap method (Pons and Latapy, 2006) are highlighted in different colors and labeled with their associated number. In addition, diseases sharing any causative gene are filled in the same color.

3.4 Werner syndrome and Keppen–Lubinsky syndrome

In this section, we carried out a detailed analysis of the modules of two specific PA diseases: Werner Syndrome (WS) (*ORPHANET* code: 902) and Keppen–Lubinsky syndrome (KLS) (*ORPHANET* code: 435628). Both disorders share some phenotypes associated to premature aging, but also display specific phenotypes. They are caused by mutations in genes involved in different molecular functions: WS is caused by mutations in the *WRN* gene, which has a major role in genome stability. KLS is caused by mutations in the *KCNJ6* gene, which encodes a potassium channel located in the plasma membrane. Applying the repRWR-MH algorithms using these diseases and their associated genes as seeds, we obtained the modules presented in Figure 4. Enrichment analyses revealed different functional annotations for each module. In the WS module, we detected a significant number of genes implicated in the cell cycle arrest and DNA mismatch repair (Figure 4). The WRN protein is located in the cell nucleus and member of the mismatch repair complex. Mutations in the corresponding gene can lead to an increased level of damaged DNA resulting on more frequent cell cycle arrest (Malumbres and Barbacid, 2009). We also identified a significant enrichment of genes involved in cholesterol metabolism. The inhibition of cholesterol biosynthesis results in cell cycle arrest under certain conditions (Singh et al., 2013), suggesting a link between sterol metabolism and cell cycle control. Furthermore, a study with rabbits showed that the activity of DNA repair pathways declined progressively when cholesterol-fed animals were placed on a normal diet (Martinet et al., 2001).

Among the genes of the Keppen–Lubinsky syndrome module, we identified a large number of genes implicated in the *mTOR* signaling pathway (4), a master regulator of cell cycle. Interestingly, some of the genes involved in the *mTOR* signaling pathway are also involved in face development. Keppen–Lubinsky syndrome patients show a peculiar face with an aged appearance (Basel-Vanagaite et al., 2009), which may arise from an abnormal behaviour of some of these genes through the perturbations originated by the mutation in the *KCNJ6* gene. Finally, we also detected a significant number of components of the extracellular matrix (4). As also shown by the high frequency of collagen genes in PA disease modules, deregulations of the organization of the extracellular matrix has been associated to aging (Frantz et al., 2010). As expected given the very different molecular bases of the two diseases, no gene nor disease belong to both modules. However, we can observed that the modules are linked on one side by interactions between the subnetwork containing some components of the nuclear lamina, such as the *LMNA* and *LMNB1* genes, in the WS module and the *MAPK* subnetwork in the KLS module. On the other side, we can also notice links between a cholesterol subnetwork in the WS module and an extracellular matrix subnetwork in the KLS module. We can hypothesize that the common phenotype of premature aging observed in these diseases might arise from the deregulations around these connected subnetworks.

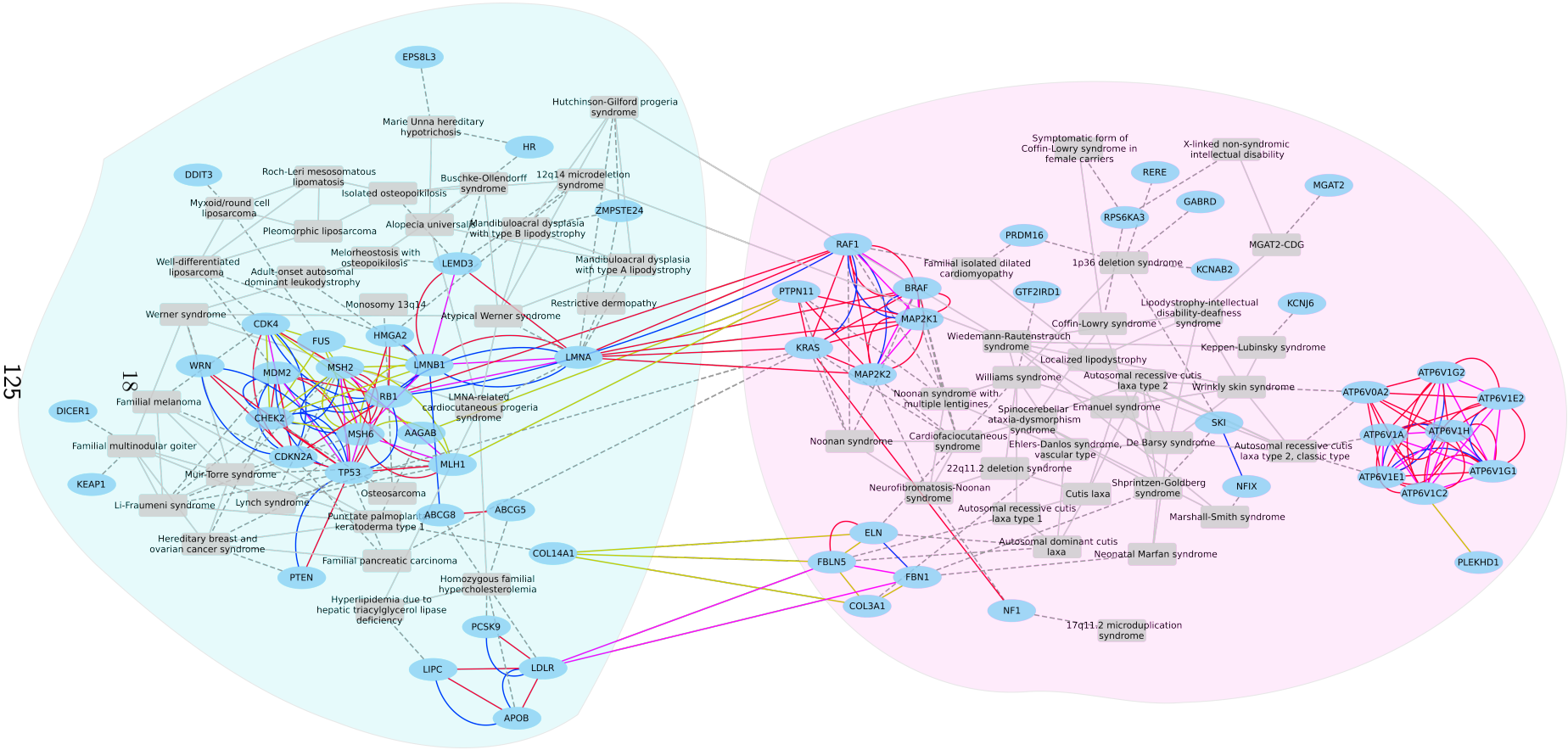


Figure 4: Network representation of the genes and disease nodes belonging to the modules of Werner syndrome (light blue) and Keppen–Lubinsky syndrome (light pink). Blue edges are PPI interactions; yellow edges are co-expression relationships; red edges are pathway interactions; grey edges link nodes belonging to the same complex. It is to note that results are represented as an aggregated network only for visualization purposes.

4 Discussion

Complex systems can be described in a more realistic way using multiplex networks, as it has been shown in articles published in several different research fields (Battiston et al., 2016; Stella et al., 2017; Cardillo et al., 2013; De Domenico et al., 2016). Multiplex networks preserve the topological features of the individual layers that make them up. This is particularly interesting for biological networks, which are built from large-scale interaction datasets derived from different experimental methods, each one of them with their weaknesses, strengths and technical-bias. Furthermore, biological networks are far from complete (eg. high-throughput methods cover less than 20% of all potential pairwise protein interactions in the human cell (Menche et al., 2015)), and their integration increases the available biological information and may reduce the incompleteness of the individual sources.

We have shown previously that, under certain conditions, the multiplex framework is more efficient than single networks or network aggregations to extract communities from biological networks (Didier et al., 2015, 2018). We also demonstrated the increased performance of the RWR on multiplex and multiplex-heterogeneous networks to recover gene-disease associations (Valdeolivas et al., 2018). Based on a extended version of this work, we presented here a seed-expanding method that generates modules starting from genes and/or diseases. To our knowledge, this is the first module generation seed-expanding algorithm in multiplex-heterogeneous networks. We applied the new method to the set of diseases that have been described to present a PA phenotype according to *ORPHANET* (Rath et al., 2012). We showed the increased ability in recovering age-associated genes and a richer biological information of the multiplex-heterogeneous generated modules than the ones built with single networks.

Then, we aimed at defining the functional landscape of PA diseases. To do so, we first analyzed the common mechanisms perturbed in the whole set of diseases. We found some well-established age-related processes, but also some attractive ones whose implication in aging is not evident. We also identified a set of candidate genes likely to be involved in human aging. We then classified the 77 PA diseases in different clusters. We functionally characterized each cluster and alluded to their link with the hallmarks of physiological aging and age-related diseases. Finally, we centered our attention in two particular PA diseases whose causative genes seem to be very distant from a molecular point of view. We described their perturbed mechanisms and pointed out to some potential processes that may account for their shared phenotypic features.

Aging is a extremely complex phenomenon where many processes interplay, so our approach leveraging different sources of information seems to be very suitable to address this topic. Nevertheless, our method can be applied to any set of diseases and/or genes sharing any kind of feature. Moreover, it can be applied to any problem where the goal is to detect seed-centered modules. In our particular case, we decided to set a given size for every module in order to cluster and classify all the PA diseases under equal conditions. This potential limitation

of the method can be overcome by adding some parameters. For instance, we can determine at each iteration whether a new node is added to the module based on measures describing how strongly connected they are (Xu et al., 2018). We should also cite the hesitation about the performance of current clustering methods on multiplex networks (Choobdar et al., 2018). However, our seed-expanding approach is quite different than others multiplex-tailored clustering methods because it just relies on the topological distances within the multiplex-heterogeneous network. We previously showed that these measures are more precise than the ones provided by single or aggregated networks (Valdeolivas et al., 2018).

Another interesting point concerns the election of the diseases database. We selected *ORPHANET*(Rath et al., 2012) rather than *OMIM* because HGPS, the classical example of PA disease, lacks the PA phenotype term in the latter according to the employed version of HPO (Köhler et al., 2017). On the other hand, we also detected some problems in *ORPHANET*(Rath et al., 2012) as some databases entries are kind of duplicated. For instance, this seems to be the case for Griscelli syndrome, which has a general entry in the database (*ORPHANET* code: 381) missing gene associations. However, three more entries for Griscelli syndrome can be found associated to their responsible genes: Type I (*ORPHANET* code: 79476) - *MYO5A*, Type II (*ORPHANET* code: 79477) - *RAB27A* and Type III (*ORPHANET*) code: 79478 - *MLPH* and *MYO5A*. Some repetitive or incoherent results can consequently arise. This kind of issues reflect the need of a careful curation and revision of biological databases.

References

A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. V. (2004). The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752.

Aunan, J. R., Cho, W. C., and Søreide, K. (2017). The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging and Disease*, 8(5):628.

Basel-Vanagaite, L., Shaffer, L., and Chitayat, D. (2009). Keppen–lubinsky syndrome: Expanding the phenotype. *American Journal of Medical Genetics Part A*, 149A(8):1827–1829.

Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(3):1–16.

Battiston, F., Nicosia, V., and Latora, V. (2016). Efficient exploration of multiplex networks. *New Journal of Physics*, 18(4):043035.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):1–12.

Bulterijs, S., Hull, R. S., Björk, V. C. E., and Roy, A. G. (2015). It is time to classify biological aging as a disease. *Frontiers in Genetics*, 6(June):1–5.

Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F., and Boccaletti, S. (2013). Emergence of network features from multiplexity. *Scientific Reports*, 3:1344 EP –.

Carrero, D., Soria-Valles, C., and López-Otín, C. (2016). Hallmarks of progeroid syndromes: lessons from mice and reprogrammed cells. *Disease Models & Mechanisms*, 9(7):719–735.

Chang, F., Flavahan, S., and Flavahan, N. A. (2017). Impaired activity of adherens junctions contributes to endothelial dilator dysfunction in ageing rat arteries. *J Physiol*, 595(15):5143–5158. PMC5538197[pmcid].

Chang, S., Multani, A. S., Cabrera, N. G., Naylor, M. L., Laud, P., Lombard, D., Pathak, S., Guarente, L., and DePinho, R. A. (2004). Essential role of limiting telomeres in the pathogenesis of werner syndrome. *Nature Genetics*, 36:877 EP –.

Choobdar, S., Ahsen, M. E., Crawford, J., Lamparter, D., Lin, J., Hescott, B., and Marbach, D. (2018). Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases. *bioRxiv*.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 70(6):6.

Cleaver, J. E. (2005). Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nature Reviews Cancer*, 5(7):564.

De Domenico, M., Granell, C., Porter, M. A., and Arenas, A. (2016). The physics of multilayer networks. *arXiv*, pages 1–22.

Del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R. C., Galeota, E., Launay, G., Goll, J., Breuer, K., Ono, K., Salwinski, L., and Hermjakob, H. (2013). A new reference implementation of the PSICQUIC web service. *Nucleic acids research*, 41(Web Server issue):601–606.

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ*, 3(0):e1525.

Didier, G., Valdeolivas, A., and Baudot, A. (2018). Identifying communities from multiplex biological networks by randomized optimization of modularity. *F1000Research*, 7(0):1042.

Dreesen, O. and Stewart, C. L. (2011). Accelerated aging syndromes, are they relevant to normal human aging? *Aging*, 3(9):889–895.

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B., and Marcotte, E. M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology*, 13(6):932.

et al. Rolland T. Tasan M., C. B. P. S. Z. Q. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1213–1226.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2016). The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487.

Fisher, R. A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87.

Folgueras, A. R., Freitas-Rodríguez, S., Velasco, G., and López-Otín, C. (2018). Mouse Models to Disentangle the Hallmarks of Human Aging. *Circulation Research*, 123(7):905–924.

Frantz, C., Stewart, K. M., and Weaver, V. M. (2010). The extracellular matrix at a glance. *J Cell Sci*, 123(Pt 24):4195–4200.

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, 29(3):150–159.

Gavrilov, L. A. and Gavrilova, N. S. (2017). "Is Aging a Disease? Biodemographers' Point of View.". *Advances in gerontology*, 4(11):841–842.

Gerald M. Rubin, G. S. M. A. C. A. B. J. A. B. S. S. D. J. T. E. M. A. H. D. P. H. L. I.-T. R. (2000). Gene ontologie: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2018). Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*, page gky973.

GJ, F., S, K., J, V., and et al (2002). Mechanisms of photoaging and chronological skin aging. *Archives of Dermatology*, 138(11):1462–1470.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761supp):C47–C52.

Hu, L., Mauro, T. M., Dang, E., Man, G., Zhang, J., Lee, D., Wang, G., Feingold, K. R., Elias, P. M., and Man, M.-Q. (2017). Epidermal dysfunction leads to an age-associated increase in levels of serum?inflammatory cytokines. *Journal of Investigative Dermatology*, 137(6):1277–1285.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(SUPPL. 1):480–484.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, 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., DeMare, L. E., Devereau, A. D., De Vries, B. B., Firth, H. V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jähn, J. A., James, R., Krause, R., 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., Vulliamy, T., Yu, J., Von Ziegenweidt, J., Zankl, A., Züchner, S., Zemojtel, T., Jacobsen, J. O., Groza, T., Smedley, D., Mungall, C. J., Haendel, M., and Robinson, P. N. (2017). The human phenotype ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–217.

Macropol, K., Can, T., and Singh, A. (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10(1):283.

Malumbres, M. and Barbacid, M. (2009). Cell cycle, cdks and cancer: a changing paradigm. *Nature Reviews Cancer*, 9:153 EP –. Review Article.

Martinet, W., Knaapen, M. W. M., Meyer, G. R. Y. D., Herman, A. G., and Kockx, M. M. (2001). Oxidative dna damage and repair in experimental atherosclerosis are reversed by dietary lipid lowering. *Circulation Research*, 88(7):733–739.

McCord, R. P., Nazario-Toole, A., Zhang, H., Chines, P. S., Zhan, Y., Erdos, M. R., Collins, F. S., Dekker, J., and Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and dna-lamin a/c interactions in hutchinson-gilford progeria syndrome. *Genome Res*, 23(2):260–269. 23152449[pmid].

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601.

Navarro, C. L., Cau, P., and Lévy, N. (2006). Molecular bases of progeroid syndromes. *Human Molecular Genetics*, 15(SUPPL. 2):151–161.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(3).

Pons, P. and Latapy, M. (2006). Journal of Graph Algorithms and Applications Computing Communities in Large Networks Using Random Walks. *arXiv*, 10(2):191–218.

Puzianowska-Kuznicka, M. and Kuznicki, J. (2005). Genetic alterations in accelerated ageing syndromes: Do they play a role in natural ageing? *The International Journal of Biochemistry & Cell Biology*, 37(5):947–960.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(3):1–11.

Rand, W. M. (1971). Journal of the American Statistical Association Objective Criteria for the Evaluation of Clustering Methods Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(December 1971).

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(suppl):W307–W315.

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *European Physical Journal: Special Topics*, 178(1):13–23.

Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). Graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20.

Santos, J. M. and Embrechts, M. (2009). On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In Alippi, C., Polycarpou, M., Panayiotou, C., and Ellinas, G., editors, *Artificial Neural Networks – ICANN 2009*, pages 175–184, Berlin, Heidelberg. Springer Berlin Heidelberg.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 1(Karp 2001):2498–2504.

Singh, P., Saxena, R., Srinivas, G., Pande, G., and Chattopadhyay, A. (2013). Cholesterol biosynthesis and homeostasis in regulation of the cell cycle. *PLOS ONE*, 8(3):1–12.

Stella, M., Andreazzi, C. S., Selakovic, S., Goudarzi, A., and Antonioni, A. (2017). Parasite spreading in spatial ecological multiplex networks. *Journal of Complex Networks*, 5(3):486–511.

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., DIana, E., Lehmann, G., Toren, D., Wang, J., Fraifeld, V. E., and De Magalhães, J. P. (2018). Human Ageing Genomic Resources: New and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090.

Tanaka, K. and Matsuda, N. (2014). Proteostasis and neurodegeneration: The roles of proteasomal degradation and autophagy. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(1):197 – 204. Ubiquitin-Proteasome System.

Tripathi, R. and Mishra, R. (2012). Aging-Associated Modulation in the Expression of Pax6 in Mouse Brain. *Cellular and Molecular Neurobiology*, 32(2):209–218.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J.,

Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419.

Vafaie, F., Yin, H., O'Neil, C., Nong, Z., Watson, A., Arpino, J. M., Chu, M. W., Wayne Holdsworth, D., Gros, R., and Pickering, J. G. (2014). Collagenase-resistant collagen promotes mouse aging and vascular cell senescence. *Aging Cell*, 13(1):121–130.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2018). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 1(August):1–9.

Westbury, S. K., Turro, E., Greene, D., Lentaigne, C., Kelly, A. M., Bariana, T. K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., Jansen, S. B., Bakchoul, T., Crisp-Hihn, A., Erber, W. N., Favier, R., Foad, N., Gattens, M., Jolley, J. D., Liesner, R., Meacham, S., Millar, C. M., Nurden, A. T., Peerlinck, K., Perry, D. J., Poudel, P., Schulman, S., Schulze, H., Stephens, J. C., Furie, B., Van Geet, C., Rendon, A., Gomez, K., Laffan, M. A., Lambert, M. P., Nurden, P., Ouwehand, W. H., Richardson, S., and Mumford, A. D. (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, 7:36.

Xu, B., Liu, Y., Lin, C., Dong, J., Liu, X., and He, Z. (2018). Reconstruction of the Protein-Protein Interaction Network for Protein Complexes Identification by Walking on the Protein Pair Fingerprints Similarity Network. *Frontiers in Genetics*, 9(July):1–10.

Zhen, X., Uryu, K., Cai, G., Johnson, G. P., and Friedman, E. (1999). Age-associated impairment in brain MAPK signal pathways and the effect of caloric restriction in Fischer 344 rats. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 54(12):539–548.

Zhong, Q., Simonis, N., Li, Q.-r., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., Dupuy, D., Brasseur, R., Hill, D. E., Cusick, M. E., and Vidal, M. (2009). Edgetic perturbation models of human inherited disorders. *Molecular Systems Biology*, 5(321).

6 Proteomic portrait of prostate cancer

During my master thesis and the beginning of my PhD thesis, I thoroughly analyzed *proteomics* and *phosphoproteomics* mass spectrometry experimental data obtained from benign, cancerous and resistant prostate cell lines. Some of my results are included in an article that is now under review. In order to introduce this work, I first briefly describe some general aspects about prostate cancer (PC). Then, I discuss about the relevance of *proteomics* and *phosphoproteomics* data, and present a concise overview about the experimental techniques used in our study.

6.1 Prostate cancer

Prostate cancer is the most common non-cutaneous cancer diagnosed in men, and a major cause of death in industrialized countries. It is estimated that each year 1.6 million men are diagnosed and $366\,000$ men die of PC (Pernar et al. 2018). The main risk factors to develop the disease are obesity, age and family history. PC can be considered as an age-related disease since it is very unusual in men younger than 45, but the number of affected patients exponentially increases after the age of 65. As a matter of fact, the average age at the time of diagnosis is 70 (Hankey et al. 1999).

A considerable increase in the overall survival of PC patients has been observed in the last decade due to a more precise identification and characterization of the disease. These achievements have been possible as a result of improved risk stratification, and thanks to the emergence of molecular biomarkers. We must also take into account the great technical advances recently accomplished in magnetic resonance and functional imaging (Litwin et al. 2017). Nevertheless, survival rates among patients with metastatic PC have not significantly contributed to this decline in mortality (Karantanos et al. 2015).

PC progresses through a series of clinical states characterized by tumor growth, hormonal status (castration-sensitive or castration-resistant) and presence or absence of metastases. Treatment of aggressive cases of PC usually involves androgen deprivation therapy (surgical castration, chemical castration or antiandrogen therapy). However, in most of the cases, PC becomes resistant to treatment after one to three years and resume growth despite androgen deprivation. This stage of the disease is therefore called Castration-Resistant Prostate-Cancer (CRPC). It is highly aggressive and incurable, jeopardizing the patient's lifespan and quality of life (Seruga et al. 2010). Progression to CRPC involves several mechanisms such as ligand-independent androgen receptor activation, adaptive

up-regulation of anti-apoptotic genes and mRNA splicing events (Katsogiannou et al. 2015). Nevertheless, many of the resistance mechanisms, as well as their interplay during disease progression, are yet to be fully deciphered.

6.2 Proteomics

The term *proteomics* was coined in analogy to *genomics*, and refers to the study of proteins on a large-scale (James 1997). Unlike the genome, which is roughly constant for an organism, the proteome evolves with time and differ from cell to cell making its study more complex. Historically, protein expression levels were assessed at large-scale by measuring the amount of RNA transcripts. Even though these *transcriptomics* technologies have provided great insights about cell functioning, a lack of strong correlation between transcript expression levels and protein content has been observed (Rogers et al. 2008). Indeed, many factors determine protein expression from mRNA transcripts, calling for the development of new techniques at the protein level. *Proteomics* technologies determine the presence/absence of a protein and can provide a direct measure of its expression levels. Consequently, *proteomics* has gained momentum over the past years thanks to the evolution and improvement of different experimental approaches.

6.2.1 Stable isotope labeling by amino acids in cell culture

Mass spectrometry-based methods are the most conventional techniques to identify proteins on a large-scale. These methods require the conversion of proteins into an ionized form in the gas phase, which can be achieved by bombarding the sample with electrons, for example. These collisions dissociates the proteins into a set of charged fragments (ionized peptides). The ionized peptides are injected into the mass spectrometer where they are accelerated in an electric or magnetic field, and separated according to their mass-to-charge ratio. In tandem mass spectrometry, the previous step is the first stage of mass spectrometry (MS1), and it is used to measure with high precision the mass-to-charge ratio of the peptides. Then, a coupled second stage of mass spectrometry (MS2) identifies the peptides at high speed and accuracy (Chait 2011).

More recently, several new methods were developed to allow the quantification of proteins, **quantitative proteomics**, based on mass spectrometry techniques. The study presented in section 6.3 took advantage of SILAC (Stable Isotope Labeling by Amino acids in Cell culture), one of the most popular methods in quantitative *proteomics*, to compare the expression levels of proteins in different prostate cell lines. In the SILAC method, two different cell cultures are required. The first cell population grows in a specialized media supplemented with heavy forms of essential amino acids (usualy lysine or arginine contain-

ing heavy stable isotopes of carbon or nitrogen). These heavy-labeled amino acids are metabolically incorporated into cells and then into proteins through the usual protein synthesis processes. The second cell culture grows in a media containing the corresponding light forms of essential amino acids. After cell lysis, equivalent amounts of protein from both cell cultures are combined and enzymatically digested in such a way that all the resulting peptides contain at least one labeled amino acid. The mass shift from the labeled sample over the non-labeled one can be detected by the mass spectrometer. The relative protein abundance is determined based on that intensity shift (Geiger, J. Cox, et al. 2010; Geiger, Wisniewski, et al. 2011). The great advantages of SILAC are its relatively simple implementation, quantitative accuracy, and reproducibility over other quantification strategies (Xiulan Chen et al. 2015). SILAC was developed to be performed in cell lines in culture (Hogrebe et al. 2018). This limitation to cell lines represents its main drawback, however, some studies have shown that SILAC can be extended to in vivo animals, such as mice (Zanivan et al. 2011).

6.2.2 Phosphoproteomics with SILAC

Proteins can experience a broad range of chemical modifications after translation. Such post-translational modifications are crucial to determine the protein's function. Phosphorylation, in particular, is the most frequent post-translational modification. Phosphorylation is a reversible process in which a kinase mediates the addition of a phosphate group to an amino acid of the protein. It regulates protein activity state, function, localization and degradation, and the phosphorylation of certain amino acids may provide meaningful clues on pathways activity. It is estimated that between $30\% - 65\%$ of all the human proteins may be phosphorylated. Furthermore, many proteins have multiple potential phosphorylation sites in their sequence (Vlastaridis et al. 2017). In this framework, a whole branch of *proteomics*, called *phosphoproteomics*, has dedicated its efforts to identify differentially phosphorylated sites in proteins under different experimental conditions.

The most prevalent quantification techniques for global *phosphoproteomics* rely also on SILAC approaches. Our collaborators applied a SILAC-based method in order to detect variations in the phosphorylation sites of proteins belonging to prostate cell lines. The experimental protocol is similar to the one described in the previous section,but includes a phosphopeptide enrichment step. The phosphopeptides enrichment is crucial due to the small amount of sites simultaneously phosphorylated. Indeed, it is estimated that only $1 - 2\%$ of all the proteins are phosphorylated at a specific moment (X.-S. Li et al. 2016). Then, mass spectrometry is used to identify phosphoserine, phosphothreonine, and phosphotyrosine-containing peptides, because the most common phosphorylation events in eukaryotic cells take place in those residues (Schmelzle et al. 2006). The additional procedures related to the phosphopeptide enrichment step

make more complicated the experimental setup, leading to some issues concerning reproducible quantification of SILAC-based *phosphoproteomics* techniques (Hogrebe et al. 2018).

6.3 Integrative proteomic and phosphoproteomic profiling of prostate cell lines

Large-scale *omics* techniques can assist us to monitor the molecular changes occurring in cells during cancer progression. In particular, mass spectrometry *proteomics* techniques can directly measure protein abundance under different conditions, allowing to point out to deregulated process and aiding in the discovery of biomarkers. Moreover, the identification of changes in phosphorylation sites through *phosphoproteomics* provide insights about pathways activity. Altogether, we considered that a global *proteomics* and *phosphoproteomics* approach can reveal the mechanisms driving the progression of PC to the incurable CRPC. In this context, we developed the study presented below, where I conducted the statistical and bioinformatics analyses of the *proteomics* and *phosphoproteomics* datasets generated to study different PC cell lines. Finally, we integrated *proteomics* and *phosphoproteomics* resistance markers in a molecular network, to highlight biological processes differentially expressed in the castration-resistant stage of PC.

Katsogiannou, M.†, Boyer, J.B.†, **Valdeolivas, A.†**, Remy, E., et al. (2018). *Integrative proteomic and phosphoproteomic profiling of prostate cell lines*. Under Review. †Equal contribution.

The supplementary material of this article can be downloaded by clicking here:
Supplementary Material

Integrative proteomic and phosphoproteomic profiling of prostate cell lines

Katsogiannou Maria[1,2]*, Boyer Jean-Baptiste[1]*, Valdeolivas Alberto[3,4,5]*, Remy Elisabeth [3], Calzone Laurence [6], Audebert Stéphane[1], Rocchi Palma[1,†], Camoin Luc[1,†], Baudot Anaïs[3,4†]

[1] Aix Marseille Univ, CNRS, INSERM, Institut Paoli-Calmettes, CRCM, Marseille, France

[2] Present address: Clinical Research Unit, Hôpital Saint Joseph, Marseille, France

[3] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

[4] Present address : Aix Marseille Univ, INSERM, MMG, Marseille, France

[5] ProGeLife, Marseille, France

[6] Institut Curie, Mines Paris Tech, INSERM, U900, PSL Research University, F-75005, Paris, France

*Equal contribution

†Correspondance to: anais.baudot@univ-amu.fr, luc.camoin@inserm.fr, palma.rocchi@inserm.fr

Acknowledgement

This work was supported by the French "Plan Cancer 2009-2013" (Systems Biology call). The authors thank Christine Brun, Andreas Zanzoni and all the partners of the Hsp27BioSys project for fruitful discussion. The differential proteomic analyses were done using the Mass Spectrometry facility of Marseille Proteomics (http://map.univmed.fr/) supported by IBISA (Infrastructures Biologie Santé et Agronomie), the Cancéropôle PACA, the Provence-Alpes-Côte d'Azur Region, the Institut Paoli-Calmettes and the Centre de Recherche en Cancérologie de Marseille.

Conflicts of Interest: The authors declare no conflicts of interest.

Integrative proteomic and phosphoproteomic profiling of prostate cell lines

Keywords: proteomics, phosphoproteomics, prostate cancer, resistance, cell lines, network integration

Abstract

Prostate cancer is a major public health issue, mainly because patients relapse after androgen deprivation therapy. Proteomic strategies, aiming to reflect the functional activity of cells, are nowadays among the leading approaches to tackle the challenges not only of better diagnosis, but also of unraveling mechanistic details related to disease etiology and progression.

We conducted here a large SILAC-based Mass Spectrometry experiment to map the proteomes and phosphoproteomes of four widely used prostate cell lines representative of different cancerous and hormonal status. Using SILAC-based mass spectrometry, we characterized housekeeping as well as cell-line specific proteins, phosphosites and functional features of each cell line. In addition, by comparing the sensitive and resistant cell lines, we identified protein and phosphosites differentially expressed in the resistance context. Further data integration in a molecular network highlighted the differentially expressed pathways in the resistant context, namely migration and invasion, RNA splicing, DNA damage repair response and transcription regulation.

Significance

PNT1A, LNCaP, DU145 and PC3 are four widely used prostate cell lines spanning different features, from non-tumorigenic to castration-sensitive to castration-resistance. We describe a joint proteome and phosphoproteome profiling of these cells, identifying more than 3000 proteins and phosphosites. Deep analyses of these data allowed us characterizing the housekeeping, cell-line, cancer and resistance proteomes and phosphoproteomes of the four cell lines. Finally, we interpreted the results at the pathway scale, and integrated the most significant features in a molecular network.

Introduction

Prostate cancer (PC) is a major public health issue in industrialized countries, mainly because patients relapse by castration-resistant disease after androgen deprivation [1, 2]. PC is associated to a panel of clinical states characterized by tumor growth, hormonal status (castration-sensitive or castration-resistant) and presence/absence of metastases. After androgen deprivation therapy, the disease usually progresses to castration-resistant prostate cancer (CRPC), which is highly aggressive and incurable, and jeopardizes the patient's lifespan and quality of life. This progression involves several molecular mechanisms such as ligand-independent androgen receptor activation or adaptive upregulation of anti-apoptotic genes (for review [3]).

Despite an existing treatment guideline for PC and novel clinical trials for CRPC [4, 5], major challenges remain to understand and treat these cancers appropriately. Large-scale *-omics* approaches, able to monitor cancer-induced changes at the cellular level, are among the most promising strategies. Proteomic strategies, by measuring the abundance and activity of proteins, have the ability to directly reflect the functional activity of cells, and to point to deregulations in the most druggable cellular components. In this context, several proteomic studies started to map the landscape of the PC proteome [6-9]. These studies identified biomarkers, such as the proneuropeptide Y [6], as well as proteomic changes associated to prostate cancer progression (e.g., increased anabolic processes and oxidative phosphorylation in primary prostate cancer [6, 7]). Overall, such analyses are valuable not only for diagnosis, but also for providing mechanistic details related to disease etiology and progression.

These proteomic approaches focused on protein quantification, but neglect protein phosphorylation, a key point in the measurement of cellular activity. Protein phosphorylation is a post-translational modification central to signal transduction, that influences cell growth, division, differentiation, cancer development and progression [10, 11]. Protein phosphosites can trigger protein activation or inactivation, and profiling the phosphorylation patterns of proteins can be a powerful tool for understanding key roles in tumor progression and/or drug resistance [12]. Technological advances in the last decade have led to the development of several high-throughput strategies to map the cellular phosphoproteome [13]. Several recent studies examined the phosphoproteome of PC, thereby informing about the activity status of signaling pathways involved in CRPC progression [14-16]. In particular, a recent study integrating phosphoproteomics with

transcriptomics and genomics data revealed the diversity of activated signaling pathways in metastatic PC patients, in relation to their resistance to the anti-androgen therapy [17]. This work further demonstrated the utility of combining -*omics* approaches to better understand PC and CRPC progression.

Here, we used a SILAC-based Mass Spectrometry approach, and identified and quantified the proteomes and phosphoproteomes of four widely used prostate cell lines representative of different cancerous and hormonal status. We first identified a common set of housekeeping proteins highly expressed in all cell lines, and enriched in biological processes related to RNA metabolism and oxidative stress. We further detected that each cell line possesses specific protein, phosphosite and functional features, in particular related to cellular metabolism, transport and protein localization. In addition, comparing the sensitive and resistant cell lines, we were able to pinpoint potential biomarkers differentially expressed or phosphorylated in the resistant context. Finally, pathway and network-level interpretation of the biomarkers reveal cellular processes associated with resistance, including, among others, an upregulation of cell migration, extracellular processes and epithelial-mesenchymal transition, and a downregulation of the cellular respiration.

Experimental Procedures

Cell culture and SILAC Labeling

We cultivated three replicates of four cell lines derived from prostate tissue. PNT1A, a non-tumorigenic SV40-immortalized human prostatic epithelial cell line (ECACC, European Collection of Cell Cultures, England), castration-sensitive (CS) LNCaP (ATCC, American Type Culture Collection (Rockville, MD, USA)) as well as castration-resistant (CR) DU145 and PC3 cell lines (ATCC). All cell lines were routinely cultured at 37°C in a humidified 5% CO_2-95% air atmosphere. They were maintained in Dulbecco's Modified Eagle's Medium (PC3) and RPMI-1640 (Roswell Park Memorial Institute) (Invitrogen, Cergy Pontoise, France), supplemented with 10% fetal bovine serum (FBS). Stable Isotope Labelling with Amino acids in Culture (SILAC) labeling of cell lines was carried out according to Geiger *et al.* [18, 19] using SILAC media with 10% dialyzed fetal bovine serum supplemented with $^{13}C_6{}^{15}N_2$-L-lysine (K8) and $^{13}C_6{}^{15}N_4$-L-arginine (R10). Before creating the reference proteome, the incorporation rate of the heavy amino acid was checked for each cell lines using LC-MS/MS and cell extracts were used if this rate reached 95% (data not shown). Additionally, the interconversion of arginine to proline

was checked and found to be negligible (data not shown). Cells were washed on ice with PBS and collected in a lysis buffer containing 4% SDS, 100 mM Tris-HCl pH7.4, 1 mM DTT (with protease and phosphatase inhibitors cocktails, EDTA-free, ROCHE, usually 1 tablet of each per 10 ml of lysis buffer). Each pellet was resuspended in the lysis buffer and heated to 95°C for 5 min. Viscous lysates were first homogenized mechanically with a syringe and DNAse was added at a 1:40 dilution (benzonase endonuclease, Sigma). Samples were left on ice for 40 min, then centrifuged at 16 000 rcf (g) for 25 min. Supernatants were collected in clean Lo-Bind Eppendorf tubes and protein quantitation was done using BCA assay. After cell lysis, the protein extracts from the four heavy cell lines were mixed in equimolar amounts (1:1:1:1), to generate the super SILAC reference proteome which was then aliquoted and stored at -80°C. For proteomics and phosphoproteomics profiling the reference proteome was mixed in equimolar amounts with protein extracts from each non-labeled cells (Figure 1, c).

Proteomes preparation

40μg of protein extract were loaded on NuPAGE 4–12% bis–Tris acrylamide gels (Life Technologies) to separate proteins, and were stained with Imperial Blue (Pierce, Rockford, IL). Each lane of the gel was cut into 20 bands that were placed in individual Eppendorf tubes. Gel pieces were submitted to an in-gel trypsin digestion using a slightly modified version of the method described by Shevchenko et al.[20]. Briefly, gel pieces were washed and destained using few steps of 100mM ammonium bicarbonate. Destained gel pieces were shrunk with 100 mM ammonium bicarbonate in 50% acetonitrile and dried at RT. Protein spots were then rehydrated using 10mM DTT in 25 mM ammonium bicarbonate pH 8.0 for 45 min at 56°C. This solution was replaced by 55 mM iodoacetamide in 25 mM ammonium bicarbonate pH 8.0 and the gel pieces were incubated for 30 min at room temperature in the dark. They were then washed twice in 25 mM ammonium bicarbonate and finally shrunk by incubation for 5 min with 25 mM ammonium bicarbonate in 50% acetonitrile. The resulting alkylated gel pieces were dried at RT. The dried gel pieces were re-swollen by incubation in 25 mM ammonium bicarbonate pH 8.0 supplemented with 12.5 ng/ml trypsin (Promega) for 1h at 4°C and then incubated overnight at 37°C. Peptides were harvested by collecting the initial digestion solution and carrying out two extractions; first in 5% formic acid and then in 5% formic acid in 60% acetonitrile. Pooled extracts were dried down in a centrifugal vacuum system.

Phosphoproteomes preparation

For each condition, 400μg of cell lysate implemented with 400μg of the reference proteome was precipitated using Acetone/Ethanol (sample/Acetone/EtOH 1/4/4 v/v/v) overnight at -20°C. The acetone-precipitated lysate was resolubilized in 50 mM ammonium bicarbonate, pH 8.0. The soluble proteins were reduced for 45 min at 56°C with 10 mM dithiothreitol (DTT), and then alkylated for 30 min at room temperature in the dark with 10mg/ml Iodoacetamide. The protein mixture was then digested with trypsin (1:50 w/w) overnight. Trypsin was quenched by acidification of the reaction mixture with TFA. The peptide mixture was desalted and concentrated on a C18-SepPak cartridge (Waters, Milford, MA) and eluted with 1x 2 mL of 75% acetonitrile (ACN) in 0.1% TFA and dried down. The phosphopeptide enrichment was performed with TiO2 beads 10 μm (Titansphere TIO, GL Sciences, Japan). Titania beads (6 mg) were prepacked in 200 μL pipet tips filled at the orifice with a C8 Empore disk (3M Empore). Prior to loading samples, the titania tips were rinsed with 200 μL of buffer A (3% TFA/70% CH3CN). Digest samples were reconstituted with 200 μl of loading buffer (buffer A + 1M Glycolic acid). After centrifugation the supernatant was slowly loaded three times onto the titania tip using centrifugation at 300 g for 10 min. The titania beads were sequentially washed with 200 μL loading buffer, twice with 200 μl of buffer A and 200 μl of 0.1% TFA. Bound peptides were eluted with 140 μL of 1% NH_4OH pH 10.5 and dried down with a vacuum concentrator.

Mass Spectrometry analysis

Samples were reconstituted in 0.1% TFA 4% acetonitrile and analyzed by liquid chromatography (LC)–tandem Mass Spectrometry (MS/MS) using Q-Exactive Mass Spectrometer (Thermo Electron, Bremen, Germany) for proteome and phosphopeptide experiments. For the phosphopeptide experiments, an LTQ-Orbitrap Velos Mass Spectrometer (Thermo Electron, Bremen, Germany) was also used. Mass Spectrometers were on line with a nanoLC Ultimate 3000 chromatography system (Dionex, Sunnyvale, CA). Peptides were separated on a Dionex Acclaim PepMap RSLC C18 column at 37°C. First, peptides were concentrated and purified on a pre-column from Dionex (C18 PepMap100, 2 cm x 100 μm I.D, 100 Å pore size, 5 μm particle size) in solution A (0.05% trifluoroacetic acid – 2% acetonitrile). In the second step, peptides were separated on a reverse phase column from Dionex (C18 PepMap100, 15 cm x 75 μm I.D, 100 Å pore size, 2μm particle size) at 300 nL/min flow rate. After column equilibration by 4% of solution B (20% water – 80% acetonitrile – 0.1% formic acid), peptides were eluted from the analytical column by a two steps linear gradient. For proteome analyses, these two steps were 4-25% acetonitrile/H_20; 0.1 % formic acid for 40 min and 25-50% acetonitrile/H_2O;

0.1 % formic acid for 10 min. For phosphopeptide analyses, these two steps were 4-20% acetonitrile/H2O; 0.1% formic acid for 90 min and 20-45% acetonitrile/H2O; 0.1% formic acid for 30 min. For peptides ionisation in the nanospray source, spray voltage was set at 1.5 kV and the capillary temperature at 275 °C. Instrument method for the Q-Exactive was set up in data dependant mode to switch consistently between MS and MS/MS. MS spectra were acquired with the Orbitrap in the range of m/z 300-1700 at a FWHM resolution of 70,000 (AGC target at 1e6, maximum IT 120 ms and 250 ms for proteomes and phosphopeptides respectively). For internal mass calibration the 445.120025 ions was used as lock mass. The 12 most intense ions per survey scan (Intensity threshold 1e5) were selected for HCD fragmentation (AGC target 5e5, NCE 25%, maximum IT 60 ms) and resulting fragments were analysed at a resolution of 17,500 in the Orbitrap. Charge state screening was enabled to exclude precursors with unassigned, 1 and >8 charge states. Fragmented precursor ions were dynamically excluded for 25 s. For phosphopeptides analysis using the LTQ-Orbitrap Velos, the Mass Spectrometer was set as above except for the following parameters. Survey spectra were acquired with a resolution of 60,000 (AGC target at 1e6, maximum IT 100 ms) and the 15 most intense precursors ions per cycle were selected for fragmentation by activation of the neutral loss ions (-48.99, -32.66, and -24.49 Thompson relative to the precursor ions) with collision induced dissociation (AGC target 3,000, NCE 35%, maximum IT 200 ms). The Mass Spectrometry proteomics data, including search result, have been deposited to the ProteomeXchange Consortium (www.proteomexchange.org) [21] via the PRIDE partner repository with datasets identifiers PXD004970 and PXD004992.

Protein identification and quantification

Relative intensity–based SILAC quantification was processed using MaxQuant computational proteomics platform, version 1.3.0.5 [22]. First the acquired raw LC Orbitrap MS data were processed using the integrated Andromeda search engine [23]. Spectra were searched against a SwissProt human database (version 2014.02; 20 284 entries). This database was supplemented with a set of 245 frequently observed contaminants. The following parameters were used for searches: (i) trypsin allowing cleavage before proline [24]; (ii) two missed cleavages were allowed; (ii) monoisotopic precursor tolerance of 20 ppm in the first search used for recalibration, followed by 6 ppm for the main search and 20 ppm for fragment ions from MS/MS; (iii) cysteine carbamidomethylation (+57.02146 Da) as a fixed modification and methionine oxidation (+15.99491 Da) and N-terminal

acetylation (+42.0106 Da) as variable modifications; (iv) a maximum of five modifications per peptide allowed; and (v) minimum peptide length was 7 amino acids. The re-quantify option was enabled to search for missing SILAC partners. The quantification was performed using a minimum ratio count of 2 (unique+razor) and the second peptide option to allow identification of two co-fragmented co-eluting peptides with similar masses. The false discovery rate (FDR) at the peptide level and protein level were set to 1% and determined by searching a reverse database. For protein grouping, all proteins that cannot be distinguished based on their identified peptides were assembled into a single entry according to the MaxQuant rules.

Phosphopeptide identification and quantification

Peptide identification was done similarly than above using MaxQuant software except that serine, threonine, and tyrosine phosphorylation (+79.96633 Da) were allowed as variable modifications.

Preliminary treatment of the datasets

Statistical analyses were done with the Perseus program (version 1.3.0.5; freely available at www.maxquant.org) from the MaxQuant environment [25]. The relative intensity-based SILAC ratio, iBAQ normalised intensities and peptide intensities were uploaded from the proteinGroups.txt and Phospho(STY)Sites.txt files for proteome and phosphoproteome studies, respectively. Proteins marked as contaminant, reverse hits, and “only identified by site” were discarded.

One DU145 cell line replicate in the phosphoproteome study was discarded due to high divergence. In all other cases, for each experiment and for each cell line, the measurements of three replicates were considered. We identified 3 219 proteins (FDR 1% for peptide and protein identification) in triplicates (Supplementary Table 1). We kept for further quantification analyses only those proteins containing at least two valid values (over the 3 replicates) in each cell line. This very conservative approach avoids imputing missing values, and ensures the results of the statistical tests. Doing so, we quantified 1 229 proteins (Supplementary Table 1), used for all subsequent analyses. For the phosphoproteomics analysis, a similar strategy allowed identifying 3 746 phosphosites, of which 563 were kept for further quantification analysis following the previously defined filters (Supplementary Table 2).

Data analyses

R statistical programming environment [26] was used for the treatment of the proteomic and

phosphoproteomic datasets. Expression ratios towards the internal standard were base-2 logarithmized and normalized using z-scores.

Clustering

Unsupervised hierarchical clustering using average method was performed for the proteomic and phosphoproteomic datasets based on Euclidean distances of the expression ratio after normalization.

Identification of the highly-expressed housekeeping proteome

The abundance of each protein in each cell line was computed as the sum of the IBAQ values of every replicate. The housekeeping proteome was obtained by selecting the 10% most abundant proteins matching across all cell lines.

Identification of differentially expressed proteins and phosphosites

We first applied a 1-way ANOVA over the four different cell lines. Benjamini & Hochberg FDR [27] was used for multiple testing corrections, and the threshold for significance was set to 0.01.

Next, to characterize cell line specific protein/phosphosite expression, a t-test was applied to compare the expression value in the three PC cell lines (LNCaP, DU145 and PC3) to the reference non-tumorigenic PNT1A cell line. Benjamini & Hochberg FDR [27] was used for multiple testing corrections, and the threshold of significance set to 0.1.

Pairwise comparisons of protein/phosphosite expression values between the castration-sensitive (CS: LNCaP) and the castration-resistant cell lines (CR: DU145 and PC3) were performed with a t-test, and the threshold of significance set to 0.1 after FDR multiple testing corrections. The results of the pairwise comparisons with the two CR cell lines were combined to define proteins/phosphosites always significantly up- or downregulated in CS as compared to CR.

It is to note that these analyses are conducted with a very stringent filter that select only proteins and phosphosites with at least two over three valid quantification values in all four cell lines. In this context, the proteins identified only in the CR resistant or only in the CS sensitive contexts were discarded, whereas they could be considered as pertinent biomarkers. We then also rescued these potential biomarkers as "CR_only" proteins and phosphosites, having at least two valid expression values in CR and strictly none in CS cell lines and "CS_only" proteins and phosphosites, having at least two valid values in the two CS cell line and strictly none in the CR cell lines.

Pathway and biological process analyses

Functional Enrichments

Enrichment Analyses were conducted with G:Profiler [28], and the significance threshold set to 0.01 after FDR multiple testing corrections. The list of 1 229 proteins used for quantification analyses was used as statistical background. Additionally, the strong filter option was selected on G: Profiler to display solely the most significant ontology in each ontological group, and reduce annotation redundancy.

ROMA

ROMA (Representation and Quantification of Module Activity) is a software focused on the quantification and representation of biological module activity using expression data [29]. The reference gene sets used for this analysis were selected from pathway databases including Reactome [30] and HALLMARK [31]. For each of these pathways, a score corresponding to the weighted sum of the protein expression was computed. The weights are based on the first principal component (PC1). ROMA quantifies the statistical significance of the amount of variance explained by the PC1, and is referred to as the gene set overdispersion. Overdispersed pathways are selected based on a p-value set to 0.01, and the resulting list of pathways can be interpreted as the pathways that contribute significantly to the total expression variance. A detailed presentation of the computational method and use of software can be found at [29]. For this study, we applied on the proteomic dataset an R implementation of ROMA (https://github.com/Albluca/rRoma), which is an improved version of the initial software. The results are presented as a heatmap where the mean value of the scores was computed by types of cancer cell lines: CS for castration-sensitive and CR for castration-resistant, and scaled between -1 and 1.

Ingenuity Pathway Analysis (IPA)

Proteomic datasets were also analyzed with Ingenuity Pathway Analysis (IPA) software (Qiagen, http://www.ingenuity.com/) to predict pathway activation or inhibition. The IPA knowledgebase, derived from literature, compute a score based on one-tailed Fisher test. The final score corresponds to the negative log of p-value, and thresholds were set to 0.01.

KSEA

In order to use the KSEA App (https://casecpb.shinyapps.io/ksea/) [32] on the phosphoproteomic datasets, we computed the fold-changes (FC) between DU145 and LNCaP, and between PC3 and LNCaP, using the mean raw expression values of the replicates. We selected the sites where the expression values are over or under-expressed in both CR cell lines in comparison with LNCaP. Finally, we computed the mean of the FC for the 337 Sites, and normalized it between 0 and 1.

We used this list of sites as input for the KSEA App. The kinases with at least 3 targeted phosphosite substrates, and a p-value smaller than 0.05 were considered as significant.

Network analyses

We constructed a network encompassing molecular complex interaction data by merging Corum complexes [33] and Hu.MAP complexes [34]. This network contains 8 653 nodes and 91 500 edges. Then, we fetched interactions between:

- Proteins significantly differentially expressed in CR versus CS;
- Proteins containing phosphosites significantly differentially expressed in CR versus CS;
- Proteins and proteins containing phosphosites identified only in CR or CS contexts (CR_only and CS_only).

Overall, the interactions between these proteins led to an interaction network of 359 nodes and 1 161 edges, including a large connected component encompassing 194 nodes and 1 098 edges represented with Cytoscape [35]. For visualization purposes, the expression values mapped on the network in figure 4 correspond to the mean of the expression of PC3 and DU145 cell lines.

Results

Proteomic and phosphoproteomic profiles of prostate cell lines

In order to elucidate prostate cancer progression and androgen escape pathway with proteomics and phosphoproteomics identification and quantification, we selected four widely exploited prostate cell lines, namely PNT1A, LNCaP, DU145 and PC3 for proteomic and phosphoproteomic profiling (Figure 1, a). These cell lines are routinely used, and are representative of normal, cancerous and castration-resistant progression of prostate cancer (Figure 1, b). The PNT1A benign prostate cell line was established by immortalizing non-tumorigenic human prostate benign epithelial cells by transfection with the SV40 large-T antigen gene [36]. The castration-sensitive LNCaP cell line was established from metastatic deposit in a lymph node and demonstrates androgen sensitivity [37]. Finally, the two castration-resistant tumor cell lines, DU145 and PC3, were established from metastatic deposits (bone/lumbar spine and central nervous system, respectively), lack the androgen receptor (AR) and are androgen-independent. Moreover, PC3 cells are more tumorigenic and have a higher metastatic potential than DU145 [38]. It is to note that the benign PNT1A cell line

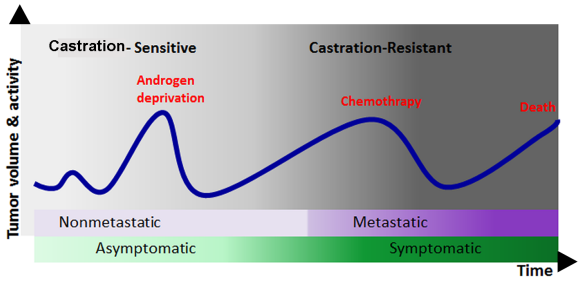
also lacks the AR. The loss of AR and prostate-associated markers (PSA and PAP) appears to be a consistent feature of immortalized cells of prostatic origin, observed in SV40 immortalized cell lines such as PNT1A [39].

We used Stable Isotope Labelling with Amino acids in Culture (SILAC) and Mass Spectrometry (MS) to identify and quantify the proteomes of these four cell lines [40, 41] (Figure 1, c, Experimental Procedures). We elected the spike-in super SILAC method described by Geiger *et al*. [18, 19]. In this protocol, the protein expression in each cell line is compared to the same reference proteome, thereby maximizing the number of detected proteins. We identified 3 219 proteins (Supplementary Table 1). We plotted the median iBAQ values considering all the cell lines to estimate the absolute abundance of the 3 219 identified proteins, and obtained the expected S-shaped distribution covering six orders of dynamic range of MS signals (Experimental procedures, Supplementary Figure 1). The most highly expressed proteins include the core histones, tubulins as well as heat shock proteins. Both the most abundant proteins detected as well as the lowest ones have been previously reported in other studies with a similar approach [42].

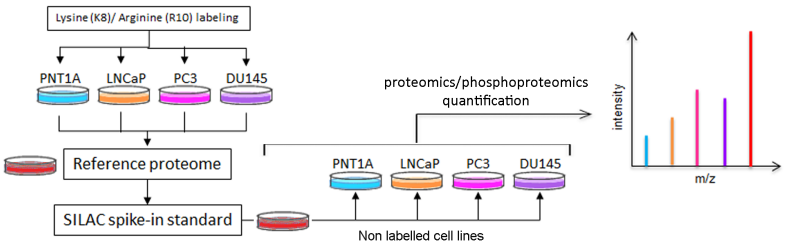
a

Cell Line Name	PNT1A	LNCaP	DU145	PC3
Characterization	SV40 immortalized **normal** prostate	**Castration-sensitive** prostate cancer	**Castration-resistant** prostate cancer	**Castration-resistant** prostate cancer
Origin	Prostate epithelium	Lymph node metastasis	Brain metastasis	Bone metastasis
Androgen Receptor status	AR-	AR+	AR-	AR-

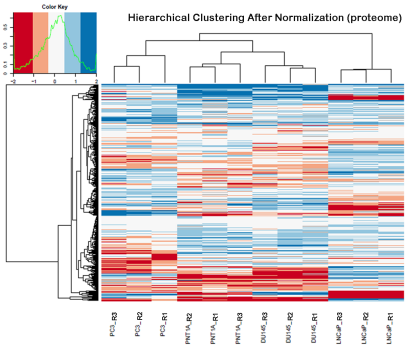
b



c



d



e

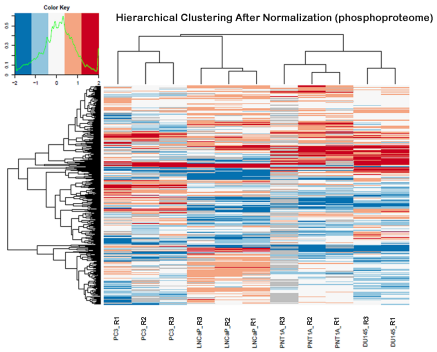


Figure 1. (a) Prostate cell lines used in the present study. AR: Androgen Receptor. (b) Prostate Cancer (PCa) progression over time, from localized asymptomatic castration-sensitive to metastatic castration-resistant disease. (c) SILAC Cell line culture preparation, Spike-in and Mass Spectrometry analysis of the proteomes and phosphoproteomes. Figure adapted from Geiger et al.(39) (d) Hierarchical clustering of the proteomes and (e) phosphoproteomes normalized expression data in the four cell lines.

We kept for further analysis only those proteins containing at least two valid quantification values over the 3 replicates in each cell line. This very conservative approach avoids imputing missing values, and ensures the results of the statistical tests. Doing so, we used for subsequent analyses the quantitative expression data of 1 229 proteins (Supplementary Table 1).

A similar strategy was used to identify and quantify phosphopeptides (Experimental procedures). We identified 3 746 phosphosites, of which 563 were kept for expression analysis considering the strong filters we defined (Supplementary Table 2). These 563 phosphosites correspond to 381 proteins. Overall, 135 proteins were associated with quantitative expression values both at the proteomic and phosphoproteomic levels, with a correlation ranging from 0.43 to 0.62 in each of the four cell lines (Supplementary Figure 2). Therefore, the level of phosphorylation of a protein is not strictly correlated to its level of expression, but might also reflect its activity status.

The unsupervised clustering of the quantified proteins and phosphosites first confirms that the cell line replicates cluster together (Figure 1, d and e). In addition, we observed that the benign PNT1A cell line clusters with the resistant DU145. The genetic instability associated with continuous propagation in culture is a particular problem with benign immortalized cell lines such as PNT1A, in which the insertion of viral DNA drives the cell to replicate continuously [43]. This might explain why its global expression patterns may be similar to that of more malignant cell lines.

The highly-expressed housekeeping proteome

A large number of proteins are essential in all the cells, suggesting that their expression is crucial for the maintenance of basic functionality and survival [44]. These proteins are often called housekeeping. We focused here on the top 10% most expressed proteins in each cell line, corresponding to 321 proteins. Among those 321 highly expressed proteins, 257 are common to the four cell lines (Experimental Procedures, Supplementary Table 3). This means that 80% of the most expressed proteins are the same in all the four cell lines studied here, and can thereby be defined as the highly-expressed housekeeping proteome.

This housekeeping proteome is enriched in functions related to RNA metabolism and response to oxidative stress (functional enrichments with G:Profiler [28], Experimental Procedures and Supplementary Table 5). It contains for instance many RNA binding proteins (mainly from the RPS family) and structural constituents of the ribosome. Eight members of the eukaryotic chaperonin TriC/CCT complex are also highly abundant in all the four cell lines studied.

LNCaP, DU145 and PC3 cancer cell lines characterization

In a second step, we focused on the differences between the cell lines. We first conducted an ANOVA analysis to identify the proteins and phosphosites with the most variation among the four cell lines (Experimental Procedures). 46 proteins and 13 phosphosites (corresponding to 13 proteins) are varying significantly among the four cell lines (FDR < 0.01, Supplementary Tables 3 and 4). Almost half of the 46 ANOVA-significant proteins play a role in stress response (e.g., DNAJB1, VDAC1, ZYX, TCEB1), several are involved in actin cytoskeleton organization (e.g., ACTN1, RHOA, PLS3), and 15 proteins are associated with RNA binding (e.g., CCT6A, NOP2, OCT3, HNRNPA2B1). Among the 13 proteins with phosphosites associated with ANOVA-significant variations in the four cell lines, five are cell-adhesion molecule binding (SEPT9, AHNAK, TNKS1BP1, SCRIB, TAGLN2). Of note, Septin-9 (SEPT9), a filament-forming cytoskeletal GTPase, presents significant variations across the cell lines both at the protein and Serine-30 phosphosite levels (Supplementary Figure 3). SEPT9 has been shown to be highly expressed in PC and positively correlates with malignant progression [45].

Interestingly, two highly expressed housekeeping proteins are associated with phosphosites differentially expressed between the four cell lines according to the ANOVA analysis. First, TAGLN2 presents a significant variation in the Serine-163 expression. In liver cancer, this protein has been reported as a putative tumor suppressor and the involvement of its phosphorylation in actin binding and cell migration has been demonstrated [46]. Second, HNRNPA1, involved in the packaging of pre-mRNA, is highly expressed in the four cell lines, but also shows a significant differential phosphorylation in the Serine-6. To our knowledge, a role for HNRNPA1 phosphorylation in PC has not been described previously.

In order to provide insights into the cellular mechanisms that are involved in cell malignant transformation, we then compared protein and phosphosite expressions of each of the three cancer cell lines (LNCaP, DU145 and PC3) to the benign PNT1A cell line (Experimental Procedures). On a global scale, LNCaP clusters apart and appears to be the most divergent cell line (Figure 1, d). LNCaP cells display 226 up- and 219 downregulated proteins as compared to PNT1A (Supplementary Table 3). Functional enrichment analyses reveal that the proteins upregulated in LNCaP are related to cellular metabolism (Figure 2, a, Supplementary Table 5). The association of tumorigenesis and metabolism is well established; it is not surprising that a cancer cell, in order to meet its increased requirements of proliferation, displays fundamental changes in pathways of energy metabolism and

nutrient uptake [47]. In contrast, the proteins downregulated in LNCaP as compared to PNT1A are enriched in cell recognition and protein/RNA localization processes. Protein and RNA localization mechanisms have shown to play pivotal roles for the presence of specific protein components in cancer cell protrusions, involved in cell migration and invasion [48]. Cell recognition is one of the ways that cells communicate with each other and their environment (adhesion proteins, surface molecules); loss of cell recognition has been shown to lead to cancer development [49]. IPA analysis (Experimental Procedures) confirmed a high metabolic activity in LNCaP, in particular an upregulation of TCA cycle for aerobic respiration. It further delineates a downregulation in the RAN signaling pathway, central to the nucleo-cytoplasmic transport, with seven downregulated proteins, including RAN and its regulator RANBP1, four importins and one exportin (Supplementary Table 5).

The resistant cell line DU145 presents 80 up- and 92 downregulated proteins as compared to PNT1A. Upregulated proteins are enriched in transport and cellular organization processes. Moreover, 61/80 proteins upregulated in DU145 are annotated as extracellular proteins. By contrast, we observed that proteins downregulated in DU145 as compared to PNT1A are enriched in cellular respiration and protein/RNA localization (Figure 2, b, Supplementary Table 5). IPA analysis confirmed an upregulation of actin and Rho signaling and a downregulation of TCA cycle for aerobic respiration.

Finally, the most tumorigenic cell line, PC3, displays 180 up- and 158 downregulated proteins as compared to PNT1A. The upregulated proteins are enriched in vesicle-mediated transport, as it is the case for the other resistant cell line DU145 (Figure 2, c, Supplementary Table 5). In recent years, several publications have proposed vesicle-mediated transport as a mechanism to explain the transfer of resistance to drugs among tumorigenic cells [50]. In addition, many proteins upregulated in PC3 are localized in the extracellular exosome. The proteins downregulated in PC3 are enriched in toxin transport and protein-RNA localization processes. These functional enrichments are complemented by the IPA analysis that revealed strong enrichment in epithelial adherence junction annotation among the upregulated proteins in PC3.

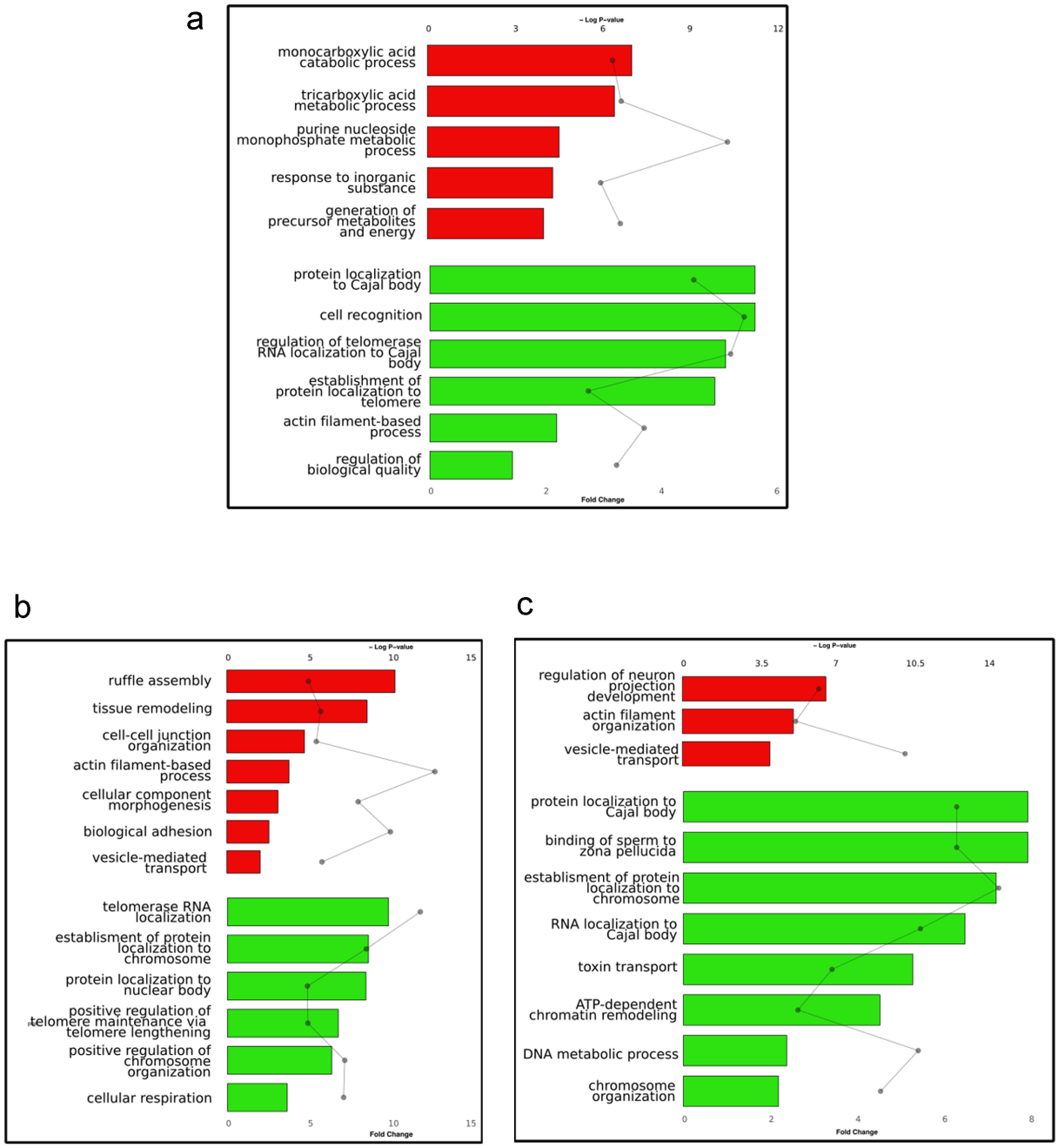


Figure 2 Functional Enrichments of proteins up- and downregulated in PC cell lines. Bar graphs represent relative fold change of Gene Ontology Biological Processes among (a) LNCaP, (b) DU145, (c) PC3 upregulated proteins (red bars) and downregulated proteins (green bars), as compared to PNT1A cells. Significance is represented in the dot plot by –log (P-values).

Overall, we identified 13 proteins upregulated and 19 proteins downregulated together in LNCaP, DU145 and PC3 cells as compared to PNT1A (Supplementary Table 3). We propose that these proteins, differentially expressed in the PC cell lines as compared to the benign cell line, could constitute markers of oncogenic transformation. The upregulated proteins are almost all annotated for secretion and exosomes (e.g., RAB5B, RAB7A, RPL36A, NES, SRI). It has been recently described that exosomes derived from PC cells modulate the prostatic tumor adjacent environment by inducing, among others, tumor-associated target cells growth [51]. Among the 19 downregulated proteins, several are annotated for regulation of protein stability and chaperone-mediated protein folding, and almost half are involved in DNA metabolism. Overall, many proteins of the chaperonin TriC/CCT folding complex, which were observed as highly abundant in all cell lines and thereby classified as housekeeping, are also underexpressed in the three cancer cell lines as compared to PNT1A. The TriC/CCT chaperonin complex directly modulates the folding and activity of as many as 10% of cytosolic client proteins [52, 53]. Recently, the TRiC/CCT complex was also shown to be required for maintaining the wild-type conformation of the tumor suppressor p53 [54]. The downregulation of this chaperone complex could promote the oncogenic functions of p53, such as cancer cell migration and invasion.

We reproduced the cell line characterization protocol for phosphosites, thereby identifying 146 up- and 98 downregulated phosphosites in LNCaP, 5 up- and 3 down in DU145, and 82 up- and 44 down in PC3, as compared to PNT1A. No functional enrichments were significant for the corresponding proteins. Nevertheless, two proteins are associated with phosphosites significantly deregulated in all three PC cell lines as compared to PNT1A. First, TP53BP1 (tumor protein p53 binding protein 1) phosphosites Serine-500 and Threonine-1056 are downregulated in LNCaP. TP53BP1 Serine-500 phosphosite is also downregulated in DU145, and the Threonine-1056 phosphosite downregulated in PC3, as compared to PNT1A. This TP53BP1 protein is well known to be involved in DNA Damage Response (DDR) and its phosphorylation could be a marker of malignant transformation [55]. Previously published studies described TP53BP1 phosphorylation necessary for recruitment to DNA double strand breaks (DSB) [56]. In this context, a downregulation of TP53BP1 phosphorylation, as we observed in the three PC cell lines, could lead to impaired DDR. Second, the DEAD-box RNA helicase 10 (DDX10) Serine-539 phosphosite is significantly upregulated in LNCaP, DU145 and PC3 as compared to PNT1A. DDX10 is an ATP-dependent RNA helicase [57], but, to our knowledge, little is known about its phosphorylation and function in cancer. Other members of the same family of RNA

helicases have been well described, and the phosphorylation of DDX p68 is reported to be associated with cancer development and cell proliferation [58]. Interestingly, the phosphosite Serine-539 that we identified as upregulated in PC cell lines is one of the known post-translational DDX modification sites [59]. Thus, our approach allowed us identifying a well-known cancer-related phosphosite, as well as another potential new candidate.

Identification of Resistance markers

One of the features of PC is, in most cases, its progression to highly aggressive and incurable castration resistant (CR) disease after androgen deprivation therapy. Identifying resistance biomarkers is essential to guide the development of new therapeutic strategies and avoid drug resistance. In order to identify proteins and processes potentially involved in resistance, we compared the protein expression levels in LNCaP cell line (castration-sensitive, CS) with DU145 and PC3, the two castration-resistant cell lines (CR). We found 135 proteins upregulated and 135 downregulated in CR as compared to CS cell lines, and propose them as resistance biomarkers (Supplementary Table 3). Protein biomarkers upregulated in the CR contexts are functionally enriched in processes related to cell-cell adhesion and external communication (Figure 3, a, Supplementary Table 5). This finding is in accordance with previously published studies demonstrating the involvement of these processes in invasion and metastasis, features for which CR cells have a higher potential [60]. Conversely, proteins downregulated in CR are enriched in cellular respiration and protein maturation processes. The downregulation of cellular respiration in the CR context could highlight the Warburg effect [61], in which castration-resistant progression would be associated with a switch from oxidative respiration to glycolysis as primary energy source. The ROMA pathway analysis tool [29] also points to a downregulation in CR cells of oxidative phosphorylation and metabolic pathways such as fatty acid metabolism, as well as signaling pathways related to p53 and apoptosis (Figure 3, b). Conversely, it reveals an upregulation of the epithelial-mesenchymal transition (EMT) and reactive oxygen species (ROS) pathways. EMT refers to the morphological and functional alterations involved in cancer invasion [62]. Finally, IPA analysis points to an upregulation of actin cytoskeleton and Rho signaling in CR cells, and further identifies an upregulation of Integrin Signaling and Calpain protease signaling.

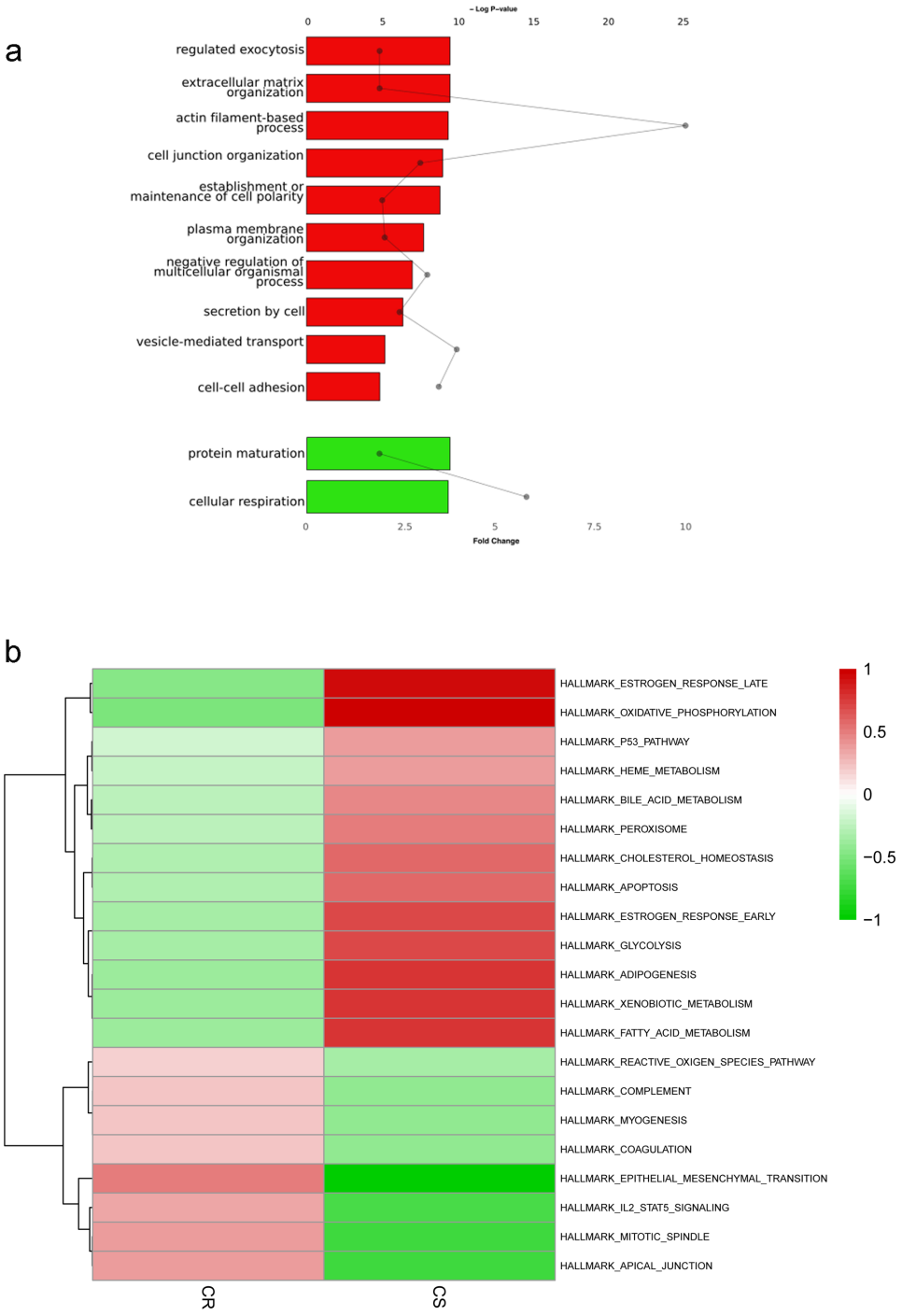


Figure 3 Functional Enrichments of protein resistance biomarkers. (a) Bar graphs represent relative fold change of Gene Ontology Biological Processes among proteins upregulated (red bars) and downregulated (green bars) in Castration Resistant cell lines DU145 and PC3 as compared to castration-sensitive LNCaP cell line. Significance is represented in the dot plot by –log (P-values). (b) Clustered heatmap of ROMA pathway analysis. The color intensities correspond to the values of the scores of each signaling pathway (red, upregulated; green, downregulated).

Phosphoproteomics data reveal 41 phosphosites upregulated and 40 downregulated in CR versus CS, which we also predict as resistance biomarkers (Supplementary Table 4). The 41 upregulated phosphosites concern essentially nuclear proteins involved in functions such as transcription regulation, genome stability and RNA processing (e.g., SMARCC1, SRRM1, SRRM2, SSB). The deregulation of these processes, and their implication in cancer development and progression, have been largely documented [63]. Moreover, 2 kinases are hyper-phosphorylated in the resistant context. First, the Serine/threonine-protein kinase N2 (PKN2), which plays a role in the regulation of cell cycle progression, actin cytoskeleton assembly, cell migration, cell adhesion, tumor cell invasion and transcription activation signaling processes. It was recently shown to be phosphorylated by the PI-3 Kinase pathway and implicated in prostate cancer progression [64]. Second, the nuclear receptor binding protein (NRBP1), which is involved in subcellular ER-Golgi trafficking. To our knowledge, a role of its phosphorylation status in prostate cancer has not been described previously.

The 40 downregulated phosphosites concern mainly proteins involved in cell migration and invasion, such as PLEC, AHNAK, ESYT1 and ZYX. A group of kinases sharing the same identified peptide and that consequently cannot be distinguished with the MS experiment (CDK2;CDK3;CDK1;CDC2) shows a decrease in phosphorylation activity in the CR context.

Kinase-Substrate Enrichment Analysis (KSEA [32], Experimental Procedure) predicted from these phosphosite expression profiles the high activity of 3 kinases, namely CDK1, MAPK13 and MAPK3, with 9, 4 and 3 targeted phosphosites that present significant changes in the CR context, respectively (Supplementary Table 4). For instance, the Serine-25 and Serine-38 of the stathmin protein (STMN1) are targets of the three kinases. The STMN1 protein displays a complex pattern of activity and phosphorylation in cancers [65]. The sequestosome 1 protein (SQSTM1) Threonine-269 and Serine-272 are targets of both CDK1 and MAPK13.

Another interesting set of putative biomarkers can be derived from the proteins and phosphosites that have been identified in the MS experiment, but that were not further considered for quantification analyses because of the strong filtering criteria we have defined. We thus rescued the proteins and phosphosites that have been identified in at least 2 replicates in the CS cell line but that are completely absents in the CR cell lines, and vice-versa (Experimental Procedures). This concerns 140 proteins and 5 phosphosites that are identified only in the CR cell lines, and 8 proteins and 108 phosphosites that are identified only in the CS cell line. Focusing particularly on kinases, 8 of them are identified only in the CR cell lines (CALM1, EGFR, EIF2AK2, EPHA2, HK2, PIK3R4, PPP4C, ROCK2). A

majority of these kinases are involved in response to stress. Two other kinases are associated with phosphosites identified only in the CR contexts (PRPF4B, TAOK1). TAOK1 is particularly appealing as it activates the Hippo pathway involved in cellular homeostasis [66].

Proteome and phosphoproteome integration in a molecular network

We finally sought to provide a larger-scale functional interpretation of resistance-associated candidate biomarkers. The separated analysis of the proteomics and phosphoproteomics datasets provided one-dimensional views of cellular processes. We expect to obtain a comprehensive perspective of cellular processes and their interplays by integrating the information about protein abundances, activation status and molecular interactions [67, 68]. Toward this goal, we devised a network-guided integration of CS and CRPC cell lines proteome and phosphoproteome, by mapping the candidate biomarkers to molecular complex interaction data (Experimental Procedures). The resulting network is composed of 356 nodes and 1 161 edges, including a large connected component encompassing 194 nodes and 1 098 edges (Figure 4). The network reveals the links between up- and downregulated proteins, up- and downregulated phosphosites and corresponding proteins, as well as the links between the proteins and phosphosites that were identified by the MS approach only in the CR or CS contexts. At-a-glance, we can observe that the network is organized around several strongly connected subnetworks.

First, we identified a cell migration/invasion subnetwork, which is composed mainly of upregulated proteins in CR cells (e.g., ANXA2, IQGAP1, ACTN4, TWF1, MYO1B, CORO1C, ARPC4) (Figure 4). It contains in particular the plectrin protein (PLEC), overexpressed and hyper-phosphorylated in CR; this protein is known to interlink cytoskeleton elements and promote cancer cell invasion and migration [69]. Indeed, it was shown that along with vimentin intermediate filaments, plectrin provide a scaffold for invadopodia formation, facilitating cancer cell invasion extravasation for metastasis [70]. Recently, Burch et al. demonstrated that upregulation of vimentin and plectrin expressions positively correlates with the invasion and metastasis of androgen-independent PC cells [71]. Another interesting member of this complex is ACTG1 (actin gamma-1), which is not identified, and thereby might be not expressed, in CS cells. ACTG1 is involved in cell motility/cytoskeleton maintenance and cancer cell migration. ACTG1 was shown to induce cancer cell migration in lung cancer cells and hepatocellular carcinoma cells [72]. To date, there is no report concerning ACTG1 involvement in PC. The subnetwork also contains components of the Arp2/3

complex (ARPC1B, ARPC3, ARPC4) involved in the regulation of actin polymerization.

A smaller subnetwork, composed of interactions between EZR, MSN, SLC9A3R1 and EGFR, is located close to the larger migration subnetwork. Ezrin (EZR) and moesin (MSN) are scaffolding proteins that are involved in crosslinking cytoskeletal and membrane proteins. Ezrin is involved in oncogenesis through these interactions [73], and it was also shown recently that Ezrin can increase the oncogenic functions of EGFR [74]. SLC9A3R1 is also a scaffold protein that connects plasma membrane proteins with members of the ezrin/moesin/radixin family linking them to the actin cytoskeleton and regulating their surface expression[75].

We also identified a small subnetwork of interacting proteins involved in actin cytoskeleton regulation (e.g., STMN1, CDC42, CRLK1). Intriguingly, we found that stathmin1 (STMN1) was both hyper- and hypo-phosphorylated in CR cells. This protein is associated with cancer metastasis and exhibits a complicated phosphorylation pattern in response to various extracellular signals [76].

We next focused on a small subnetwork composed of proteins underexpressed in the CR context. It contains prohibitin (PHB), a putative tumor suppressor protein involved in the inhibition of DNA synthesis and regulating proliferation, and prohibitin-2 (PHB2), a mediator of transcriptional repression by nuclear receptors, also potentially involved in mitochondrial respiration. Indeed, the subnetwork also contains the VDAC1 mitochondrial membrane and plasma membrane channel, involved in apoptosis. The role of this subnetwork is unclear, but the proteins are depicted as members of the same complexes in the Hu.map dataset [34]. The subnetwork is tightly linked to another subnetwork containing many mitochondrial membrane ATP synthase proteins (e.g., ATP5F1, ATP5B, ATP5H), also downregulated in CR cell lines.

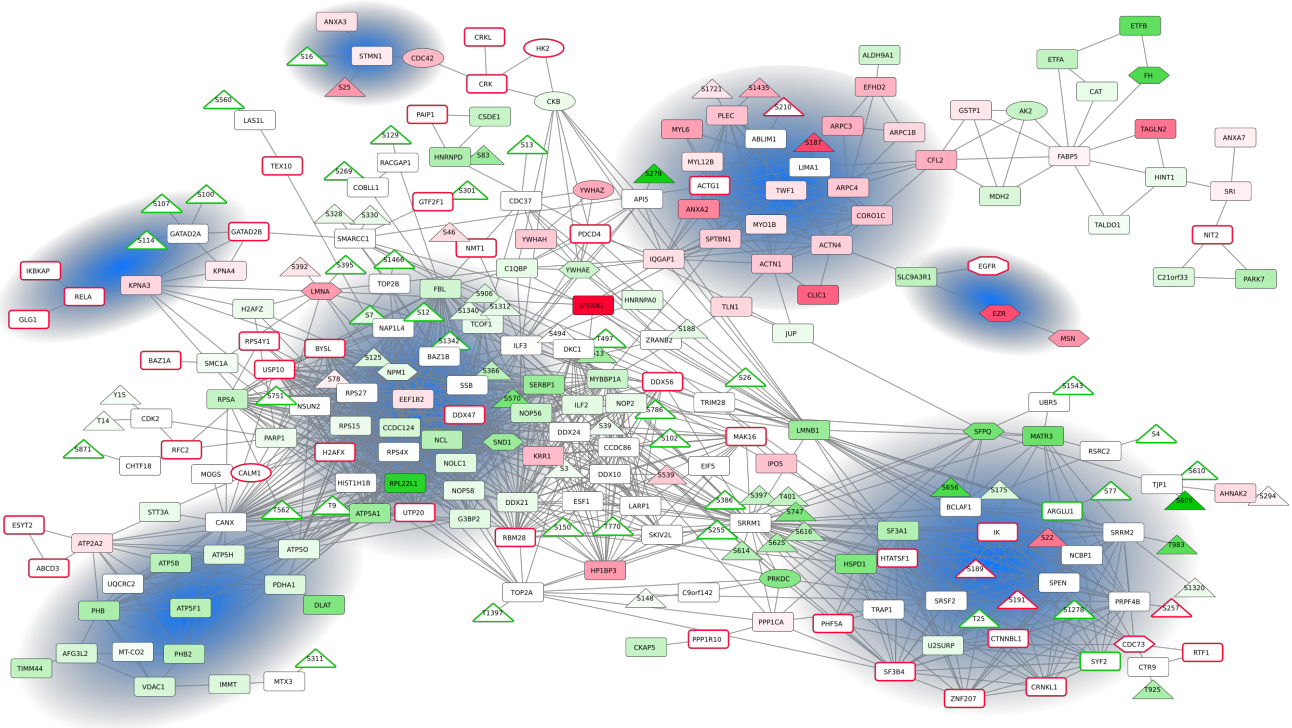


Figure 4 Network of CR biomarker interactions. Proteins (boxes) and phosphosites (triangles) significantly upregulated or downregulated in the CR contexts are mapped in red or green, respectively, with color intensities related to fold-changes. For visualization purposes, the expression values correspond to the mean of the expression of PC3 and DU145 cell lines. Proteins and phosphosites identified only in CR (DU145 and PC3) or CS (LNCaP) cells lines are squared in red and green, respectively.

A heterogeneous subnetwork is composed of many proteins involved in splicing and RNA processing, that are either up- or downregulated in CR cells (Figure 4). Splicing events control gene expression and their alterations have been shown to play a role in cancer [77] and specifically in PC [78]. Fine regulation of expression and/or phosphorylation status determines whether a splicing

factor functions as a splicing repressor or activator [79, 80]. The subnetwork contains the hypo-phosphorylated splicing factors SRRM1 (a highly phosphorylated protein under normal conditions [81]) and SRRM2. It also contains NCBP1, which is identified as hyper-phosphorylated, and PRPF4B kinase and SPEN that were both hypo-phosphorylated. The subnetwork also incorporates pre-mRNA splicing factor SYF2, absent in CR cells, and SF3A1, TRAP1 and HSPD1 proteins that are downregulated in CR cells. The protein phosphatase 1 (PPP1CA) is contrarily upregulated. Interestingly, we can also observe many proteins identified in the CR cell lines and absent in the CS cell line, all involved in RNA processing and splicing (IK, ZNF207, CTNNBL1, CRNKL1, SF3B4, HTATSF1, PHF5A, PPP1R10). PPP1R10, the Ser/Thr-protein phosphatase-1 regulatory subunit 10 is only expressed in CR cells and is absent in CS cells. It has been shown that certain Ser/Thr-specific protein phosphatases are required for catalytic steps of pre-mRNA splicing [82].

We then emphasize a large and highly connected component (Figure 4) composed of proteins implicated in DNA damage response. It contains protein biomarkers downregulated in CR cells (NPM1, NOLC1, RPL22L1, FBL, G3BP2), but also several proteins identified only in the CR cell lines, namely H2AFX, kinase CALM1, DDX47, UTP20, USP10, BYSL. All these proteins interact with single-strand DNA-binding protein and are involved in DNA repair and genome stability [83]. DNA repair and DNA damage response are known to be defective in PC and lead to genome instability [84, 85]. Interestingly, several of the proteins of this subnetwork (e.g., UTP20, BYSL, RPL22L1, NOLC1) are known for their role in RNA processing. There is an increasing number of studies demonstrating the involvement of RNA processing factors in DNA damage response [86, 87]. For instance, NOLC1 (nuclear and coiled-body phosphoprotein-1) is a regulator of RNA polymerase I and has been recently shown to regulate the nucleolar retention of TERF2, inducing telomeric DNA damage [88].

A closer look into this molecular network allowed us to pinpoint several interesting smaller subnetworks. For instance, we noticed a small subnetwork composed of interacting proteins RELA, IKBKAP, GLG1, KPNA3, KPNA4. Importin subunits alpha-4 (KPNA3) and alpha-3 (KPNA4) are involved in nuclear transport of NF kappa B [89], and an elevated activity of the NF-kappa B signaling in CRPC is positively correlated with poor prognosis in CRPC [90]. Close to this subnetwork, GATAD2B is known to form a homodimer with GATAD2A and the complex is part of a highly conserved chromatin-remodeling complex, the NuRD complex associated with DNA damage-induced transcription repression but also metastasis and EMT [91, 92]. This subnetwork is also linked to the SWI/SNF complex subunit SMARCC1, which contains downregulated phosphosites in PC3 and DU145 cells.

SMARCC1 positively regulates transcription and was previously shown to induce PC survival [93]. It interacts with proteins associated with phosphosites only detected in CS cells (transcriptional elongation factor TRIM28, transcription kinase BAZ1B and TOP2B), as well as with proteins only expressed in PC3 and DU145 cells (e.g., GATAD2B, GTF2F1), all involved directly or indirectly in transcription regulation. The transcriptional reprogramming in PC progression has been extensively studied, as it is one of the hallmarks of CRPC [3, 94-96].

Conclusion

The complex nature of PC, due to its clinical and molecular heterogeneities, makes it difficult to determine a perfect model representing tumor development, and precludes easy correlation of carcinoma cell lines with specific stages of PC. Nevertheless PC cell lines routinely used for the last three decades have provided a valuable resources for understanding important functional molecular mechanisms involved in this disease. In the present study, we used four cell lines that constitute a gold standard for pre-clinical studies of PC progression, hypothesizing that they would be a more homogeneous system as compared to tumor tissues [97]. We conducted a large SILAC-based Mass Spectrometry identification and quantification of peptides and phosphopeptides of prostate benign, castration-sensitive (CS) and castration-resistant (CR) cells, and characterized housekeeping, cell line, cancer and resistance associated proteomes and phosphoproteomes.

Supplementary Information

- Supplementary Table 1

Proteins identified in the MS experiment, and subset of filtered proteins associated with at least 2 valid quantification values in all four cell lines, which were kept for expression analyses.

- Supplementary Table 2

Phosphosites identified in the MS experiment, and subset of filtered phosphosites associated with at least 2 valid quantification values in all four cell lines, which were kept for expression analyses.

- Supplementary Table 3

Subdatasets of interest in proteomic expression analyses. It contains the ANOVA-significant proteins, the proteins up- and downregulated in the three Prostate Cancer (PCa) cell lines as compared to the benign PNT1A cell line, the proteins up- and downregulated in the castration-resistant (CR: DU145 and PC3) cell lines as compared to the castration-sensitive (CS: LNCaP) cell line, and the proteins

identified only in the CR or CS contexts (CR_only, CS_only).

- Supplementary Table 4

Subdatasets of interest in phosphoproteomic expression analyses. It contains the ANOVA-significant phosphosites, the phosphosites up- and downregulated in the three Prostate Cancer (PCa) cell lines as compared to the benign PNT1A cell line, the phosphosites up- and downregulated in the castration-resistant (CR: DU145 and PC3) cell lines as compared to the castration-sensitive (CS: LNCaP) cell line, and the phosphosites identified only in the CR or CS contexts (CR_only, CS_only). It further contains the results of the KSEA analysis.

- Supplementary Table 5

Raw results of the functional enrichment analyses with G:profiler and Ingenuity Pathway Analyses (IPA) [28].

- Supplementary Figure 1. Dynamic range of the prostate cancer proteome. (a) Ranking of the absolute abundance using the IBAQ intensity. The expression values of every protein in the three replicates of the four studied cell lines were considered. (b) Zoom on the left box in (a) displaying the 25 less abundant proteins. (c) Zoom on the right box in (a) displaying the 25 most abundant proteins.

- Supplementary Figure 2. Correlation between proteomic and phosphoproteomic expression values. We computed for each cell line the correlation between the expression values of the 135 proteins that were quantified both at the proteomic and the phosphoproteomic levels. For proteomics data, we computed the mean of the three replicated. For phosphoproteomics data, we computed the mean for all the phosphosites belonging to the same protein.

- Supplementary Figure 3. Expression Profiles associated with Septin-9 (SEPT9). (a) Boxplot showing the SEPT9 protein expression values in the four cell lines under study. (b) Boxplot revealing the SEPT9 Serine-30 phosphosite expression values in the four cell lines under study.

References

[1] A. Fusi, G. Procopio, S. Della Torre, R. Ricotta, G. Bianchini, R. Salvioni, L. Ferrari, A. Martinetti, G. Savelli, S. Villa, E. Bajetta, Treatment options in hormone-refractory metastatic prostate carcinoma, Tumori 90(6) (2004) 535-46.
[2] T. Karantanos, C.P. Evans, B. Tombal, T.C. Thompson, R. Montironi, W.B. Isaacs, Understanding the mechanisms of androgen deprivation resistance in prostate cancer at the molecular level, Eur Urol 67(3) (2015) 470-9.
[3] M. Katsogiannou, H. Ziouziou, S. Karaki, C. Andrieu, M. Henry de Villeneuve, P. Rocchi, The hallmarks of castration-resistant prostate cancers, Cancer Treat Rev 41(7) (2015) 588-97.
[4] M.D. Galsky, A.C. Small, C.K. Tsao, W.K. Oh, Clinical development of novel therapeutics for castration-resistant prostate cancer: historic challenges and recent successes, CA Cancer J Clin 62(5) (2012) 299-308.
[5] D.L. Suzman, E.S. Antonarakis, Castration-resistant prostate cancer: latest evidence and

therapeutic implications, Ther Adv Med Oncol 6(4) (2014) 167-79.
[6] D. Iglesias-Gato, P. Wikstrom, S. Tyanova, C. Lavallee, E. Thysell, J. Carlsson, C. Hagglof, J. Cox, O. Andren, P. Stattin, L. Egevad, A. Widmark, A. Bjartell, C.C. Collins, A. Bergh, T. Geiger, M. Mann, A. Flores-Morales, The Proteome of Primary Prostate Cancer, Eur Urol (2015).
[7] H. Kuruma, S. Egawa, M. Oh-Ishi, Y. Kodera, T. Maeda, Proteome analysis of prostate cancer, Prostate Cancer Prostatic Dis 8(1) (2005) 14-21.
[8] H. Kuruma, S. Egawa, M. Oh-Ishi, Y. Kodera, M. Satoh, W. Chen, H. Okusa, K. Matsumoto, T. Maeda, S. Baba, High molecular mass proteome of androgen-independent prostate cancer, Proteomics 5(4) (2005) 1097-112.
[9] D.K. Ornstein, D.R. Tyson, Proteomics for the identification of new prostate cancer biomarkers, Urol Oncol 24(3) (2006) 231-6.
[10] L.N. Johnson, The regulation of protein phosphorylation, Biochem Soc Trans 37(Pt 4) (2009) 627-41.
[11] C. Cans, R. Mangano, D. Barila, G. Neubauer, G. Superti-Furga, Nuclear tyrosine phosphorylation: the beginning of a map, Biochem Pharmacol 60(8) (2000) 1203-15.
[12] E. Lopez Villar, L. Madero, A.L.-P. J, C.C. W, Study of phosphorylation events for cancer diagnoses and treatment, Clin Transl Med 4(1) (2015) 59.
[13] H.C. Harsha, A. Pandey, Phosphoproteomics in cancer, Mol Oncol 4(6) (2010) 482-95.
[14] R.M. Lescarbeau, D.L. Kaplan, Quantitative analysis of castration resistant prostate cancer progression through phosphoproteome signaling, BMC Cancer 14 (2014) 325.
[15] N. Jiang, K. Hjorth-Jensen, O. Hekmat, D. Iglesias-Gato, T. Kruse, C. Wang, W. Wei, B. Ke, B. Yan, Y. Niu, J.V. Olsen, A. Flores-Morales, In vivo quantitative phosphoproteomic profiling identifies novel regulators of castration-resistant prostate cancer growth, Oncogene 34(21) (2015) 2764-76.
[16] X. Wang, P.A. Stewart, Q. Cao, Q.X. Sang, L.W. Chung, M.R. Emmett, A.G. Marshall, Characterization of the phosphoproteome in androgen-repressed human prostate cancer cells by Fourier transform ion cyclotron resonance mass spectrometry, J Proteome Res 10(9) (2011) 3920-8.
[17] J.M. Drake, E.O. Paull, N.A. Graham, J.K. Lee, B.A. Smith, B. Titz, T. Stoyanova, C.M. Faltermeier, V. Uzunangelov, D.E. Carlin, D.T. Fleming, C.K. Wong, Y. Newton, S. Sudha, A.A. Vashisht, J. Huang, J.A. Wohlschlegel, T.G. Graeber, O.N. Witte, J.M. Stuart, Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer, Cell 166(4) (2016) 1041-54.
[18] T. Geiger, J. Cox, P. Ostasiewicz, J.R. Wisniewski, M. Mann, Super-SILAC mix for quantitative proteomics of human tumor tissue, Nat Methods 7(5) (2010) 383-5.
[19] T. Geiger, J.R. Wisniewski, J. Cox, S. Zanivan, M. Kruger, Y. Ishihama, M. Mann, Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics, Nat Protoc 6(2) (2011) 147-57.
[20] A. Shevchenko, M. Wilm, O. Vorm, O.N. Jensen, A.V. Podtelejnikov, G. Neubauer, P. Mortensen, M. Mann, A strategy for identifying gel-separated proteins in sequence databases by MS alone, Biochem Soc Trans 24(3) (1996) 893-6.
[21] E.W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D.S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R.L. Moritz, J.J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, J.A. Vizcaino, The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition, Nucleic Acids Res 45(D1) (2017) D1100-D1106.
[22] J. Cox, I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J.V. Olsen, M. Mann, A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics, Nat Protoc 4(5) (2009) 698-705.
[23] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide

search engine integrated into the MaxQuant environment, J Proteome Res 10(4) (2011) 1794-805.
[24] J.V. Olsen, S.E. Ong, M. Mann, Trypsin cleaves exclusively C-terminal to arginine and lysine residues, Mol Cell Proteomics 3(6) (2004) 608-14.
[25] J. Cox, M. Mann, 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data, BMC Bioinformatics 13 Suppl 16 (2012) S12.
[26] RCoreTeam, R: A language and environment for statistical computing. , in: R.F.f.S. Computing (Ed.) Vienna, Austria., 2015.
[27] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. , Journal of the Royal Statistical Society Series B (Methodological) Vol. 57(N°1) (1995) 289-300.
[28] J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, J. Vilo, g:Profiler-a web server for functional interpretation of gene lists (2016 update), Nucleic Acids Res 44(W1) (2016) W83-9.
[29] L. Martignetti, L. Calzone, E. Bonnet, E. Barillot, A. Zinovyev, ROMA: Representation and Quantification of Module Activity from Target Expression Data, Front Genet 7 (2016) 18.
[30] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The Reactome pathway Knowledgebase, Nucleic Acids Res 44(D1) (2016) D481-7.
[31] A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J.P. Mesirov, P. Tamayo, The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell Syst 1(6) (2015) 417-425.
[32] P. Casado, J.C. Rodriguez-Prados, S.C. Cosulich, S. Guichard, B. Vanhaesebroeck, S. Joel, P.R. Cutillas, Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells, Sci Signal 6(268) (2013) rs6.
[33] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H.W. Mewes, CORUM: the comprehensive resource of mammalian protein complexes--2009, Nucleic Acids Res 38(Database issue) (2010) D497-501.
[34] K. Drew, C. Lee, R.L. Huizar, F. Tu, B. Borgeson, C.D. McWhite, Y. Ma, J.B. Wallingford, E.M. Marcotte, Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes, Mol Syst Biol 13(6) (2017) 932.
[35] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res 13(11) (2003) 2498-504.
[36] C. Avances, V. Georget, B. Terouanne, F. Orio, O. Cussenot, N. Mottet, P. Costa, C. Sultan, Human prostatic cell line PNT1A, a useful tool for studying androgen receptor transcriptional activity and its differential subnuclear localization in the presence of androgens and antiandrogens, Mol Cell Endocrinol 184(1-2) (2001) 13-24.
[37] P.J. Russell, E.A. Kingsley, Human prostate cancer cell lines, Methods Mol Med 81 (2003) 21-39.
[38] M.M. Webber, D. Bello, S. Quader, Immortalized and tumorigenic adult human prostatic epithelial cell lines: characteristics and applications Part 2. Tumorigenic cell lines, Prostate 30(1) (1997) 58-64.
[39] S. Mitchell, P. Abel, M. Ware, G. Stamp, E. Lalani, Phenotypic and genotypic characterization of commonly used human prostatic cell lines, BJU Int 85(7) (2000) 932-44.
[40] M. Mann, Functional and quantitative proteomics using SILAC, Nat Rev Mol Cell Biol 7(12) (2006) 952-8.

[41] S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, Mol Cell Proteomics 1(5) (2002) 376-86.
[42] N. Nagaraj, J.R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Paabo, M. Mann, Deep proteome and transcriptome mapping of a human cancer cell line, Mol Syst Biol 7 (2011) 548.
[43] A. Degeorges, F. Hoffschir, O. Cussenot, C. Gauville, A. Le Duc, B. Dutrillaux, F. Calvo, Recurrent cytogenetic alterations of prostate carcinoma and amplification of c-myc or epidermal growth factor receptor in subclones of immortalized PNT1 human prostate epithelial cell line, Int J Cancer 62(6) (1995) 724-31.
[44] M. Uhlen, E. Bjorling, C. Agaton, C.A. Szigyarto, B. Amini, E. Andersen, A.C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergstrom, H. Brumer, D. Cerjan, M. Ekstrom, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M.G. Bjorklund, K. Gumbel, A. Halimi, I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Odling, P. Oksvold, I. Olsson, E. Oster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, A. Sivertsson, A. Skollermo, J. Steen, M. Stenvall, F. Sterky, S. Stromberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Walden, J. Wan, H. Wernerus, J. Westberg, K. Wester, U. Wrethagen, L.L. Xu, S. Hober, F. Ponten, A human protein atlas for normal and cancer tissues based on antibody proteomics, Mol Cell Proteomics 4(12) (2005) 1920-32.
[45] R. Gilad, K. Meir, I. Stein, L. German, E. Pikarsky, N.J. Mabjeesh, High SEPT9_i1 protein expression is associated with high-grade prostate cancers, PLoS One 10(4) (2015) e0124251.
[46] W.K. Leung, A.K. Ching, A.W. Chan, T.C. Poon, H. Mian, A.S. Wong, K.F. To, N. Wong, A novel interplay between oncogenic PFTK1 protein kinase and tumor suppressor TAGLN2 in the control of liver cancer cell motility, Oncogene 30(44) (2011) 4464-75.
[47] K. Garber, Energy deregulation: licensing tumors to grow, Science 312(5777) (2006) 1158-9.
[48] K.R. Jakobsen, E. Sorensen, K.K. Brondum, T.F. Daugaard, R. Thomsen, A.L. Nielsen, Direct RNA sequencing mediated identification of mRNA localized in protrusions of human MDA-MB-231 metastatic breast cancer cells, J Mol Signal 8(1) (2013) 9.
[49] R. Mayor, C. Carmona-Fontaine, Keeping in touch with contact inhibition of locomotion, Trends Cell Biol 20(6) (2010) 319-28.
[50] A.S. Azmi, B. Bao, F.H. Sarkar, Exosomes in cancer development, metastasis, and drug resistance: a comprehensive review, Cancer Metastasis Rev 32(3-4) (2013) 623-42.
[51] L. Rauschenberger, D. Staar, K. Thom, C. Scharf, S. Venz, G. Homuth, R. Schluter, L.O. Brandenburg, P. Ziegler, U. Zimmermann, W. Weitschies, U. Volker, U. Lendeckel, R. Walther, M. Burchardt, M.B. Stope, Exosomal particles secreted by prostate cancer cells are potent mRNA and protein vehicles for the interference of tumor and tumor environment, Prostate 76(4) (2016) 409-24.
[52] A.Y. Yam, Y. Xia, H.T. Lin, A. Burlingame, M. Gerstein, J. Frydman, Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies, Nat Struct Mol Biol 15(12) (2008) 1255-62.
[53] S.H. Roh, M. Kasembeli, D. Bakthavatsalam, W. Chiu, D.J. Tweardy, Contribution of the Type II Chaperonin, TRiC/CCT, to Oncogenesis, Int J Mol Sci 16(11) (2015) 26706-20.
[54] A.G. Trinidad, P.A. Muller, J. Cuellar, M. Klejnot, M. Nobis, J.M. Valpuesta, K.H. Vousden, Interaction of p53 with the CCT complex promotes protein folding and wild-type p53 activity, Mol Cell 50(6) (2013) 805-17.
[55] J. Bi, A. Huang, T. Liu, T. Zhang, H. Ma, Expression of DNA damage checkpoint 53BP1 is correlated with prognosis, cell proliferation and apoptosis in colorectal cancer, Int J Clin Exp Pathol 8(6) (2015) 6070-82.

[56] S.M. Harding, R.G. Bristow, Discordance between phosphorylation and recruitment of 53BP1 in response to DNA double-strand breaks, Cell Cycle 11(7) (2012) 1432-44.
[57] K. Savitsky, Y. Ziv, A. Bar-Shira, S. Gilad, D.A. Tagle, S. Smith, T. Uziel, S. Sfez, J. Nahmias, A. Sartiel, R.L. Eddy, T.B. Shows, F.S. Collins, Y. Shiloh, G. Rotman, A human gene (DDX10) encoding a putative DEAD-box RNA helicase at 11q22-q23, Genomics 33(2) (1996) 199-206.
[58] L. Yang, C. Lin, Z.R. Liu, Phosphorylations of DEAD box p68 RNA helicase are associated with cancer development and cell proliferation, Mol Cancer Res 3(6) (2005) 355-63.
[59] E.A. Gustafson, G.M. Wessel, DEAD-box helicases: posttranslational regulation and function, Biochem Biophys Res Commun 395(1) (2010) 1-6.
[60] K. Lassi, N.A. Dawson, Update on castrate-resistant prostate cancer: 2010, Curr Opin Oncol 22(3) (2010) 263-7.
[61] M.V. Liberti, J.W. Locasale, The Warburg Effect: How Does it Benefit Cancer Cells?, Trends Biochem Sci 41(3) (2016) 211-8.
[62] J.P. Thiery, J.P. Sleeman, Complex networks orchestrate epithelial-mesenchymal transitions, Nat Rev Mol Cell Biol 7(2) (2006) 131-42.
[63] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144(5) (2011) 646-74.
[64] C.S. Yang, T.A. Melhuish, A. Spencer, L. Ni, Y. Hao, K. Jividen, T.E. Harris, C. Snow, H.F. Frierson, Jr., D. Wotton, B.M. Paschal, The protein kinase C super-family member PKN is regulated by mTOR and influences differentiation during prostate cancer progression, Prostate 77(15) (2017) 1452-1467.
[65] K. Williams, R. Ghosh, P.V. Giridhar, G. Gu, T. Case, S.M. Belcher, S. Kasper, Inhibition of stathmin1 accelerates the metastatic process, Cancer Res 72(20) (2012) 5407-17.
[66] S.W. Plouffe, Z. Meng, K.C. Lin, B. Lin, A.W. Hong, J.V. Chun, K.L. Guan, Characterization of Hippo Pathway Components by Gene Inactivation, Mol Cell 64(5) (2016) 993-1008.
[67] J.-M. Park, J.-H. Park, D.-G. Mun, J. Bae, J.H. Jung, S. Back, H. Lee, H. Kim, H.-J. Jung, H.K. Kim, H. Lee, K.P. Kim, D. Hwang, S.-W. Lee, Integrated analysis of global proteome, phosphoproteome, and glycoproteome enables complementary interpretation of disease-related protein networks, 5 (2015) 18189.
[68] H. Tan, K. Yang, Y. Li, T.I. Shaw, Y. Wang, D.B. Blanco, X. Wang, J.H. Cho, H. Wang, S. Rankin, C. Guy, J. Peng, H. Chi, Integrative Proteomics and Phosphoproteomics Profiling Reveals Dynamic Signaling Networks and Bioenergetics Pathways Underlying T Cell Activation, Immunity 46(3) (2017) 488-503.
[69] K. Katada, T. Tomonaga, M. Satoh, K. Matsushita, Y. Tonoike, Y. Kodera, T. Hanazawa, F. Nomura, Y. Okamoto, Plectin promotes migration and invasion of cancer cells and is a novel prognostic marker for head and neck squamous cell carcinoma, J Proteomics 75(6) (2012) 1803-15.
[70] M. Sutoh Yoneyama, S. Hatakeyama, T. Habuchi, T. Inoue, T. Nakamura, T. Funyu, G. Wiche, C. Ohyama, S. Tsuboi, Vimentin intermediate filament and plectin provide a scaffold for invadopodia, facilitating cancer cell invasion and extravasation for metastasis, Eur J Cell Biol 93(4) (2014) 157-69.
[71] T.C. Burch, M.T. Watson, J.O. Nyalwidhe, Variable metastatic potentials correlate with differential plectin and vimentin expression in syngeneic androgen independent prostate cancer cells, PLoS One 8(5) (2013) e65005.
[72] Y. Luo, F. Kong, Z. Wang, D. Chen, Q. Liu, T. Wang, R. Xu, X. Wang, J.Y. Yang, Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to suppress cancer cell migration, Mol Med Rep 9(2) (2014) 387-94.
[73] A. Bretscher, K. Edwards, R.G. Fehon, ERM proteins and merlin: integrators at the cell cortex,

Nat Rev Mol Cell Biol 3(8) (2002) 586-99.
[74] Y. Saygideger-Kont, T.Z. Minas, H. Jones, S. Hour, H. Celik, I. Temel, J. Han, N. Atabey, H.V. Erkizan, J.A. Toretsky, A. Uren, Ezrin Enhances EGFR Signaling and Modulates Erlotinib Sensitivity in Non-Small Cell Lung Cancer Cells, Neoplasia 18(2) (2016) 111-20.
[75] D. Soave, H. Corvol, N. Panjwani, J. Gong, W. Li, P.Y. Boelle, P.R. Durie, A.D. Paterson, J.M. Rommens, L.J. Strug, L. Sun, A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways, Am J Hum Genet 97(1) (2015) 125-38.
[76] X.Y. Kuang, H.S. Jiang, K. Li, Y.Z. Zheng, Y.R. Liu, F. Qiao, S. Li, X. Hu, Z.M. Shao, The phosphorylation-specific association of STMN1 with GRP78 promotes breast cancer metastasis, Cancer Lett 377(1) (2016) 87-96.
[77] S. Germann, L. Gratadou, M. Dutertre, D. Auboeuf, Splicing programs and cancer, J Nucleic Acids 2012 (2012) 269570.
[78] J. Munkley, K. Livermore, P. Rajan, D.J. Elliott, RNA splicing and splicing regulator changes in prostate cancer pathology, Hum Genet (2017).
[79] M. Chen, J.L. Manley, Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches, Nat Rev Mol Cell Biol 10(11) (2009) 741-54.
[80] C. Naro, C. Sette, Phosphorylation-Mediated Regulation of Alternative Splicing in Cancer, Int J Cell Biol 2013 (2013) 151839.
[81] H. Molina, D.M. Horn, N. Tang, S. Mathivanan, A. Pandey, Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry, Proc Natl Acad Sci U S A 104(7) (2007) 2199-204.
[82] J.E. Mermoud, P. Cohen, A.I. Lamond, Ser/Thr-specific protein phosphatases are required for both catalytic steps of pre-mRNA splicing, Nucleic Acids Res 20(20) (1992) 5263-9.
[83] S. Broderick, K. Rehmet, C. Concannon, H.P. Nasheuer, Eukaryotic single-stranded DNA binding proteins: central factors in genome stability, Subcell Biochem 50 (2010) 143-63.
[84] R. Fan, T.S. Kumaravel, F. Jalali, P. Marrano, J.A. Squire, R.G. Bristow, Defective DNA strand break repair after DNA damage in prostate cancer cells: implications for genetic instability and prostate cancer progression, Cancer Res 64(23) (2004) 8526-33.
[85] M. Dhawan, C.J. Ryan, A. Ashworth, DNA Repair Deficiency Is Common in Advanced Prostate Cancer: New Therapeutic Opportunities, Oncologist 21(8) (2016) 940-5.
[86] A. Montecucco, G. Biamonti, Pre-mRNA processing factors meet the DNA damage response, Front Genet 4 (2013) 102.
[87] X. Jacq, M. Kemp, N.M. Martin, S.P. Jackson, Deubiquitylating enzymes and DNA damage response pathways, Cell Biochem Biophys 67(1) (2013) 25-43.
[88] F. Yuan, G. Li, T. Tong, Nucleolar and coiled-body phosphoprotein 1 (NOLC1) regulates the nucleolar retention of TRF2, Cell Death Discov 3 (2017) 17043.
[89] R. Fagerlund, L. Kinnunen, M. Kohler, I. Julkunen, K. Melen, NF-{kappa}B is transported into the nucleus by importin {alpha}3 and importin {alpha}4, J Biol Chem 280(16) (2005) 15942-51.
[90] R.J. Jin, Y. Lho, L. Connelly, Y. Wang, X. Yu, L. Saint Jean, T.C. Case, K. Ellwood-Yen, C.L. Sawyers, N.A. Bhowmick, T.S. Blackwell, F.E. Yull, R.J. Matusik, The nuclear factor-kappaB pathway controls the progression of prostate cancer to androgen-independent growth, Cancer Res 68(16) (2008) 6762-9.
[91] A.Y. Lai, P.A. Wade, Cancer biology and NuRD: a multifaceted chromatin remodelling complex, Nat Rev Cancer 11(8) (2011) 588-96.
[92] C.G. Spruijt, M.S. Luijsterburg, R. Menafra, R.G. Lindeboom, P.W. Jansen, R.R. Edupuganti, M.P. Baltissen, W.W. Wiegant, M.C. Voelker-Albert, F. Matarese, A. Mensinga, I. Poser, H.R. Vos, H.G. Stunnenberg, H. van Attikum, M. Vermeulen, ZMYND8 Co-localizes with NuRD on Target Genes and Regulates Poly(ADP-Ribose)-Dependent Recruitment of GATAD2A/NuRD to Sites of DNA

Damage, Cell Rep 17(3) (2016) 783-798.
[93] S. Heeboll, M. Borre, P.D. Ottosen, C.L. Andersen, F. Mansilla, L. Dyrskjot, T.F. Orntoft, N. Torring, SMARCC1 expression is upregulated in prostate cancer and positively correlated with tumour recurrence and dedifferentiation, Histol Histopathol 23(9) (2008) 1069-76.
[94] R. Hu, C. Lu, E.A. Mostaghel, S. Yegnasubramanian, M. Gurel, C. Tannahill, J. Edwards, W.B. Isaacs, P.S. Nelson, E. Bluemn, S.R. Plymate, J. Luo, Distinct transcriptional programs mediated by the ligand-dependent full-length androgen receptor and its splice variants in castration-resistant prostate cancer, Cancer Res 72(14) (2012) 3457-62.
[95] Q. Wang, W. Li, Y. Zhang, X. Yuan, K. Xu, J. Yu, Z. Chen, R. Beroukhim, H. Wang, M. Lupien, T. Wu, M.M. Regan, C.A. Meyer, J.S. Carroll, A.K. Manrai, O.A. Janne, S.P. Balk, R. Mehra, B. Han, A.M. Chinnaiyan, M.A. Rubin, L. True, M. Fiorentino, C. Fiore, M. Loda, P.W. Kantoff, X.S. Liu, M. Brown, Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer, Cell 138(2) (2009) 245-56.
[96] E.J. Faivre, D. Wilcox, X. Lin, P. Hessler, M. Torrent, W. He, T. Uziel, D.H. Albert, K. McDaniel, W. Kati, Y. Shen, Exploitation of Castration-Resistant Prostate Cancer Transcription Factor Dependencies by the Novel BET Inhibitor ABBV-075, Mol Cancer Res 15(1) (2017) 35-44.
[97] E. Lundberg, L. Fagerberg, D. Klevebring, I. Matic, T. Geiger, J. Cox, C. Algenas, J. Lundeberg, M. Mann, M. Uhlen, Defining the transcriptome and proteome in three functionally different human cell lines, Mol Syst Biol 6 (2010) 450.

Discussion

Proteins can be considered as the functional units of cells. The information about isolated protein function is highly valuable, but is not enough to understand cellular functioning. Indeed, proteins interact between them to execute all the cellular functions. In this context, the application of graph-theory methods to mine the information contained in networks of protein interactions can improve our knowledge about how cells work. In addition, proteins, but also genes and other biological molecules, interact with each other thanks to different types of relationships, which are identified at large-scale by different *-omics* technologies, each one with its own strengths and weaknesses. Consequently, we propose the use of multiplex networks to better describe and analyze biological systems.

I conduct the discussion of my work following an example that recapitulates the main results presented in my PhD thesis. To this goal, I applied the Random Walk with Restart algorithm on Multiplex-Heterogeneous network (RWR-MH) (Valdeolivas et al. 2018) using as seeds Hutchinson-Gilford progeria syndrome (HGPS), the most representative premature aging (PA) disease, and its causative gene *LMNA*. We identified and explored the top 50 ranked diseases and genes, together with their interactions and associations (Figure 6.1).

Among the top 50 ranked diseases, we found well known progeroid syndromes, such as Werner Syndrome (MIM code: 277700), Néstor-Guillermo progeria syndrome (MIM code: 614008) and Rothmund Thomson syndrome (MIM code: 268400). We also retrieve diseases sharing non-PA phenotypes with HGPS. For example, the Melnick-needles syndrome (MIM code: 309350) and the otopalatodigital syndromes I and II (MIM code: 311300, 304120) share a large spectrum of skeletal anomalies with HPGS. Moreover, mutations in the *LMNA* gene are causing disorders with very diverse phenotypes, in addition to HGPS. Consequently, other groups of diseases (e.g. cardiomyopathies and muscular dystrophies) emerge among the top ranked diseases, even if they only share a reduced number of clinical features with HPGS.

Focusing on the top 50 ranked genes, we can observe a group containing several genes encoding integrin proteins (*ITGA3*, *ITGB1*, *ITGB3* and *ITGB5*). These proteins have an important role in various cellular functions such as cell adhesion, migration, proliferation and cell differentiation (D. Cox et al. 2010). Tumour cells use this integrin-mediated processes to promote tumour growth and metastatic spread (Hamidi et al. 2016). The identification of integrins starting from LMNA was quite unexpected. Indeed, lamins are component of the nuclear cytoskeleton and lamina. The nuclear stiffness and deformability appear to be controlled by the expression of the main components of the nuclear lamina, A-type lamins and B-type lamins (products of the *LMNA* and *LMNB1* genes). A reduction in the expression of the A-type lamins leads to an increase in the de-

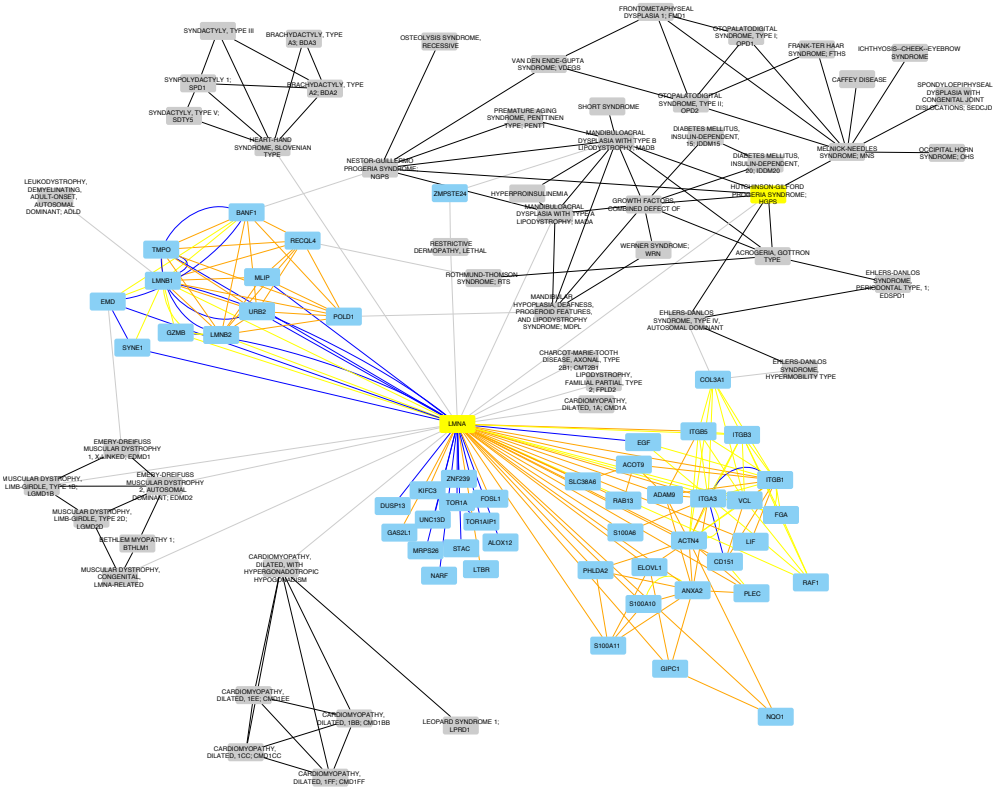


Figure 6.1 – Network representation of the top 50 ranked genes and top 50 ranked diseases when the RWR-MH algorithm is applied using HGPS as seed diseases and *LMNA* as seed gene (yellow nodes). Grey elliptical nodes are diseases; Turquoise rectangles are genes/proteins. Black edges are bipartite gene-disease associations from OMIM (Hamosh et al. 2005). Grey edges are the phenotypic similarity links in the disease-disease network; Blue edges are PPI interactions; Orange edges are co-expression relationships; Yellow edges are pathway interactions. It is to note that results are represented as an aggregated network only for visualization purposes.

formability of the nucleus, which in turn can facilitate the migration of cancer cells promoting the metastatic process (Guinde et al. 2018). However, it is interesting to note that some recent studies have revealed a potential role of integrins in the modulation of the physical structure of the nucleus through functional links with nuclear lamina components. In this context, it has been suggested that integrins might mediate nuclear stiffness during cancer cell migration (Madrazo et al. 2017).

Another important point is that the identification of integrins in the top-scoring genes is due to the direct co-expression link between *LMNA* and some integrins, but also results from the contribution of indirect relationships. Indeed, HGPS is connected with Ehlers-Danlos syndrome type IV (MIM code: 130050), whose causative gene *COL3A1* is highly connected with the group of integrins, because they participate in extracellular matrix organization. Even though physical interactions between A-type lamins and integrins are not present in the databases used in this study, our method was able to highlight potential relations between them, supported by recent literature.

The integration of diverse sources of biological information into a multiplex network, and its subsequent exploration with random walks enabled this interesting prediction. The importance of integrating multiple network sources is also shown by the presence of *ZMPST24* in the top scoring genes. *ZMPSTE24* is a protease implicated in the maturation process of lamin A, but even if its physical interaction with lamin A is well-known, it is not described in databases, and is therefore absent from our multiplex molecular network. However, the integration of the disease-disease network thank to our multiplex-heterogeneous framework captures *ZMPSTE24* via its involvement in Restrictive Dermopathy, a disease also linked to *LMNA*. We therefore state that a suitable combination of multiple data sources can help to overcome the bias and missing data of individual sources. Large-scale biological networks suffer from those limitations: the human interactome is far from complete partially because high-throughput technologies systematically fail to detect specific interactions. For instance, yeast two hybrids screens are not able to detect neither interactions involving membrane proteins nor interactions depending on post-translational modifications mediated by enzymes.

On the left hand side of figure 6.1, we can observe a group of genes tightly connected through different types of interactions. The group contain genes coding proteins mainly located in the cell nucleus. Noteworthy, many of these genes are annotated as related with human aging in the GenAge database (Tacutu et al. 2018), namely *EMD*, *LMNB1*, *RECQL4*, *BANF1* and *POLD1*. We can hypothesize, following the guilt-by-association approach, that some of the remaining genes of the group could be also associated with human aging. We carefully inspected the literature and several biological databases searching for the most suitable candidate to undertake experimental validations, and selected the *URB2* gene for experimental validations. The protein encoded by this gene is widely

expressed and interacts with the nuclear envelope proteins, but its functions are poorly understood (Vlcek et al. 2004). My *ProGeLife* colleagues performed a set of experimental validations detailed in the next paragraphs.

The expression level of several proteins expression was tested after 48h or 72h URB2 silencing with a siRNA, in control and HGPS cell lines. These proteins include major lamins types, as URB2 was predicted to be part of a process involving A and B types lamins (Figure 6.2). In addition, they tested p21, a key protein of the cell cycle progression, involved in senescence, as well as p53 recetly described to be an interactor of URB2 in Zebrafish (Cai et al. 2018) (Figure 6.2). Except a slight increase of p21 at 48h in HGPS cell line, and at 72h in control fibroblasts (needing to be reproduced), all other protein expression levels were not altered after 48h or 72h of URB2 silencing.

Then, they explored lamins A/C and B1 localization 48h after URB2 silencing with siURB2 in both control (data not shown) and HGPS cell lines (Figure 6.3). No abnormal distribution of lamins A/C (Figure 6.3A) or lamin B1 (Figure 6.3B) was observed in both cell lines.

Finally, they performed viability and proliferation assays on both control and HGPS fibroblasts using CyQuant™ and Prestoblue™, respectively (Figure 6.4). An effect of URB2 inhibition with siURB2 was observed at 72h on HGPS cell line compared to control, but it is not clear for now if this is clearly specific as the non-specific siRNA tended to show also an effect. Additional studies using other specific and negative siRNAs could allow us concluding regarding this observation.

These negative or non-conclusive results show the exploratory character of our computational biology approach. Most of bioinformatic methods are a good approach to select potential candidates. Nevertheless, biological experimental validations are needed to assess the predictions and to unveil new biological knowledge.

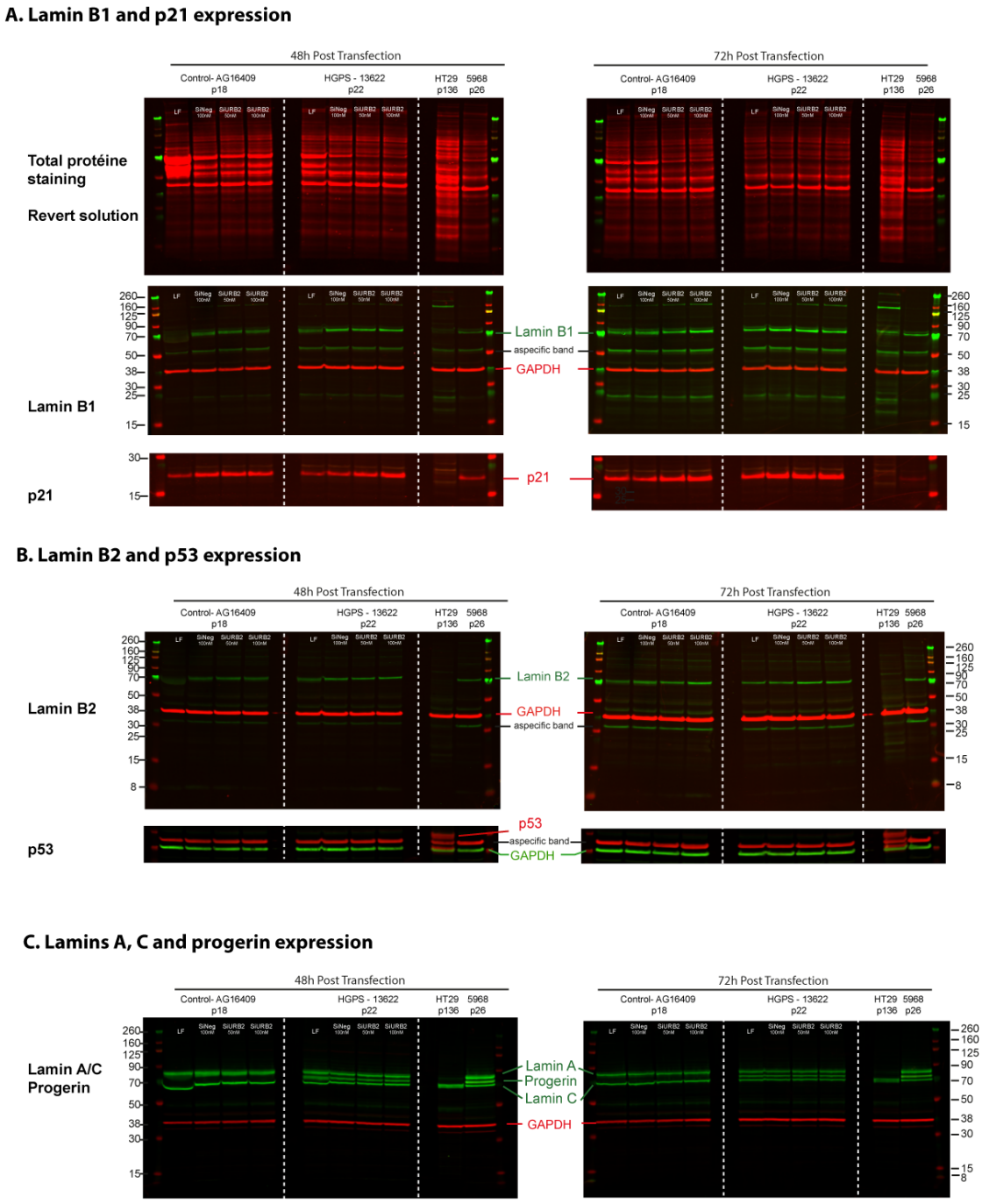


Figure 6.2 – **Western blot analyses following URB2 silencing in control and HGPS fibroblasts**. Proteins were extracted 48h (left) or 72h (right) post SiURB2 transfection. Lanes were loaded with $40\mu g$ of protein. Each western blot shows from left to right: control fibroblast - AG16409 (lanes 1 to 4) or HGPS fibroblast -13622 (lanes 6 to 9). Fibroblasts were transfected with lipofectamin alone (LF, lanes 1 & 6), a nonspecific siRNA $[100nM]$ (SiNeg 100nM, lanes 2 & 7) or a SiURB2 $[50nM]$ lanes 3 & 8 and $[100nM]$ lanes 4 & 9. Lanes 5 and 10 are free of protein deposit and two last lanes correspond to specific controls: a cancer colon cell line - HT29 (lane 11) and a high progerin expression cell line HGPS - 5968, at late passage p26 (lane 12). Smaller lanes from left and right are Chameleon duo protein ladder. **(A)** Western blot analysis of lamin B1 expression and p21. Total protein stain (Revert^{TM}) control is shown on the top panel. Lamin B1 is from Abcam (ab16048, 1/1000, cf: $1\mu g/ml$) and GAPDH from Millipore (MAB374, 1/40000, cf: $0.025\mu g/ml$ - middle panel); reincubation on the same membrane with antibody p21 from Abcam (ab109520, 1/5000, cf: $0,17\mu g/mL$ - lower panel). **(B)** Western blot analysis of Lamin B2 expression and p53. Lamin B2 is from Sigma (AV46356, 1/500, cf: $1\mu g/ml$) and GAPDH from Millipore (MAB374, 1/40000, cf: $0,025\mu g/ml$ - upper panel); reincubation on the same membrane with antibody p53 from Abcam (ab179477, 1/1500, cf: $1,59\mu g/mL$ - lower panel). **(C)** Western blot analysis of A types lamins. Lamin A, lamin C and progerin were detected with an antibody against lamin A/C from Santa-Cruz (Sc-20681, 1/200, $1\mu g/ml$).

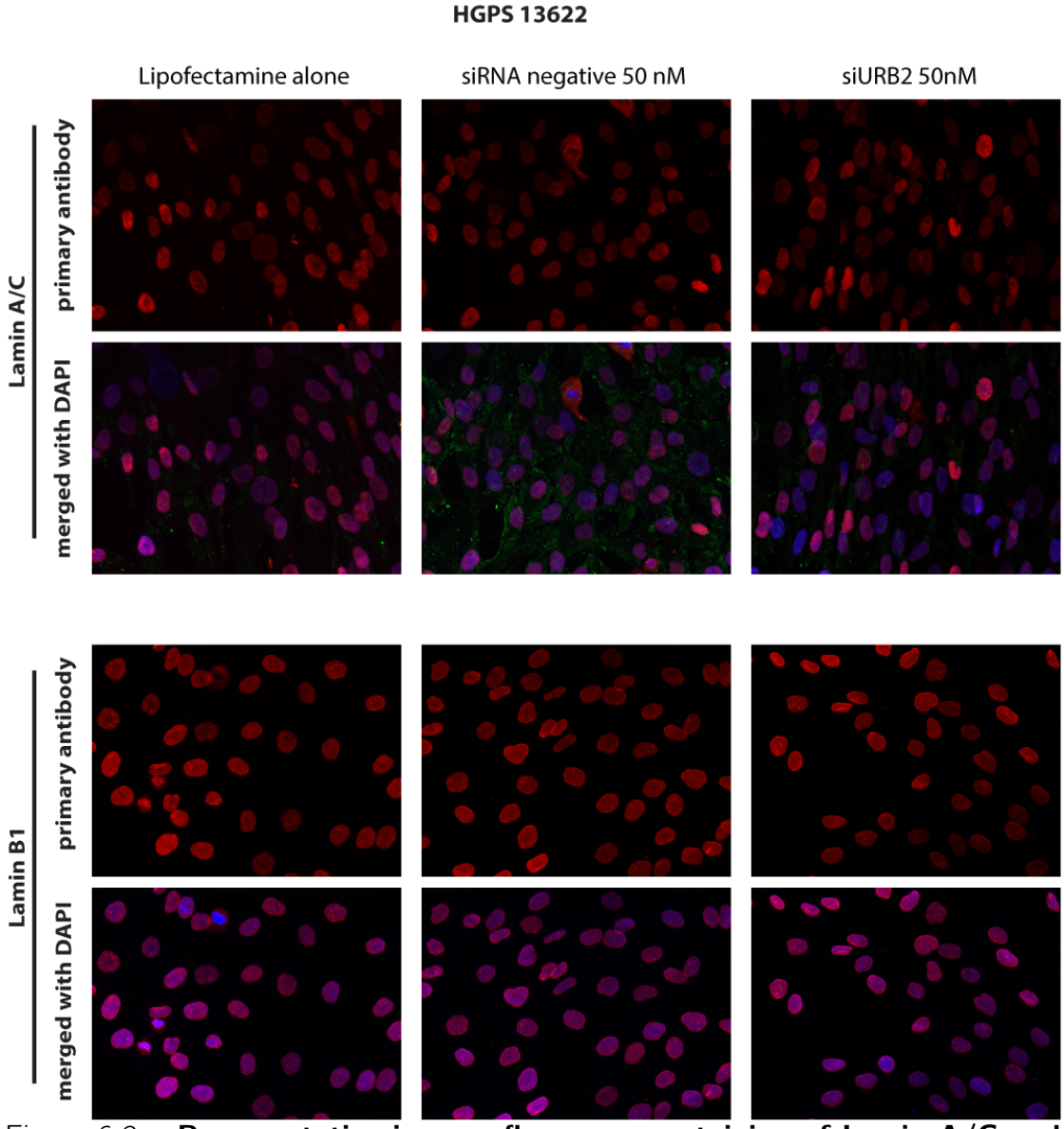
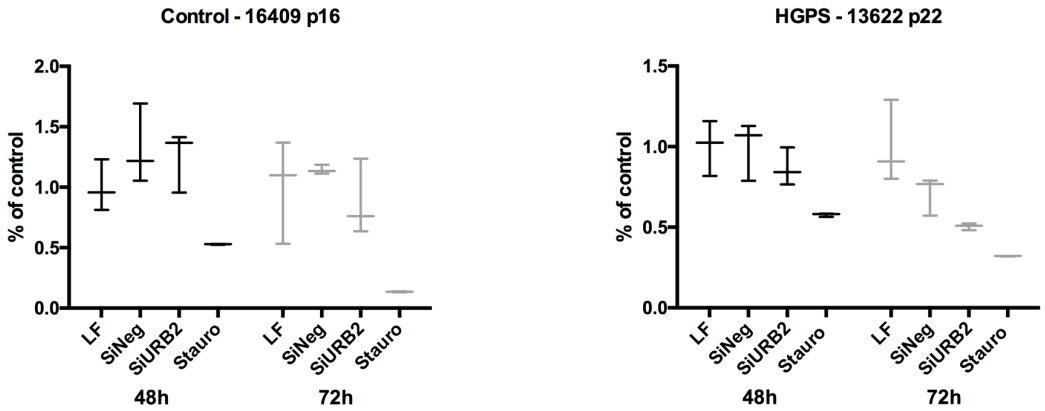


Figure 6.3 – **Representative immunofluorescence staining of Lamin A/C and Lamin B1 in HGPS fibroblast, following 48h URB2 silencing**. Top panels (a, b,c) show primary antibody staining and lower panels (d,e,f) show merged pictures with DAPI stained DNA. **(2A)** Immunofluorescence with rabbit antibody lamin A/C (Santa Cruz, Sc- 20681, cf: $2\mu g/mL$). **(2B)** Immunofluorescence with rabbit antibody lamin B1 (Abcam, ab16048, cf: 10 $\mu g/mL$). Secondary antibody is from Life Technology (DAR Alexa 555) and was used at $5\mu g/mL$. All pictures were acquired on Apotome (Zeiss) at Microscope objective 40X.

A. Viability assay - CyQuant™



B. Proliferation assay - PrestoBlue™

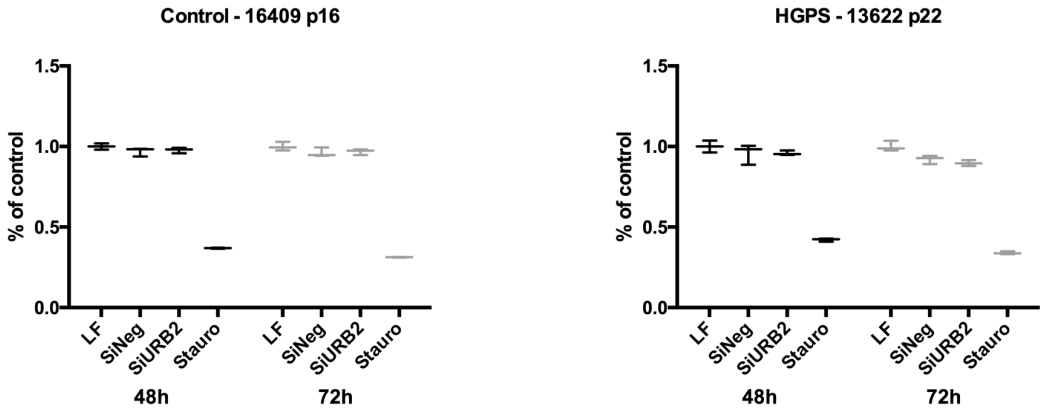


Figure 6.4 – **Viability and proliferation assay after 48h or 72h of URB2 silencing.** Control AG16409 fibroblast (A, C) and HGPS 13622 fibroblast (B, D) were exposed to silencing of URB2 at a final concentration of $50nM$ for 48h or 72h. Each assay included conditions Lipofectamin alone (LF), Si negative at $50nM$ (SiNeg) and a positive control treated with Staurosporine at $1nM$ during 24h (Stauro).

Conclusion

The work presented in this manuscript is the outcome of my Ph.D. studies, which were first conducted at the *Institut de Mathématiques de Marseille, (I2M, UMR7373, Campus de Luminy)*, and then at *Centre de Génétique Médicale, (MMG, UMR S910)*. They were possible thanks to the funding and the close collaboration with the biotechnology R&D company *ProGeLife*. My results have been gathered and composed into a Ph.D. thesis entitled: *"Approaches to Explore Multiplex Biological Networks and Application to Premature Aging Diseases"*

This thesis describes my contribution to some lines of research that are currently gaining the attention of the community. Indeed, the utilization of multiplex networks to describe real-world systems is now expanding to very diverse scientific areas, but few studies have employed multiplex networks in computational biology so far. In addition, recent improvements in the experimental techniques that produce *proteomic* and *phosphoproteomic* data at large-scale have brought a notable increase in the number of studies analyzing this type of datasets. On the application side, we used these approaches to study diseases that are in the research spotlight at the present time: *(i)* premature aging diseases are capturing attention by themselves, and due to their potential molecular links with physiological aging; and *(ii)* prostate cancer is widely studied because is a major cause of death in industrialized countries.

As far as I know, random walk with restart had not previously been extended to multiplex and to multiplex-heterogeneous networks. Moreover, we developed the first module generation seed-expanding algorithm in multiplex and multiplex-heterogeneous networks. These novel methods were applied to study both progeria and the global set of diseases showing a PA phenotype. However, such approaches are not limited to disease studies; they can be applied to networks describing any type of complex system. Noteworthy, the results obtained in this work are predictive and still requiring experimental validations to state further biological conclusions.

Concerning the prostate cancer project, I conducted a more classical statistical and bioinfomatics analyses of *-omics* data. The originality of this work comes, along with its experimental design, from the integration in a network of information about protein abundances, phosphorylation status and molecular interactions. This permitted to obtain a comprehensive perspective of some of the cellular processes deregulated in castration-resistant prostate cancer.

Overall, this thesis introduces innovative methods to tackle relevant biological questions, but they are extensible to further biological topics, and even to other research fields. I strongly believe that my work can be beneficial for future research from both the methodological and the application side.

Bibliography

[14] *About Rare Diseases*. Sept. 2014. URL: https://www.eurordis.org/about-rare-diseases (cit. on p. 70).

[AB15] Eyal Akiva et al. "Evolutionary Reprograming of Protein-Protein Interaction Specificity". In: *Cell* 163.3 (2015), pp. 535–537. ISSN: 10974172. DOI: 10.1016/j.cell.2015.10.010. URL: http://dx.doi.org/10.1016/j.cell.2015.10.010 (cit. on p. 37).

[Alo+16] Diego Alonso-López et al. "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". In: *Nucleic Acids Res* 44.W1 (July 2016). gkw363[PII], W529–W535. ISSN: 1362-4962. DOI: 10.1093/nar/gkw363. URL: https://www.ncbi.nlm.nih.gov/pubmed/27131791 (cit. on p. 42).

[Ami+07] Takeshi Amino et al. "Redefining the disease locus of 16q22.1-linked autosomal dominant cerebellar ataxia". In: *Journal Of Human Genetics* 52 (July 2007). Original Article, 643 EP -. URL: https://doi.org/10.1007/s10038-007-0154-1 (cit. on p. 72).

[Ami+04] Gil Amitai et al. "Network Analysis of Protein Structures Identifies Functional Residues". In: *Journal of Molecular Biology* 344.4 (2004), pp. 1135–1146. ISSN: 0022-2836. URL: http://www.sciencedirect.com/science/article/pii/S0022283604013592 (cit. on p. 57).

[ASW09] Rozalyn M. Anderson et al. "Caloric Restriction and Aging: Studies in Mice and Monkeys". In: *Toxicologic Pathology* 37.1 (Jan. 2009), pp. 47–51. ISSN: 0192-6233. DOI: 10.1177/0192623308329476. URL: https://doi.org/10.1177/0192623308329476 (cit. on p. 102).

[And+09] Bill Andreopoulos et al. "A roadmap of clustering algorithms: finding a match for a biomedical application". In: *Briefings in Bioinformatics* 10.3 (2009), pp. 297–314. ISSN: 14675463. DOI: 10.1093/bib/bbn058 (cit. on pp. 67, 68).

[Ani03] Vladimir N. Anisimov. "Insulin/IGF-1 signaling pathway driving aging and cancer as a target for pharmacological intervention". In: *Experimental Gerontology* 38.10 (2003), pp. 1041–1049. ISSN: 0531-5565. URL: http://www.sciencedirect.com/science/article/pii/S0531556503001694 (cit. on p. 105).

[Atk+09] Holly J. Atkinson et al. “Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies”. In: *PLOS ONE* 4.2 (Feb. 2009), e4345. DOI: 10.1371/journal.pone.0004345. URL: https://doi.org/10.1371/journal.pone.0004345 (cit. on p. 56).

[AS05] Daniel Auerbach et al. “Yeast Two-Hybrid Protein-Protein Interaction Networks”. In: *Proteomics and Protein-Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design*. Ed. by Gabriel Waksman. Boston, MA: Springer US, 2005, pp. 19–31. ISBN: 978-0-387-24532-4. DOI: 10.1007/0-387-24532-4_2. URL: https://doi.org/10.1007/0-387-24532-4%7B%5C_%7D2 (cit. on p. 41).

[Azm+10] Asfar S. Azmi et al. “Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations”. In: *Mol Cancer Ther* 9.12 (Dec. 2010). 1535-7163.MCT-10-0642[PII], pp. 3137–3144. ISSN: 1538-8514. DOI: 10.1158/1535-7163.MCT-10-0642. URL: https://www.ncbi.nlm.nih.gov/pubmed/21041384 (cit. on p. 75).

[Bad+01] G. D. Bader et al. “BIND–The Biomolecular Interaction Network Database”. In: *Nucleic Acids Res* 29.1 (Jan. 2001). PMC29820[pmcid], pp. 242–245. ISSN: 1362-4962. URL: https://www.ncbi.nlm.nih.gov/pubmed/11125103 (cit. on p. 41).

[BH02] Gary D. Bader et al. “Analyzing yeast protein-protein interaction data obtained from different sources”. In: *Nature Biotechnology* 20 (Oct. 2002), 991 EP -. URL: https://doi.org/10.1038/nbt1002-991 (cit. on p. 46).

[BVG15] S. Ballouz et al. “Guidance for RNA-seq co-expression network construction and analysis: safety in numbers”. In: *Bioinformatics* 31.13 (2015), pp. 2123–2130. DOI: 10.1093/bioinformatics/btv118. eprint: /oup/backfile/content_public/journal/bioinformatics/31/13/10.1093_bioinformatics_btv118/3/btv118.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btv118 (cit. on p. 53).

[Bar18] Albert László Barabási. *Love is All You Need. Clauset’s fruitless search for scale-free networks*. Mar. 2018. URL: https://www.barabasilab.com/post/love-is-all-you-need (cit. on p. 63).

[BGL11] Albert László Barabási, Natali Gulbahce, et al. “Network medicine: a network-based approach to human disease”. In: *Nat Rev Genet* 12.1 (Jan. 2011). nrg2918[PII], pp. 56–68. ISSN: 1471-0064. DOI: 10.1038/nrg2918. URL: https://www.ncbi.nlm.nih.gov/pubmed/21164525 (cit. on pp. 65, 72).

[BO04] Albert László Barabási and Zoltán N. Oltvai. “Network biology: Understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5.2 (2004), pp. 101–113. ISSN: 14710056. DOI: 10.1038/nrg1272. arXiv: NIHMS150003 (cit. on pp. 58–60, 62).

[Bar07] Albert-László Barabási. “Network Medicine – From Obesity to the "Diseasome"”. In: *New England Journal of Medicine* 357.4 (July 2007), pp. 404–407. ISSN: 0028-4793. DOI: 10.1056/NEJMe078114. URL: https://doi.org/10.1056/NEJMe078114 (cit. on p. 74).

[BA99] Albert-László Barabási and Réka Albert. “Emergence of Scaling in Random Networks”. In: *Science* 286.5439 (Oct. 1999), p. 509. DOI: 10.1126/science.286.5439.509. URL: http://science.sciencemag.org/content/286/5439/509.abstract (cit. on p. 59).

[Bar+15] Florian Barthélémy et al. “Truncated prelamin A expression in HGPS-like patients: a transcriptional study”. In: *Eur J Hum Genet* 23.8 (Aug. 2015). ejhg2014239[PII], pp. 1051–1061. ISSN: 1476-5438. DOI: 10.1038/ejhg.2014.239. URL: https://www.ncbi.nlm.nih.gov/pubmed/25649378 (cit. on p. 106).

[Bas07] Jordi Bascompte. “Networks in ecology”. In: *Basic and Applied Ecology* 8.6 (2007), pp. 485–490. DOI: 10.1016/j.baae.2007.06.003 (cit. on p. 23).

[Bas+11] George W Bassel et al. “Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions”. In: *Proceedings of the National Academy of Sciences* 108.23 (2011), pp. 9709–9714. ISSN: 0027-8424. DOI: 10.1073/pnas.1100958108. URL: http://www.pnas.org/content/108/23/9709 (cit. on p. 54).

[Bat+16] Federico Battiston et al. “Emergence of Multiplex Communities in Collaboration Networks”. In: *PLOS ONE* 11.1 (Jan. 2016), e0147451. DOI: 10.1371/journal.pone.0147451. URL: https://doi.org/10.1371/journal.pone.0147451 (cit. on p. 78).

[Ben+15] Laura Bennett et al. “Detection of composite communities in multiplex biological networks”. In: *Sci Rep* 5 (May 2015). srep10345[PII], pp. 10345–10345. ISSN: 2045-2322. DOI: 10.1038/srep10345. URL: https://www.ncbi.nlm.nih.gov/pubmed/26012716 (cit. on p. 79).

[BLJ07] Tord Berggård et al. “Methods for the detection and analysis of protein–protein interactions”. In: *PROTEOMICS* 7.16 (2007), pp. 2833–2842. DOI: 10.1002/pmic.200700131. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200700131 (cit. on p. 43).

[BCC15] Gloria Bertoli et al. “MicroRNAs: new biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer”. In: *Theranostics* 5.10 (2015), p. 1122 (cit. on p. 73).

[BS05] Zhang Bin et al. “Statistical Applications in Genetics and Molecular Biology”. In: vol. 4. 1. 2005. Chap. A General Framework for Weighted Gene Co-Expression Network Analysis. DOI: 10.2202/1544-6115.1128. URL: https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1128/sagmb.2005.4.1.1128.xml (cit. on p. 54).

[Bla+13] David R Blair et al. “A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk”. In: *Cell* 155.1 (2013), pp. 70–80 (cit. on p. 71).

[Bla07] Maria A. Blasco. “Telomere length, stem cells and aging”. In: *Nature Chemical Biology* 3 (Sept. 2007). Review Article, 640 EP -. URL: https://doi.org/10.1038/nchembio.2007.38 (cit. on p. 103).

[BS16] Charles Blatti et al. “Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks”. In: *Bioinformatics* 32.14 (July 2016). btw151[PII], pp. 2167–2175. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw151. URL: https://www.ncbi.nlm.nih.gov/pubmed/27153592 (cit. on p. 66).

[Bon+13] Vincenzo Bonnici et al. “A subgraph isomorphism algorithm and its application to biochemical data”. In: *BMC Bioinformatics* 14.SUPPL7 (2013), S13. ISSN: 14712105. DOI: 10.1186/1471-2105-14-S7-S13. URL: http://www.biomedcentral.com/1471-2105/14/S7/S13 (cit. on p. 29).

[Bor+09] Stephen P. Borgatti et al. “Network Analysis in the Social Sciences”. In: *Science* 323.5916 (2009), pp. 892–895. ISSN: 0036-8075. DOI: 10.1126/science.1165821. eprint: http://science.sciencemag.org/content/323/5916/892.full.pdf. URL: http://science.sciencemag.org/content/323/5916/892 (cit. on p. 23).

[BD10] Ralph A. Bradshaw et al. *Handbook of cell signaling*. Academic, 2010 (cit. on p. 50).

[Bra+08] U. Brandes et al. “On Modularity Clustering”. In: *IEEE Transactions on Knowledge and Data Engineering* 20.2 (2008), pp. 172–188. ISSN: 1041-4347. DOI: 10.1109/TKDE.2007.190689 (cit. on p. 68).

[Bra+03] Alvis Brazma et al. “ArrayExpress–a public repository for microarray gene expression data at the EBI”. In: *Nucleic Acids Res* 31.1 (Jan. 2003). PMC165538[pmcid], pp. 68–71. ISSN: 1362-4962. URL: https://www.ncbi.nlm.nih.gov/pubmed/12519949 (cit. on p. 55).

[BP98] S Brin et al. “The anatomy of a large scale hypertextual Web search engine”. In: *Computer Networks and ISDN Systems* 30.1/7 (1998), pp. 107–17. ISSN: 01697552. DOI: 10.1.1.109.4049. arXiv: 1111.6189v1 (cit. on p. 33).

[Bro+00] Gary L Brodsky et al. “Lamin A/C gene mutation associated with dilated cardiomyopathy with variable skeletal muscle involvement”. In: *Circulation* 101.5 (2000), pp. 473–476 (cit. on p. 71).

[Bro18] Tom Brody. “FDA’s Drug Review Process and the Package Label”. In: Academic Press, 2018. Chap. Chapter 7 - Drug-Drug Interactions: Part One (Small Molecule Drugs), pp. 255–335. ISBN: 978-0-12-814647-7. URL: http://www.sciencedirect.com/science/article/pii/B9780128146477000075 (cit. on p. 56).

[BC18] Anna D. Broido et al. “Scale-free networks are rare”. In: *arXiv* (2018). ISSN: 07423225. arXiv: 1801.03400. URL: http://arxiv.org/abs/1801.03400 (cit. on p. 63).

[BJ05] Kevin R. Brown et al. “Online Predicted Human Interaction Database”. In: *Bioinformatics* 21.9 (2005), pp. 2076–2082. DOI: 10.1093/bioinformatics/bti273. eprint: /oup/backfile/content_public/journal/bioinformatics/21/9/10.1093/bioinformatics/bti273/2/bti273.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/bti273 (cit. on p. 42).

[Brü+09] Anna Brückner et al. “Yeast two-hybrid, a powerful tool for systems biology”. In: *International Journal of Molecular Sciences* 10.6 (2009), pp. 2763–2788. ISSN: 14220067. DOI: 10.3390/ijms10062763 (cit. on pp. 39–41).

[BK00] A. J. Butte et al. In: 0. WORLD SCIENTIFIC, Jan. 2000. Chap. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, pp. 418–429. ISBN: 978-981-02-4188-9. DOI: 10.1142/9789814447331_0040. URL: https://doi.org/10.1142/9789814447331_0040 (cit. on pp. 52–54).

[Cab+11] Rubén Cabanillas et al. “Néstor–Guillermo progeria syndrome: A novel premature aging condition with early onset and chronic development caused by BANF1 mutations”. In: *American Journal of Medical Genetics Part A* 155.11 (2011), pp. 2617–2625. DOI: 10.1002/ajmg.a.34249. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajmg.a.34249. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.34249 (cit. on p. 106).

[Cai+18] Pengcheng Cai et al. “Ribosome biogenesis protein Urb2 regulates hematopoietic stem cells development via P53 pathway in zebrafish”. In: *Biochemical and Biophysical Research Communications* 497.2 (2018), pp. 776–782. ISSN: 0006-291X. URL: http://www.sciencedirect.com/science/article/pii/S0006291X18303905 (cit. on p. 175).

[Cam+06] L. M. Camargo et al. “Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia”. In: *Molecular Psychiatry* 12 (Oct. 2006). Original Article, 74 EP -. URL: https://doi.org/10.1038/sj.mp.4001880 (cit. on p. 72).

[18] *Cancer Screening Guidelines | Detecting Cancer Early*. May 2018. URL: https://www.cancer.org/healthy/find-cancer-early/cancer-screening-guidelines/american-cancer-society-guidelines-for-the-early-detection-of-cancer.html (cit. on p. 70).

[Can+15] Laura Cantini et al. “Detection of gene communities in multi-networks reveals cancer drivers”. In: *Scientific Reports* 5 (Dec. 2015). Article, 17386 EP -. URL: https://doi.org/10.1038/srep17386 (cit. on p. 79).

[Car+13] Alessio Cardillo et al. “Emergence of network features from multiplexity”. In: *Scientific Reports* 3 (Feb. 2013). Article, 1344 EP -. URL: https://doi.org/10.1038/srep01344 (cit. on p. 78).

[Car+05] Francesca Cargnin et al. “PHOX2B Regulates Its Own Expression by a Transcriptional Auto-regulatory Mechanism”. In: *Journal of Biological Chemistry* 280.45 (Nov. 2005), pp. 37439–37448. DOI: 10.1074/jbc.M508368200. URL: http://www.jbc.org/content/280/45/37439 (cit. on p. 24).

[Car17] Stephan C. Carlson. *Graph theory*. May 2017. URL: https://www.britannica.com/topic/graph-theory?oasmId=68671 (cit. on p. 23).

[CSL16] Dido Carrero et al. “Hallmarks of progeroid syndromes: lessons from mice and reprogrammed cells”. In: *Dis Model Mech* 9.7 (July 2016). 9/7/719[PII], pp. 719–735. ISSN: 1754-8411. DOI: 10.1242/dmm.024711. URL: https://www.ncbi.nlm.nih.gov/pubmed/27482812 (cit. on pp. 102, 106, 107).

[Cer+11] Ethan G. Cerami et al. “Pathway Commons, a web resource for biological pathway data”. In: *Nucleic Acids Research* 39.suppl$_1$ (2011), pp. D685–D690. DOI: 10.1093/nar/gkq1039. eprint: /oup/backfile/content_public/journal/nar/39/suppl_1/10.1093_nar_gkq1039/2/gkq1039.pdf. URL: http://dx.doi.org/10.1093/nar/gkq1039 (cit. on pp. 48, 51).

[Cha11] Brian T. Chait. “Mass Spectrometry in the Postgenomic Era”. In: *Annual Review of Biochemistry* 80.1 (2011). PMID: 21675917, pp. 239–246. DOI: 10.1146/annurev-biochem-110810-095744. eprint: https://doi.org/10.1146/annurev-biochem-110810-095744. URL: https://doi.org/10.1146/annurev-biochem-110810-095744 (cit. on p. 135).

[CZ12] Gary Chartrand et al. *A first course in graph theory*. Dover Publications, 2012 (cit. on pp. 26, 28).

[Cha+17] Andrew Chatr-Aryamontri et al. “The BioGRID interaction database: 2017 update”. In: *Nucleic Acids Res* 45.D1 (Jan. 2017). gkw1102[PII], pp. D369–D379. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1102. URL: https://www.ncbi.nlm.nih.gov/pubmed/27980099 (cit. on pp. 39, 41).

[Cha+07] Andrew Chatr-aryamontri et al. “MINT: the Molecular INTeraction database”. In: *Nucleic Acids Res* 35.Database issue (Jan. 2007). gkl950[PII], pp. D572–D574. ISSN: 1362-4962. DOI: 10.1093/nar/gkl950. URL: https://www.ncbi.nlm.nih.gov/pubmed/17135203 (cit. on pp. 41, 42).

[Che97] Chen. *Graph theory and its engineering applications*. World Scientific, 1997 (cit. on p. 26).

[Che+15a] Hailin Chen et al. “Network-Based Inference Methods for Drug Repositioning”. In: *Computational and Mathematical Methods in Medicine* 2015 (2015), pp. 1–7. DOI: 10.1155/2015/130620 (cit. on p. 76).

[Che+16a] Xiaowen Chen et al. “Large-scale identification of adverse drug reaction-related proteins through a random walk model”. In: *Scientific Reports* 6.August (2016), p. 36325. ISSN: 2045-2322. DOI: 10.1038/srep36325. URL: http://www.nature.com/articles/srep36325 (cit. on pp. 56, 65, 75).

[CLY12] Xing Chen, Ming-Xi Liu, et al. “Drug–target interaction prediction by random walk on the heterogeneous network”. In: *Molecular BioSystems* 8.7 (2012), p. 1970. ISSN: 1742-206X. DOI: 10.1039/c2mb00002d (cit. on pp. 58, 65).

[Che+16b] Xing Chen, Zhu-Hong You, et al. “IRWRLDA: improved random walk with restart for lncRNA-disease association prediction”. In: *Oncotarget* 7.36 (Aug. 2016). 11141[PII], pp. 57919–57931. ISSN: 1949-2553. URL: https://www.ncbi.nlm.nih.gov/pubmed/27517318 (cit. on p. 65).

[Che+15b] Xiulan Chen et al. “Quantitative proteomics using SILAC: Principles, applications, and developments”. In: *PROTEOMICS* 15.18 (2015), pp. 3175–3192. DOI: 10.1002/pmic.201500108. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201500108. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201500108 (cit. on p. 136).

[Cho+18] Chih-Hung Chou et al. “miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions”. In: *Nucleic Acids Res* 46.D1 (Jan. 2018). 4595852[PII], pp. D296–D302. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1067. URL: https://www.ncbi.nlm.nih.gov/pubmed/29126174 (cit. on p. 56).

[Cla+01] David J. Clancy et al. “Extension of Life-Span by Loss of CHICO, a em $Drosophila/em\ InsulinReceptorSubstrateProtein$”. In: *Science* 292.5514 (Apr. 2001), p. 104. DOI: 10.1126/science.1057991. URL: http://science.sciencemag.org/content/292/5514/104.abstract (cit. on p. 102).

[Cle05] James E Cleaver. “Cancer in xeroderma pigmentosum and related disorders of DNA repair”. In: *Nature Reviews Cancer* 5.7 (2005), p. 564 (cit. on p. 71).

[CH03] Reuven Cohen et al. “Scale-Free Networks Are Ultrasmall”. In: *Phys. Rev. Lett.* 90 (5 Feb. 2003), p. 058701. DOI: 10.1103/PhysRevLett.90.058701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.90.058701 (cit. on p. 60).

[CBS07] Manuel Collado et al. “Cellular Senescence in Cancer and Aging”. In: *Cell* 130.2 (July 2007), pp. 223–233. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.07.003. URL: https://doi.org/10.1016/j.cell.2007.07.003 (cit. on p. 105).

[Cow+17] Lenore Cowen et al. “Network propagation: a universal amplifier of genetic associations”. In: *Nature Reviews Genetics* 18.9 (2017), pp. 551–562. ISSN: 1471-0056. DOI: 10.1038/nrg.2017.38. URL: http://www.nature.com/doifinder/10.1038/nrg.2017.38 (cit. on pp. 65, 66).

[Cow+14] Glenn S. Cowley et al. “Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies”. In: *Scientific Data* 1 (Sept. 2014). Data Descriptor, 140035 EP -. URL: https://doi.org/10.1038/sdata.2014.35 (cit. on p. 56).

[Cow+12] Mark J. Cowley et al. “PINA v2.0: mining interactome modules”. In: *Nucleic Acids Res* 40.Database issue (Jan. 2012). gkr967[PII], pp. D862–D865. ISSN: 1362-4962. DOI: 10.1093/nar/gkr967. URL: https://www.ncbi.nlm.nih.gov/pubmed/22067443 (cit. on p. 42).

[CBM10] Dermot Cox et al. “Integrins as therapeutic targets: lessons and opportunities”. In: *Nature Reviews Drug Discovery* 9 (Oct. 2010). Review Article, 804 EP -. URL: https://doi.org/10.1038/nrd3266 (cit. on p. 172).

[Cus+05] Michael E. Cusick et al. “Interactome: Gateway into systems biology”. In: *Human Molecular Genetics* 14.SUPPL. 2 (2005), pp. 171–181. ISSN: 09646906. DOI: 10.1093/hmg/ddi335 (cit. on pp. 39, 44).

[CDS95] Dragoš M. Cvetković et al. *Spectra of graphs: theory and applications*. Johann Ambrosius Barth, 1995 (cit. on p. 26).

[Dau+02] Matthew Daugherty et al. “Complete Reconstitution of the Human Coenzyme A Biosynthetic Pathway via Comparative Genomics”. In: *Journal of Biological Chemistry* 277.24 (2002), pp. 21431–21439. DOI: 10.1074/jbc.M201708200. eprint: http://www.jbc.org/content/277/24/21431.full.pdf+html. URL: http://www.jbc.org/content/277/24/21431.abstract (cit. on p. 60).

[DSA16] Manlio De Domenico et al. “Mapping Multiplex Hubs in Human Functional Brain Networks”. In: *Frontiers in Neuroscience* 10 (2016), p. 326. ISSN: 1662-453X. DOI: 10.3389/fnins.2016.00326. URL: https://www.frontiersin.org/article/10.3389/fnins.2016.00326 (cit. on p. 78).

[dF10] Javier de Las Rivas et al. “Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks”. In: *PLoS Computational Biology* 6.6 (2010), pp. 1–8. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000807 (cit. on p. 46).

[De +03] Annachiara De Sandre-Giovannoli et al. “Lamin a truncation in Hutchinson-Gilford progeria.” In: *Science (New York, N.Y.)* 300.5628 (2003), p. 2055. ISSN: 0036-8075. DOI: 10.1126/science.1084125 (cit. on p. 105).

[Dea+02] Charlotte M Deane et al. “Protein Interactions: two Methods for Assessment of the Reliability of High Throughput Observations ”. In: *Molecular & Cellular Proteomics* 1.5 (2002), pp. 349–356. ISSN: 1535-9476. DOI: 10.1074/mcp.M100037-MCP200. URL: http://www.mcponline.org/content/1/5/349 (cit. on p. 41).

[Deh11] Matthias Dehmer. *Applied statistics for network biology: methods in systems biology*. Wiley-Blackwell, 2011 (cit. on pp. 53, 54).

[del+13] Noemi del-Toro et al. "A new reference implementation of the PSICQUIC web service". In: *Nucleic Acids Res* 41.Web Server issue (July 2013). gkt392[PII], W601–W606. ISSN: 1362-4962. DOI: 10.1093/nar/gkt392. URL: https://www.ncbi.nlm.nih.gov/pubmed/23671334 (cit. on p. 42).

[Dem+10] Emek Demir et al. "The BioPAX community standard for pathway data sharing". In: *Nature Biotechnology* 28 (Sept. 2010), 935 EP -. URL: https://doi.org/10.1038/nbt.1666 (cit. on p. 52).

[Deo17] Narsingh Deo. *Graph theory: with applications to engineering and computer science*. 2017 (cit. on p. 23).

[Dep+06] Bart Deplancke et al. "A Gene-Centered C. elegans Protein-DNA Interaction Network". In: *Cell* 125.6 (June 2006), pp. 1193–1205. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.04.038. URL: https://doi.org/10.1016/j.cell.2006.04.038 (cit. on p. 60).

[DBB15] Gilles Didier et al. "Identifying Communities from Multiplex Biological Networks". In: *PeerJ* (2015), pp. 1–9. ISSN: 2167-8359. DOI: 10.7717/peerj.1525 (cit. on pp. 79, 91).

[Dit+14] Travis a Dittmer et al. "Systematic identification of pathological lamin A interactors." In: *Molecular biology of the cell* 25.9 (2014), pp. 1493–510. ISSN: 1939-4586. DOI: 10.1091/mbc.E14-02-0733. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4004598%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract (cit. on p. 71).

[DMM14] Miguel J. Divo et al. "Ageing and the epidemiology of multimorbidity". In: *Eur Respir J* 44.4 (Oct. 2014). 09031936.00059814[PII], pp. 1055–1068. ISSN: 1399-3003. DOI: 10.1183/09031936.00059814. URL: https://www.ncbi.nlm.nih.gov/pubmed/25142482 (cit. on p. 74).

[Do+18] Duy N. Do et al. "Integration of miRNA weighted gene co-expression network and miRNA-mRNA co-expression analyses reveals potential regulatory functions of miRNAs in calf rumen development". In: *Genomics* (2018). ISSN: 0888-7543. URL: http://www.sciencedirect.com/science/article/pii/S0888754318301253 (cit. on p. 56).

[Dre+17] Kevin Drew et al. "Integration of over 9,000 mass spectrometry experiments builds a global map of human protein˜complexes". In: *Molecular Systems Biology* 13.6 (2017). DOI: 10.15252/msb.20167490. URL: http://msb.embopress.org/content/13/6/932 (cit. on p. 47).

[Dur+18] Miquel Duran-Frigola et al. “Formatting biological big data for modern machine learning in drug discovery”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1408 (2018). DOI: 10.1002/wcms.1408. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1408 (cit. on p. 79).

[EDL02] Ron Edgar et al. “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository”. In: *Nucleic Acids Research* 30.1 (Jan. 2002), pp. 207–210. ISSN: 0305-1048. DOI: 10.1093/nar/30.1.207. URL: http://dx.doi.org/10.1093/nar/30.1.207 (cit. on p. 55).

[ER60] Paul Erdös et al. “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1960), pp. 17–61 (cit. on p. 59).

[Est+96] Martin Ester et al. “A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. Portland, Oregon: AAAI Press, 1996, pp. 226–231. URL: http://dl.acm.org/citation.cfm?id=3001460.3001507 (cit. on p. 68).

[Est13] E. Estrada. “Graph and Network Theory in Physics”. In: *ArXiv e-prints* (Feb. 2013). arXiv: 1302.4378 [math-ph] (cit. on p. 23).

[Eul36] Leonhard Euler. “Solutio problematis ad geometriam situs pertinentis”. In: *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8 (1736), pp. 128–140 (cit. on p. 24).

[Fab+18] Antonio Fabregat et al. “The Reactome Pathway Knowledgebase”. In: *Nucleic Acids Res* 46.D1 (Jan. 2018). 4626770[PII], pp. D649–D655. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1132. URL: https://www.ncbi.nlm.nih.gov/pubmed/29145629 (cit. on pp. 50, 51).

[Fan+14] Rosa Faner et al. “Network medicine, multimorbidity and the lung in the elderly”. In: *European Respiratory Journal* 44.3 (Sept. 2014), p. 775. DOI: 10.1183/09031936.00078714. URL: http://erj.ersjournals.com/content/44/3/775.abstract (cit. on p. 74).

[Fer+14] Patricia Fernandez et al. “Transformation resistance in a premature aging disorder identifies a tumor-protective function of BRD4”. In: *Cell reports* 9.1 (2014), pp. 248–260 (cit. on p. 71).

[FS89] Stanley Fields et al. “A novel genetic system to detect protein-protein interactions”. In: *Nature* 342 (1989), pp. 340–342. ISSN: 0028-0836. DOI: 10.1038/340301a0. arXiv: nature.vol.342.30nov1989 (cit. on p. 39).

[Fil10] Katja Filippova. “Multi-sentence Compression: Finding Shortest Paths in Word Graphs”. In: COLING ’10 (2010), pp. 322–330. URL: http://dl.acm.org/citation.cfm?id=1873781.1873818 (cit. on p. 23).

[Fla12] Thomas Flatt. “A new definition of aging?” In: *Front Genet* 3 (Aug. 2012). PMC3425790[pmcid], pp. 148–148. ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00148. URL: https://www.ncbi.nlm.nih.gov/pubmed/22936945 (cit. on p. 102).

[Fol+18] Alicia R. Folgueras et al. “Mouse Models to Disentangle the Hallmarks of Human Aging”. In: *Circulation Research* 123.7 (2018), pp. 905–924. ISSN: 0009-7330. DOI: 10.1161/CIRCRESAHA.118.312204. URL: https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.118.312204 (cit. on p. 102).

[For+13] Paola Forabosco et al. “Insights into TREM2 biology by network analysis of human brain gene expression data”. In: *Neurobiology of aging* 34.12 (2013), pp. 2699–2714 (cit. on p. 72).

[Fra+06] L Franke et al. “Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes”. In: *Am J Hum Genet* 78.6 (2006), pp. 1011–1025. ISSN: 0002-9297. DOI: 10.1086/504300. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=16685651 (cit. on p. 65).

[Fra+02] H B Fraser et al. “Evolutionary rate in the protein interaction network”. In: *Science* 296.April (2002), pp. 750–752. ISSN: 1095-9203. DOI: 10.1126/science.1068696\r296/5568/750[pii] (cit. on p. 62).

[Fre77] Linton C. Freeman. “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1 (1977), p. 35. DOI: 10.2307/3033543 (cit. on p. 34).

[FD07] Brendan J. Frey et al. “Clustering by Passing Messages Between Data Points”. In: *Science* 315.5814 (Feb. 2007), p. 972. DOI: 10.1126/science.1136800. URL: http://science.sciencemag.org/content/315/5814/972.abstract (cit. on p. 68).

[FKZ09] Caroline C Friedel et al. “Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast”. In: *Journal of Computational Biology* 16.8 (2009), pp. 971–987. DOI: 10.1089/cmb.2009.0023. URL: https://doi.org/10.1089/cmb.2009.0023 (cit. on p. 44).

[FJ88] D. B. Friedman et al. “A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility”. In: *Genetics* 118.1 (Jan. 1988). PMC1203268[pmcid], pp. 75–86. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pubmed/8608934 (cit. on p. 102).

[Fux12] Monika Fuxreiter. “Fuzziness: linking regulation to protein dynamics”. In: *Molecular BioSystems* 8.1 (2012), pp. 168–177. ISSN: 1742-206X. DOI: 10.1039/C1MB05234A. URL: http://dx.doi.org/10.1039/C1MB05234A (cit. on p. 38).

[Gan+06] T. K. B. Gandhi et al. “Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets”. In: *Nature Genetics* 38 (Feb. 2006), 285 EP -. URL: https://doi.org/10.1038/ng1747 (cit. on p. 39).

[Gav+06] Anne-Claude Gavin, Patrick Aloy, et al. “Proteome survey reveals modularity of the yeast cell machinery”. In: *Nature* 440 (Jan. 2006). Article, 631 EP -. URL: https://doi.org/10.1038/nature04532 (cit. on p. 44).

[Gav+02] Anne-Claude Gavin, Markus Bösche, et al. “Functional organization of the yeast proteome by systematic analysis of protein complexes”. In: *Nature* 415 (Jan. 2002). Article, 141 EP -. URL: https://doi.org/10.1038/415141a (cit. on p. 44).

[Gei+10] Tamar Geiger, Juergen Cox, et al. “Super-SILAC mix for quantitative proteomics of human tumor tissue”. In: *Nature Methods* 7 (Apr. 2010), 383 EP -. URL: https://doi.org/10.1038/nmeth.1446 (cit. on p. 136).

[Gei+11] Tamar Geiger, Jacek R. Wisniewski, et al. “Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics”. In: *Nature Protocols* 6 (Feb. 2011), 147 EP -. URL: https://doi.org/10.1038/nprot.2010.192 (cit. on p. 136).

[GP13] David Gems et al. “Genetics of Longevity in Model Organisms: Debates and Paradigm Shifts”. In: *Annual Review of Physiology* 75.1 (2013). PMID: 23190075, pp. 621–644. DOI: 10.1146/annurev-physiol-030212-183712. eprint: https://doi.org/10.1146/annurev-physiol-030212-183712. URL: https://doi.org/10.1146/annurev-physiol-030212-183712 (cit. on p. 102).

[Geo+06] Richard A. George et al. “Analysis of protein sequence and interaction data for candidate disease gene prediction”. In: *Nucleic Acids Research* 34.19 (2006). ISSN: 03051048. DOI: 10.1093/nar/gkl707 (cit. on p. 65).

[Giu+18] Madalina Giurgiu et al. “CORUM: the comprehensive resource of mammalian protein complexes—2019”. In: *Nucleic Acids Research* (2018), gky973. DOI: 10.1093/nar/gky973. eprint: /oup/backfile/content_public/journal/nar/pap/10.1093_nar_gky973/1/gky973.pdf. URL: http://dx.doi.org/10.1093/nar/gky973 (cit. on pp. 43, 47).

[GR01] Chris Godsil et al. *Algebraic Graph Theory*. Vol. 207. Jan. 2001 (cit. on p. 27).

[Goe+04] Heike Goehler et al. “A Protein Interaction Network Links GIT1, an Enhancer of Huntingtin Aggregation, to Huntington’s Disease”. In: *Molecular Cell* 15.6 (Sept. 2004), pp. 853–865. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2004.09.016. URL: https://doi.org/10.1016/j.molcel.2004.09.016 (cit. on p. 72).

[Goh+07] Kwang-Il Goh et al. “The human disease network”. In: *Proceedings of the National Academy of Sciences* 104.21 (May 2007), p. 8685. DOI: 10.1073/pnas.0701361104. URL: http://www.pnas.org/content/104/21/8685.abstract (cit. on p. 57).

[Gos+15] Marko Gosak et al. “Multilayer network representation of membrane potential and cytosolic calcium concentration dynamics in beta cells”. In: *Chaos, Solitons & Fractals* 80 (2015), pp. 76–82 (cit. on p. 80).

[Gro12] Eitan Gross. “Statistical mechanics of scale-free gene expression networks”. In: *EPL (Europhysics Letters)* 100.5 (2012), p. 58004. URL: http://stacks.iop.org/0295-5075/100/i=5/a=58004 (cit. on p. 60).

[Gro18] Jonathan Gross. *Graph theory and its applications*. CHAPMAN and HALL CRC, 2018 (cit. on pp. 24, 26–28).

[GW02] Björn Grünenfelder et al. “Treasures and traps in genome-wide data sets: case examples from yeast”. In: *Nature Reviews Genetics* 3 (Sept. 2002). Review Article, 653 EP -. URL: https://doi.org/10.1038/nrg886 (cit. on p. 45).

[Gu+17] Changlong Gu et al. “Global network random walk for predicting potential human lncRNA-disease associations”. In: *Scientific Reports* 7.1 (2017), p. 12442. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12763-z. URL: https://doi.org/10.1038/s41598-017-12763-z (cit. on p. 66).

[Gua90] John Guare. *Six degrees of separation: a play*. Random House, 1990 (cit. on p. 60).

[Gui+18] Julien Guinde et al. “Lamins in Lung Cancer: Biomarkers and Key Factors for Disease Progression through miR-9 Regulation?” In: *Cells* 7.7 (July 2018). PMC6071028[pmcid], p. 78. ISSN: 2073-4409. DOI: 10.3390/cells7070078. URL: https://www.ncbi.nlm.nih.gov/pubmed/30012957 (cit. on pp. 71, 174).

[Gül+05] U. Güldener et al. “CYGD: the Comprehensive Yeast Genome Database”. In: *Nucleic Acids Res* 33.Database issue (Jan. 2005). 33/suppl_1/D364[PII], pp. D364–D368. ISSN: 1362-4962. DOI: 10.1093/nar/gki053. URL: https://www.ncbi.nlm.nih.gov/pubmed/15608217 (cit. on p. 43).

[HRO05] Luke Hakes, David L. Robertson, and Stephen G. Oliver. “Effect of dataset selection on the topological interpretation of protein interaction networks”. In: *BMC Genomics* 6 (Sept. 2005). 1471-2164-6-131[PII], pp. 131–131. ISSN: 1471-2164. DOI: 10.1186/1471-2164-6-131. URL: https://www.ncbi.nlm.nih.gov/pubmed/16174296 (cit. on p. 46).

[Hak+07] Luke Hakes, David L. Robertson, Stephen G. Oliver, and Simon C. Lovell. “Protein interactions from complexes: a structural perspective”. In: *Comp Funct Genomics* 2007 (2007). PMC1838958[pmcid], pp. 49356–49356. ISSN: 1531-6912. DOI: 10.1155/2007/49356. URL: https://www.ncbi.nlm.nih.gov/pubmed/17538689 (cit. on p. 46).

[Hal+17] Arda Halu et al. “The multiplex network of human diseases”. In: *bioRxiv* (Jan. 2017), p. 100370. DOI: 10.1101/100370. URL: http://biorxiv.org/content/early/2017/01/18/100370.abstract (cit. on pp. 58, 79).

[HPI16] Hellyeh Hamidi et al. “The complexity of integrins in cancer and new scopes for therapeutic targeting”. In: *Br J Cancer* 115.9 (Oct. 2016). bjc2016312[PII], pp. 1017–1023. ISSN: 1532-1827. DOI: 10.1038/bjc.2016.312. URL: https://www.ncbi.nlm.nih.gov/pubmed/27685444 (cit. on p. 172).

[Ham+05] Ada Hamosh et al. “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders”. In: *Nucleic Acids Research* 33.DATABASE ISS. (2005), pp. 514–517. ISSN: 03051048. DOI: 10.1093/nar/gki033 (cit. on p. 173).

[Han+99] Benjamin F. Hankey et al. “Cancer Surveillance Series: Interpreting Trends in Prostate Cancer—Part I: Evidence of the Effects of Screening in Recent Prostate Cancer Incidence, Mortality, and Survival Rates”. In: *JNCI: Journal of the National Cancer Institute* 91.12 (1999), pp. 1017–1024. DOI: 10.1093/jnci/91.12.1017. eprint: /oup/backfile/content_public/journal/jnci/91/12/10.1093_

jnci_91.12.1017/2/1017.pdf. URL: http://dx.doi.org/10.1093/jnci/91.12.1017 (cit. on p. 134).

[HR05] Robert A. Hanneman et al. *Introduction to social networks*. University of California, 2005 (cit. on p. 32).

[Har+99] Leland H. Hartwell et al. "From molecular to modular cell biology". In: *Nature* 402.6761supp (1999), pp. C47–C52. ISSN: 0028-0836. DOI: 10.1038/35011540. URL: http://www.nature.com/doifinder/10.1038/35011540 (cit. on pp. 61, 66).

[Hav+12] Pierre C. Havugimana et al. "A census of human soluble protein complexes". In: *Cell* 150.5 (Aug. 2012). S0092-8674(12)01006-9[PII], pp. 1068–1081. ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.08.011. URL: https://www.ncbi.nlm.nih.gov/pubmed/22939629 (cit. on p. 44).

[Hei+15] Marco Y Hein et al. "A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances". In: *Cell* 163.3 (Oct. 2015), pp. 712–723. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.09.053. URL: https://doi.org/10.1016/j.cell.2015.09.053 (cit. on p. 47).

[Hid+09] César A. Hidalgo et al. "A Dynamic Network Approach for the Study of Human Phenotypes". In: *PLOS Computational Biology* 5.4 (Apr. 2009), e1000353. DOI: 10.1371/journal.pcbi.1000353. URL: https://doi.org/10.1371/journal.pcbi.1000353 (cit. on pp. 57, 73).

[HB15] Daniel S. Himmelstein et al. "Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes". In: *PLoS Computational Biology* 11.7 (2015), pp. 1–27. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004259 (cit. on p. 56).

[Hog+18] Alexander Hogrebe et al. "Benchmarking common quantification strategies for large-scale phosphoproteomics". In: *Nature Communications* 9.1 (2018), p. 1045. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03309-6. URL: https://doi.org/10.1038/s41467-018-03309-6 (cit. on pp. 136, 137).

[Hop08] Andrew L Hopkins. "Network pharmacology: the next paradigm in drug discovery". In: *Nature chemical biology* 4.11 (2008), p. 682 (cit. on p. 75).

[Hru+08] Tomas Hruz et al. "Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes". In: *Advances in Bioinformatics* 2008 (2008), pp. 1–5. DOI: 10.1155/2008/420747 (cit. on p. 55).

[HTB16] Jessica Xin Hu et al. “Network biology concepts in complex disease comorbidities”. In: *Nature Reviews Genetics* 17 (Aug. 2016). Review Article, 615 EP -. URL: https://doi.org/10.1038/nrg.2016.87 (cit. on p. 74).

[Hut86] J. Hutchinson. “Congenital Absence of Hair and Mammary Glands with Atrophic Condition of the Skin and its Appendages, in a Boy whose Mother had been almost wholly Bald from Alopecia Areata from the age of Six”. In: *Med Chir Trans* 69 (1886). PMC2121576[pmcid], pp. 473–477. ISSN: 0959-5287. URL: https://www.ncbi.nlm.nih.gov/pubmed/20896687 (cit. on p. 105).

[Hut+15] Edward L. Huttlin et al. “The BioPlex Network: A Systematic Exploration of the Human Interactome”. In: *Cell* 162.2 (July 2015). S0092-8674(15)00768-0[PII], pp. 425–440. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.06.043. URL: https://www.ncbi.nlm.nih.gov/pubmed/26186194 (cit. on p. 47).

[Iva+09] Joseph Ivanic et al. “Influence of Protein Abundance on High-Throughput Protein-Protein Interaction Detection”. In: *PLOS ONE* 4.6 (June 2009), e5815. DOI: 10.1371/journal.pone.0005815. URL: https://doi.org/10.1371/journal.pone.0005815 (cit. on p. 62).

[Jam97] Peter James. “Protein identification in the post-genome era: the rapid rise of proteomics”. In: *Quarterly Reviews of Biophysics* 30.4 (1997), pp. 279–331 (cit. on p. 135).

[Jeo+01] H. Jeong et al. “Lethality and centrality in protein networks”. In: *Nature* 411 (May 2001), 41 EP -. URL: https://doi.org/10.1038/35075138 (cit. on p. 62).

[JV94] N. Johnsson et al. “Split ubiquitin as a sensor of protein interactions in vivo”. In: *Proceedings of the National Academy of Sciences* 91.22 (Oct. 1994), p. 10340. DOI: 10.1073/pnas.91.22.10340. URL: http://www.pnas.org/content/91/22/10340.abstract (cit. on p. 41).

[Jos+05] G. Joshi-Tope et al. “Reactome: a knowledgebase of biological pathways”. In: *Nucleic Acids Res* 33.Database issue (Jan. 2005). 33/suppl_1/D428[PII], pp. D428–D432. ISSN: 1362-4962. DOI: 10.1093/nar/gki072. URL: https://www.ncbi.nlm.nih.gov/pubmed/15608231 (cit. on p. 50).

[Kam+13] Atanas Kamburov et al. “The ConsensusPathDB interaction database: 2013 update”. In: *Nucleic Acids Research* 41.D1 (2013), pp. D793–D800. DOI: 10.1093/nar/gks1055. eprint: /oup/backfile/content_public/journal/nar/41/d1/10.1093_nar_gks1055/2/gks1055.pdf. URL: http://dx.doi.org/10.1093/nar/gks1055 (cit. on p. 51).

[Kan+10] Kumaran Kandasamy et al. “NetPath: a public resource of curated signal transduction pathways”. In: *Genome Biol* 11.1 (Jan. 2010). gb-2010-11-1-r3[PII], R3–R3. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-1-r3. URL: https://www.ncbi.nlm.nih.gov/pubmed/20067622 (cit. on p. 51).

[Kan+17] Minoru Kanehisa, Miho Furumichi, et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Res* 45.D1 (Jan. 2017). gkw1092[PII], pp. D353–D361. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1092. URL: https://www.ncbi.nlm.nih.gov/pubmed/27899662 (cit. on pp. 50, 51).

[KG00] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic Acids Res* 28.1 (Jan. 2000). gkd027[PII], pp. 27–30. ISSN: 0305-1048. URL: https://www.ncbi.nlm.nih.gov/pubmed/10592173 (cit. on p. 50).

[Kar+15] Theodoros Karantanos et al. “Understanding the mechanisms of androgen deprivation resistance in prostate cancer at the molecular level”. In: *Eur Urol* 67.3 (Mar. 2015). S0302-2838(14)01000-8[PII], pp. 470–479. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2014.09.049. URL: https://www.ncbi.nlm.nih.gov/pubmed/25306226 (cit. on p. 134).

[Kat+15] Maria Katsogiannou et al. “The hallmarks of castration-resistant prostate cancers”. In: *Cancer Treatment Reviews* (2015). ISSN: 03057372. DOI: 10.1016/j.ctrv.2015.05.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S0305737215000985 (cit. on p. 135).

[KR87] Leonard Kaufmann et al. *Clustering by Means of Medoids*. Data Analysis based on the L1-Norm and Related Methods. Jan. 1987, pp. 405–416 (cit. on p. 67).

[KGP86] L Keegan et al. “Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein”. In: *Science* 231.4739 (1986), pp. 699–704. ISSN: 0036-8075. DOI: 10.1126/science.3080805. eprint: http://science.sciencemag.org/content/231/4739/699.full.pdf. URL: http://science.sciencemag.org/content/231/4739/699 (cit. on p. 39).

[Kel01] Paul Kellam. “Microarray gene expression database: progress towards an international repository of gene expression data”. In: *Genome Biology* 2.5 (2001), reports4011.1. ISSN: 1474-760X. DOI: 10.1186/gb-2001-2-5-reports4011. URL: https://doi.org/10.1186/gb-2001-2-5-reports4011 (cit. on p. 54).

[Ker+12] Samuel Kerrien et al. “The IntAct molecular interaction database in 2012”. In: *Nucleic Acids Research* 40.D1 (2012), pp. 841–846. ISSN: 03051048. DOI: 10.1093/nar/gkr1088. arXiv: 1303.0213 (cit. on pp. 39, 41, 42).

[Kes+09] T. S. Keshava Prasad et al. “Human Protein Reference Database—2009 update”. In: *Nucleic Acids Research* 37.suppl$_1$ (2009), pp. D767–D772. DOI: 10.1093/nar/gkn892. eprint: /oup/backfile/content_public/journal/nar/37/suppl_1/10.1093/nar/gkn892/2/gkn892.pdf. URL: http://dx.doi.org/10.1093/nar/gkn892 (cit. on p. 41).

[KUS18] Arif Khan et al. “Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression”. In: *International Journal of Medical Informatics* 115 (2018), pp. 1–9. ISSN: 1386-5056. URL: http://www.sciencedirect.com/science/article/pii/S1386505618303071 (cit. on p. 57).

[KW06] Raya Khanin et al. “How Scale-Free Are Biological Networks”. In: *Journal of Computational Biology* 13.3 (2006), pp. 810–818. DOI: 10.1089/cmb.2006.13.810 (cit. on p. 63).

[Kiv+14] Mikko Kivelä et al. “Multilayer networks”. In: *Journal of Complex Networks* 2.3 (2014), pp. 203–271. ISSN: 20511329. DOI: 10.1093/comnet/cnu016. arXiv: 1309.7233 (cit. on p. 78).

[Kla18] Erica Klarreich. *Scant Evidence of Power Laws Found in Real-World Networks*. Feb. 2018. URL: https://www.quantamagazine.org/scant-evidence-of-power-laws-found-in-real-world-networks-20180215/ (cit. on p. 63).

[KU07] Manfred Koegl et al. “Improving yeast two-hybrid screening systems”. In: *Briefings in Functional Genomics* 6.4 (2007), pp. 302–312. DOI: 10.1093/bfgp/elm035. eprint: /oup/backfile/content_public/journal/bfg/6/4/10.1093/bfgp/elm035/2/elm035.pdf. URL: http://dx.doi.org/10.1093/bfgp/elm035 (cit. on p. 41).

[Köh+08] Sebastian Köhler et al. “Walking the Interactome for Prioritization of Candidate Disease Genes”. In: *AJHG* 82.April (2008), pp. 949–958. DOI: 10.1016/j.ajhg.2008.02.013.. URL: http://www.cell.com/AJHG/abstract/S0002-9297(08)00172-9 (cit. on p. 65).

[Koh82] Teuvo Kohonen. “Self-organized formation of topologically correct feature maps”. In: *Biological Cybernetics* 43.1 (1982), pp. 59–69. ISSN: 1432-0770. DOI: 10.1007/BF00337288. URL: https://doi.org/10.1007/BF00337288 (cit. on p. 68).

[Kro+06] Nevan J. Krogan et al. “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae”. In: *Nature* 440 (Mar. 2006). Article, 637 EP -. URL: https://doi.org/10.1038/nature04670 (cit. on p. 44).

[Kut+16] Martina Kutmon et al. “WikiPathways: capturing the full diversity of pathway knowledge”. In: *Nucleic Acids Res* 44.D1 (Jan. 2016). gkv1024[PII], pp. D488–D494. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1024. URL: https://www.ncbi.nlm.nih.gov/pubmed/26481357 (cit. on p. 51).

[La +13] Jorge de La Rosa et al. “Prelamin A causes progeria through cell-extrinsic mechanisms and prevents cancer invasion”. In: *Nature communications* 4 (2013), p. 2268 (cit. on p. 71).

[Lac+15] Daniel H. Lackner et al. “A generic strategy for CRISPR-Cas9-mediated gene tagging”. In: *Nature Communications* 6 (Dec. 2015). Article, 10237 EP -. URL: https://doi.org/10.1038/ncomms10237 (cit. on p. 45).

[Lal+08] Sylvie Lalonde et al. “Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations”. In: *The Plant Journal* 53.4 (2008), pp. 610–635. DOI: 10.1111/j.1365-313X.2007.03332.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2007.03332.x (cit. on pp. 41–43).

[LF12] Andrea Lancichinetti et al. “Consensus clustering in complex networks”. In: *Scientific Reports* 2 (Mar. 2012). Article, 336 EP -. URL: https://doi.org/10.1038/srep00336 (cit. on p. 79).

[LZH08] Peter Langfelder et al. “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R”. In: *Bioinformatics* 24.5 (Mar. 2008), pp. 719–720. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm563. URL: http://dx.doi.org/10.1093/bioinformatics/btm563 (cit. on p. 67).

[Lat97] David S. Latchman. “Transcription factors: An overview”. In: *The International Journal of Biochemistry & Cell Biology* 29.12 (1997), pp. 1305–1312. ISSN: 1357-2725. URL: http://www.sciencedirect.com/science/article/pii/S135727259700085X (cit. on p. 38).

[Lee+08] D.-S. Lee et al. “The implications of human metabolic network topology for disease comorbidity”. In: *Proceedings of the National Academy of Sciences* 105.29 (July 2008), p. 9880. DOI: 10.1073/pnas.0802208105. URL: http://www.pnas.org/content/105/29/9880.abstract (cit. on p. 75).

[Lef+08] Vincent D. Blondel Lefebvre et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. ISSN: 1742-5468. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008 (cit. on p. 68).

[LA08] Nicolas Lentze et al. “Membrane-Based Yeast Two-Hybrid System to Detect Protein Interactions”. In: *Current Protocols in Protein Science* 52.1 (2008), pp. 19.17.1–19.17.28. DOI: 10.1002/0471140864.ps1917s52. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471140864.ps1917s52. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471140864.ps1917s52 (cit. on p. 41).

[Li+18a] Hongyang Li et al. “Network Propagation Predicts Drug Synergy in Cancers”. In: *Cancer Research* (Jan. 2018). DOI: 10.1158/0008-5472.CAN-18-0740. URL: http://cancerres.aacrjournals.org/content/early/2018/07/27/0008-5472.CAN-18-0740.abstract (cit. on p. 56).

[Li+13] Meiyi Li et al. “Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis”. In: *Briefings in bioinformatics* 15.2 (2013), pp. 229–243 (cit. on p. 73).

[Li+04] Qiang Li et al. “A modified mammalian tandem affinity purification procedure to prepare functional polycystin-2 channel”. In: *FEBS Letters* 576.1-2 (2004), pp. 231–236. DOI: 10.1016/j.febslet.2004.09.017. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/j.febslet.2004.09.017 (cit. on p. 45).

[LYF16] Xiao-Shui Li et al. “Recent advances in phosphopeptide enrichment: Strategies and techniques”. In: *TrAC Trends in Analytical Chemistry* 78 (2016), pp. 70–83. ISSN: 0165-9936. URL: http://www.sciencedirect.com/science/article/pii/S016599361500271X (cit. on p. 136).

[LL12] Yongjin Li and Jinyan Li. “Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data.” In: *BMC genomics* 13 Suppl 7.Suppl 7 (2012), S27. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-S7-S27. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3521411%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract (cit. on pp. 56, 65).

[LP10] Yongjin Li and Jagdish C. Patra. “Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network”. In: *Bioinformatics* 26.9 (2010), pp. 1219–1224. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq108 (cit. on pp. 58, 65).

[Li+18b] Y. Li et al. “Bayesian Hidden Markov Tree Models for Clustering Genes with Shared Evolutionary History”. In: *ArXiv e-prints* (Aug. 2018). arXiv: 1808.06109 [stat.AP] (cit. on p. 56).

[Lim+06] Janghoo Lim et al. “A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration”. In: *Cell* 125.4 (2006), pp. 801–814. ISSN: 00928674. DOI: 10.1016/j.cell.2006.03.032 (cit. on p. 72).

[LH09] Gipsi Lima-Mendez et al. “The powerful law of the power law and other myths in network biology”. In: *Molecular BioSystems* 5.12 (2009), pp. 1482–1493. ISSN: 1742-206X. DOI: 10.1039/B908681A. URL: http://dx.doi.org/10.1039/B908681A (cit. on p. 63).

[Lin+17] Hui-Heng Lin et al. “Network Analysis of Drug-target Interactions: A Study on FDA-approved New Molecular Entities Between 2000 to 2015”. In: *Scientific Reports* 7.1 (2017), p. 12230. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12061-8. URL: https://doi.org/10.1038/s41598-017-12061-8 (cit. on p. 56).

[LT17] Litwin et al. “The diagnosis and treatment of prostate cancer: A review”. In: *JAMA* 317.24 (2017), pp. 2532–2542. DOI: 10.1001/jama.2017.7248. eprint: /data/journals/jama/936324/jama_litwin_2017_rv_170003.pdf. URL: +%20http://dx.doi.org/10.1001/jama.2017.7248 (cit. on p. 134).

[Liu+16] Hui Liu et al. “Screening lifespan-extending drugs in Caenorhabditis elegans via label propagation on drug-protein networks”. In: *BMC Systems Biology* 10.Suppl 4 (2016). ISSN: 17520509. DOI: 10.1186/s12918-016-0362-4. URL: http://dx.doi.org/10.1186/s12918-016-0362-4 (cit. on p. 65).

[Liu+14] Rui Liu et al. “Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers”. In: *Medicinal research reviews* 34.3 (2014), pp. 455–478 (cit. on p. 73).

[Liu+13] Xiaoping Liu et al. “Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers”. In: *BMC medical genomics* 6.2 (2013), S8 (cit. on p. 73).

[Liu+17] Xiaoyun Liu et al. “Histone Deacetylase AtSRT1 Links Metabolic Flux and Stress Response in Arabidopsis”. In: *Molecular Plant* 10.12 (Dec. 2017), pp. 1510–1522. ISSN: 1674-2052. DOI: 10.1016/j.molp.2017.10.010. URL: https://doi.org/10.1016/j.molp.2017.10.010 (cit. on p. 41).

[Lod16] Harvey F. Lodish. *Molecular cell biology*. Freeman, 2016 (cit. on pp. 49, 50).

[Lóp+13] Carlos López-Otín et al. “The hallmarks of aging.” In: *Cell* 153.6 (2013), pp. 1194–217. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.05.039. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3836174%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract (cit. on pp. 102, 103).

[Lot+18] Maryam Lotfi Shahreza et al. “A review of network-based approaches to drug repositioning”. In: *Briefings in Bioinformatics* 19.5 (Sept. 2018), pp. 878–892. ISSN: 1467-5463. DOI: 10.1093/bib/bbx017. URL: http://dx.doi.org/10.1093/bib/bbx017 (cit. on pp. 56, 76).

[Mac67] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297. URL: https://projecteuclid.org/euclid.bsmsp/1200512992 (cit. on p. 67).

[MCR17] Elena Madrazo et al. “Inside the Cell: Integrins as New Governors of Nuclear Alterations?” In: *Cancers (Basel)* 9.7 (July 2017). PMC5532618[pmcid], p. 82. ISSN: 2072-6694. DOI: 10.3390/cancers9070082. URL: https://www.ncbi.nlm.nih.gov/pubmed/28684679 (cit. on p. 174).

[Mae+10] Shinji Maegawa et al. “Widespread and tissue specific age-related DNA methylation changes in mice”. In: *Genome Res* 20.3 (Mar. 2010). gr.096826.109[PII], pp. 332–340. ISSN: 1549-5469. DOI: 10.1101/gr.096826.109. URL: https://www.ncbi.nlm.nih.gov/pubmed/20107151 (cit. on p. 104).

[Mar18] Daniel Marbach. “Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases”. In: *bioRxiv* (2018) (cit. on pp. 56, 61).

[Mar00] Stijn Marinus van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, Center for Math and Computer Science (CWI). May 2000 (cit. on p. 68).

[Mas+15] Andrea Masotti et al. “Keppen-Lubinsky syndrome is caused by mutations in the inwardly rectifying K+ channel encoded by KCNJ6”. In: *Am J Hum Genet* 96.2 (Feb. 2015). S0002-9297(14)00517-5[PII], pp. 295–300. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2014.12.011. URL: https://www.ncbi.nlm.nih.gov/pubmed/25620207 (cit. on p. 107).

[Mat+12] Julie A. Mattison et al. “Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study”. In: *Nature* 489.7415 (Sept. 2012). nature11432[PII], pp. 318–321. ISSN: 1476-4687. DOI: 10.1038/nature11432. URL: https://www.ncbi.nlm.nih.gov/pubmed/22932268 (cit. on p. 104).

[MSB09] Mark D. McDowall et al. “PIPs: human protein-protein interaction prediction database”. In: *Nucleic Acids Res* 37.Database issue (Jan. 2009). gkn870[PII], pp. D651–D656. ISSN: 1362-4962. DOI: 10.1093/nar/gkn870. URL: https://www.ncbi.nlm.nih.gov/pubmed/18988626 (cit. on p. 42).

[Men+15] Jörg Menche et al. “Disease networks. Uncovering disease-disease relationships through the incomplete interactome.” In: *Science* 347.6224 (2015), p. 1257601. ISSN: 1095-9203. DOI: 10.1126/science.1116608. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1257601%7B%5C%%7D5Cnpapers3://publication/doi/10.1126/science.1257601 (cit. on p. 43).

[MLL15] Jun Meng et al. “Inferring plant microRNA functional similarity using a weighted protein-protein interaction network”. In: *BMC Bioinformatics* 16 (Nov. 2015). PMC4634583[pmcid], pp. 361–361. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0789-4. URL: https://www.ncbi.nlm.nih.gov/pubmed/26538106 (cit. on p. 56).

[Mer+02] Christian von Mering et al. “Comparative assessment of large-scale data sets of protein-protein interactions”. In: *Nature* 417 (May 2002), 399 EP -. URL: https://doi.org/10.1038/nature750 (cit. on p. 46).

[Mi+17] Huaiyu Mi et al. “PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D183–D189. DOI: 10.1093/nar/gkw1138. eprint: /oup/backfile/content_public/journal/nar/45/d1/10.1093_nar_gkw1138/3/gkw1138.pdf. URL: http://dx.doi.org/10.1093/nar/gkw1138 (cit. on p. 51).

[Mil67] S. Milgram. “The small world problem”. In: *Psychology Today* 1 (May 1967), pp. 61–67 (cit. on p. 60).

[Mil+02a] R. Milo et al. “Network Motifs: Simple Building Blocks of Complex Networks”. In: *Science* 298.5594 (2002), pp. 824–827. ISSN: 0036-8075. DOI: 10.1126/science.298.5594.824. eprint: http://science.sciencemag.org/content/298/5594/824.full.pdf. URL: http://science.sciencemag.org/content/298/5594/824 (cit. on p. 29).

[Mil+02b] R Milo et al. “Network Motifs: Simple Building Blocks of Complex Networks”. In: *Science* 298.5594 (2002), pp. 824–827. ISSN: 0036-8075. DOI: 10.1126/science.298.5594.824. URL: http://science.sciencemag.org/content/298/5594/824 (cit. on p. 61).

[Mos+13] Alexey A. Moskalev et al. “The role of DNA damage and repair in aging through the prism of Koch-like criteria”. In: *Ageing Research Reviews* 12.2 (2013), pp. 661–684. ISSN: 1568-1637. URL: http://www.sciencedirect.com/science/article/pii/S1568163712000086 (cit. on p. 103).

[Mot06] Arno G. Motulsky. “Genetics of complex diseases”. In: *J Zhejiang Univ Sci B* 7.2 (Feb. 2006). PMC1363767[pmcid], pp. 167–168. ISSN: 1673-1581. DOI: 10.1631/jzus.2006.B0167. URL: https://www.ncbi.nlm.nih.gov/pubmed/16421979 (cit. on p. 70).

[NM13] Dorogovtsev S. N. et al. *Evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, 2013 (cit. on p. 58).

[Nav+05] Claire L. Navarro, Juan Cadiñanos, et al. “Loss of ZMPSTE24 (FACE-1) causes autosomal recessive restrictive dermopathy and accumulation of Lamin A precursors”. In: *Human Molecular Genetics* 14.11 (June 2005), pp. 1503–1513. ISSN: 0964-6906. DOI: 10.1093/hmg/ddi159. URL: http://dx.doi.org/10.1093/hmg/ddi159 (cit. on p. 106).

[NCL06] Claire L. Navarro, Pierre Cau, et al. “Molecular bases of progeroid syndromes”. In: *Human Molecular Genetics* 15.SUPPL. 2 (2006), pp. 151–161. ISSN: 09646906. DOI: 10.1093/hmg/ddl214 (cit. on pp. 106, 107).

[New06] Mark Newman. “Modularity and community structure in networks”. In: *Proc Natl Acad Sci U S A* 103.23 (June 2006). 0601602103[PII], pp. 8577–8582. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103. URL: https://www.ncbi.nlm.nih.gov/pubmed/16723398 (cit. on p. 32).

[New10] Mark Newman. “Mathematics of networks”. In: *Networks* (2010). DOI: 10.1093/acprof:oso/9780199206650.003.0006 (cit. on p. 32).

[New18] Mark Newman. *Networks an introduction*. Oxford University Press, 2018 (cit. on pp. 32–34).

[NG04] Mark Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69 (2 Feb. 2004), p. 026113. DOI: 10.1103/PhysRevE.69.026113. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.026113 (cit. on pp. 31, 35).

[NJW01] Andrew Y. Ng et al. "On Spectral Clustering: Analysis and an algorithm". In: *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. MIT Press, 2001, pp. 849–856 (cit. on p. 67).

[NT03] Irene M.A. Nooren et al. "Diversity of protein-protein interactions". In: *The EMBO Journal Vol.* 22.14 (2003), pp. 3486–3492 (cit. on p. 38).

[Orc+07] Sandra Orchard et al. "The minimum information required for reporting a molecular interaction experiment (MIMIx)". In: *Nature Biotechnology* 25 (Aug. 2007), 894 EP -. URL: https://doi.org/10.1038/nbt1324 (cit. on p. 42).

[Oti+06] M Oti et al. "Predicting disease genes using protein-protein interactions." In: *Journal of medical genetics* 43.8 (2006), pp. 691–8. ISSN: 1468-6244. DOI: 10.1136/jmg.2006.041376. URL: http://www.ncbi.nlm.nih.gov/pubmed/16611749%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2564594 (cit. on p. 65).

[Ou+13] S R Ou et al. "Does Alzheimer's disease protect against cancers? A nationwide population-based study." In: *Neuroepidemiology* 40 1 (2013), pp. 42–9 (cit. on p. 74).

[Pad+17] Kanchana Padmanabhan et al. "Toward Personalized Network Biomarkers in Alzheimer's Disease: Computing Individualized Genomic and Protein Crosstalk Maps". In: *Front Aging Neurosci* 9 (Sept. 2017). PMC5649142[pmcid], pp. 315–315. ISSN: 1663-4365. DOI: 10.3389/fnagi.2017.00315. URL: https://www.ncbi.nlm.nih.gov/pubmed/29085293 (cit. on p. 73).

[Pal+08] Jacqueline Palmari et al. "Estimating the size of the human interactome". In: *PNAS* 105.19 (2008), pp. 6959–6964. ISSN: 09218912. DOI: 10.1155/2000/356749 (cit. on p. 43).

[Pan+13] Alexander Panossian et al. "Synergy and Antagonism of Active Constituents of ADAPT-232 on Transcriptional Level of Metabolic Regulation of Isolated Neuroglial Cells". In: *Front Neurosci* 7 (Feb. 2013). PMC3576868[pmcid], pp. 16–16. ISSN: 1662-4548. DOI: 10.3389/

fnins.2013.00016. URL: https://www.ncbi.nlm.nih.gov/pubmed/23430930 (cit. on p. 75).

[PSS03] Romualdo Pastor-Satorras et al. "Evolving protein interaction networks through gene duplication". In: *Journal of Theoretical Biology* 222.2 (2003), pp. 199–210. ISSN: 00225193. DOI: 10.1016/S0022-5193(03)00028-6 (cit. on p. 62).

[Per+10] James R. Perkins et al. "Transient Protein-Protein Interactions: Structural, Functional, and Network Properties". In: *Structure* 18.10 (2010), pp. 1233–1243. ISSN: 0969-2126. URL: http://www.sciencedirect.com/science/article/pii/S0969212610003035 (cit. on p. 47).

[Per+18] Claire H. Pernar et al. "The Epidemiology of Prostate Cancer". In: *Cold Spring Harbor Perspectives in Medicine* 8.12 (Aug. 2018). DOI: 10.1101/cshperspect.a030361 (cit. on p. 134).

[Poo+16] Paramasivan Poornima et al. "Network pharmacology of cancer: From understanding of complex interactomes to the design of multi-target specific therapeutics from nature". In: *Pharmacological Research* 111 (2016), pp. 290–302. ISSN: 1043-6618. URL: http://www.sciencedirect.com/science/article/pii/S1043661815301870 (cit. on p. 75).

[Pra+15] Dexter Pratt et al. "NDEx, the Network Data Exchange". In: *Cell Systems* 1.4 (Oct. 2015), pp. 302–305. ISSN: 2405-4712. DOI: 10.1016/j.cels.2015.10.001. URL: https://doi.org/10.1016/j.cels.2015.10.001 (cit. on p. 51).

[Pui+01] Oscar Puig et al. "The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification". In: *Methods* 24.3 (2001), pp. 218–229. ISSN: 1046-2023. DOI: https://doi.org/10.1006/meth.2001.1183. URL: http://www.sciencedirect.com/science/article/pii/S1046202301911831 (cit. on p. 44).

[Puj+07] Miguel Angel Pujana et al. "Network modeling links breast cancer susceptibility and centrosome dysfunction". In: *Nature Genetics* 39 (Oct. 2007). Article, 1338 EP -. URL: https://doi.org/10.1038/ng.2007.2 (cit. on p. 72).

[Pus+18] Sudeep Pushpakom et al. "Drug repurposing: progress, challenges and recommendations". In: *Nature Reviews Drug Discovery* (2018) (cit. on p. 70).

[Qu+16] Miao Qu et al. "Qualitative and quantitative characterization of protein biotherapeutics with liquid chromatography mass spectrometry". In: *Mass Spectrometry Reviews* 36.6 (2016), pp. 734–754. DOI: 10.1002/mas.21500. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21500 (cit. on p. 45).

[RE06] J. Rankin et al. “The laminopathies: a clinical review”. In: *Clinical Genetics* 70.4 (Aug. 2006), pp. 261–274. ISSN: 0009-9163. DOI: 10.1111/j.1399-0004.2006.00677.x. URL: https://doi.org/10.1111/j.1399-0004.2006.00677.x (cit. on p. 71).

[Rao+14] V. Srinivasa Rao et al. “Protein-Protein Interaction Detection: Methods and Analysis”. In: *International Journal of Proteomics* 2014.ii (2014), pp. 1–12. ISSN: 2090-2166. DOI: 10.1155/2014/147648. URL: http://www.hindawi.com/journals/ijpro/2014/147648/ (cit. on pp. 41, 45).

[Rav+02] E Ravasz et al. “Hierarchical Organization of Modularity in Metabolic Networks”. In: *Science* 297.5586 (2002), pp. 1551–1555. ISSN: 0036-8075. DOI: 10.1126/science.1073374. URL: http://science.sciencemag.org/content/297/5586/1551 (cit. on p. 61).

[RC09] Oliver Renner et al. “Mouse Models to Decipher the PI3K Signaling Network in Human Cancer”. In: *Current Molecular Medicine* 9.5 (Jan. 2009), pp. 612–625. DOI: 10.2174/156652409788488766 (cit. on p. 104).

[Rib+18] Diogo M. Ribeiro et al. “Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs”. In: *Nucleic Acids Res* 46.2 (Jan. 2018). 4641912[PII], pp. 917–928. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1169. URL: https://www.ncbi.nlm.nih.gov/pubmed/29165713 (cit. on p. 38).

[Ris13] H Rischbieth. “Progeria And Ateleiosis.” In: *The Lancet* 181.4684 (1913), p. 1625. DOI: 10.1016/s0140-6736(01)42940-0 (cit. on p. 105).

[RS11] Michael Ristow et al. “Extending life span by increasing oxidative stress”. In: *Free Radical Biology and Medicine* 51.2 (2011), pp. 327–336. ISSN: 0891-5849. URL: http://www.sciencedirect.com/science/article/pii/S0891584911003121 (cit. on p. 104).

[RTT15] Liat Rockah-Shmuel et al. “Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations”. In: *PLOS Computational Biology* 11.8 (Aug. 2015), e1004421. DOI: 10.1371/journal.pcbi.1004421. URL: https://doi.org/10.1371/journal.pcbi.1004421 (cit. on p. 57).

[Rog+08] Simon Rogers et al. “Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models”. In: *Bioinformatics* 24.24 (Dec. 2008). btn553[PII], pp. 2894–2900. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn553. URL: https://www.ncbi.nlm.nih.gov/pubmed/18974169 (cit. on p. 135).

[RMK11] David?C Rubinsztein et al. “Autophagy and Aging”. In: *Cell* 146.5 (Sept. 2011), pp. 682–695. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.07.030. URL: https://doi.org/10.1016/j.cell.2011.07.030 (cit. on p. 104).

[Rue+08] Andreas Ruepp et al. “CORUM: the comprehensive resource of mammalian protein complexes”. In: *Nucleic Acids Res* 36.Database issue (Jan. 2008). gkm936[PII], pp. D646–D650. ISSN: 1362-4962. DOI: 10.1093/nar/gkm936. URL: https://www.ncbi.nlm.nih.gov/pubmed/17965090 (cit. on p. 47).

[Ryu+17] Hojin Ryu et al. “MAP Kinase-Mediated Negative Regulation of Symbiotic Nodule Formation in Medicago truncatula”. In: *Mol Cells* 40.1 (Jan. 2017). molcells.2017.2211[PII], pp. 17–23. ISSN: 0219-1032. URL: https://www.ncbi.nlm.nih.gov/pubmed/28152300 (cit. on p. 41).

[Sab66] Gert Sabidussi. “The centrality index of a graph”. In: *Psychometrika* 31.4 (Dec. 1966), pp. 581–603. ISSN: 1860-0980. DOI: 10.1007/BF02289527. URL: https://doi.org/10.1007/BF02289527 (cit. on p. 33).

[Sal+04] Lukasz Salwinski et al. “The Database of Interacting Proteins: 2004 update”. In: *Nucleic Acids Res* 32.Database issue (Jan. 2004), pp. D449–D451. ISSN: 1362-4962. DOI: 10.1093/nar/gkh086. URL: https://www.ncbi.nlm.nih.gov/pubmed/14681454 (cit. on pp. 41, 42).

[SP15] Jose A Santiago et al. “Network-based metaanalysis identifies HNF4A and PTBP1 as longitudinally dynamic biomarkers for Parkinson’s disease”. In: *Proceedings of the National Academy of Sciences* 112.7 (2015), pp. 2257–2262 (cit. on p. 73).

[SS08] Alberto Sanz et al. “The Mitochondrial Free Radical Theory of Aging: A Critical View”. In: *Current Aging Sciencee* 1.1 (Jan. 2008), pp. 10–21. DOI: 10.2174/1874609810801010010 (cit. on p. 104).

[Sch09] Eric E. Schadt. “Molecular networks as sensors and drivers of common human diseases”. In: *Nature* 461.7261 (2009), pp. 218–223. ISSN: 00280836. DOI: 10.1038/nature08454. arXiv: 0808.3099 (cit. on p. 64).

[Sch+09] Carl F. Schaefer et al. “PID: the Pathway Interaction Database”. In: *Nucleic Acids Research* 37.suppl$_1$ (2009), pp. D674–D679. DOI: 10.1093/nar/gkn653. eprint: /oup/backfile/content_public/journal/nar/37/suppl_1/10.1093_nar_gkn653/2/gkn653.pdf. URL: http://dx.doi.org/10.1093/nar/gkn653 (cit. on p. 51).

[Sch+11] Sven-Eric Schelhorn et al. “Inferring physical protein contacts from large-scale purification data of protein complexes”. In: *Mol Cell Proteomics* 10.6 (June 2011). M110.004929[PII], pp. M110.004929–M110.004929. ISSN: 1535-9484. DOI: 10.1074/mcp.M110.004929. URL: https://www.ncbi.nlm.nih.gov/pubmed/21451165 (cit. on p. 46).

[SLP00] Christophe H Schilling et al. “Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective”. In: *Journal of Theoretical Biology* 203.3 (2000), pp. 229–248. ISSN: 0022-5193. DOI: https://doi.org/10.1006/jtbi.2000.1073. URL: http://www.sciencedirect.com/science/article/pii/S0022519300910737 (cit. on p. 48).

[SW06] Katrin Schmelzle et al. “Phosphoproteomic approaches to elucidate cellular signaling networks”. In: *Current Opinion in Biotechnology* 17.4 (2006), pp. 406–414. ISSN: 0958-1669. URL: http://www.sciencedirect.com/science/article/pii/S0958166906000887 (cit. on p. 136).

[Ser+16] Elise A. R. Serin et al. “Learning from Co-expression Networks: Possibilities and Challenges”. In: *Frontiers in Plant Science* 7.April (2016), pp. 1–18. ISSN: 05717256. DOI: 10.1007/BF02114000. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00444/abstract (cit. on pp. 53, 55).

[SOT10] Bostjan Seruga et al. “Drug resistance in metastatic castration-resistant prostate cancer”. In: *Nature Reviews Clinical Oncology* 8 (Sept. 2010). Review Article, 12 EP -. URL: https://doi.org/10.1038/nrclinonc.2010.136 (cit. on p. 134).

[SUS07] Roded Sharan et al. “Network-based prediction of protein function”. In: *Molecular Systems Biology* 3.1 (Jan. 2007). DOI: 10.1038/msb4100129. URL: http://msb.embopress.org/content/3/1/88.abstract (cit. on p. 69).

[SA17] Asuda Sharma et al. “Analysis of clustering algorithms in biological networks”. In: *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017* 2017-January (2017), pp. 2303–2305. ISSN: 0177-798X. DOI: 10.1109/BIBM.2017.8218036 (cit. on p. 67).

[Sha+10] Albert C. Shaw et al. “Aging of the innate immune system”. In: *Curr Opin Immunol* 22.4 (Aug. 2010). S0952-7915(10)00088-9[PII], pp. 507–513. ISSN: 1879-0372. DOI: 10.1016/j.coi.2010.05.003. URL: https://www.ncbi.nlm.nih.gov/pubmed/20667703 (cit. on p. 105).

[SBZ12] Mingguang Shi et al. “A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients”. In: *PloS one* 7.7 (2012), e41292 (cit. on p. 73).

[Shi12] Rob Shields. “Cultural Topology: The Seven Bridges of Königsburg, 1736”. In: *Theory, Culture & Society* 29.4-5 (2012), pp. 43–57. DOI: 10.1177/0263276412451161. URL: https://doi.org/10.1177/0263276412451161 (cit. on p. 24).

[SVE10] S.G. Shirinivas et al. “APPLICATIONS OF GRAPH THEORY IN COMPUTER SCIENCE AN OVERVIEW”. In: *International Journal of Engineering Science and Technology* 2.9 (2010), pp. 4610–4621. ISSN: 1092-0684. DOI: 10.3171/2016.12.FOCUS16475. URL: http://thejns.org/doi/10.3171/2016.12.FOCUS16475 (cit. on p. 26).

[SA10] Oren Shoval et al. “SnapShot: Network Motifs”. In: *Cell* 143.2 (Oct. 2010), 326–326.e1. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.09.050. URL: https://doi.org/10.1016/j.cell.2010.09.050 (cit. on p. 61).

[Sme+15] Damian Smedley, Julius O B Jacobsen, et al. “Next-generation diagnostics and disease-gene discovery with the Exomiser.” In: *Nature protocols* 10.12 (2015), pp. 2004–2015. ISSN: 1750-2799. DOI: 10.1038/nprot.2015.124. URL: http://www.ncbi.nlm.nih.gov/pubmed/26562621 (cit. on pp. 66, 72).

[Sme+14] Damian Smedley, Sebastian Köhler, et al. “Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases”. In: *Bioinformatics* 30.22 (2014), pp. 3215–3222. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu508 (cit. on p. 66).

[Smy+08] Deborah J. Smyth et al. “Shared and distinct genetic variants in type 1 diabetes and celiac disease”. In: *N Engl J Med* 359.26 (Dec. 2008). NEJMoa0807917[PII], pp. 2767–2777. ISSN: 1533-4406. DOI: 10.1056/NEJMoa0807917. URL: https://www.ncbi.nlm.nih.gov/pubmed/19073967 (cit. on p. 73).

[SS09] Jimin Song et al. “How and when should interactome-derived clusters be used to predict functional modules and protein function?” In: *Bioinformatics* 25.23 (Dec. 2009). btp551[PII], pp. 3143–3150. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp551. URL: https://www.ncbi.nlm.nih.gov/pubmed/19770263 (cit. on p. 69).

[SLH12] Lin Song et al. “Comparison of co-expression measures: mutual information, correlation, and model based indices”. In: *BMC Bioinformatics* 13 (Dec. 2012). 1471-2105-13-328[PII], pp. 328–328. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-328. URL: https://www.ncbi.nlm.nih.gov/pubmed/23217028 (cit. on p. 53).

[SA14] Dietrich Stauffer et al. *Introduction To Percolation Theory Revised Second Edition*. Taylor and Francis, 2014 (cit. on p. 61).

[Ste+17] Massimo Stella et al. "Parasite spreading in spatial ecological multiplex networks". In: *Journal of Complex Networks* 5.3 (July 2017), pp. 486–511. ISSN: 2051-1310. DOI: 10.1093/comnet/cnw028. URL: http://dx.doi.org/10.1093/comnet/cnw028 (cit. on p. 78).

[Sto18] James K Stoller. "The Challenge of Rare Diseases". In: *Chest* 153.6 (2018), pp. 1309–1314 (cit. on p. 70).

[Str01] Steven H. Strogatz. "Exploring complex networks". In: *Nature* 410 (Mar. 2001), 268 EP -. URL: https://doi.org/10.1038/35065725 (cit. on p. 58).

[Str+18] Sylwia Struk et al. "Exploring the protein–protein interaction landscape in plants". In: *Plant, Cell & Environment* (2018). DOI: 10.1111/pce.13433. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pce.13433. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pce.13433 (cit. on pp. 39, 45).

[Sty+12] Bram Stynen et al. "Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system". In: *Microbiol Mol Biol Rev* 76.2 (June 2012). 76/2/331[PII], pp. 331–382. ISSN: 1098-5557. DOI: 10.1128/MMBR.05021-11. URL: https://www.ncbi.nlm.nih.gov/pubmed/22688816 (cit. on p. 41).

[Sun+14] Jie Sun et al. "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network". In: *Molecular BioSystems* 10.8 (2014), pp. 2074–2081. ISSN: 1742-206X. DOI: 10.1039/C3MB70608G. URL: http://dx.doi.org/10.1039/C3MB70608G (cit. on p. 58).

[Sze+07] Ildiko Szeverenyi et al. "The Human Intermediate Filament Database: comprehensive information on a gene family involved in many human diseases". In: *Human Mutation* 29.3 (Nov. 2007), pp. 351–360. ISSN: 1059-7794. DOI: 10.1002/humu.20652. URL: https://doi.org/10.1002/humu.20652 (cit. on p. 71).

[SJ15] Damian Szklarczyk and Lars Juhl Jensen. "Protein-Protein Interaction Databases". In: *Protein-Protein Interactions: Methods and Applications*. Ed. by Cheryl L Meyerkord et al. New York, NY: Springer New York, 2015, pp. 39–56. ISBN: 978-1-4939-2425-7. DOI: 10.1007/978-1-4939-2425-7_3. URL: https://doi.org/10.1007/978-1-4939-2425-7%7B%5C_%7D3 (cit. on pp. 41, 43, 50, 51).

[Szk+17] Damian Szklarczyk, John H. Morris, et al. “The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible”. In: *Nucleic Acids Res* 45.D1 (Jan. 2017). gkw937[PII], pp. D362–D368. ISSN: 1362-4962. DOI: 10.1093/nar/gkw937. URL: https://www.ncbi.nlm.nih.gov/pubmed/27924014 (cit. on p. 42).

[Tab+11] Rafael Tabarés-Seisdedos et al. “No paradox, no progress: inverse cancer comorbidity in people with other complex diseases”. In: *The Lancet Oncology* 12.6 (2011), pp. 604–608. ISSN: 1470-2045. URL: http://www.sciencedirect.com/science/article/pii/S1470204511700419 (cit. on p. 74).

[Tac+18] Robi Tacutu et al. “Human Ageing Genomic Resources: New and updated databases”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D1083–D1090. ISSN: 13624962. DOI: 10.1093/nar/gkx1042 (cit. on p. 174).

[Tak+17] Takako Takeda et al. “Predicting drug-drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge”. In: *J Cheminform* 9 (Mar. 2017). 200[PII], pp. 16–16. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0200-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/28316654 (cit. on p. 56).

[Tom+12] Utano Tomaru et al. “Decreased Proteasomal Activity Causes Age-Related Phenotypes and Promotes the Development of Metabolic Abnormalities”. In: *The American Journal of Pathology* 180.3 (Mar. 2012), pp. 963–972. ISSN: 0002-9440. DOI: 10.1016/j.ajpath.2011.11.012. URL: https://doi.org/10.1016/j.ajpath.2011.11.012 (cit. on p. 104).

[TWE18] Vincent Traag et al. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Arvix* (Oct. 2018). arXiv: 1810.08473. URL: http://arxiv.org/abs/1810.08473 (cit. on p. 68).

[TK14] Tien-Dzung Tran et al. “Hierarchical closeness efficiently predicts disease genes in a directed signaling network”. In: *Computational Biology and Chemistry* 53 (2014), pp. 191–197. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927114001030 (cit. on p. 33).

[Tri83] Nenad Trinajstić. *Chemical graph theory*. CRC Press, 1983 (cit. on p. 23).

[Udr+16] Lucretia Udrescu et al. “Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing”. In: *Scientific Reports* 6 (Sept. 2016). Article, 32745 EP -. URL: https://doi.org/10.1038/srep32745 (cit. on p. 56).

[Uhl+15] M. Uhlen et al. “Tissue-based map of the human proteome”. In: *Science* 347.6220 (2015), pp. 1260419–1260419. ISSN: 0036-8075. DOI: 10.1126/science.1260419. arXiv: 0208024 [gr-qc]. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1260419 (cit. on pp. 54, 55).

[Usa+09] Björn Usadel et al. “Co-expression tools for plant biology: opportunities for hypothesis generation and caveats”. In: *Plant, Cell & Environment* 32.12 (Nov. 2009), pp. 1633–1651. ISSN: 0140-7791. DOI: 10.1111/j.1365-3040.2009.02040.x. URL: https://doi.org/10.1111/j.1365-3040.2009.02040.x (cit. on p. 54).

[Val+18] Alberto Valdeolivas et al. “Random walk with restart on multiplex and heterogeneous biological networks”. In: *Bioinformatics* August (2018), pp. 1–9. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty637. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty637/5055408 (cit. on p. 172).

[Val+09] Ravishankar R. Vallabhajosyula et al. “Identifying Hubs in Protein Interaction Networks”. In: *PLOS ONE* 4.4 (Apr. 2009), e5344. DOI: 10.1371/journal.pone.0005344. URL: https://doi.org/10.1371/journal.pone.0005344 (cit. on p. 60).

[Van+14] Jelle Van Leene et al. “An improved toolbox to unravel the plant cellular machinery by tandem affinity purification of Arabidopsis protein complexes”. In: *Nature Protocols* 10 (Dec. 2014), 169 EP -. URL: https://doi.org/10.1038/nprot.2014.199 (cit. on p. 44).

[Van+10] Oron Vanunu et al. “Associating genes and protein complexes with disease via network propagation”. In: *PLoS Computational Biology* 6.1 (2010). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000641 (cit. on p. 65).

[Váz+02] Alexei Vázquez et al. “Modeling of Protein Interaction Networks”. In: *Complexus* 1.1 (Nov. 2002), pp. 38–44. DOI: 10.1159/000067642 (cit. on p. 62).

[VVR06] Pernette J. Verschure et al. “Step out of the Groove: Epigenetic Gene Control Systems and Engineered Transcription Factors”. In: vol. 56. Advances in Genetics. Academic Press, 2006, pp. 163–204. DOI: https://doi.org/10.1016/S0065-2660(06)56005-5. URL: http://www.sciencedirect.com/science/article/pii/S0065266006560055 (cit. on p. 39).

[Ver07] M. Verwoerd. “Graph theory and networks in Biology”. English. In: *IET Systems Biology* 1 (2 Mar. 2007), 89–119(30). ISSN: 1751-8849. URL: http://digital-library.theiet.org/content/journals/10.1049/iet-syb_20060038 (cit. on p. 23).

[VCB11] Marc Vidal et al. “Interactome networks and human disease”. In: *Cell* 144.6 (Mar. 2011). S0092-8674(11)00130-9[PII], pp. 986–998. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.02.016. URL: https://www.ncbi.nlm.nih.gov/pubmed/21414488 (cit. on pp. 60–62).

[VSB15] Karthik Viswanathan et al. “Amino acid interaction networks provide a new lens for therapeutic antibody discovery and anti-viral drug optimization”. In: *Current Opinion in Virology* 11 (2015), pp. 122–129. ISSN: 1879-6257. URL: http://www.sciencedirect.com/science/article/pii/S1879625715000528 (cit. on p. 57).

[Vla+17] Panayotis Vlastaridis et al. “Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes”. In: *Gigascience* 6.2 (Jan. 2017). 2865214[PII], pp. 1–11. ISSN: 2047-217X. DOI: 10.1093/gigascience/giw015. URL: https://www.ncbi.nlm.nih.gov/pubmed/28327990 (cit. on p. 136).

[VFW04] Sylvia Vlcek et al. “Lco1 is a novel widely expressed lamin-binding protein in the nuclear interior”. In: *Experimental Cell Research* 298.2 (2004), pp. 499–511. ISSN: 0014-4827. URL: http://www.sciencedirect.com/science/article/pii/S0014482704002484 (cit. on p. 175).

[WF01] A. Wagner et al. “The small world inside large metabolic networks”. In: *Proc Biol Sci* 268.1478 (Sept. 2001). PMC1088812[pmcid], pp. 1803–1810. ISSN: 0962-8452. DOI: 10.1098/rspb.2001.1711. URL: https://www.ncbi.nlm.nih.gov/pubmed/11522199 (cit. on p. 60).

[Wan+15] Cuihong Wan et al. “Panorama of ancient metazoan macromolecular complexes”. In: *Nature* 525 (Sept. 2015). Article, 339 EP -. URL: https://doi.org/10.1038/nature14877 (cit. on p. 47).

[WC11] Yu-Chao Wang et al. “A network-based biomarker approach for molecular investigation and diagnosis of lung cancer”. In: *BMC medical genomics* 4.1 (2011), p. 2 (cit. on p. 73).

[Wan+11] Zheng Wang et al. “A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny”. In: *PLOS ONE* 6.3 (Mar. 2011), e17906. DOI: 10.1371/journal.pone.0017906. URL: https://doi.org/10.1371/journal.pone.0017906 (cit. on p. 56).

[War+10] David Warde-Farley et al. “The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function”. In: *Nucleic Acids Res* 38.Web Server issue (July 2010). gkq537[PII], W214–W220. ISSN: 1362-4962. DOI: 10.1093/nar/gkq537. URL: https://www.ncbi.nlm.nih.gov/pubmed/20576703 (cit. on p. 42).

[WS98] Duncan J. Watts et al. “Collective dynamics of ’small-world’ networks”. In: *Nature* 393 (June 1998), 440 EP -. URL: https://doi.org/10.1038/30918 (cit. on p. 60).

[Wei11] Matthew T. Weirauch. “Gene Coexpression Networks for the Analysis of DNA Microarray Data”. In: *Applied Statistics for Network Biology*. John Wiley Sons, Ltd, 2011. Chap. 11, pp. 215–250. ISBN: 9783527638079. DOI: 10.1002/9783527638079.ch11. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527638079.ch11. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527638079.ch11 (cit. on p. 52).

[WIC13] Jukka Westermarck et al. “Identification of Protein Interactions Involved in Cellular Signaling”. In: *Molecular & Cellular Proteomics* 12.7 (2013), pp. 1752–1763. ISSN: 1535-9476. DOI: 10.1074/mcp.R113.027771. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.R113.027771 (cit. on p. 37).

[Wil+12] John E. Wilkinson et al. “Rapamycin slows aging in mice”. In: *Aging Cell* 11.4 (Aug. 2012). PMC3434687[pmcid], pp. 675–682. ISSN: 1474-9726. DOI: 10.1111/j.1474-9726.2012.00832.x. URL: https://www.ncbi.nlm.nih.gov/pubmed/22587563 (cit. on p. 102).

[WJL76] Robin J. Wilson et al. *Graph theory, 1736-1936*. 1976 (cit. on p. 24).

[WBR15] Christian Wiwie et al. “Comparing the performance of biomedical clustering methods”. In: *Nature Methods* 12 (Sept. 2015), 1033 EP -. URL: https://doi.org/10.1038/nmeth.3583 (cit. on pp. 67, 68).

[WKB05] Cecily J. Wolfe et al. “Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks”. In: *BMC Bioinformatics* 6 (Sept. 2005). 1471-2105-6-227[PII], pp. 227–227. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-227. URL: https://www.ncbi.nlm.nih.gov/pubmed/16162296 (cit. on p. 52).

[WS12] Guanming Wu et al. “A network module-based method for identifying cancer prognostic signatures”. In: *Genome biology* 13.12 (2012), R112 (cit. on p. 73).

[WCW14] Xiaodan Wu et al. “Network biomarkers, interaction networks and dynamical network biomarkers in respiratory diseases”. In: *Clinical and translational medicine* 3.1 (2014), p. 16 (cit. on p. 73).

[WOB03] S. Wuchty et al. “Evolutionary conservation of motif constituents in the yeast protein interaction network”. In: *Nature Genetics* 35 (Sept. 2003), 176 EP -. URL: https://doi.org/10.1038/ng1242 (cit. on p. 61).

[XLG06] Yu Xia et al. “Integrated Prediction of the Helical Membrane Protein Interactome in Yeast”. In: *Journal of Molecular Biology* 357.1 (2006), pp. 339–349. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2005.12.067. URL: http://www.sciencedirect.com/science/article/pii/S0022283605016438 (cit. on p. 41).

[Xie+15] MaoQiang Xie et al. “Network-based phenome-genome association prediction by bi-random walk”. In: *PLoS ONE* 10.5 (2015), pp. 1–18. ISSN: 19326203. DOI: 10.1371/journal.pone.0125138 (cit. on p. 65).

[Xu06] Xiaojie Xu. “New concepts and approaches for drug discovery based on traditional Chinese medicine”. In: *Drug Discovery Today: Technologies* 3.3 (2006), pp. 247–253. ISSN: 1740-6749. URL: http://www.sciencedirect.com/science/article/pii/S1740674906000564 (cit. on p. 75).

[YK18] Ayako Yachie-Kinoshita et al. “Cell Modeling and Simulation”. In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by Shoba Ranganathan et al. Oxford: Academic Press, 2018, pp. 864–873. ISBN: 978-0-12-811432-2. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20294-X. URL: http://www.sciencedirect.com/science/article/pii/B978012809633820294X (cit. on p. 50).

[Yan+18] Long Yang et al. “InsP3R-SEC5 interaction on phagosomes modulates innate immunity to Candida albicans by promoting cytosolic Ca2+ elevation and TBK1 activity”. In: *BMC biology* 16.1 (Apr. 2018), p. 46. ISSN: 1741-7007. DOI: 10.1186/s12915-018-0507-6. URL: http://europepmc.org/articles/PMC5921305 (cit. on p. 41).

[YC14] Hye Eun Yoon et al. “The renin-angiotensin system and aging in the kidney”. In: *Korean J Intern Med* 29.3 (May 2014), pp. 291–295. ISSN: 2005-6648. DOI: 10.3904/kjim.2014.29.3.291. URL: https://www.ncbi.nlm.nih.gov/pubmed/24851061 (cit. on p. 105).

[YBL06] Jeongah Yoon et al. “An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality”. In: *Bioinformatics* 22.24 (Dec. 2006), pp. 3106–3108. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl533. URL: http://dx.doi.org/10.1093/bioinformatics/btl533 (cit. on p. 35).

[Yu+09] Haiyuan Yu et al. “High Quality Binary Protein Interaction Map of the Yeast Interactome Network”. In: *Science* 322.5898 (2009), pp. 104–110. ISSN: 1350-1763. DOI: 10.1126/science.1158684.High (cit. on p. 62).

[Yua+14] Jiao Yuan et al. “NPInter v2.0: an updated database of ncRNA interactions”. In: *Nucleic Acids Res* 42.Database issue (Jan. 2014). gkt1057[PII], pp. D104–D108. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1057. URL: https://www.ncbi.nlm.nih.gov/pubmed/24217916 (cit. on p. 56).

[ZKM11] Sara Zanivan et al. “In Vivo Quantitative Proteomics: The SILAC Mouse”. In: *Integrin and Cell Adhesion Molecules Methods in Molecular Biology* (2011), pp. 435–450. DOI: 10.1007/978-1-61779-166-6_25 (cit. on p. 136).

[Zha+16] Fan Zhang et al. “A network medicine approach to build a comprehensive atlas for the prognosis of human cancer”. In: *Brief Bioinform* 17.6 (Nov. 2016). bbw076[PII], pp. 1044–1059. ISSN: 1477-4054. DOI: 10.1093/bib/bbw076. URL: https://www.ncbi.nlm.nih.gov/pubmed/27559151 (cit. on p. 73).

[Zha03] Jianzhi Zhang. “Evolution by gene duplication: an update”. In: *Trends in Ecology & Evolution* 18.6 (2003), pp. 292–298. ISSN: 0169-5347. URL: http://www.sciencedirect.com/science/article/pii/S0169534703000338 (cit. on p. 62).

[Zha+15a] Xiao-Fei Zhang et al. “Identifying binary protein-protein interactions from affinity purification mass spectrometry data”. In: *BMC Genomics* 16 (Oct. 2015). PMC4595009[pmcid], pp. 745–745. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1944-z. URL: https://www.ncbi.nlm.nih.gov/pubmed/26438428 (cit. on p. 46).

[Zha+11] Jing Zhao et al. “Ranking Candidate Disease Genes from Gene Expression and Protein Interaction: A Katz-Centrality Based Approach”. In: *PLOS ONE* 6.9 (Sept. 2011), e24306. DOI: 10.1371/journal.pone.0024306. URL: https://doi.org/10.1371/journal.pone.0024306 (cit. on p. 33).

[Zha+14] Shanrong Zhao et al. “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells”. In: *PLOS ONE* 9.1 (Jan. 2014), e78644. DOI: 10.1371/journal.pone.0078644. URL: https://doi.org/10.1371/journal.pone.0078644 (cit. on p. 55).

[Zha+15b] Zhi Qin Zhao et al. “Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization”. In: *Computational Biology and Chemistry* 57 (2015), pp. 21–28. ISSN: 14769271. DOI: 10.1016/j.compbiolchem.2015.02.008. URL: http://dx.doi.org/10.1016/j.compbiolchem.2015.02.008 (cit. on p. 65).

[Zho+09] Quan Zhong et al. “Edgetic perturbation models of human inherited disorders”. In: *Molecular Systems Biology* 5.321 (2009). DOI: 10.1038/msb.2009.80 (cit. on p. 64).

[Zho+16] Meng Zhou et al. “Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer”. In: *Oncotarget* 7.35 (2016), p. 56383 (cit. on p. 73).

[Zho+14] XueZhong Zhou et al. “Human symptoms-disease network”. In: *Nature Communications* 5 (June 2014). Article, 4212 EP -. URL: https://doi.org/10.1038/ncomms5212 (cit. on p. 57).

[ZL17] Marinka Zitnik et al. “Predicting multicellular function through multi-layer tissue networks”. In: *Bioinformatics* 33.14 (2017), pp. i190–i198. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx252. arXiv: 1707.04638 (cit. on p. 79).

[Zuo+18a] Huali Zuo et al. “A network pharmacology-based approach to analyse potential targets of traditional herbal formulas: An example of Yu Ping Feng decoction”. In: *Scientific Reports* 8.1 (2018), p. 11418. ISSN: 2045-2322. DOI: 10.1038/s41598-018-29764-1. URL: https://doi.org/10.1038/s41598-018-29764-1 (cit. on p. 75).

[Zuo+18b] Lingkun Zuo et al. “Lamin A/C might be involved in the EMT signalling pathway”. In: *Gene* 663 (2018), pp. 51–64. ISSN: 0378-1119. URL: http://www.sciencedirect.com/science/article/pii/S0378111918304074 (cit. on p. 71).

Appendices

A Random walk with restart on multiplex networks

A.1 Random walk with restart on multiplex and heterogeneous biological networks: supplementary materials

Random Walk with Restart on Multiplex and Heterogeneous Biological Networks: Supplementary Materials

Alberto Valdeolivas[1,2*], Laurent Tichit[1], Claire Navarro[3], Sophie Perrin[3], Gaëlle Odelin[2], Nicolas Levy[3], Pierre Cau[2,3], Elisabeth Remy[1], Anaïs Baudot[1*]

[1]*Aix-Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Marseille, France,*
[2]*ProGeLife, 8 Rue Sainte Barbe 13001, Marseille, France and*
[3]*Aix-Marseille Université, INSERM, UMR_S910, Faculté de Médecine, Marseille, France.*

* alberto.valdeolivas@etu.univ-amu.fr, anais.baudot@univ-amu.fr

Contents

1 Supplementary Methods

The source codes to generate networks and run the RWR algorithms are available on GitHub at: https://github.com/alberto-valdeolivas/RWR-MH.

1.1 Construction of the Biological Networks

We constructed three biological networks containing genes or proteins as nodes. First, a protein-protein interaction (PPI) network of physical direct binary interactions, selected from the PSICQUIC portal (Del-Toro *et al.*, 2013), and merged with the CCSB Interactome database (Rolland *et al.*, 2014). Second, a network connecting proteins according to pathway interaction data, using the R package graphite (Sales *et al.*, 2012). Finally, a co-expression network from RNA-Seq data downloaded from the Human Protein Atlas (http://www.proteinatlas.org) (Uhlen *et al.*, 2015). We computed Spearman correlations of TPM expression data from 32 tissues and 45 cell lines, and selected correlation having an absolute value $\geqslant 0.7$ to be included in the network. The technical details and source code to recreate these datasets are available on GitHub.

1.2 Construction of the Disease-Disease similarity network

A disease can be described as a set of phenotypes. We downloaded the annotation file *phenotype_annotation.tab*, containing diseases and their associated phenotypes, from the Human Phenotype Ontology (HPO), together with the HPO ontology graph structure (Köhler *et al.*, 2014) on November, 2016. We kept only disease records from OMIM (Hamosh *et al.*, 2005), and for each disease, we extracted its minimal set of HPO terms. A set of phenotypes is minimal if it describes a disease without redundancy: we considered only the deepest (i.e., the most precise) nodes in the directed ontology structure, as described by Greene *et al.*, 2016.

The phenotype similarity between a pair of diseases can be computed by counting the number of shared phenotypes. However, some phenotypes are more relevant than others. We indeed want to consider as more similar two diseases sharing a very rare phenotype, than two diseases sharing a very common phenotype, as proposed by Westbury *et al.*, 2015. To this goal, we estimated the relevance of each phenotype based on its frequency in the HPO database, and used the relative information content (IC), defined as follows:

$$IC(i) = -log(f_i) \tag{S-6}$$

where f_i is the frequency of the phenotype i within our set of HPO diseases. The similarity between phenotypes i and j is then computed as:

$$sim(i,j) = \max_{t \in anc(i) \cap anc(j)} IC(t) \tag{S-7}$$

where $anc(i)$ indicates the ancestors of the phenotype i in the ontology graph. Finally, the phenotype similarity between a pair of diseases D_a and D_b, corresponding to two sets of HPO phenotypes, is measured by the total IC of their shared phenotypes, as described in Resnik, 1999:

$$sim(D_a, D_b) = \frac{1}{|D_a|} \sum_{i \in D_a} \underset{j \in D_b}{max}(sim(i,j)) + \frac{1}{|D_b|} \sum_{j \in D_b} \underset{i \in D_a}{max}(sim(j,i)) \tag{S-8}$$

The similarity score between all pairs of diseases is computed according to Equation S-8. The disease-disease similarity network is built by linking every disease to its five nearest diseases according to this similarity score, as in Li and Patra, 2010. The resulting disease-disease similarity network is composed of 6 947 diseases connected by 28 246 edges (Supplementary Table S-1).

1.3 RWR-H: Extension of RWR to heterogeneous graphs

Following the approach proposed by Li and Patra, 2010, let us consider the graphs G_V, G_U and G_B, which form a heterogeneous graph. We define $A_{P(n\times n)}$, $A_{D(m\times m)}$ and $B_{(n\times m)}$ as their respective adjacency matrices. These matrices can be considered here as the adjacency matrices of the PPI network, the disease-disease similarity network and the bipartite network, respectively. Therefore, the adjacency matrix of the heterogeneous network can be represented as: $A = \begin{bmatrix} A_P & B \\ B^T & A_D \end{bmatrix}$, with B^T the transpose of the matrix B.

We then compute the different transition probabilities of the random walk with restart on heterogeneous graphs (RWR-H). Let $H = \begin{bmatrix} H_{PP} & H_{PD} \\ H_{DP} & H_{DD} \end{bmatrix}$ denotes the matrix of transitions on the heterogeneous graph, where H_{PP} and H_{DD} describe the walks within a network, and H_{PD}, H_{DP} describe the jumps between networks. For a given node, if a bipartite association exists, the particle can either jump between graphs or stay in the current graph with a probability given by the parameter $\lambda \in [0, 1]$. The closer λ is to one, the higher is the probability of jumping between networks.

Let a particle be located at the protein node $p_i \in V$. At the next step, the particle can either walk to a protein $p_j \in V$ with the following transition probability:

$$H_{PP}(i,j) = \begin{cases} A_P(i,j)/\sum_{k=1}^{n} A_P(i,k), \text{ if } \sum_{k=1}^{m} B(i,k) = 0 \\ (1-\lambda)A_P(i,j)/\sum_{k=1}^{n} A_P(i,k), \text{ otherwise} \end{cases} \quad \text{(S-1)}$$

or jump through a bipartite association to the disease $d_b \in U$ with a probability:

$$H_{PD}(i,b) = \begin{cases} \lambda B(i,b)/\sum_{k=1}^{m} B(i,k), \text{ if } \sum_{k=1}^{m} B(i,k) \neq 0 \\ 0, \text{ otherwise} \end{cases} \quad \text{(S-2)}$$

The same situation arises if a particle is located at the disease $d_a \in U$. It can walk to the disease $d_b \in U$:

$$H_{DD}(a,b) = \begin{cases} A_D(a,b)/\sum_{k=1}^{m} A_D(a,k), \text{ if } \sum_{k=1}^{n} B(k,b) = 0 \\ (1-\lambda)A_D(a,b)/\sum_{k=1}^{m} A_D(a,k), \text{ otherwise} \end{cases} \quad \text{(S-3)}$$

or jump to the protein $p_j \in V$:

$$H_{DP}(a,j) = \begin{cases} \lambda B(j,a) / \sum_{k=1}^{n} B(k,a), \text{ if } \sum_{k=1}^{n} B(k,a) \neq 0 \\ 0, \text{ otherwise} \end{cases} \quad \text{(S-4)}$$

Therefore, we can write the RWR-H equation on a heterogeneous graph as:

$$\widetilde{\mathbf{p}}_{t+1}^{T} = (1-r) H \widetilde{\mathbf{p}}_{t}^{T} + r \widetilde{\mathbf{p}}_{RS}^{T} \quad \text{(S-5)}$$

The vectors $\widetilde{\mathbf{p}}_{t+1}$, $\widetilde{\mathbf{p}}_{t}$ and $\widetilde{\mathbf{p}}_{RS}$ are now of dimension $n+m$, since the RWR-H algorithm is ranking proteins and diseases at the same time. Importantly, after a restart step, the particle can go back either to a seed protein or to a seed disease. It is to note that it is possible to tune the importance of each network by defining $\widetilde{\mathbf{p}}_{RS} = \begin{bmatrix} (1-\eta)\mathbf{v}_0 \\ \eta \mathbf{u}_0 \end{bmatrix}$, where $\mathbf{v}_0$ and $\mathbf{u}_0$ represent the initial probability distributions on the PPI and the disease-disease similarity networks given by their seed nodes. The parameter $\eta \in [0,1]$ controls the probability of restarting in each network (PPI or disease-disease). If $\eta < 0.5$ the particle will be more likely to restart in one of the seed proteins than in one of the seed diseases. In our work, we set both parameters λ and η to 0.5.

1.4 Leave-one-out cross validation

In order to evaluate the performances of the different RWR algorithms, we designed a Leave-One-Out Cross-Validation (LOOCV) strategy. We downloaded diseases and associated genes from OMIM (Hamosh *et al.*, 2005) (downloaded on December, 2016) and DisGeNET v4.0 (Piñero *et al.*, 2016) (associations with a score greater than or equal to 0.15, downloaded on December, 2016). Only diseases associated to at least two genes are considered for the LOOCV procedure. For each disease, each associated gene is removed one-by-one, and considered as the *left-out gene*. The remaining genes associated to this disease are used as seed(s) for the RWR algorithms. It is to note that different subsets of gene-disease associations, from DisGeNET or from Omim, are used to evaluate the different versions of the algorithms, as detailed in the sections 1.4.1 and 1.4.2 below.

All the network nodes are then scored and ranked according to their proximity to the seed(s). The rank of the disease-gene that was left-out in the current run is recorded. This rank is always between one and the total number of scored genes, minus the number of seeds used for the disease under evaluation. Finally, the Cumulative Distribution Function (CDF) of the ranks of the left-out genes is plotted, as in Mordelet and Vert, 2011. It displays the percentage of left-out genes that are ranked within the top k genes. CDFs are used to evaluate and compare the performance of the different algorithms. The plots are focused on the top 60 ranked genes.

1.4.1 Leave-one-out cross-validations on monoplex, aggregated and multiplex networks

For these networks, the seed(s) used in the RWR algorithms are the gene/protein nodes only. To maximize the size of the test dataset, we ran the LOOCV with gene-disease associations extracted from DisGeNET v4.0 (Piñero *et al.*, 2016). The DisGeNET dataset contains 6 565 gene-disease associations, corresponding to 4 148 different diseases.

1.4.2 Leave-one-out cross-validation on heterogeneous and multiplex-heterogeneous networks

For these networks, the seeds used in the RWR algorithms are the gene/protein nodes, but also the disease nodes. Given that the nodes in the disease-disease network are OMIM diseases (Hamosh *et al.*, 2005), it is mandatory to use gene-disease associations from OMIM for the LOOCV. The OMIM dataset contains 4 996 gene-disease associations, corresponding to 4 188 different diseases. It is to note that in order to simulate an unknown gene-disease association, we also removed the bipartite association linking the left-out gene and the disease of the current run. Doing so, we avoid the artificial top ranking of the left-out genes.

2 Supplementary Results

2.1 Random walk with restart on random networks

To compare the performances of the RWR algorithm on real biological networks with random networks, we generated randomized versions of the 3 monoplex networks (PPI, Pathways and Co-expression), keeping the same degree distribution for each network (i.e., the number of nodes with a given degree remains identical, but the node names are randomized). The process was repeated ten times for each monoplex network. We then carried out the Leave-One-Out-Cross-Validation (LOOCV), measuring the ability of the RWR algorithm to retrieve known gene-disease associations. We selected the best performer among the 10 randomized networks.

We can observe that the RWR algorithm is not able to retrieve known gene-disease associations from the randomized networks (Figure S-2). In addition, the Cumulative Distribution Functions (CDF) of the rank position retrieved for each tested gene by LOOCV are significantly different between the monoplex networks and their randomized version (Kolmogorov–Smirnov tests, p-values $< 2.2 \times 10^{-16}$).

It is also interesting to note that the performances on the different random networks are very similar whilst their topological features are quite different. This further emphasizes that the ability of networks to allow retrieving known gene-disease associations depends on the biological content rather than on their topological features.

2.2 Random walk with restart on Pathway-Disease and Co-expression-Disease heterogeneous networks

In the main text, we tested a heterogeneous graph built by connecting the PPI network and the disease-disease similarity network, because we intent to compare our method to the existing RWR-H built by Li and Patra (2010). In order to check the performances of the algorithm on other heterogeneous networks, we also constructed a Pathway-Disease heterogeneous network by integrating the disease-disease similarity network with the Pathways network, and a Co-expression-Disease heterogeneous network in a similar way. We ran the LOOCV to test the performances of the RWR-H in retrieving known-gene disease associations from these heterogeneous networks. For the Pathway-Disease heterogeneous network, we used a test set of diseases related to at least two genes in the set of 10 534 genes/proteins present in the Pathways network. This dataset contains 245 different diseases and 840 gene-disease associations. Similarly, we created a test set containing 157 different diseases and 536 gene-disease associations for the Co-expression-Disease heterogeneous network. The performances of RWR-H strongly depend on the monoplex network used to construct the heterogeneous network, as expected from the results of RWR on monoplex networks (Figure S-3). Overall, random walk with restart on Multiplex-heterogeneous networks (RWR-MH) obtains always a remarkable better performance than the other approaches.

2.3 Effect of parameters on the RWR-MH

We checked the influence of the parameters involved in the RWR-MH algorithm, using the LOOCV strategy. In this case, we created the test set with diseases associated to at least two genes in the total of 17 559 nodes corresponding to the union of the nodes of the three networks. The test set contains 276 diseases and 1 101 gene-disease associations.

In the applications of the RWR algorithm described previously, the restart parameter was set as $r = 0.7$, as in earlier publications (Li and Patra, 2010; Li and Li, 2012; Zhao *et al.*, 2015; Blatti and Sinha, 2016). Changes in this parameter only slightly affect the results (Figure S-4A).

We then studied the effect of the parameters related to the random walks in multiplex networks, δ and τ. The parameter δ quantifies the probability that the particle jumps from the current node to the same node in a different layer, after a non-restart step. If $\delta = 0$ the particle will always stay in the same layer, and if $\delta = 1$ the particle will jump to a different layer at each step. However, we did not observe notable changes with moderate variations in this parameter, as displayed in Figure S-4B. Large variations of δ are needed to alter the RWR-MH performances (Figure S-5A). The parameter $\boldsymbol{\tau}$ controls the probability of restart in the different layers of the multiplex network. Theoretically, this would allow exploiting our knowledge about the performance of the RWR on the monoplex networks. For instance, it could seem reasonable to favor the restart in the Pathway network and to hinder it in the Co-expression network. However, Fig-

ure S-4C does not show notable differences in the performances of the LOOCV associated to variations of this parameter. Even with radical variations of this parameter, the performances do not decrease drastically (Figure S-5B). It is to note that the particle continues exploring the different network layers thanks to the jumps, and can still leverage the combined biological information, even if it does not restart in the seeds of one of the layers.

The parameters used for on heterogeneous networks are λ and η. The parameter λ quantifies the probability of jumping between the multiplex and the disease-disease similarity network, using the bipartite gene-disease associations. The larger the value of λ, the higher the probability of jumping. If $\lambda = 0$, the particle does not exploit the bipartite associations between the disease-disease similarity network and the multiplex network. Contrarily, if $\lambda = 1$, the bipartite gene-disease associations dominate the walks, and the particle is not allowed to explore the topology of each individual network. Moderate variations in this parameter show only small changes in the performances (Figure S-4D). However, the RWR-MH performance decreases when the value of λ tends to zero or one (Figure S-5C). The parameter η quantifies the probability of restart in the multiplex or in the disease-disease similarity network. If $\eta = 0$, the particle will always restart in the multiplex network. In this case, variations in the parameter slightly influence the performances of the algorithm (Figure S-4E and Figure S-5D). Here also, the particle can still explore both the multiplex and heterogeneous network thanks to the jumps, even if it does not restart on the seeds of one of the networks.

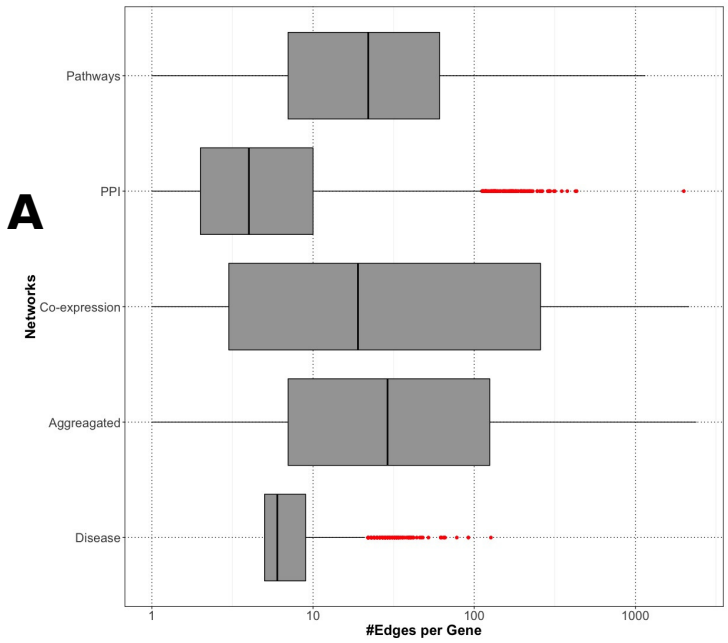
Overall, the RWR-MH is a very robust algorithm since moderate variations in the parameters do not lead to large variations in the ranking performances.

3 Supplementary Tables and Figures

Table S-1: Networks used in this study, number of nodes, edges and network densities.

Network	Number of nodes	Number of edges	Density
Pathways	10 534	254 766	4.59×10^{-3}
PPI	12 621	66 971	8.41×10^{-4}
Co-expression	10 458	1 337 347	2.45×10^{-2}
Aggregated (unique)	17 559	1 659 084	1.08×10^{-2}
Disease-disease similarity	6 947	28 246	1.17×10^{-3}

Figure S-1: Global features of the biological networks used in this study: **A)** Degree distributions, **B)** Venn Diagram displaying the network overlapping nodes and **C)** Venn Diagram displaying the network overlapping edges



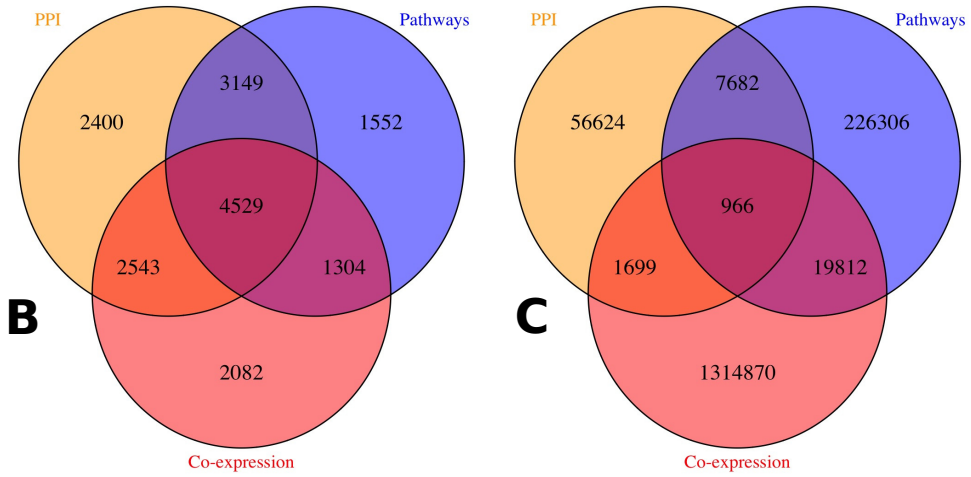


Figure S-2: Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with classical RWR algorithm applied to monoplex networks, and to randomized versions of the monoplex networks. For the randomized network, the best result over 10 randomizations is displayed.

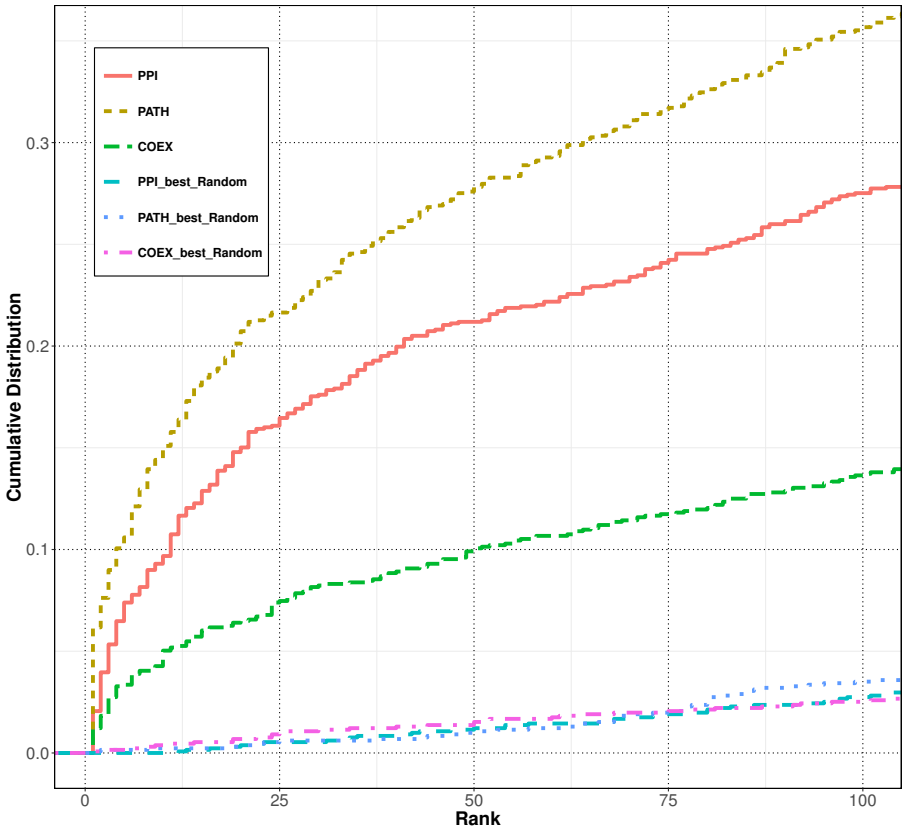


Figure S-3: Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms: **A)** classical RWR on monoplex Pathway network, and RWR-H on the Pathway-Disease heterogeneous network, and **B)** classical RWR on monoplex Co-expression network, and RWR-H on Co-expression-Disease heterogeneous network and, in both cases, RWR-M and RWR-MH on the multiplex and the multiplex-heterogeneous networks.

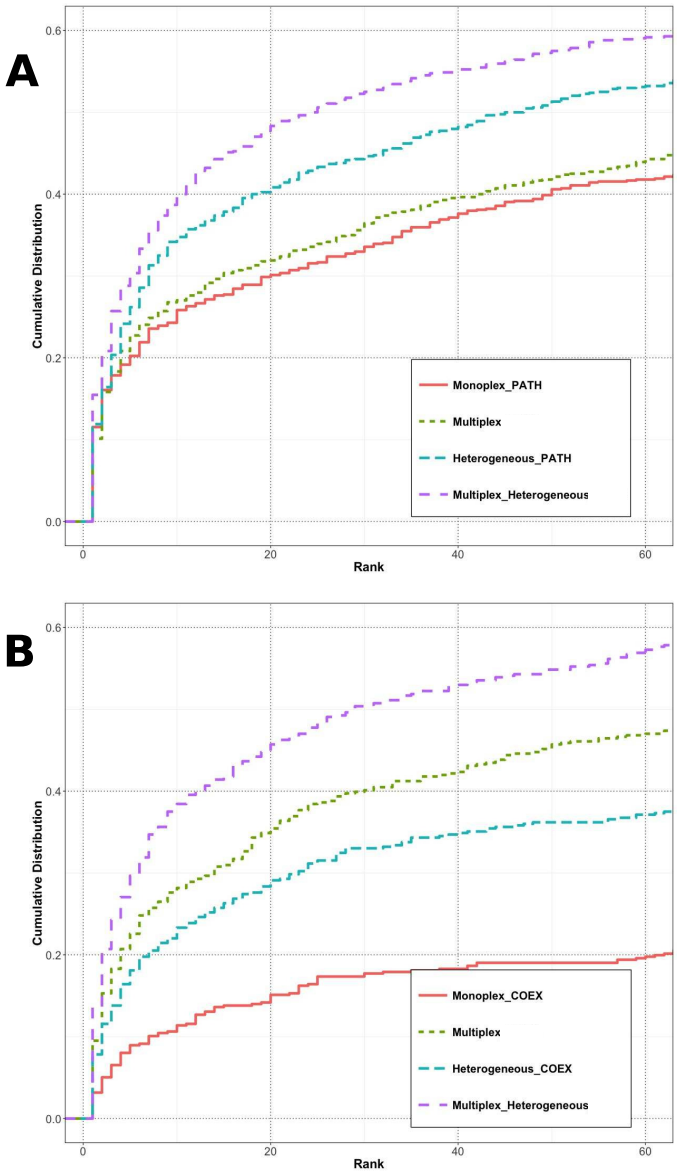


Figure S-4: Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV when running RWR-MH with moderate variations of the parameters. When one parameter changes, the other parameters remain with their default values. Variations are tested in: **A)** parameter r, **B)** parameter δ, **C)** parameter τ for (PPI, Pathway, Co-expression) Layers , **D)** parameter λ and **E)** parameter η.

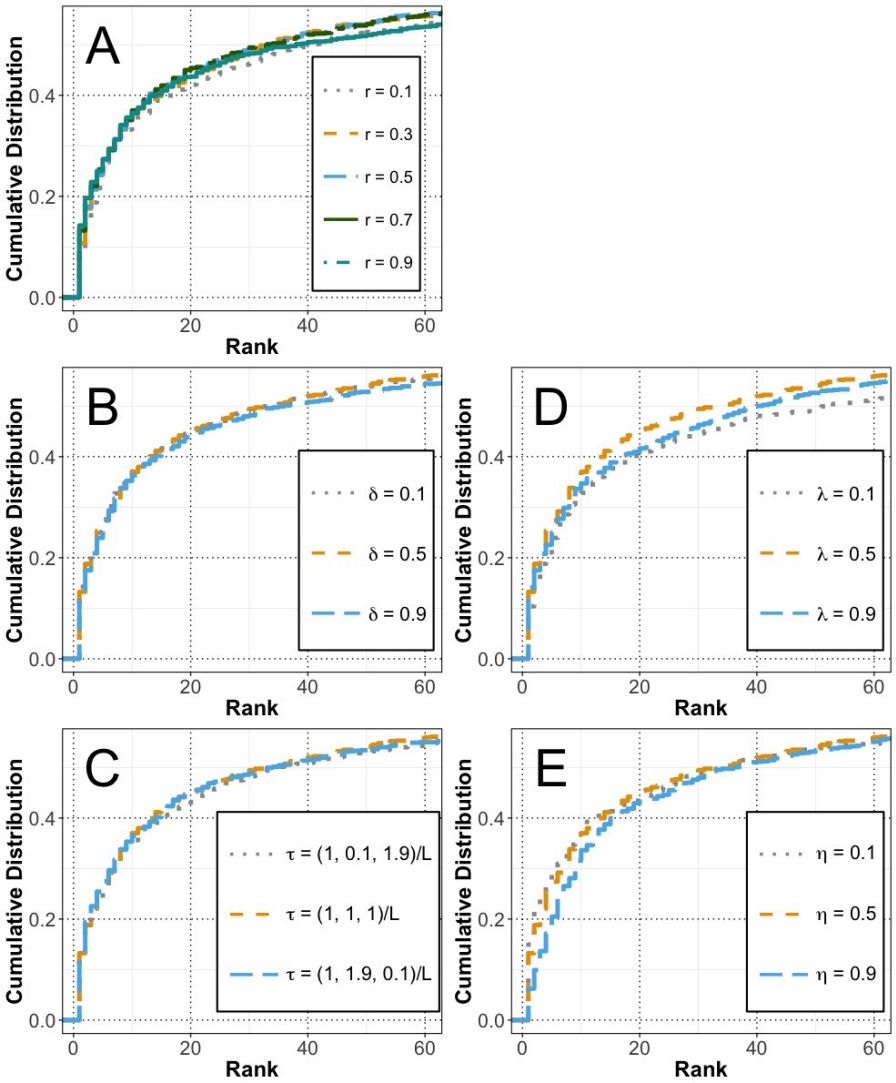


Figure S-5: Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV when running RWR-MH with large variations of the parameters. When one parameter changes, the other parameters remain with their default values. Variations are tested in: **A)** parameter δ, **B)** parameter τ for (PPI, Pathway, Co-expression) Layers , **C)** parameter λ and **D)** parameter η.

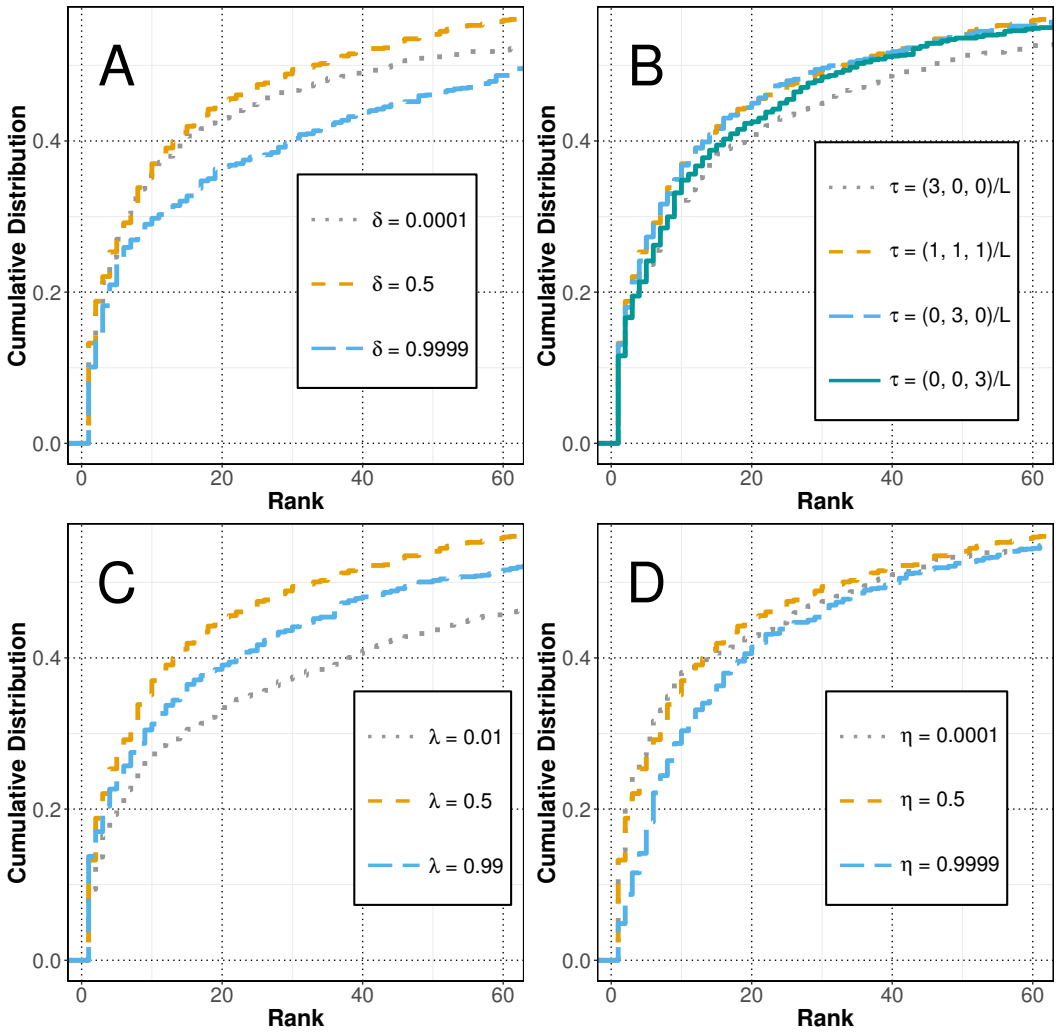


Figure S-6: Network representation of the top 25 ranked genes and diseases when RWR-MH is executed using SS as seed disease and *PIK3R1* as seed gene (yellow nodes). Grey elliptical nodes are diseases; Turquoise rectangles are genes/proteins. Black edges are bipartite gene-disease associations from OMIM (Hamosh *et al.*, 2005); Grey edges are the similarity links in the disease-disease network; Blue edges are PPI interactions; Yellow edges are co-expression relationships; Red edges are pathway interactions. It is to note that results are represented as an aggregated network only for visualization purposes.

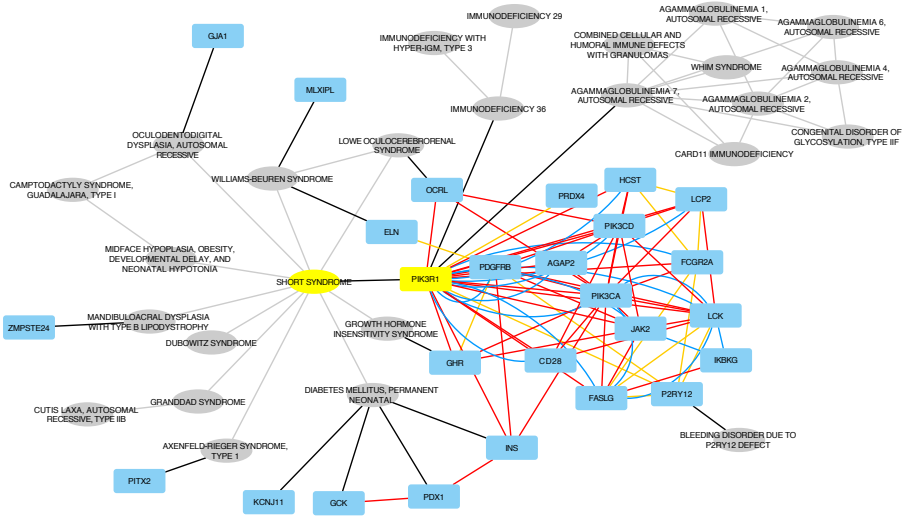


Table S-2: RWR-MH Scores of the Top 25 genes and Top 25 diseases retrieved when the RWR-MH algorithm is executed using the disease node "Wiedemann-Rautenstrauch neonatal progeriod Syndrome" as seed.

Top 25 Genes		Top 25 Diseases		
Gene Symbol	Score	MIM Code	Disease Name	Score
FIG4	1.19×10^{-3}	194190	WOLF-HIRSCHHORN SYNDROME	2.22×10^{-2}
INSR	1.18×10^{-3}	607872	CHROMOSOME 1P36 DELETION SYNDROME	2.21×10^{-2}
RNF113A	1.15×10^{-3}	300953	TRICHOTHIODYSTROPHY 5, NONPHOTOSENSITIVE	2.17×10^{-2}
LMNA	1.10×10^{-3}	616489	GROWTH RESTRICTION, SEVERE, WITH DISTINCTIVE FACIES	2.16×10^{-2}
IGF2	1.10×10^{-3}	216340	YUNIS-VARON SYNDROME	2.15×10^{-2}
FGFRL1	2.48×10^{-4}	246200	DONOHUE SYNDROME	2.14×10^{-2}
LETM1	2.40×10^{-4}	221790	DERMATOLEUKODYSTROPHY	2.13×10^{-2}
CPLX1	2.36×10^{-4}	248770	MARFANOID MENTAL RETARDATION SYNDROME, AUTOSOMAL	2.13×10^{-2}
CTBP1	2.28×10^{-4}	248370	MANDIBULOACRAL DYSPLASIA WITH TYPE A LIPODYSTROPHY	2.13×10^{-2}
WHSC1	2.23×10^{-4}	251240	MICROCEPHALY WITH CHEMOTACTIC DEFECT AND TRANSIENT HYPOGAMMAGLOBULINEMIA	2.12×10^{-2}
MBTPS2	7.38×10^{-5}	210720	MICROCEPHALIC OSTEODYSPLASTIC PRIMORDIAL DWARFISM, TYPE II	1.34×10^{-3}
SOX10	7.19×10^{-5}	223370	DUBOWITZ SYNDROME	1.33×10^{-3}
PCNT	6.76×10^{-5}	210600	SECKEL SYNDROME 1	1.32×10^{-3}
ATR	6.70×10^{-5}	216411	COCKAYNE SYNDROME, TYPE III	1.31×10^{-3}
B4GALT7	5.88×10^{-5}	300918	PALMOPLANTAR KERATODERMA, MUTILATING, WITH PERIORIFICIAL KERATOTIC PLAQUES, X-LINKED	1.30×10^{-3}
INS	5.59×10^{-5}	609136	PERIPHERAL DEMYELINATING NEUROPATHY, CENTRAL DYSMYELINATION, WAARDENBURGSYNDROME, AND HIRSCHSPRUNG DISEASE	1.30×10^{-3}
PTDSS1	3.80×10^{-5}	311450	PALLISTER W SYNDROME	1.29×10^{-3}
ERCC6	3.38×10^{-5}	169170	PATTERSON PSEUDOLEPRECHAUNISM SYNDROME	1.29×10^{-3}
CDC6	3.37×10^{-5}	211910	CAMPTODACTYLY SYNDROME, GUADALAJARA, TYPE I	1.17×10^{-3}
ZMPSTE24	3.36×10^{-5}	227330	FACIODIGITOGENITAL SYNDROME, AUTOSOMAL RECESSIVE	1.11×10^{-3}
ERCC8	3.19×10^{-5}	148050	KBG SYNDROME	1.11×10^{-3}
RPS6KA3	3.16×10^{-5}	130070	EHLERS-DANLOS SYNDROME WITH SHORT STATURE AND LIMB ANOMALIES	1.10×10^{-3}
POLD1	2.94×10^{-5}	259690	OSTEOPENIA AND SPARSE HAIR	1.08×10^{-3}
PARN	2.84×10^{-5}	616214	HYPERPROINSULINEMIA	1.03×10^{-3}
BANF1	2.83×10^{-5}	151050	LENZ-MAJEWSKI HYPEROSTOTIC DWARFISM	7.36×10^{-4}

Table S-3: RWR-MH scores for the Top 25 genes and Top 25 diseases when the SHORT Syndrome and its causative gene, *PIK3R1*, are used as seeds.

Top 25 Genes		Top 25 Diseases		
Gene Symbol	Score	MIM Code	Disease Name	Score
PRDX4	9.83×10^{-4}	615214	AGAMMAGLOBULINEMIA 7, AUTOSOMAL RECESSIVE	2.10×10^{-2}
OCRL	3.10×10^{-4}	616005	IMMUNODEFICIENCY 36	2.11×10^{-2}
PITX2	3.08×10^{-4}	194050	WILLIAMS-BEUREN SYNDROME	5.78×10^{-3}
ZMPSTE24	3.05×10^{-4}	309000	LOWE OCULOCEREBRORENAL SYNDROME	5.70×10^{-3}
GHR	3.04×10^{-4}	608612	MANDIBULOACRAL DYSPLASIA WITH TYPE B LIPODYSTROPHY	5.68×10^{-3}
GJA1	2.93×10^{-4}	262500	GROWTH HORMONE INSENSITIVITY SYNDROME	5.67×10^{-3}
MLXIPL	1.55×10^{-4}	138920	GRANDDAD SYNDROME	5.65×10^{-3}
ELN	1.49×10^{-4}	180500	AXENFELD-RIEGER SYNDROME, TYPE 1	5.64×10^{-3}
P2RY12	1.34×10^{-4}	223370	DUBOWITZ SYNDROME	5.64×10^{-3}
INS	1.01×10^{-4}	606176	DIABETES MELLITUS, PERMANENT NEONATAL	5.64×10^{-3}
KCNJ11	8.07×10^{-5}	257850	OCULODENTODIGITAL DYSPLASIA, AUTOSOMAL RECESSIVE	5.63×10^{-3}
PDX1	7.65×10^{-5}	608624	MIDFACE HYPOPLASIA, OBESITY, DEVELOPMENTAL DELAY, AND NEONATAL HYPOTONIA	5.62×10^{-3}
GCK	7.32×10^{-5}	609821	BLEEDING DISORDER DUE TO P2RY12 DEFECT	8.01×10^{-4}
HCST	7.13×10^{-5}	613500	GAMMAGLOBULINEMIA 2, AUTOSOMAL RECESSIVE	4.46×10^{-4}
PDGFRB	5.61×10^{-5}	613502	AGAMMAGLOBULINEMIA 4, AUTOSOMAL RECESSIVE	4.40×10^{-4}
PIK3CD	5.28×10^{-5}	612692	AGAMMAGLOBULINEMIA 6, AUTOSOMAL RECESSIVE	4.36×10^{-4}
PIK3CA	5.07×10^{-5}	233650	COMBINED CELLULAR AND HUMORAL IMMUNE DEFECTS WITH GRANULOMAS	4.27×10^{-4}
AGAP2	4.99×10^{-5}	603585	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE IIF	4.20×10^{-4}
FCGR2A	4.85×10^{-5}	193670	WHIM SYNDROME	4.18×10^{-4}
CD28	4.69×10^{-5}	601495	AGAMMAGLOBULINEMIA 1, AUTOSOMAL RECESSIVE	4.17×10^{-4}
JAK2	4.42×10^{-5}	615206	CARD11 IMMUNODEFICIENCY	4.13×10^{-4}
LCK	4.29×10^{-5}	211910	CAMPTODACTYLY SYNDROME, GUADALAJARA, TYPE I	3.70×10^{-4}
IKBKG	4.22×10^{-5}	612940	CUTIS LAXA, AUTOSOMAL RECESSIVE, TYPE IIB	3.62×10^{-4}
FASLG	4.21×10^{-5}	606843	IMMUNODEFICIENCY WITH HYPER-IGM, TYPE 3	3.60×10^{-4}
LCP2	3.98×10^{-5}	614890	IMMUNODEFICIENCY 29	3.45×10^{-4}

4 Supplementary Discussion

We extended here the RWR algorithm by designing the RWR-M and RWR-MH algorithms able to leverage Multiplex and Multiplex-Heterogeneous networks. We performed the evaluations with a Leave-One-Out Cross Validation (LOOCV) strategy, using a cumulative distribution function (CDF) to display the results. As compared to a more classical Receiver Operating Curve (AUC), the CDF ranks all the nodes in the networks, thereby resulting in a more general validation Mordelet and Vert, 2011. It is to note that our CDF approach also allows us evaluating the absolute performances of the different algorithms. For instance, the average ranking position of the 880 left-out genes (section 3.3, Figure 4 of the manuscript) in the Monoplex PPI is 3 286, in the Multiplex is 2 025, in the PPI-Heterogeneous is 1 864, and in the Multiplex-Heterogeneous is 1 237. The RWR-MH on a Multiplex-Heterogeneous network obtains the best absolute performance. In addition, even if the average ranking is 1 237, we have to consider that it's 1 237 over 12 621 ranked nodes (all network proteins), thereby being in the top 10% of the ranked proteins. However, more importantly, our approach is able to estimate the percentage of left-out genes ranked within the top x ranked genes. Indeed, guilt-by-association algorithms are usually used either i) to focus on the top x ranked genes for further study or ii) to prioritize within a reduced list of candidate genes. For instance, previous efforts used networks to rank a small candidate list of genes extracted from linkage analysis (Köhler *et al.*, 2008). This is why the LOOCV procedures described in previous literature were creating artificial linkage intervals containing a reduced list of the 100 closest genes on the chromosome from the disease gene (Köhler *et al.*, 2008; Li and Patra, 2010; Li and Li, 2012; Zhao *et al.*, 2015). In this context, even if all network nodes are ranked by our CDF function, we focused our analyses on the top 60 ranked nodes.

It is to note that the global curves of the LOOCV CDF do not change significantly when moderate variations of the parameters are applied to the RWR-MH algorithm (Supplementary Figures S-4). The performance of the RWR-MH decreases notably when radical variations are introduced on the parameters lambda and delta (Supplementary Figures S-5). On the other hand, RWR-MH performance is just slightly affected when large changes are introduced solely on the parameters tau and eta (Supplementary Figures S-5). This is due to the fact that the particle keeps exploring the multiplex-heterogeneous network thanks to the jumps, even if it does not restart on the seeds of one of the layers. However, it is important to note that, even if the overall ranking of the nodes in the LOOCV are stable, a focused analysis and network representation of the top 25 ranked genes and diseases in a real-case applications would reveal variations, even with moderate variations of the parameters. In these applied cases, changes in parameters can be used to tune the output. For instance, the parameter $\boldsymbol{\tau}$ would allow giving more emphasis on some input network layers, based on prior knowledge related to their biological relevance.

Random walks with restart in biology have been applied to predict disease-associated genes (Köhler *et al.*, 2008; Li and Patra, 2010; Li and Li, 2012; Zhao

et al., 2015; Xie *et al.*, 2015), but also to predict drug-target interactions (Chen *et al.*, 2012; Liu *et al.*, 2016) and adverse drug reactions (Chen *et al.*, 2016), and to identify clusters from PPI Networks (Macropol *et al.*, 2009). Smedley *et al.*, 2014, 2015 developed Exomiser, where RWR is applied to prioritize genes and variants in the context of whole-exome sequencing. We applied here our advanced version of the random walk with restart algorithm, RWR-MH, to two real-case biological examples. In the first one, we predicted candidate genes that could be associated to the WSR syndrome, whose responsible gene(s) remain to be described. We hereby demonstrate the usefulness of the approach to study disease etiology and help diagnose patients. The next step will be to validate these predictions, for instance using exome-sequencing data. We also applied the RWR-MH algorithm to study the network vicinity of a disease, the SHORT syndrome, and its associated gene, *PIK3R1*. We show that the disease is sharing phenotype with other syndromes, which are caused by genes in the neighborhood of *PIK3R1* when multiple interaction types are considered. This is an additional example of the fact that mutations in genes participating to the same pathway, or more generally biological processes, lead to diseases with similar phenotypes (Oti *et al.*, 2006).

The main underlying hypothesis of the work presented here is that the integration of multiple interaction sources, each having its own features and biases, will improve the results of the random walks by providing complementary data. For instance, in the application of the RWR-MH to the WRS syndrome, we retrieved as top candidates the *LMNA* and *ZMPSTE24* genes. The *ZMPSTE24* gene codes a peptidase acting during the post-translation modifications of the prelamin A, coded by *LMNA*, to undergo the complete maturation to lamin A. It is interesting to note that the direct interaction between the products of *LMNA* and *ZMPSTE24* is missing in the databases we used to construct the multiplex network. However, the *ZMPSTE24* node is retrieved through different trajectories in the random walk. Hence, the combination of multiple network sources in this case allow completing missing interaction data.

The highly connected nodes, called hubs, can be genes or proteins highly connected and central in the cells, but can also result from biased biological experiments studying "fashion" proteins, such as *TP53* in cancer or *APP* in Alzheimer. RWR algorithms and other network propagation algorithms are biased towards highly connected proteins, as demonstrated by Erten *et al.*, 2011. In this context, poorly-connected and unwell-known nodes, which are also potentially relevant for diseases, are more complicated to find than highly-connected and well-known proteins. To address this issue, biased random walks have been developed to favor the walk of the particle according to network topological features (Battiston *et al.*, 2016). In the simplest case, the transition probability depends on the degree of the neighbors of the current node: the walk of the particle can be tuned towards less connected nodes (Bonaventura *et al.*, 2014). Such a degree-biased random walk could be applied to the RWR-MH algorithm in the future.

This is why overall the combination of different interaction sources is more informative. For instance, PPI interactions might have only sparse number of

interaction for membrane proteins, that are difficult to detect from Yeast-2-hybrid experiments. Co-expression interaction are functional relationships but many of them might not correspond to physical interactions. It is much larger and much noisy than the others. Pathway interaction will focus on well-known proteins and dismiss information about less studied proteins. All these biases are then different, and the combination of the different sources, each considered as different realizations of the true underlying comprehensive biological network, will offer the best possible view on gene and proteins functional relationships.

In addition, for the sake of simplicity, all the networks considered in this study are unweighted. Nevertheless, the extension to weighted networks is straightforward, as pointed out in methods. The use of weighted networks could improve the prioritization results because we can assign larger transition probabilities to the most confident interactions or to the more similar diseases. For instance, STRING database stores scored protein-protein interactions indicating its confidence based on the evidences (Szklarczyk *et al.*, 2015). The edges in our Co-expression network are established based on threshold imposed on the value of the computed correlation coefficient. This coefficient can be included into the Co-expression network to favor the transitions between the proteins whose expressions are more correlated. In addition, we built the disease-disease similarity network according to the similarity scores between every pair of diseases. This score could be introduced into the corresponding edges.

References

Battiston, F. *et al.* (2016). Efficient exploration of multiplex networks. *New Journal of Physics*, **18**(4), 043035.

Blatti, C. and Sinha, S. (2016). Characterizing Gene Sets using Discriminative Random Walks with Restart on Heterogeneous Biological Networks. *Bioinformatics*, **32**(March), 1–9.

Bonaventura, M. *et al.* (2014). Characteristic times of biased random walks on complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **89**(1), 1–18.

Chen, X. *et al.* (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, **8**(7), 1970.

Chen, X. *et al.* (2016). Large-scale identification of adverse drug reaction-related proteins through a random walk model. *Scientific Reports*, **6**(August), 36325.

Del-Toro, N. *et al.* (2013). A new reference implementation of the PSICQUIC web service. *Nucleic acids research*, **41**(Web Server issue), 601–606.

Erten, S. *et al.* (2011). DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData mining*, **4**(1), 19.

Greene, D. *et al.* (2016). Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. *The American Journal of Human Genetics*, pages 1–10.

Hamosh, A. *et al.* (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**(DATABASE ISS.), 514–517.

Köhler, S. *et al.* (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, **42**(D1), 966–974.

Köhler, S. *et al.* (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *AJHG*, **82**(April), 949–958.

Li, Y. and Li, J. (2012). Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC genomics*, **13 Suppl 7**(Suppl 7), S27.

Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**(9), 1219–1224.

Liu, H. *et al.* (2016). Screening lifespan-extending drugs in Caenorhabditis elegans via label propagation on drug-protein networks. *BMC Systems Biology*, **10**(Suppl 4).

Macropol, K. *et al.* (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, **10**(1), 283.

Mordelet, F. and Vert, J.-P. (2011). ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC bioinformatics*, **12**(1), 389.

Oti, M. *et al.* (2006). Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, **43**(8), 691–8.

Piñero, J. *et al.* (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, page gkw943.

Resnik, P. (1999). Semantic Similarity i n a T axonomy: An Information-Based Measure and its Application to Problems of Ambiguity i n Natural Language. *Journal of Artiicial Intelligence Research Submitted*, **11**(3398), 95–130.

Rolland, T. *et al.* (2014). A proteome-scale map of the human interactome network. *Cell*, **159**(5), 1213–1226.

Sales, G. *et al.* (2012). Graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**(1), 20.

Smedley, D. *et al.* (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**(22), 3215–3222.

Smedley, D. *et al.* (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*, **10**(12), 2004–2015.

Szklarczyk, D. *et al.* (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**(D1), D447–D452.

Uhlen, M. *et al.* (2015). Tissue-based map of the human proteome. *Science*, **347**(6220), 1260419–1260419.

Westbury, S. K. *et al.* (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, **7**, 36.

Xie, M. *et al.* (2015). Network-based phenome-genome association prediction by bi-random walk. *PLoS ONE*, **10**(5), 1–18.

Zhao, Z. Q. *et al.* (2015). Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Computational Biology and Chemistry*, **57**, 21–28.

A.2 RandomWalkRestartMH: random walk with restart on multiplex and heterogeneous Networks. R package

Random Walk with Restart on Multiplex and Heterogeneous Network

Alberto Valdeolivas Urbelz[*1, 2]

[1]MMG, Marseille Medical Genetics U 1251, Faculte de Medecine, France
[2]ProGeLife, 8 Rue Sainte Barbe 13001, Marseille, France

[*]alvaldeolivas@gmail.com

October 30, 2018

Abstract

This vignette describes how to use the *RandomWalkRestartMH* package to run Random Walk with Restart algorithms on monoplex, multiplex, heterogeneous and multiplex-heterogeneous networks. It is based on the work we presented on the following article:

https://www.biorxiv.org/content/early/2017/08/30/134734

Package

RandomWalkRestartMH 1.2.0

Report issues to alvaldeolivas@gmail.com

Contents

1 Introduction

RandomWalkRestartMH (Random Walk with Restart on Multiplex and Heterogeneous Networks) is an R package built to provide easy access to the Random Walk with Restart (RWR) algorithm on different types of networks: i) Monoplex networks, ii) Multiplex networks, iii) Heterogeneous networks and iv) Multiplex-Heterogeneous networks. It is based on the work we presented in the article: https://www.biorxiv.org/content/early/2017/08/30/134734.

RWR simulates an imaginary particle that starts on a seed(s) node(s) and follows randomly the edges of a network. At each step, there is a restart probability, r, meaning that the particle can come back to the seed(s) [1]. This imaginary particle can explore the following types of networks:

- A monoplex or single network, which contains solely nodes of the same nature. In addition, all the edges belong to the same category.
- A multiplex network, defined as a collection of monoplex networks considered as layers of the multiplex network. In a multiplex network, the different layers share the same set of nodes, but the edges represent relationships of different nature [2]. In this case, the RWR particle can jump from one node to its counterparts on different layers.
- A heterogeneous network, which is composed of two monoplex networks containing nodes of different nature. These different kind of nodes can be connected thanks to bipartite edges, allowing the RWR particle to jump between the two networks.
- A multiplex and heterogeneous network, which is built by linking the nodes in every layer of a multiplex network to nodes of different nature thanks to bipartite edges. The RWR particle can now explore the full multiplex-heterogeneous network.

The user can introduce up to six single networks (monoplex networks) to create a multiplex network. The multiplex network can also be integrated, thanks to bipartite relationships, with a network containing nodes of different nature. Proceeding this way, a network both multiplex and heterogeneous will be generated.

Please note that this first version of the package deals only with undirected and unweighted networks. New functionalities will be included in future updated versions of *RandomWalkRestartMH*.

2 Installation of the *RandomWalkRestartMH* package

First of all, you need a current version of *R* (www.r-project.org). *RandomWalkRestartMH* is a freely available package deposited on http://bioconductor.org/. You can install it by running the following commands on an *R* console:

```
> if (!requireNamespace("BiocManager", quietly=TRUE))
+     install.packages("BiocManager")
> BiocManager::install("RandomWalkRestartMH")
```

3 A Detailed Workflow

In the following paragraphs, we describe how to use the *RandomWalkRestartMH* package to perform RWR on different types of biological networks. Concretely, we use a protein-protein interaction (PPI) network, a pathway network, a disease-disease similarity network and combinations thereof. These networks are obtained as detailed in [3]. The PPI and the Pathway network were reduced by only considering genes/proteins expressed in the adipose tissue, in order to reduce the computation time of this vignette.

The goal in the example presented here is, as described in [3], to find candidate genes potentially associated with diseases by a guilt-by-association approach. This is based on the fact that genes/proteins with similar functions or similar phenotypes tend to lie closer in biological networks. Therefore, the larger the RWR score of a gene, the more likely it is to be functionally related with the seeds.

We focus on a real biological example: the SHORT syndrome (MIM code: 269880) and its causative gene *PIK3R1* as described in [3]. We will see throughout the following paragraphs how the RWR results evolve due to the the integration and exploration of additional networks.

3.1 Random Walk with Restart on a Monoplex Network

RWR has usually been applied within the framework of single PPI networks in bioinformatics[4]. A gene or a set of genes, so-called seed(s), known to be implicated in a concrete function or in a specific disease, are chosen as the starting point(s) of the algorithm. The RWR particle explores the neighbourhood of the seeds and the algorithm computes a score for all the nodes of the network. The larger it is the score of a node, the closer it is to the seed(s).

Let us generate an object of the class *Multiplex*, even if it is a monoplex network, with our PPI network.

```
> library(RandomWalkRestartMH)
> library(igraph)
> data(PPI_Network) # We load the PPI_Network
> ## We create a Multiplex object composed of 1 layer (It's a Monoplex Network)
> ## and we display how it looks like
> PPI_MultiplexObject <- create.multiplex(PPI_Network,Layers_Name=c("PPI"))
> PPI_MultiplexObject

Number of Layers:
[1] 1

Number of Nodes:
[1] 4317

IGRAPH 1573a63 UN-- 4317 18062 --
+ attr: name (v/c), type (e/c)
+ edges from 1573a63 (vertex names):
 [1] AAMP   --VPS52     AAMP   --BHLHE40   AAMP   --GABARAPL2 AAMP   --MAP1LC3B
 [5] VPS52  --TXN2      VPS52  --DDX6      VPS52  --MFAP1     VPS52  --PRKAA1
 [9] VPS52  --LMO4      VPS52  --STX11     VPS52  --KANK2     VPS52  --PPP1R18
[13] VPS52  --TXLNA     VPS52  --KIAA1217  VPS52  --VPS28     VPS52  --ATP6V1D
[17] VPS52  --TPM3      VPS52  --KIF5B     VPS52  --NOP2      VPS52  --RNF41
```

```
[21] VPS52  --WTAP      VPS52  --MAPK3     VPS52  --ZMAT2     VPS52  --VPS51
[25] BHLHE40--AES       BHLHE40--PRKAA1    BHLHE40--CCNK      BHLHE40--RBPMS
[29] BHLHE40--COX5B     BHLHE40--UBE2I     BHLHE40--MAGED1    BHLHE40--PLEKHB2
+ ... omitted several edges
```

To apply the RWR on a monoplex network, we need to compute the adjacency matrix of the network and normalize it by column [4], as follows:

```
> AdjMatrix_PPI <- compute.adjacency.matrix(PPI_MultiplexObject)
> AdjMatrixNorm_PPI <- normalize.multiplex.adjacency(AdjMatrix_PPI)
```

Then, we need to define the seed(s) before running the RWR algorithm on this PPI network. As commented above, we are focusing on the example of the SHORT syndrome. Therefore, we take the *PIK3R1* gene as seed, and we execute RWR.

```
> SeedGene <- c("PIK3R1")
> ## We launch the algorithm with the default parameters (See details on manual)
> RWR_PPI_Results <- Random.Walk.Restart.Multiplex(AdjMatrixNorm_PPI,
+                         PPI_MultiplexObject,SeedGene)
> # We display the results
> RWR_PPI_Results

Top 10 ranked Nodes:
   NodeNames       Score
1       GRB2 0.006845881
2       EGFR 0.006169129
3        CRK 0.005674261
4       ABL1 0.005617041
5        FYN 0.005611086
6      CDC42 0.005594680
7       SHC1 0.005577900
8       CRKL 0.005509182
9    KHDRBS1 0.005443541
10     TYRO3 0.005441887

Seed Nodes used:
[1] "PIK3R1"
```

Finally, we can create a network (an *igraph* object) with the top scored genes. Visualize the top results within their interaction network is always a good idea in order to prioritize genes, since we can have a global view of all the potential candidates. The results are presented in Figure 1.

```
> ## In this case we selected to induce a network with the Top 15 genes.
> TopResults_PPI <-
+     create.multiplexNetwork.topResults(RWR_PPI_Results,PPI_MultiplexObject,
+                                        k=15)
```

```
> par(mar=c(0.1,0.1,0.1,0.1))
> plot(TopResults_PPI, vertex.label.color="black",vertex.frame.color="#ffffff",
+     vertex.size= 20, edge.curved=.2,
+     vertex.color = ifelse(igraph::V(TopResults_PPI)$name == "PIK3R1","yellow",
```

```
+     "#00CCFF"), edge.color="blue",edge.width=0.8)
```

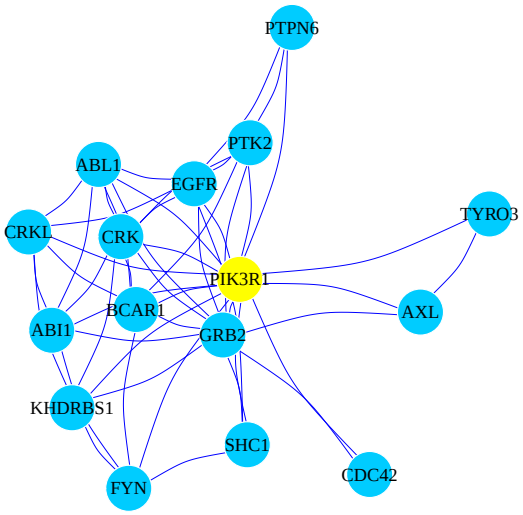


Figure 1: RWR on a monoplex PPI Network
Network representation of the top 15 ranked genes when the RWR algorithm is executed using the *PIK3R1* gene as seed (yellow node). Blue edges represent PPI interactions.

3.2 Random Walk with Restart on a Heterogeneous Network

A RWR on a heterogeneous (RWR-H) biological network was described by [5]. They connected a PPI network with a disease-disease similarity network using known gene-disease associations. In this case, genes and/or diseases can be used as seed nodes for the algorithm. In the following example, we also use a heterogeneous network integrating a PPI and a disease-disease similarity network. However, the procedure to obtain these networks is different to the one proposed in [5], and the details are described in our article [3].

To generate a PPI-disease heterogeneous network object, we load the disease-disease network, and combine it with our previously defined *Multiplex* object containing the PPI network, thanks to the gene-diseases associations obtained from OMIM [6]. A *MultiplexHet* object will be created, even if we are dealing with a monoplex-heterogeneous network.

```
> data(Disease_Network) # We load our disease Network
> ## We load a data frame containing the gene-disease associations.
> ## See ?create.multiplexHet for details about its format
> data(GeneDiseaseRelations)
> ## We keep gene-diseases associations where genes are present in the PPI
> ## network
> GeneDiseaseRelations_PPI <-
+     GeneDiseaseRelations[which(GeneDiseaseRelations$hgnc_symbol %in%
+     PPI_MultiplexObject$Pool_of_Nodes),]
> ## We create the MultiplexHet object.
> PPI_Disease_Net <- create.multiplexHet(PPI_MultiplexObject,
```

```
+     Disease_Network, GeneDiseaseRelations_PPI, c("Disease"))
> ## The results look like that
> PPI_Disease_Net

Number of Layers:
[1] 1

Number of Nodes Multiplex:
[1] 4317

IGRAPH 1573a63 UN-- 4317 18062 --
+ attr: name (v/c), type (e/c)
+ edges from 1573a63 (vertex names):
 [1] AAMP   --VPS52     AAMP   --BHLHE40   AAMP   --GABARAPL2 AAMP   --MAP1LC3B
 [5] VPS52  --TXN2      VPS52  --DDX6      VPS52  --MFAP1     VPS52  --PRKAA1
 [9] VPS52  --LMO4      VPS52  --STX11     VPS52  --KANK2     VPS52  --PPP1R18
[13] VPS52  --TXLNA     VPS52  --KIAA1217  VPS52  --VPS28     VPS52  --ATP6V1D
[17] VPS52  --TPM3      VPS52  --KIF5B     VPS52  --NOP2      VPS52  --RNF41
[21] VPS52  --WTAP      VPS52  --MAPK3     VPS52  --ZMAT2     VPS52  --VPS51
[25] BHLHE40--AES       BHLHE40--PRKAA1    BHLHE40--CCNK      BHLHE40--RBPMS
[29] BHLHE40--COX5B     BHLHE40--UBE2I     BHLHE40--MAGED1    BHLHE40--PLEKHB2
+ ... omitted several edges

Number of Nodes of the second network:
[1] 6947

Second Network
IGRAPH 3eed1b5 UN-- 6947 28246 --
+ attr: name (v/c), type (e/c)
+ edges from 3eed1b5 (vertex names):
 [1] 100050--122470 100050--227330 100050--259200 100050--305400 100050--601803
 [6] 100070--105800 100070--105805 100070--107550 100070--120000 100070--130090
[11] 100070--132900 100070--154750 100070--180300 100070--192310 100070--208060
[16] 100070--210050 100070--219100 100070--252350 100070--277175 100070--300537
[21] 100070--309520 100070--600459 100070--604308 100070--606519 100070--608967
[26] 100070--609192 100070--610168 100070--610380 100070--611788 100070--613780
[31] 100070--613834 100070--614042 100070--614437 100070--614980 100070--610655
[36] 100070--615436 100070--616166 100100--192350 100100--236700 100100--236730
+ ... omitted several edges
```

To apply the RWR-H on a heterogeneous network, we need to compute a matrix that accounts for all the possible transitions of the RWR particle within that network [5].

```
> PPIHetTranMatrix <- compute.transition.matrix(PPI_Disease_Net)
```

Before running RWR-H on this PPI-disease heterogeneous network, we need to define the seed(s). As in the previous paragraph, we take *PIK3R1* as a seed gene. In addition, we can now set the SHORT syndrome itself as a seed disease.

```
> SeedDisease <- c("269880")
> ## We launch the algorithm with the default parameters (See details on manual)
> RWRH_PPI_Disease_Results <-
```

```
+     Random.Walk.Restart.MultiplexHet(PPIHetTranMatrix,
+     PPI_Disease_Net,SeedGene,SeedDisease)
> # We display the results
> RWRH_PPI_Disease_Results

Top 10 ranked Multiplex nodes:
     NodeNames       Score
1479      GRB2 0.001965500
1081      EGFR 0.001754048
797        CRK 0.001636329
603      CDC42 0.001630575
19        ABL1 0.001623304
798       CRKL 0.001605543
1360       FYN 0.001598567
3405      SHC1 0.001597720
2583    PDGFRB 0.001596830
4027     TYRO3 0.001589911

Multiplex Seed Nodes used:
[1] "PIK3R1"
Top 10 ranked Second Network Nodes:
     SecondNet_node       Score
6352         615214 0.020817435
6705         616005 0.020785895
1699         194050 0.005868407
3625         309000 0.005687206
2901         262500 0.005681162
686          138920 0.005655559
2150         223370 0.005655117
4770         608612 0.005649291
4464         606176 0.005641777
1411         180500 0.005639919

Second Network Seed Nodes used:
[1] "269880"
```

Finally, we can create a heterogeneous network (an *igraph* object) with the top scored genes and the top scored diseases. The results are presented in Figure 2.

```
> ## In this case we select to induce a network with the Top 10 genes
> ## and the Top 10 diseases.
> TopResults_PPI_Disease <-
+     create.multiplexHetNetwork.topResults(RWRH_PPI_Disease_Results,
+     PPI_Disease_Net, GeneDiseaseRelations_PPI, k=10)
```

```
> par(mar=c(0.1,0.1,0.1,0.1))
> plot(TopResults_PPI_Disease, vertex.label.color="black",
+     vertex.frame.color="#ffffff",
+     vertex.size= 20, edge.curved=.2,
+     vertex.color = ifelse(V(TopResults_PPI_Disease)$name == "PIK3R1"
+     | V(TopResults_PPI_Disease)$name == "269880","yellow",
+     ifelse(V(TopResults_PPI_Disease)$name %in% PPI_Disease_Net$Pool_of_Nodes,
```

```
+     "#00CCFF","Grey75")),
+     edge.color=ifelse(E(TopResults_PPI_Disease)$type == "PPI","blue",
+         ifelse(E(TopResults_PPI_Disease)$type == "Disease","black","grey50")),
+     edge.width=0.8,
+     edge.lty=ifelse(E(TopResults_PPI_Disease)$type == "bipartiteRelations",
+         2,1),
+     vertex.shape= ifelse(V(TopResults_PPI_Disease)$name %in%
+         PPI_Disease_Net$Pool_of_Nodes,"circle","rectangle"))
```

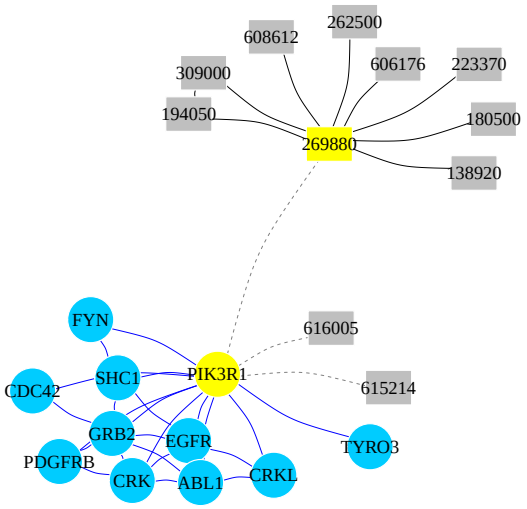


Figure 2: **RWR-H on a heterogeneous PPI-Disease Network**
Network representation of the top 10 ranked genes and the top 10 ranked diseases when the RWR-H algorithm is executed using the *PIK3R1* gene and the SHORT syndrome disease (MIM code: 269880) as seeds (yellow nodes). Circular nodes represent genes and rectangular nodes show diseases. Blue edges are PPI interactions and black edges are similarity links between diseases. Dashed edges are the bipartite gene-disease associations.

3.3 Random Walk with Restart on a Multiplex Network

Some limitations can arise when single networks are used to represent and describe systems whose entities can interact through more than one type of connections [2]. This is the case of social interactions, transportation networks or biological systems, among others. The Multiplex framework provides an appealing approach to describe these systems, since they are able to integrate this diversity of data while keeping track of the original features and topologies of the different sources.

Consequently, algorithms able to exploit the information stored on multiplex networks should improve the results provided by methods operating on single networks. In this context, we extended the random walk with restart algorithm to multiplex networks (RWR-M) [3].

In the following example, we create a multiplex network integrated by our PPI network and a network derived from pathway databases [3].

```
> data(Pathway_Network) # We load the Pathway Network
> ## We create a 2-layers Multiplex object
> PPI_PATH_Multiplex <- create.multiplex(PPI_Network,Pathway_Network,
+                         Layers_Name=c("PPI","PATH"))
> PPI_PATH_Multiplex

Number of Layers:
[1] 2

Number of Nodes:
[1] 4899

IGRAPH ed22f9b UN-- 4899 18062 --
+ attr: name (v/c), type (e/c)
+ edges from ed22f9b (vertex names):
 [1] AAMP   --VPS52     AAMP   --BHLHE40   AAMP   --GABARAPL2 AAMP   --MAP1LC3B
 [5] VPS52  --TXN2      VPS52  --DDX6      VPS52  --MFAP1     VPS52  --PRKAA1
 [9] VPS52  --LMO4      VPS52  --STX11     VPS52  --KANK2     VPS52  --PPP1R18
[13] VPS52  --TXLNA     VPS52  --KIAA1217  VPS52  --VPS28     VPS52  --ATP6V1D
[17] VPS52  --TPM3      VPS52  --KIF5B     VPS52  --NOP2      VPS52  --RNF41
[21] VPS52  --WTAP      VPS52  --MAPK3     VPS52  --ZMAT2     VPS52  --VPS51
[25] BHLHE40--AES       BHLHE40--PRKAA1    BHLHE40--CCNK      BHLHE40--RBPMS
[29] BHLHE40--COX5B     BHLHE40--UBE2I     BHLHE40--MAGED1    BHLHE40--PLEKHB2
+ ... omitted several edges

IGRAPH 71919ca UN-- 4899 62602 --
+ attr: name (v/c), type (e/c)
+ edges from 71919ca (vertex names):
 [1] BANF1--PSIP1   BANF1--HMGA1   BANF1--PPP2R1A BANF1--PPP2CA  BANF1--KPNA1
 [6] BANF1--TPR     BANF1--NUP62   BANF1--NUP153  BANF1--RANBP2  BANF1--NUP54
[11] BANF1--POM121  BANF1--NUP85   BANF1--PPP2R2A BANF1--EMD     BANF1--LEMD2
[16] BANF1--ANKLE2  PSIP1--HMGA1   PSIP1--KPNA1   PSIP1--TPR     PSIP1--NUP62
[21] PSIP1--NUP153  PSIP1--RANBP2  PSIP1--NUP54   PSIP1--POM121  PSIP1--NUP85
[26] HMGA1--TP53    HMGA1--KPNA1   HMGA1--TPR     HMGA1--NUP62   HMGA1--NUP153
[31] HMGA1--RANBP2  HMGA1--NUP54   HMGA1--POM121  HMGA1--NUP85   HMGA1--MYC
[36] HMGA1--RB1     HMGA1--MAX     HMGA1--RPS6KB1 XRCC6--XRCC5   XRCC6--IRF3
+ ... omitted several edges
```

Afterwards, as in the monoplex case, we have to compute and normalize the adjacency matrix of the multiplex network.

```
> AdjMatrix_PPI_PATH <- compute.adjacency.matrix(PPI_PATH_Multiplex)
> AdjMatrixNorm_PPI_PATH <- normalize.multiplex.adjacency(AdjMatrix_PPI_PATH)
```

Then, we set again as seed the *PIK3R1* gene and we perform RWR-M on this new multiplex network.

```
> ## We launch the algorithm with the default parameters (See details on manual)
> RWR_PPI_PATH_Results <- Random.Walk.Restart.Multiplex(AdjMatrixNorm_PPI_PATH,
+                         PPI_PATH_Multiplex,SeedGene)
> # We display the results
> RWR_PPI_PATH_Results
```

```
Top 10 ranked Nodes:
   NodeNames       Score
1       GRB2 0.001662893
2        FYN 0.001517786
3       HCST 0.001506594
4       EGFR 0.001494459
5       SHC1 0.001492456
6       PTK2 0.001430282
7       JAK2 0.001401867
8       HRAS 0.001396854
9       CRKL 0.001389391
10    PDGFRB 0.001369533

Seed Nodes used:
[1] "PIK3R1"
```

Finally, we can create a multiplex network (an *igraph* object) with the top scored genes. The results are presented in Figure 3.

```
> ## In this case we select to induce a multiplex network with the Top 15 genes.
> TopResults_PPI_PATH <-
+     create.multiplexNetwork.topResults(RWR_PPI_PATH_Results,
+                                        PPI_PATH_Multiplex, k=15)
```

```
> par(mar=c(0.1,0.1,0.1,0.1))
> plot(TopResults_PPI_PATH, vertex.label.color="black",
+     vertex.frame.color="#ffffff", vertex.size= 20,
+     edge.curved= ifelse(E(TopResults_PPI_PATH)$type == "PPI",
+                         0.4,0),
+     vertex.color = ifelse(igraph::V(TopResults_PPI_PATH)$name == "PIK3R1",
+                         "yellow","#00CCFF"),edge.width=0.8,
+     edge.color=ifelse(E(TopResults_PPI_PATH)$type == "PPI",
+                           "blue","red"))
```

3.4 Random Walk with Restart on a Multiplex-Heterogeneous Network

RWR-H and RWR-M remarkably improve the results obtained by classical RWR on monoplex networks, as we demonstrated in the particular case of retrieving known gene-disease associations [3]. Therefore, an algorithm able to execute a random walk with restart on both, multiplex and heterogeneous networks, is expected to achieve an even better performance. We extended our RWR-M approach to heterogeneous networks, defining a random walk with restart on multiplex-heterogeneous networks (RWR-MH) [3].

Let us integrate all the networks described previously (PPI, Pathways and disease-disease similarity) into a multiplex and heterogeneous network. To do so, we connect genes in both multiplex layers (PPI and Pathways) to the disease network, if a bipartite gene-disease relation exists.

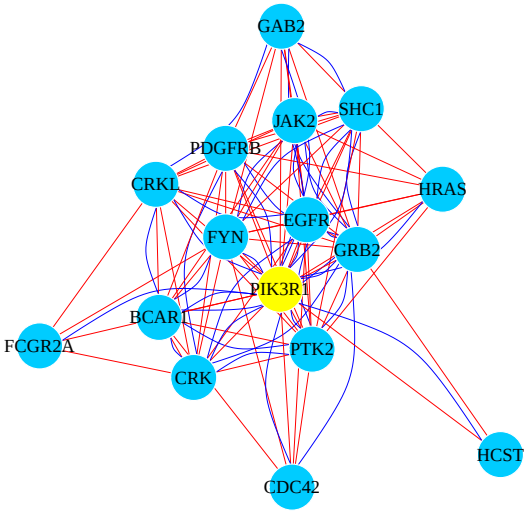


Figure 3: **RWR-M on a multiplex PPI-Pathway Network**
Network representation of the top 15 ranked genes when the RWR-M algorithm is executed using the *PIK3R1* gene (yellow node). Blue curved edges are PPI interactions and red straight edges are Pathways links. All the interactions are aggregated into a monoplex network only for visualization purposes.

```
> ## We keep gene-diseases associations where genes are present in the PPI
> ## or in the pathway network
> GeneDiseaseRelations_PPI_PATH <-
+     GeneDiseaseRelations[which(GeneDiseaseRelations$hgnc_symbol %in%
+     PPI_PATH_Multiplex$Pool_of_Nodes),]
> ## We create the MultiplexHet object.
> PPI_PATH_Disease_Net <- create.multiplexHet(PPI_PATH_Multiplex,
+     Disease_Network, GeneDiseaseRelations_PPI_PATH, c("Disease"))
> ## The results look like that
> PPI_PATH_Disease_Net

Number of Layers:
[1] 2

Number of Nodes Multiplex:
[1] 4899

IGRAPH ed22f9b UN-- 4899 18062 --
+ attr: name (v/c), type (e/c)
+ edges from ed22f9b (vertex names):
 [1] AAMP   --VPS52     AAMP   --BHLHE40   AAMP   --GABARAPL2 AAMP   --MAP1LC3B
 [5] VPS52  --TXN2      VPS52  --DDX6      VPS52  --MFAP1     VPS52  --PRKAA1
 [9] VPS52  --LMO4      VPS52  --STX11     VPS52  --KANK2     VPS52  --PPP1R18
[13] VPS52  --TXLNA     VPS52  --KIAA1217  VPS52  --VPS28     VPS52  --ATP6V1D
[17] VPS52  --TPM3      VPS52  --KIF5B     VPS52  --NOP2      VPS52  --RNF41
[21] VPS52  --WTAP      VPS52  --MAPK3     VPS52  --ZMAT2     VPS52  --VPS51
[25] BHLHE40--AES       BHLHE40--PRKAA1    BHLHE40--CCNK      BHLHE40--RBPMS
[29] BHLHE40--COX5B     BHLHE40--UBE2I     BHLHE40--MAGED1    BHLHE40--PLEKHB2
```

```
+ ... omitted several edges

IGRAPH 71919ca UN-- 4899 62602 --
+ attr: name (v/c), type (e/c)
+ edges from 71919ca (vertex names):
 [1] BANF1--PSIP1   BANF1--HMGA1   BANF1--PPP2R1A BANF1--PPP2CA  BANF1--KPNA1
 [6] BANF1--TPR     BANF1--NUP62   BANF1--NUP153  BANF1--RANBP2  BANF1--NUP54
[11] BANF1--POM121  BANF1--NUP85   BANF1--PPP2R2A BANF1--EMD     BANF1--LEMD2
[16] BANF1--ANKLE2  PSIP1--HMGA1   PSIP1--KPNA1   PSIP1--TPR     PSIP1--NUP62
[21] PSIP1--NUP153  PSIP1--RANBP2  PSIP1--NUP54   PSIP1--POM121  PSIP1--NUP85
[26] HMGA1--TP53    HMGA1--KPNA1   HMGA1--TPR     HMGA1--NUP62   HMGA1--NUP153
[31] HMGA1--RANBP2  HMGA1--NUP54   HMGA1--POM121  HMGA1--NUP85   HMGA1--MYC
[36] HMGA1--RB1     HMGA1--MAX     HMGA1--RPS6KB1 XRCC6--XRCC5   XRCC6--IRF3
+ ... omitted several edges

Number of Nodes of the second network:
[1] 6947

Second Network
IGRAPH 04739c8 UN-- 6947 28246 --
+ attr: name (v/c), type (e/c)
+ edges from 04739c8 (vertex names):
 [1] 100050--122470 100050--227330 100050--259200 100050--305400 100050--601803
 [6] 100070--105800 100070--105805 100070--107550 100070--120000 100070--130090
[11] 100070--132900 100070--154750 100070--180300 100070--192310 100070--208060
[16] 100070--210050 100070--219100 100070--252350 100070--277175 100070--300537
[21] 100070--309520 100070--600459 100070--604308 100070--606519 100070--608967
[26] 100070--609192 100070--610168 100070--610380 100070--611788 100070--613780
[31] 100070--613834 100070--614042 100070--614437 100070--614980 100070--610655
[36] 100070--615436 100070--616166 100100--192350 100100--236700 100100--236730
+ ... omitted several edges
```

To apply the RWR-MH on a multiplex and heterogeneous network, we need to compute a matrix that accounts for all the possible transitions of the RWR particle within this network [3].

```
> PPI_PATH_HetTranMatrix <- compute.transition.matrix(PPI_PATH_Disease_Net)
```

As in the RWR-H situation, we can take as seeds both, the *PIK3R1* gene and the the SHORT syndrome disease.

```
> ## We launch the algorithm with the default parameters (See details on manual)
> RWRH_PPI_PATH_Disease_Results <-
+     Random.Walk.Restart.MultiplexHet(PPI_PATH_HetTranMatrix,
+     PPI_PATH_Disease_Net,SeedGene,SeedDisease)
> # We display the results
> RWRH_PPI_PATH_Disease_Results

Top 10 ranked Multiplex nodes:
     NodeNames        Score
1616       GHR 0.0005153889
4867  ZMPSTE24 0.0004781314
```

```
1706      GRB2 0.0004642946
1625      GJA1 0.0004457777
1781      HCST 0.0004408101
1249      EGFR 0.0004239882
1564       FYN 0.0004193109
3883      SHC1 0.0004161564
2990    PDGFRB 0.0004079884
3361      PTK2 0.0004069179

Multiplex Seed Nodes used:
[1] "PIK3R1"
Top 10 ranked Second Network Nodes:
     SecondNet_node       Score
6352         615214 0.020797695
6705         616005 0.020775842
1699         194050 0.005868202
4770         608612 0.005691321
2901         262500 0.005691027
3625         309000 0.005683228
686          138920 0.005653962
2150         223370 0.005653358
4464         606176 0.005639004
1411         180500 0.005631733

Second Network Seed Nodes used:
[1] "269880"
```

Finally, we can create a multiplex and heterogeneous network (an *igraph* object) with the top scored genes and the top scored diseases. The results are presented in Figure 4.

```
> ## In this case we select to induce a network with the Top 10 genes.
> ## and the Top 10 diseases.
> TopResults_PPI_PATH_Disease <-
+     create.multiplexHetNetwork.topResults(RWRH_PPI_PATH_Disease_Results,
+     PPI_PATH_Disease_Net, GeneDiseaseRelations_PPI_PATH, k=10)
```

```
> par(mar=c(0.1,0.1,0.1,0.1))
> plot(TopResults_PPI_PATH_Disease, vertex.label.color="black",
+     vertex.frame.color="#ffffff",
+     vertex.size= 20,
+     edge.curved=ifelse(E(TopResults_PPI_PATH_Disease)$type == "PATH",
+                     0,0.3),
+     vertex.color = ifelse(V(TopResults_PPI_PATH_Disease)$name == "PIK3R1"
+     | V(TopResults_PPI_Disease)$name == "269880","yellow",
+     ifelse(V(TopResults_PPI_PATH_Disease)$name %in%
+                 PPI_PATH_Disease_Net$Pool_of_Nodes,
+     "#00CCFF","Grey75")),
+     edge.color=ifelse(E(TopResults_PPI_PATH_Disease)$type == "PPI","blue",
+     ifelse(E(TopResults_PPI_PATH_Disease)$type == "PATH","red",
+     ifelse(E(TopResults_PPI_PATH_Disease)$type == "Disease","black","grey50"))),
+     edge.width=0.8,
```

```
+     edge.lty=ifelse(E(TopResults_PPI_PATH_Disease)$type ==
+         "bipartiteRelations", 2,1),
+     vertex.shape= ifelse(V(TopResults_PPI_PATH_Disease)$name %in%
+         PPI_PATH_Disease_Net$Pool_of_Nodes,"circle","rectangle"))
```

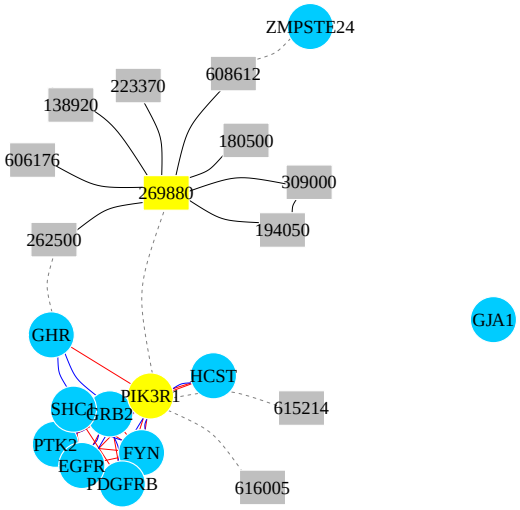


Figure 4: **RWR-MH on a multiplex and heterogeneous network (PPI-Pathway-Disease)**
Network representation of the top 10 ranked genes and the top 10 ranked diseases when the RWR-H algorithm is executed using the *PIK3R1* gene and the SHORT syndrome disease (MIM code: 269880) as seeds (yellow nodes). Circular nodes represent genes and rectangular nodes show diseases. Blue curved edges are PPI interactions and red straight edges are Pathways links. Black edges are similarity links between diseases. Dashed edges are the bipartite gene-disease associations. Multiplex interactions are aggregated into a monoplex network only for visualization purposes.

References

[1] Jia-yu Pan, Hyung Yang, Pinar Duygulu, and Christos Faloutsos. Automatic Multimedia Cross-modal Correlation Discovery. pages 653–658, 2004.

[2] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(3):1–16, 2014. arXiv:1308.3182, doi:10.1103/PhysRevE.89.032804.

[3] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anais Baudot. Random Walk With Restart On Multiplex And Heterogeneous Biological Networks. *bioRxiv*, pages 1–31, 2017. URL: http://biorxiv.org/content/early/2017/05/05/134734?rss=1, doi:10.1101/134734.

[4] Sebastian Kohler, Sebastian Bauer, Denise Horn, and Peter N Robinson. AJHG - Walking the Interactome for Prioritization of Candidate Disease Genes. (April):949–958, 2008. URL: http://www.cell.com/AJHG/abstract/S0002-9297(08)00172-9, doi:10.1016/j.ajhg.2008.02.013.

[5] Yongjin Li and Jagdish C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010. doi:10.1093/bioinformatics/btq108.

[6] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.):514–517, 2005. doi:10.1093/nar/gki033.

A Session info

- R version 3.5.1 Patched (2018-07-12 r74967), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Ubuntu 16.04.5 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.8-bioc/R/lib/libRblas.so`
- LAPACK: `/home/biocbuild/bbs-3.8-bioc/R/lib/libRlapack.so`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: RandomWalkRestartMH 1.2.0, igraph 1.2.2
- Loaded via a namespace (and not attached): BiocGenerics 0.28.0, BiocManager 1.30.3, BiocStyle 2.10.0, MASS 7.3-51, Matrix 1.2-14, Rcpp 0.12.19, Rgraphviz 2.26.0, ape 5.2, backports 1.1.2, compiler 3.5.1, digest 0.6.18, dnet 1.1.4, evaluate 0.12, graph 1.60.0, grid 3.5.1, hexbin 1.27.2, htmltools 0.3.6, knitr 1.20, lattice 0.20-35, magrittr 1.5, nlme 3.1-137, parallel 3.5.1, pkgconfig 2.0.2, rmarkdown 1.10, rprojroot 1.3-2, stats4 3.5.1, supraHex 1.20.0, tools 3.5.1, yaml 2.2.0

B Open community challenge reveals molecular network modules with key roles in diseases

Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases

Sarvenaz Choobdar,[1,2,X] Mehmet E. Ahsen,[3,X] Jake Crawford,[4,X] Mattia Tomasoni,[1,2] David Lamparter,[1,2,5] Junyuan Lin,[6] Benjamin Hescott,[7] Xiaozhe Hu,[6] Johnathan Mercer,[8,9] Ted Natoli,[10] Rajiv Narayan,[10] The DREAM Module Identification Challenge Consortium,[11] Aravind Subramanian,[10] Gustavo Stolovitzky,[3,12] Zoltán Kutalik,[2,13] Kasper Lage,[8,9,14] Donna K. Slonim,[4] Julio Saez-Rodriguez,[15,16] Lenore J. Cowen,[4,6] Sven Bergmann,[1,2,17,Y,*] and Daniel Marbach[1,2,18,Y,Z,*]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

[3]Icahn Institute for Genomics and Multiscale Biology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[4]Department of Computer Science, Tufts University, Medford, MA, USA.

[5]Verge Genomics, San Francisco, CA, USA.

[6]Department of Mathematics, Tufts University, Medford, MA, USA.

[7]College of Computer and Information Science, Northeastern University, Boston, MA, USA.

[8]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

[9]Stanley Center at the Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[10]Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[11]Full list of members appears at the end of the paper.

[12]IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA.

[13]Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland.

[14]Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark.

[15]European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge, UK

[16]RWTH Aachen University, Faculty of Medicine, Joint Research Center for Computational Biomedicine, Aachen, Germany

[17]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa

[18]Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland.

[X]Co-first author

[Y]Co-last author

[Z]Lead contact

*Correspondence: sven.bergmann@unil.ch (S.B.), daniel.marbach.dm1@roche.com (D.M.)

Table of contents

Summary

Identification of modules in molecular networks is at the core of many current analysis methods in biomedical research. However, how well different approaches identify disease-relevant modules in different types of networks remains poorly understood. We launched the “Disease Module Identification DREAM Challenge”, an open competition to comprehensively assess module identification methods across diverse gene, protein and signaling networks. Predicted network modules were tested for association with complex traits and diseases using a unique collection of 180 genome-wide association studies (GWAS). While a number of approaches were successful in terms of discovering complementary trait-associated modules, consensus predictions derived from the challenge submissions performed best. We find that most of these modules correspond to core disease-relevant pathways, which often comprise therapeutic targets and correctly prioritize candidate disease genes. This community challenge establishes benchmarks, tools and guidelines for molecular network analysis to study human disease biology (https://synapse.org/modulechallenge).

Keywords

- Network biology
- Module identification
- Community detection algorithms
- Pathway analysis
- Genome-wide association studies
- Crowdsourced challenge
- Open science

Highlights

- Crowdsourced challenge enables critical assessment of module identification methods
- Top approaches recover complementary disease modules in diverse molecular networks
- Community-established benchmarks, user guidelines and tools for network analysis
- Molecular network modules reveal core pathways underlying complex traits and diseases

Introduction

Understanding the mechanisms and pathways underlying complex human diseases remains a difficult problem, hindering the development of targeted therapeutics. Complex diseases involve many genes and molecules that interact within context-specific cellular networks (Califano et al., 2012). These densely interconnected networks sense and propagate perturbations from genetic variants and environmental factors, giving rise to disease states that may be difficult to understand at the level of individual genes (Schadt, 2009). Indeed, it has become apparent that the majority of genetic variants underlying complex traits and diseases lie in noncoding regions of the genome where they presumably disrupt gene regulatory networks (Pickrell, 2014), lending further support to the long-recognized importance of molecular network analysis for understanding disease biology (Ideker and Sharan, 2008; Vidal et al., 2011).

Experimental and computational techniques for mapping molecular networks, including physical interaction networks (e.g., protein-protein interaction, signaling and regulatory networks) as well as functional gene networks (e.g., co-expression and genetic interaction networks), have been a major focus of systems biology. Recent studies have further introduced comprehensive collections of tissue-specific networks (Greene et al., 2015; Marbach et al., 2016). Network-based approaches are now widely used for systems-level analyses in diverse fields ranging from oncology (Chen et al., 2014; Tsherniak et al., 2017) to cell differentiation (Cahan et al., 2014; Ciofani et al., 2012). A key problem in biological network analysis is the identification of functional units, called modules or pathways. It is well known that molecular networks have a high degree of modularity (i.e., subsets of nodes are more densely connected than expected by chance), and that the corresponding modules often comprise genes or proteins that are involved in the same biological functions (Hartwell et al., 1999). Moreover, biological networks are typically too large to be examined and visualized as a whole. Consequently, module identification is often a crucial step to gain biological insights from network data (Chen et al., 2008; Langfelder and Horvath, 2008; Padi and Quackenbush, 2017; Pe'er et al., 2001).

Module identification, also called community detection or graph clustering, is a key problem in network science for which a wide range of methods have been proposed (Fortunato and Hric, 2016). These methods are typically assessed on *in silico* generated benchmark graphs (Girvan and Newman, 2002). However, how well different approaches uncover biologically relevant modules in real molecular networks remains poorly understood. Crowdsourced open-data competitions (known as challenges) have proven an effective means to rigorously assess methods and, in the process, foster collaborative communities and open innovation. The

Dialogue on Reverse Engineering and Assessment (DREAM) is a community-driven initiative promoting open-data challenges in systems biology and translational medicine (http://dreamchallenges.org). DREAM challenges have established standardized resources and robust methodologies for diverse problems, including the inference of gene regulatory and signaling networks (Hill et al., 2016; Marbach et al., 2012). But, so far there has been no community effort addressing the downstream analysis of molecular networks.

Here we present the results of the Disease Module Identification DREAM Challenge (**Fig. 1**). The aim of this challenge is to comprehensively assess module identification methods across diverse molecular networks. Six research groups contributed unpublished molecular networks and over 400 participants from all over the world developed and applied module identification methods. Teams predicted disease-relevant modules both within individual networks (Sub-challenge 1) and across multiple, layered networks (Sub-challenge 2). In the final round, 75 submissions, including method descriptions and code, were made across the two sub-challenges, providing a broad sampling of state-of-the-art methods. We employed a novel approach to assess the performance of these methods based on the number of discovered modules associated with complex traits or diseases. In this paper, we discuss the top-performing approaches, show that they recover complementary modules, and introduce a method to generate robust consensus modules. Finally, we explore the biology and therapeutic relevance of trait-associated network modules.

All challenge data, including the networks, GWAS datasets, team submissions and code are available as a community resource at https://www.synapse.org/modulechallenge.

Results

A crowdsourced challenge for empirical assessment of module identification methods

We developed a panel of diverse, human molecular networks for the challenge, including custom versions of two protein-protein interaction and a signaling network extracted from the STRING (Szklarczyk et al., 2015), InWeb (Li et al., 2017) and OmniPath (Türei et al., 2016) databases, a co-expression network inferred from 19,019 tissue samples from the GEO repository (Barrett et al., 2011), a network of genetic dependencies derived from genome-scale loss-of-function screens in 216 cancer cell lines (Cowley et al., 2014; Tsherniak et al., 2017), and a homology-based network built from phylogenetic patterns across 138 eukaryotic species

(Li et al., 2014) (Methods). These networks have varying size, link density and structural properties, making a heterogeneous benchmark resource (**Fig. 1A**).

Each network was generated specifically for the challenge and released in anonymized form (i.e., we did not disclose the gene names and the identity of the networks). Using unpublished networks made it impossible for participants to infer the gene identities, thus enabling rigorous "blinded" assessment. That is, participants could only use the provided network structures, without having access to any additional information such as known disease genes.

We solicited participation in two types of module identification challenges (**Fig. 1B**). In Sub-challenge 1, solvers were asked to run module identification on each of the provided networks individually (single-network module identification). Thus, they were asked to submit one set of modules for each of the six networks. This is a typical problem in biomedical research, where one is often presented with a single network derived from a given dataset. In Sub-challenge 2, the networks were re-anonymized in a way that the same gene identifier represented the same gene across all six networks. Solvers were then asked to identify a single set of non-overlapping modules by sharing information across the six networks (multi-network module identification). This is also common problem, as network-based approaches are often used to integrate disparate molecular datasets (Krishnan et al., 2016). In both sub-challenges, predicted modules had to be non-overlapping and comprise between 3 and 100 genes (modules with over one hundred genes are typically less useful to gain specific biological insights).

We developed a framework to empirically assess module identification methods based on the number of predicted modules that show significant association with complex traits and diseases (called trait-associated modules, **Fig. 1C**). To this end, predicted modules were scored on GWAS data using the Pascal tool (Lamparter et al., 2016), which takes into account confounders such as linkage disequilibrium within and between genes (Methods). Since we are employing a large collection of 180 GWAS datasets ranging over diverse disease-related human phenotypes (**Table S1**), this approach covers a broad spectrum of molecular processes. In contrast to evaluation of module enrichment using existing gene and pathway annotations, where it is sometimes difficult to ascertain that annotations were not derived from similar data types as the networks, the GWAS-based approach provides an orthogonal means to assess disease-relevant modules.

The challenge was run using the open-science Synapse platform (Derry et al., 2012). Over a two-month period, teams could make repeated submissions and see their performance on a

real-time leaderboard to iteratively improve their methods. The total number of leaderboard submissions per team was limited to 25 and 41 for the two sub-challenges, respectively. In the final round, teams could make a single submission for each sub-challenge, which had to include detailed method descriptions and code for reproducibility. The scoring of the final submissions was based on a separate set of GWAS data sets that were not used during the leaderboard round (Methods).

Community-based collection of module identification methods

The community contributed 42 single-network and 33 multi-network module identification methods in the final round of the two sub-challenges. Single-network module identification methods are listed in **Table 1**, top-performing approaches are detailed in Methods, and full descriptions and code of all methods are available on the Synapse platform (https://www.synapse.org/modulechallenge). In the following sections we first discuss the single-network methods (Sub-challenge 1).

We grouped methods into seven broad categories: (i) kernel clustering, (ii) modularity optimization, (iii) random-walk based, (iv) local methods, (v) ensemble methods, (vi) hybrid methods and (vii) other methods (**Fig. 2A**, **Table 1**). While many teams adapted existing algorithms for community detection, other teams -- including the best performers -- developed novel approaches.

Top methods from different categories achieve comparable performance

In Sub-challenge 1, teams submitted a separate set of predicted modules for each of the six networks. We scored these predictions based on the number of trait-associated modules at 5% false discovery rate (FDR; Methods). The overall score used to rank methods in the challenge was defined as the total number of trait-associated modules across the six networks. (Module predictions, scoring scripts and full results are available in on the challenge website.)

The top five methods achieved comparable performance with scores between 55 and 60, while the remaining methods did not get to scores above 50 (**Fig. 2B**). To assess the robustness of the challenge ranking, we further scored all methods on 1,000 subsamples of the GWAS hold-out set (Methods). This analysis revealed a significant difference between the top-scoring method *K1* (method IDs are defined in **Table 1**) and the remaining methods (**Fig. 2C**). In addition, we repeated the scoring using four different FDR cutoffs: method *K1* ranked 1st in

each case, while the performance of other methods varied (**Fig. S1A**). Moreover, method *K1* also obtained the top score in the leaderboard round. We conclude that although the final scores of the top 5 methods are close, method *K1* performed more robustly in diverse settings.

The top teams used different approaches: the best performers (*K1*) developed a novel kernel approach leveraging a diffusion-based distance metric (Cao et al., 2013, 2014) and spectral clustering (Ng et al., 2001); the runner-up team (*M1*) extended different modularity optimization methods with a resistance parameter that controls the granularity of modules (Arenas et al., 2008); and the third-ranking team (*R1*) used a random-walk method based on multi-level Markov clustering with locally adaptive granularity to balance module sizes (Satuluri et al., 2010). Interestingly, teams employing the widely-used Weighted Gene Co-expression Network Analysis tool (WGCNA) (Langfelder and Horvath, 2008), which relies on hierarchical clustering to detect modules, did not perform competitively in this challenge (rank 35, 37 and 41).

Four different method categories are represented among the top five performers, suggesting that no single approach is inherently superior for module identification in molecular networks. Rather, performance depends on the specifics of each individual method, including the strategy used to define the resolution of the modular decomposition (the number and size of modules). Most teams used the leaderboard round to determine an appropriate resolution to capture disease-relevant pathways. Notably, the two runner-up teams (*M1* and *R1*) both used methods specifically designed to control the resolution of modules, and the top three teams all subdivided large modules (>100 genes) by recursively applying their methods to the corresponding subnetworks. Pre-processing steps also affected performance: many of the top teams first sparsified the networks by discarding weak edges. A notable exception is the top method (*K1*), which performed robustly without any pre-processing of the networks.

The challenge also allows us to explore how informative different types of molecular networks are for finding modules underlying complex traits. In absolute numbers, methods recovered the most trait-associated modules in the co-expression and protein-protein interaction networks (**Fig. S1B**). However, relative to the network size, the signaling network contained the most trait-associated modules (**Fig. 2D**). The cancer-related and homology-based networks, on the other hand, were less informative for the considered traits. These results are consistent with the importance of signaling pathways for many of the considered traits and diseases.

Consensus predictions outperform individual methods

Integration of multiple team submissions sometimes leads to winning predictions in crowdsourced challenges (Marbach et al., 2012). We therefore developed an ensemble approach to derive consensus modules from team submissions. To this end, module predictions from different methods were integrated in a consensus matrix *C*, where each element c_{ij} is proportional to the number of methods that put gene *i* and *j* together in the same module. The consensus matrix was then clustered using the top-performing module identification method from the challenge (**Fig. S2A**, Methods).

When applied to the top 50% of methods from the leaderboard round, the consensus indeed leads to a new best-scoring prediction (**Fig. 2B,C**). However, when applied to fewer methods, the performance of the consensus drops (**Fig. S2C**), suggesting that further work is needed to make this approach practical outside of a challenge context.

Complementarity of different module identification approaches

We next asked whether predictions from different methods and networks tend to capture the same or complementary modules. To this end, we developed a pairwise similarity metric for module predictions, which we applied to the complete set of 252 module predictions from Sub-challenge 1 (42 methods x 6 networks, Methods). We find that similarity of module predictions is primarily driven by the underlying network and not the method category (**Fig. 3A**). When comparing module predictions of different methods across networks, we find that the top-performing methods produce dissimilar clusterings, suggesting that they capture complementary functional modules (**Fig. S3A**).

These observations can be confirmed by evaluating the overlap between trait-associated modules from different methods. Within the same network, only 46% of trait modules are recovered by multiple methods with good agreement (high overlap or submodules, **Fig. 3B**). Across different networks, the number of recovered modules with substantial overlap is even lower (17%). Thus, the majority of trait modules are method- and network-specific. This suggests that users should not rely on a single method or network to find trait-relevant modules.

The modules produced by different methods also vary in terms of their structural properties. For example, the average module size ranges from 7 to 66 genes across methods and does not correlate with performance in the challenge (**Figs. 3C**, **S3B-D**). This implies that trait-relevant pathways can be captured at different levels of granularity (indeed, 26% of trait modules are

submodules of larger trait modules, **Fig. 3B**). Topological quality metrics of modules such as modularity showed only modest correlation with the challenge score (**Fig. 3D**), highlighting the need to empirically assess module identification methods for a given task.

Multi-network module identification methods did not provide added power

In Sub-challenge 2, teams submitted a single modularization of the genes, for which they could leverage information from all six networks together. While some teams developed dedicated multi-network (multi-layer) community detection methods (De Domenico et al., 2015; Didier et al., 2015), the majority of teams first merged the networks in some way and then applied single-network methods.

It turned out to be very difficult to effectively leverage complementary networks for module identification. While three teams achieved marginally higher scores than single-network module predictions, the difference is not significant (**Figs. 3E**, **S1C**). Moreover, the best-scoring team simply merged the two protein interaction networks (the two most similar networks, **Fig. S2E**), discarding the other types of networks. Since no significant improvement over single-network methods was achieved, the winning position of Sub-challenge 2 was declared vacant.

We nevertheless also applied our consensus method to integrate team submissions across networks. The exact same consensus method as we employed for Sub-challenge 1 was used, except that a cross-network consensus matrix was formed by taking the sum of the six network-specific consensus matrices (**Fig. S2B**, Methods). This resulted in the best-scoring module prediction of Sub-challenge 2 (**Fig. 3E**), the only multi-network prediction that significantly outperforms single-network predictions, thus confirming the robustness of the consensus method and demonstrating that the multi-network methods can be further improved.

Network modules reveal shared pathways between traits

We next sought to explore biological properties of trait-associated modules discovered by the challenge participants. In what follows, we focus on the single-network predictions from Sub-challenge 1. The most trait-associated modules were found for immune-related, psychiatric, blood cholesterol and anthropometric traits, for which high-powered GWAS are available that are known to show strong pathway enrichment (**Fig. 4A**).

Significant GWAS loci often show association to multiple traits. Across our GWAS compendium, we found that 46% of trait-associated genes but only 28% of trait-associated modules are associated with multiple traits (**Fig. 4B**). Thus, mapping genes onto network modules may help disentangling trait-specific pathways at shared loci.

We further asked which traits are similar in terms of the implicated network components. To this end, we considered the union of all genes within network modules associated with a given trait (called “trait-module genes”). We then evaluated the pairwise similarity of traits based on the significance of the overlap between the respective trait-module genes (Methods). Trait relationships thus inferred are consistent with known biology and comorbidities between the considered traits and diseases (**Fig. 4C**). For example, consistent with its pathophysiological basis, age-related macular degeneration shares network components with cholesterol and immune traits, while coronary artery disease shows similarity with established risk factors (cholesterol levels, body mass index) and osteoporosis, which is epidemiologically and biologically linked (atherosclerotic calcification and bone mineralization involve related pathways).

Trait-associated modules implicate core disease genes and pathways

Trait-associated modules typically include many genes that do not show any signal in the respective GWAS. A key question is whether modules correctly predict such genes as being relevant for that trait or disease. We first consider a module from the consensus method that shows association to height -- a classic polygenic trait -- as an example. In the GWAS that was used to identify this module there are only three module genes that show association to height, while the remaining genes are predicted to play a role in height solely because they are members of this module (**Fig. 5A**). We sought to evaluate such candidate genes for height as well as other traits using higher-powered GWASs, ExomeChip data, monogenic disease genes and functional annotations.

There are eight traits for which we have both an older (lower-powered) and more recent (higher-powered) GWAS in our hold-out set: height, schizophrenia, ulcerative colitis, Crohn’s disease, rheumatoid arthritis, and three blood lipid traits (**Fig. S4A**). We can thus identify trait modules and candidate genes using the lower-powered GWAS and then evaluate how well they are supported in higher-powered GWAS (a common approach used to assess methods for GWAS gene prioritization, see Methods). Indeed, while only 3 genes in the height module introduced above are associated to height in the lower-powered GWAS (Randall et al., 2013), 13 module

genes are confirmed in the higher-powered GWAS (Wood et al., 2014) and 6 module genes further comprise coding variants associated to height in an independent ExomeChip study (Marouli et al., 2017) (**Fig 5B**). Similar results are obtained when evaluating module predictions from all challenge methods across the eight above-mentioned traits: a substantial fraction of module genes that do not show any signal and are located far from any significant locus in the lower-powered GWAS are subsequently confirmed by the higher-powered GWAS (**Fig. 5C**). This demonstrates that modules are predictive for trait-associated genes and could thus be used to prioritize candidate genes for follow-up studies, for instance.

We next explored the biological function and clinical relevance of identified trait modules. For example, the height module discussed above consists of two submodules comprising extracellular matrix proteins responsible for, respectively, collagen fibril and elastic fibre formation -- pathways that are essential for growth (**Fig. 5D**). Indeed, mutations of homologous genes in mouse lead to abnormal elastic fiber morphology (**Table S2**) and one out of four module genes are known to cause monogenic skeletal growth disorders in human (**Fig. 5D**). For example, the module gene *BMP1* (*Bone Morphogenic Protein 1*) causes osteogenesis imperfecta, which is associated with short stature. Interestingly, *BMP1* does not show association to height in current GWAS and ExomeChip studies (**Fig. 5A,B**), demonstrating how network modules can implicate additional disease-relevant pathway genes (see **Fig. S4B** for a systematic comparison of trait modules with independent disease gene sets from the literature).

To evaluate more generally whether trait-associated modules correspond to generic or disease-specific pathways, we visualized and tested modules for functional enrichment of Gene Ontology (GO) annotations, mouse mutant phenotypes, and diverse pathway databases. In order to account for annotation bias of well-studied genes (Glass and Girvan, 2014), we employed a noncentral hypergeometric test (Methods). We find that the majority of trait modules reflect core disease-specific pathways. For example, in the first protein-protein interaction network only 33% of trait modules from the consensus method have generic functions, such as epigenetic gene silencing for modules associated with schizophrenia and body mass index; the remaining 66% of trait modules correspond to core disease-specific pathways, some of which are therapeutic targets (**Fig. 6** and **Tables S3, S4**). Examples include a module associated with rheumatoid arthritis that comprises the B7:CD28 costimulatory pathway required for T cell activation, which is blocked by an approved drug (**Fig. 6A**); a module associated with inflammatory bowel disease corresponding to cytokine signalling pathways mediated by Janus kinases (JAKs), which are therapeutically being targeted at multiple levels (**Fig. 6B**); and a module associated with myocardial infarction that includes the NO/cGMP signaling cascade,

which plays a key role in cardiovascular pathophysiology and therapeutics (**Fig. 6C**). We further applied our pipeline to a GWAS on IgA nephropathy (IgAN) obtained after the challenge, a disease with poorly understood etiology and no effective therapy (Kiryluk et al., 2014). IgAN is an autoimmune disorder that manifests itself by deposition of immune complexes in the kidney's glomeruli, triggering inflammation (glomerulonephritis) and tissue damage. The best-performing challenge method (*K1*) revealed one IgAN-specific module. The module implicates complement and coagulation cascades, pointing to the chemokine *PF4V1* as a novel candidate gene (**Fig. 6D**). In support of the function of this module in IgAN, top enriched mouse mutant phenotypes for module gene homologs are precisely "glomerulonephritis" and "abnormal blood coagulation" (**Fig. S5**).

Discussion

Large-scale network data are becoming pervasive in many areas ranging from the digital economy to the life sciences. While analysis goals vary across fields, robust detection of network communities remains an essential task in many applications of interest. We have conducted a critical assessment of module identification methods on real-world networks, providing much-needed guidance for users. The community-based challenge enabled comprehensive and impartial assessment, avoiding the "self-assessment trap" that leads researchers to consciously or unconsciously overestimate performance when evaluating their own algorithms (Norel et al., 2011). While it is important to keep in mind that the exact ranking of methods -- as in any benchmark -- is specific to the task and datasets considered, we believe that the resulting collection of top-performing module identification tools and methodological insights will be broadly useful for modular analysis of complex networks in biology and other domains.

In addition to providing a cross section of established approaches, the collection of contributed methods also includes novel algorithms that further advance the state-of-the-art (notably, the best-performing method). Kernel clustering, modularity optimization, random-walk-based and local methods were all represented among the top performers, suggesting that no single type of approach is inherently superior. In contrast, basic approaches such as hierarchical clustering, which is widely used for gene network analysis, did not perform competitively. Consensus modules obtained by integrating multiple team submissions achieved the top score, demonstrating that method performance can be further improved. However, this strategy was only successful when integrating predictions from over twenty methods, explaining why ensemble approaches applied by individual teams, which integrated only few methods, did not

perform well. Indeed, our analysis showed that top-performing methods produced very different modular decompositions, capturing complementary pathways at varying resolutions that may be difficult to merge in a single consensus prediction.

Published studies in biology that apply network analysis tools typically rely on a single clustering method. The results of this challenge call for a different approach. We recommend that users apply top methods from several categories, enabling the detection of different types of modules and making results less prone to biases of any single approach. We find that the top four challenge methods (*K1*, *M1*, *R1* and *M2*) already offer substantial diversity (**Fig. S3E**). The generated modules should be considered as is, without forming a consensus prediction. It should be noted that the larger number of modules also results in a higher multiple testing burden in any subsequent analyses (e.g., functional enrichment testing) and that modules from different methods may overlap. When a single non-overlapping partition is needed, the best-performing challenge method (*K1*) is a good choice as it functioned robustly in diverse settings (notably, it was also used to cluster the consensus matrices, leading to the top-scoring consensus predictions in both sub-challenges).

The challenge also emphasized the importance of the resolution (size and number of modules), which critically affected results. Biological networks typically have a hierarchical modular structure, which implies that disease-relevant pathways can be captured at different levels (Ravasz et al., 2002). Our results showed that the optimal resolution is method- and network-specific (**Fig. S3B-D**). Top-performing challenge methods allowed the resolution to be tuned. Although setting the “right” resolution can be challenging for users, this critical point should not be sidestepped. We recommend that users experiment with different resolutions and use the settings optimized by teams for the different types of networks as guidance.

Our analysis showed that signaling, protein-protein interaction and co-expression networks comprise complementary trait-relevant modules (**Fig. 3A,B**). Considering different types of networks is thus clearly advantageous. However, multi-network module identification methods that attempted to reveal integrated modules across these networks failed to significantly improve predictions compared to methods that considered each network individually. Possibly, the networks of the challenge were not sufficiently related -- multi-network methods may perform better on networks from the same tissue- and disease-context (Krishnan et al., 2016).

The benchmark datasets and results of the challenge provide a reference point for future method improvements. We see many promising avenues for future work, such as: (i) top-

performing challenge methods can potentially be further enhanced with ensemble approaches that sample multiple partitions of the same method to generate stable results (Lancichinetti and Fortunato, 2012); (ii) top teams recursively broke down large "supermodules" by iteratively applying their clustering methods, a heuristic that worked well, but more principled approaches to globally balance module sizes may improve accuracy (exemplified by method *R1*); and (iii) methods for detection of overlapping modules (Ihmels et al., 2002) may also be assessed using the benchmarks of this challenge.

An important observation about these results is that the module identification tasks were performed on completely blinded networks; gene identities and even the type of relationship captured was unknown to challenge participants. The fact that meaningful modules can be identified in such a context is perhaps surprising, revealing how much functional information is present strictly in the topological structure of biological networks. It remains to be seen whether an un-blinded approach that allows integration of prior knowledge about gene functions, relationships, and the source of network edges might further improve the quality of inferred modules, especially when integrating data from multiple types of networks.

The collective effort of over 400 challenge participants resulted in a unique compendium of modules for the different types of molecular networks considered. By leveraging the "wisdom of crowds" we generated robust consensus modules, which captured disease-relevant pathways better than any individual method. While most modules partly reflect known pathways or functional gene categories, which they reorganize and expand with additional genes, other modules may correspond to yet uncharacterized pathways. The consensus modules (gene sets) thus constitute a novel data-driven pathway collection, which may complement existing pathway collections in a range of applications (e.g., for interpretation of gene expression data using gene set enrichment analysis).

There is continuing debate over the value of GWASs for revealing disease mechanisms and therapeutic targets. Indeed, the number of GWAS hits continues to grow as sample sizes increase, but the bulk of these hits may not correspond to core genes with specific roles in disease etiology. An "omnigenic" model recently proposed by Boyle et al. (2017) explains this observation by the high interconnectivity of molecular networks, which implies that most of the expressed genes in a disease-relevant tissue are likely to be at least weakly connected to core genes and may thus have non-zero effects on that disease. Indeed, disease-associated genes tend to coalesce in regulatory networks of tissues that are specific to that disease (Marbach et al., 2016). Our analysis of 180 GWAS datasets across six molecular networks demonstrated

that, although thousands of genes may show association for a given disease, at the network level specific disease modules comprising only dozens of genes can be identified. We have shown that these modules are more disease-specific than individual genes, reveal pathway-level similarity between diseases, accurately prioritize candidate genes, and correspond to core disease pathways in the majority of cases. These results are consistent with the omnigenic model and the robustness of biological networks: presumably, the many genes that influence disease indirectly are broadly distributed across network modules, while core disease genes cluster in specific pathways underlying pathophysiological processes (Sullivan and Posthuma, 2015). Our analysis also demonstrated that GWASs with larger sample size are extremely useful for the identification of key core modules and SNP effect size (explained variance) is not necessarily an indicator of core-ness.

In this study we used global networks because the focus was on method assessment across diverse disorders. Global networks mostly comprise pathways that are either broadly expressed or specific to well-studied tissues, such as blood or immune cells. In the near future, we expect much more detailed maps of cell- and tissue-specific networks, along with diverse high-powered genetic datasets, to become available. We hope that the challenge resources will be instrumental in dissecting these networks and will provide a solid foundation for developing integrative methods to reveal the cell types and causal circuits implicated in human disease.

Consortia

The contributing members of the DREAM Module Identification Challenge Consortium are:

Fabian Aicheler,[1] Nicola Amoroso,[2,3] Alex Arenas,[4] Karthik Azhagesan,[5-,7] Aaron Baker,[8-10] Michael Banf,[11] Serafim Batzoglou,[12] Anaïs Baudot,[13] Roberto Bellotti,[2,3,14] Sven Bergmann,[15,16] Keith A. Boroevich,[17] Christine Brun,[18-19] Stanley Cai,[20,93,94] Michael Caldera,[21] Alberto Calderone,[22] Gianni Cesareni,[22] Weiqi Chen,[23] Christine Chichester,[24] Sarvenaz Choobdar,[15-16] Lenore Cowen,[25-26] Jake Crawford,[25] Hongzhu Cui,[27] Phuong Dao,[46] Manlio De Domenico,[4,29] Andi Dhroso,[27] Gilles Didier,[13] Mathew Divine,[1] Antonio del Sol,[36] Xuyang Feng,[30] Jose C. Flores-Canales,[31-32] Santo Fortunato,[33] Anthony Gitter,[8,9,10] Anna Gorska,[34] Yuanfang Guan,[35] Alain Guénoche,[13] Sergio Gómez,[4] Hatem Hamza,[24] András Hartmann,[36] Shan He,[23] Anton Heijs,[37] Julian Heinrich,[1] Benjamin Hescott,[38] Xiaozhe Hu,[26] Ying Hu,[39] Xiaoqing Huang,[46] V. Keith Hughitt,[40-41] Minji Jeon,[42] Lucas Jeub,[33] Nathan Johnson,[27] Keehyoung Joo,[32,43] InSuk Joung,[31-32] Sascha Jung,[36] Susana G. Kalko,[36] Piotr J. Kamola,[17] Jaewoo Kang,[42,44] Benjapun Kaveelerdpotjana,[23] Minjun Kim,[45] Yoo-Ah Kim,[46] Oliver Kohlbacher,[1,47-48] Dmitry Korkin,[27,49-50] Kiryluk Krzysztof,[51] Khalid Kunji,[52] Zoltàn Kutalik,[16,53] Kasper Lage,[54-56] David Lamparter,[15-16,57] Sean Lang-Brown,[58] Thuc Duy Le,[59-60] Jooyoung Lee,[31-32] Sunwon Lee,[42] Juyong Lee,[61] Dong Li,[23] Jiuyong Li,[60] Junyuan Lin,[26] Lin Liu,[60] Antonis Loizou,[62] Zhenhua Luo,[63] Artem Lysenko,[17] Tianle Ma,[64] Raghvendra Mall,[52] Daniel Marbach,[15-16] Tomasoni Mattia,[15-16] Mario Medvedovic,[65] Jörg Menche,[21] Johnathan Mercer,[54,56] Elisa Micarelli,[22] Alfonso Monaco,[3] Felix Müller,[21] Rajiv Narayan,[66] Oleksandr Narykov,[50] Ted Natoli,[66] Thea Norman,[67] Sungjoon Park,[42] Livia Perfetto,[22] Dimitri Perrin,[68] Stefano Pirrò,[22] Teresa M. Przytycka,[46] Xiaoning Qian,[69] Karthik Raman,[5-7] Daniele Ramazzotti,[12] Balaraman Ravindran,[70,6,7] Philip Rennert,[71] Julio Saez-Rodriguez,[7-,73] Charlotta Schärfe ,[1] Roded Sharan,[74] Ning

Shi,[23] Wonho Shin,[44] Hai Shu,[75] Himanshu Sinha,[5,6,7] Donna K. Slonim,[25] Lionel Spinelli,[18] Suhas Srinivasan,[49] Aravind Subramanian,[66] Christine Suver,[76] Damian Szklarczyk,[77] Sabina Tangaro,[3] Suresh Thiagarajan,[78] Laurent Tichit,[13] Thorsten Tiede,[1] Beethika Tripathi,[70,6,7] Aviad Tsherniak,[66] Tatsuhiko Tsunoda,[17,79,80] Dénes Türei,[72] Ehsan Ullah,[52] Golnaz Vahedi,[20,93,94] Alberto Valdeolivas,[13,82] Jayaswal Vivek,[83] Christian von Mering,[77] Andra Waagmeester,[37] Bo Wang,[12] Yijie Wang,[46] Barbara A. Weir,[84-85] Shana White,[65] Sebastian Winkler,[1] Ke Xu,[86] Taosheng Xu,[87] Chunhua Yan,[39] Liuqing Yang,[88] Kaixian Yu,[75] Xiangtian Yu,[89] Gaia Zaffaroni,[36] Mikhail Zaslavskiy,[90] Tao Zeng,[89] Lu Zhang,[12] Weijia Zhang,[60] Lixia Zhang,[65] Xinyu Zhang,[86] Junpeng Zhang,[91] Xin Zhou,[12] Jiarui Zhou,[23] Hongtu Zhu,[75] Junjie Zhu,[92] Guido Zuccon,[68]

[1]Applied Bioinformatics, Center for Bioinformatics, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany. [2]Department of Physics 'Michelangelo Merlin', University of Bari 'Aldo Moro', Via G. Amendola 173, 70126 Bari, Italy. [3]INFN, Sezione di Bari, Via A. Orabona 4, 70125 Bari, Italy. [4]Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Spain. [5]Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India. [6]Initiative for Biological Systems Engineering (IBSE), Indian Institute of Technology Madras. [7]Robert Bosch Centre for Data Science and Artificial Intelligence(RBC-DSAI), Indian Institute of Technology Madras. [8]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, USA. [9]Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA. [10]Morgridge Institute for Research, Madison, Wisconsin, USA. [11]Department of Plant Biology, Carnegie Institution for Science, Stanford, USA. [12]Department of Computer Science, Stanford University, USA. [13]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, UMR 7373, Marseille, France. [14]Centro TIRES, Via G. Amendola 173, 70126 Bari, Italy. [15]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [16]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [17]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [18]Aix Marseille Univ, INSERM, TAGC, UMR1090, Marseille, France. [19]CNRS, Marseille, France. [20]Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA. [21]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. [22]Bioinformatics and Computational Biology Unit, Department of Biology, Tor Vergata University, Italy. [23]School of Computer Science, The University of Birmingham, Birmingham, UK. [24]Nestle Institute of Health Sciences, Lausanne, Switzerland.[25]Department of Computer Science, Tufts University, Medford, MA, USA. [26]Department of Mathematics, Tufts University, Medford, MA, USA. [27]Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA, USA. [29]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy. [30]Department of Cancer Biology, University of Cincinnati, Cincinnati, OH, USA. [31]Center for In Silico Protein Science, Korea Institute for Advanced Study, Seoul, Korea. [32]School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Korea. [33]School of Informatics, Computing and Engineering, Indiana University, Bloomington, USA. [34]Algorithms in Bioinformatics, Center for Bioinformatics, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany. [35]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109. [36]LCSB - Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [37]Micelio, 2180 Antwerp, Belgium. [38]College of Computer and Information Science, Northeastern University, Boston, MA, USA. [39]National Cancer Institute, Center for Biomedical Informatics & Information Technology, 9609 Medical Center Drive, Bethesda, MD 20850, USA. [40]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [41]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland, USA. [42]Department of Computer Science and Engineering, Korea University, Seoul, Korea. [43]Center for Advanced Computation, Korea Institute for Advanced Study, Seoul, Korea. [44]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul, Korea. [45]Community High School, 401 N Division St, Ann Arbor, MI, 48104. [46]National Center for Biotechnology Information, National Institute of Health (NCBI/NLM/NIH), USA. [47]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Spemannstr. 38, 72076 Tuebingen, Germany. [48]Quantitative Biology

Center, University of Tuebingen, Auf der Morgenstelle 8, 72076 Tuebingen, Germany. [49]Data Science Program, Worcester Polytechnic Institute, Worcester, MA, USA. [50]Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA. [51]Department of Medicine, College of Physicians & Surgeons, Columbia University, New York, NY, USA. [52]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar. [53]Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland. [54]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [55]Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark. [56]Stanley Center at the Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [57]Verge Genomics, San Francisco, CA, USA. [58]Division of Geriatrics, Department of Medicine, University of California, San Francisco, USA. [59]Centre for Cancer Biology, University of South Australia. [60]School of Information Technology and Mathematical Sciences, University of South Australia. [61]Department of Chemistry, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon, 24341, Republic of Korea. [62]BlueSkyIt, Amsterdam, the Netherlands. [63]The Liver Care Center and Divisions of Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [64]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA. [65]Dept. of Env. Health, Division of Biostatistics and Bioinformatics, University of Cincinnati, OH, USA. [66]Broad Institute of Harvard and MIT, Cambridge, MA. [67]Bill and Melinda Gates Foundation. [68]School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia. [69]Dept. of Electrical & Computer Engineering, Texas A&M University, USA. [70]Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India. [71]Rockville, MD, USA (No affiliation). [72]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK. [73]RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine, 52057 Aachen, Germany. [74]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [75]Department of Biostatistics, the University of Texas MD Anderson Cancer Center, Houston, TX, USA. [76]Sage Bionetworks, Seattle, Washington 98109, USA. [77]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. [78]Memphis, TN, USA (No affiliation). [79]CREST, JST, Tokyo, Japan. [80]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [82]ProGeLife, Marseille, France. [83]Disease Science & Technology, Biocon Bristol-Myers Squibb Research Centre, Bangalore, India. [84]Broad Institute of Harvard and MIT, Cambridge, MA. [85]Janssen Research and Development. [86]Department of Psychiatry, Yale School of Medicine, West Haven, CT, USA. [87]Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui, China. [88]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [89]Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. [90]Computational biology consulting, avenue Kleber 100, Paris, France. [91]School of Engineering, Dali University. [92]Department of Electrical Engineering, Stanford University, USA. [93]Institute for Immunology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA. [94]Epigenetics Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Author contributions

S.C., D.L., Z.K., G.S., J.M., K.L., J.S.-R., S.B. and D.M. conceived the challenge; S.C., G.S., J.S.-R., S.B. and D.M. organized the challenge; S.C. and D.M. performed team scoring; S.C., M.E.A., J.C., M.T., D.K.S., L.J.C. and D.M. analyzed results; J.M., T.N., R.N., A.S., K.L. and J.S.-R. constructed networks; J.C., J.L., B.H., X.H., D.K.S. and L.J.C. designed the top-performing method; the DREAM Module Identification Consortium provided data and performed

module identification; S.B. and D.M. designed the study; and D.M. prepared the manuscript. All authors discussed the results and implications, and commented on the manuscript at all stages.

Acknowledgments

The challenge was hosted on Sage Bionetwork's Synapse platform (https://synapse.org/). The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1 to S.B. and grant FN 31003A-169929 to Z.K.), SystemsX.ch (grant SysGenetiX to S.B. and grant AgingX to Z.K.), the Swiss Institute of Bioinformatics (Z.K. and S.B.) and the Leenaards Foundation (Z.K.).

References

Arenas, A., Fernández, A., and Gómez, S. (2008). Analysis of the structure of complex networks at different resolution levels. New J. Phys. *10*, 053039.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. *39*, D1005-1010.

Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., and Bult, C.J. (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Res. *45*, D723–D729.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: Network Biology Applied to Stem Cell Engineering. Cell *158*, 903–915.

Cai, J., Candès, E., and Shen, Z. (2010). A Singular Value Thresholding Algorithm for Matrix Completion. SIAM J. Optim. *20*, 1956–1982.

Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat. Genet. *44*, 841–847.

Cao, M., Zhang, H., Park, J., Daniels, N.M., Crovella, M.E., Cowen, L.J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. PloS One *8*, e76339.

Cao, M., Pietras, C.M., Feng, X., Doroschak, K.J., Schaffner, T., Park, J., Zhang, H., Cowen, L.J., and Hescott, B.J. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. Bioinforma. Oxf. Engl. *30*, i219-227.

Chen, J.C., Alvarez, M.J., Talos, F., Dhruv, H., Rieckhof, G.E., Iyer, A., Diefes, K.L., Aldape, K., Berens, M., Shen, M.M., et al. (2014). Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. Cell *159*, 402–414.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., et al. (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429–435.

Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., Agarwal, A., Huang, W., Parkurst, C.N., Muratet, M., et al. (2012). A validated regulatory network for th17 cell specification. Cell *151*, 289–303.

Clauset, A., Moore, C., and Newman, M.E.J. (2008). Hierarchical structure and the prediction of missing links in networks. Nature *453*, 98–101.

Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali, L.D., Gerath, W.F., Pantel, S.E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci. Data *1*, 140035.

De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems. Phys. Rev. X *5*, 011027.

Derry, J.M.J., Mangravite, L.M., Suver, C., Furia, M.D., Henderson, D., Schildwachter, X., Bot, B., Izant, J., Sieberts, S.K., Kellen, M.R., et al. (2012). Developing predictive molecular maps of human disease through community-based modeling. Nat. Genet. *44*, 127–130.

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

Fortunato, S., and Hric, D. (2016). Community detection in networks: A user guide. Phys. Rep. *659*, 1–44.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature *536*, 41–47.

Girvan, M., and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. *99*, 7821–7826.

Glass, K., and Girvan, M. (2014). Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets. Sci. Rep. *4*.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47-52.

Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al. (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. Nat. Methods *13*, 310–318.

Ideker, T., and Sharan, R. (2008). Protein networks in disease. Genome Res. *18*, 644–652.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. Nat. Genet. *31*, 370–377.

Jiang, P., and Singh, M. (2010). SPICi: a fast clustering algorithm for large biological networks. Bioinforma. Oxf. Engl. *26*, 1105–1111.

Jin, J. (2015). Fast community detection by SCORE. Ann. Stat. *43*, 57–89.

Kiryluk, K., Li, Y., Scolari, F., Sanna-Cherchi, S., Choi, M., Verbitsky, M., Fasel, D., Lata, S., Prakash, S., Shapiro, S., et al. (2014). Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens. Nat. Genet. *46*, 1187–1196.

Kondor, R.I., and Lafferty, J.D. (2002). Diffusion Kernels on Graphs and Other Discrete Input Spaces. In Proceedings of the Nineteenth International Conference on Machine Learning, (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.), pp. 315–322.

Kraehling, J.R., and Sessa, W.C. (2017). Contemporary Approaches to Modulating the Nitric Oxide-cGMP Pathway in Cardiovascular Disease. Circ. Res. *120*, 1174–1182.

Krishnan, A., Taroni, J.N., and Greene, C.S. (2016). Integrative Networks Illuminate Biological Factors Underlying Gene–Disease Associations. Curr. Genet. Med. Rep. *4*, 155–162.

Lamparter, D., Marbach, D., Rico, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol *12*, e1004714.

Lancichinetti, A., and Fortunato, S. (2012). Consensus clustering in complex networks. Sci. Rep. *2*, srep00336.

de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.-G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. *49*, 256–261.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lee, J., Gross, S., and Lee, J. (2012). Mod-CSA: Modularity optimization by conformational space annealing. Phys Rev E *85*.

Li, D., He, S., Pan, Z., and Hu, G. (2016). Active modules for multilayer weighted gene co-expression networks: a continuous optimization approach. BioRxiv 056952.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods *14*, 61–64.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S., and Mootha, V.K. (2014). Expansion of biological pathways based on evolutionary inference. Cell *158*, 213–225.

Madhusudhan, T., Kerlin, B.A., and Isermann, B. (2016). The emerging role of coagulation proteases in kidney disease. Nat. Rev. Nephrol. *12*, 94–109.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods *13*, 366–370.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M., et al. (2017). Rare and low-frequency coding variants alter human adult height. Nature *542*, 186–190.

Neurath, M.F. (2017). Current and emerging therapeutic targets for IBD. Nat. Rev. Gastroenterol. Hepatol. *14*, 269–278.

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

Ng, A.Y., Jordan, M.I., and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems, (MIT Press), pp. 849–856.

Norel, R., Rice, J.J., and Stolovitzky, G. (2011). The self-assessment trap: can we all be better than average? Mol. Syst. Biol. *7*, 537.

Padi, M., and Quackenbush, J. (2017). Phenotype-Driven Transitions In Regulatory Network Structure. BioRxiv 142281.

Parisi, F., Strino, F., Nadler, B., and Kluger, Y. (2014). Ranking and combining multiple predictors without labeled data. Proc. Natl. Acad. Sci. 201219097.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. Bioinforma. Oxf. Engl. *17 Suppl 1*, S215-224.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). DeepWalk: Online Learning of Social Representations. ArXiv14036652 Cs 701–710.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

Pons, P., and Latapy, M. (2005). Computing communities in large networks using random walks (long version). ArXiv:Physics/0512106.

Randall, J.C., Winkler, T.W., Kutalik, Z., Berndt, S.I., Jackson, A.U., Monda, K.L., Kilpeläinen, T.O., Esko, T., Mägi, R., Li, S., et al. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet. *9*, e1003500.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.L. (2002). Hierarchical organization of modularity in metabolic networks. Science *297*, 1551–1555.

Rosvall, M., Axelsson, D., and Bergstrom, C.T. (2009). The map equation. Eur. Phys. J. Spec. Top. *178*, 13–23.

Satuluri, V., Parthasarathy, S., and Ucar, D. (2010). Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability. In Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, (New York, NY, USA: ACM), pp. 247–256.

Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218–223.

Shao, J., Yang, Q., Liu, J., and Kramer, S. (2016). Graph Clustering with Density-Cut. ArXiv160600950 Phys.

Shiokawa, H., Fujiwara, Y., and Onizuka, M. (2015). SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-scale Graphs. Proc VLDB Endow *8*, 1178–1189.

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., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437–1452.e17.

Sullivan, P.F., and Posthuma, D. (2015). Biological pathways and networks implicated in psychiatric disorders. Curr. Opin. Behav. Sci. *2*, 58–68.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447-452.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). LINE: Large-scale Information Network Embedding. In Proceedings of the 24th International Conference on World Wide Web, (Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee), pp. 1067–1077.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564–576.e16.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods *13*, 966–967.

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

Wang, Y., and Qian, X. (2017). Finding low-conductance sets with dense interactions (FLCD) for better protein complex prediction. BMC Syst. Biol. *11*.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

Table 1

Table 1. Module identification methods

ID[a]	Description	Score[b]	Pre- / post-processing
Kernel clustering: (i) the weighted adjacency matrix is transformed into a gene similarity matrix; (ii) a clustering algorithm is applied.			
K1	(i) Diffusion State Distance metric (Cao et al., 2013); (ii) spectral clustering.	60	R
K2	(i) Singular Value Thresholding (Cai et al., 2010) maps the graph into a latent feature space; (ii) hierarchical clustering using Ward's method.	48	W, R
K3	(i) Large-scale Information Network Embedding (LINE) (Tang et al., 2015); (ii) K-means clustering.	46	-
K4	(i) Extension of Spectral Clustering On Ratios-of-Eigenvectors (SCORE) (Jin, 2015) allowing for weighted networks and hierarchical structure of submodules; (ii) spectral clustering.	42	R
K5	(i) SCORE (Jin, 2015); (ii) spectral clustering.	38	-
K6	(i) Diffusion kernel is applied to graph Laplacian (Kondor and Lafferty, 2002); (ii) Weighted Gene Coexpression Network Analysis (WGCNA) (Langfelder and Horvath, 2008).	30	M
Modularity optimization: search algorithms are employed to find modules that maximize a modularity quality function.			
M1	Modularity optimization algorithms are extended with a multiresolution technique (Arenas et al., 2008).	60	S, R
M2	Louvain community detection algorithm (Blondel et al., 2008).	56	S,W,R,M
M3	Extension of a multi-network module identification method (Didier et al., 2015), here applied to single-layer networks.	48	R
M4	PageRank algorithm is used to create an initial partition for the Louvain method.	44	W, R
M5	A hierarchical module tree is generated using the Louvain method, optimal partitions are selected using modularity, conductance and connectivity metrics.	42	W,R,M,F
M6	Greedy agglomerative clustering approach optimizes a score based on total weight of intra-module edges and module size.	40	S,W, M
M7	Fast greedy clustering algorithm (Clauset et al., 2008) that iteratively divides modules to optimize the modularity.	40	-
M8	Modularity optimization by Conformational Space Annealing (Mod-CSA) (Lee et al., 2012) using the weighted adjacency matrix.	38	S, R
M9	Louvain algorithm is used for optimization of a generalized modularity metric with a resolution parameter.	37	R
M10	Louvain algorithm.	33	R
Random-walk-based: modules are identified using diffusion processes over the network.			
R1	Multi-level Markov clustering is extended with a regularization matrix to balance module sizes (Satuluri et al., 2010).	58	S, W, R
R2	Walktrap algorithm (Pons and Latapy, 2005), output modules are filtered based on the median node degree.	44	S, R
R3	Walktrap algorithm.	43	S, R
R4	A machine learning approach for predicting disease genes from graph features is combined with the Infomap algorithm (Rosvall et al., 2009) for community detection.	40	S,R,F
R5	Walktrap algorithm with varying number of steps.	39	S, F, M
R6	Infomap algorithm, Markov-time parameter is optimized to yield maximum number of modules of valid size.	38	R,M
R7	Markov clustering, output modules are filtered based on conductance and module size.	36	S, w
R8	Recursive local graph sparsification and clustering using Infomap for scalable community detection.	36	S, R
R9	Walktrap is used for the first network, Infomap for the remaining networks.	28	R
R10	Modules detected using Walktrap and Infomap are combined.	20	S
Local methods: agglomerative algorithms that grow modules from seed nodes.			
L1	Topological overlap matrix is clustered using the fast agglomerative SPICi (Jiang and Singh, 2010) and SCAN++ algorithms (Shiokawa et al., 2015).	55	S, W,R
L2	Basic agglomerative approach assigning genes to connected modules until the module size limit is reached.	31	W,R,M
L3	Local method that grows modules from seed nodes using a novel Triangle based Community Expansion (TCE) method.	30	M
Ensemble clustering: alternative clusterings sampled either from stochastic runs or from a set of different methods are merged.			
E1	Various clustering methods are applied on network embeddings created using DeepWalk (Perozzi et al., 2014), consensus modules are obtained using a bagging method.	46	S,W,M
E2	Consensus modules are derived from two flat clustering algorithms: ClusterOne and Finding Low-Conductance set with Dense interactions (FLCD) (Wang and Qian, 2017).	41	S,W,F
E3	Ensemble approach applied to integrate multiple Markov clustering runs.	24	S,R
Hybrid methods: different clustering methods are selected for each network based on leaderboard performance or structural quality scores.			
H1	Either Louvain, Infomap, or a continuous optimization method (Li et al., 2016) are selected for each network.	50	R, F
H2	Either Louvain, Infomap, SPICi, or DCut (Shao et al., 2016) are selected for each network.	50	W,R
H3	Up to five different methods are applied to cluster networks, followed by filtering of modules based on structural quality metrics.	40	W,R, M, F
H4	Up to nine different methods are applied in different combinations, followed by module filtering and post-processing steps.	37	
H5	Up to seven different methods are applied including an ensemble approach, followed by filtering and post-processing steps.	31	S,W,R,M,F
H6	WGCNA followed by fast greedy community detection to refine modules.	19	R
H7	No detailed description provided.	14	-
Others			
O1	Agglomerative algorithm that joins clusters based on the number of shared neighbors and the cluster sizes.	36	W,F
O2	Two-way modules (dense bipartite subgraphs) are mined using a heuristic algorithm.	33	W,F
O3	No detailed description provided.	12	-

(legend on next page)

Table 1. The 42 module identification methods applied in Sub-challenge 1 grouped by category (see **Fig. 2A**).

[a]Identifier (ID) of the method used throughout the paper.

[b]Overall score of the method as defined in **Fig. 2B**.

[c]Common pre- and post-processing steps. Pre-processing steps are coded as: (S) sparsification of networks and (W) rescaling of edge weights. Post-processing steps are coded as: (R) recursive break-down of large modules, (M) merging modules of invalid size followed by re-modularization, and (F) filtering modules according to a quality metric.

Figure 1

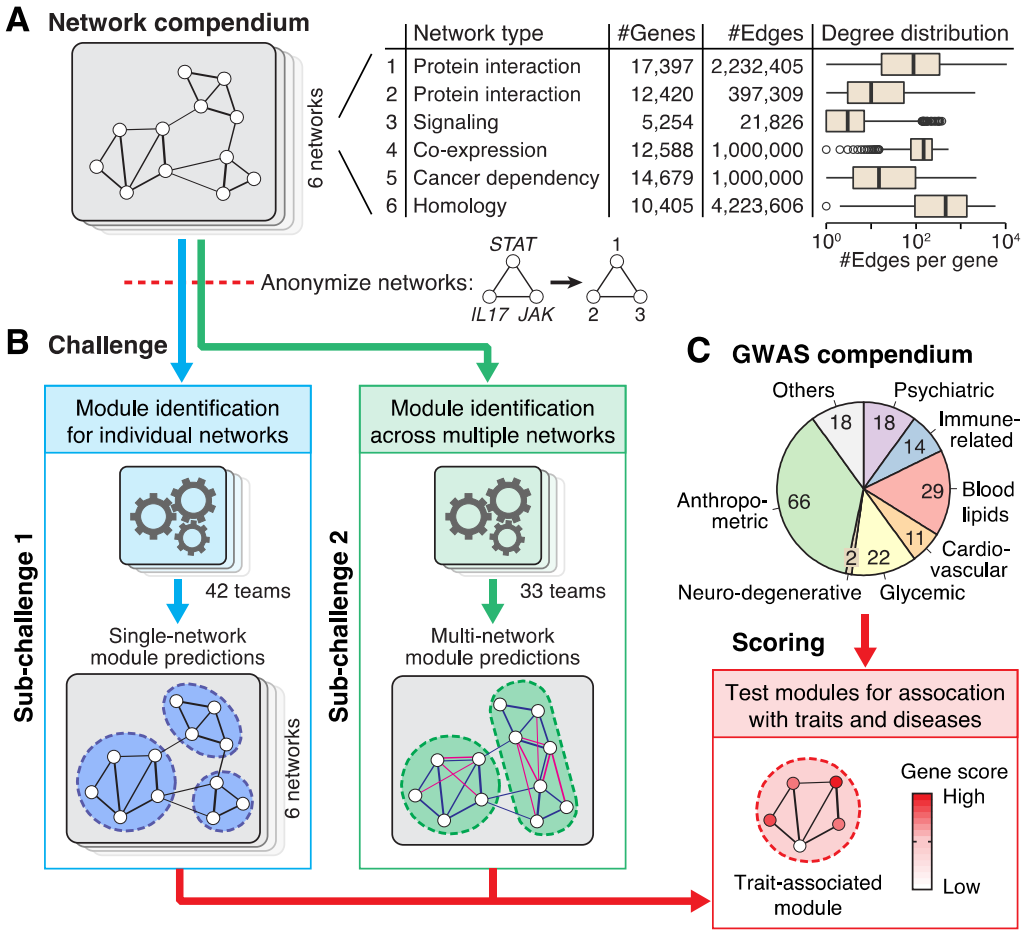


Figure 1: The Disease Module Identification DREAM Challenge.
We launched an open-participation community challenge, where teams competed to predict groups of functionally related genes (i.e., modules) within diverse molecular networks.

(A) The challenge comprised six networks, including protein-protein interaction, signaling, co-expression, cancer dependency, and homology-based gene networks. As the networks were all unpublished, we could anonymize them by removing the gene labels. This prevented participants from using existing knowledge of gene functions, thus enabling rigorous, blinded assessment.

(B) The aim of the challenge was to identify disease-relevant modules within the provided networks. Teams could participate in either or both sub-challenges: 42 teams predicted modules for individual networks (Sub-challenge 1) and 33 teams predicted integrated modules across multiple networks (Sub-challenge 2).

(C) The submitted modules were tested for association with complex traits and diseases using a comprehensive collection of 180 GWAS datasets. The final score for each method was the number of trait-associated modules that it discovered. Since GWAS are based on data completely different from those used to construct the networks, they can provide independent support for biologically relevant modules.

Figure 2

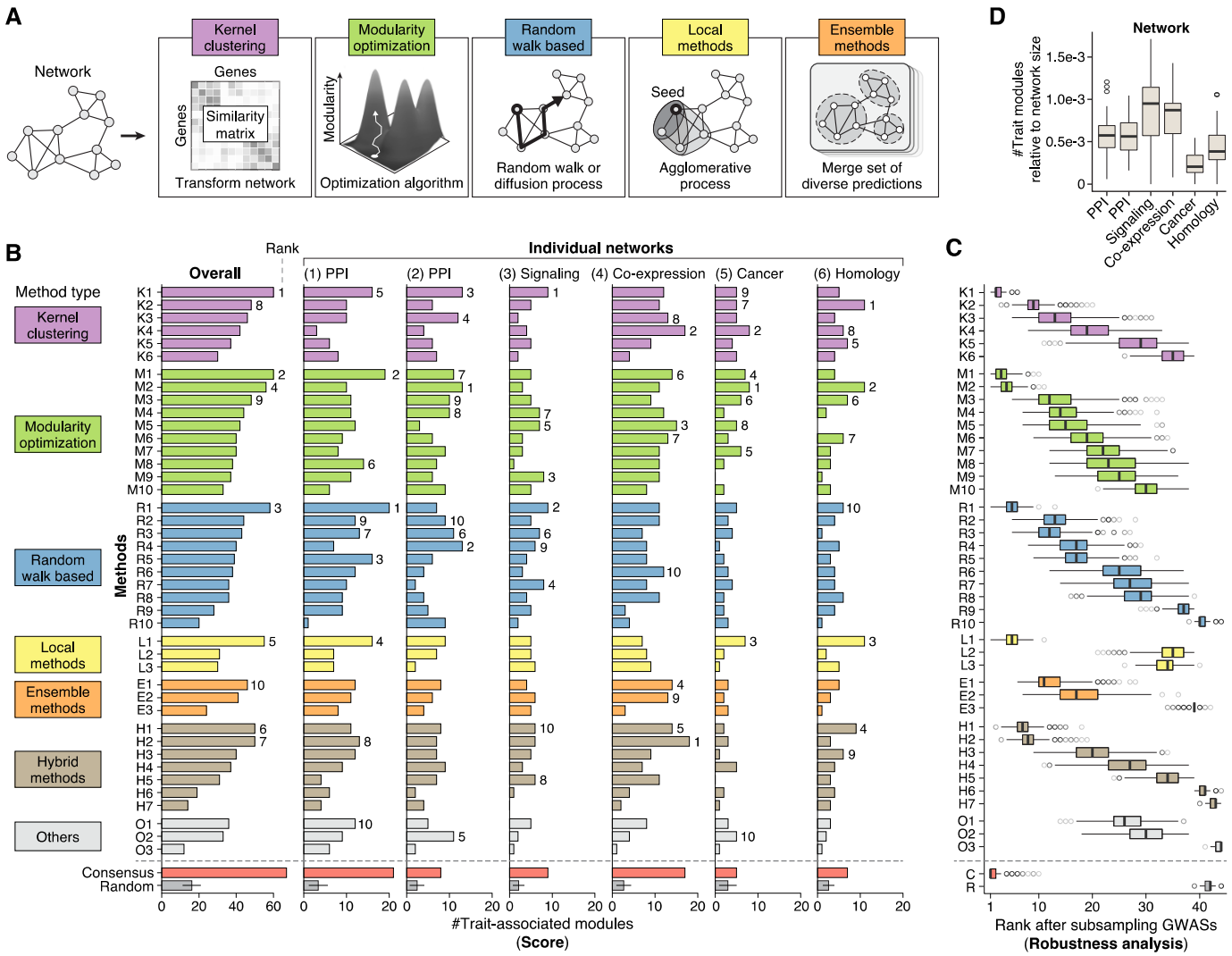


Figure 2: Assessment of module identification methods.

(A) Main types of module identification approaches used in the challenge: kernel clustering methods transform and cluster the network adjacency matrix; modularity optimization methods rely on search algorithms to find modular decompositions that maximize a structural quality metric; random-walk-based methods take inspiration from diffusion processes over the network; local methods use agglomerative processes to grow modules from seed nodes; and ensemble methods merge alternative clusterings sampled either from stochastic runs of a given method or from a set of different methods. In addition, hybrid methods employ more than one of the above approaches and then pick the best modules according to a quality metric. See also **Table 1**.

(B) Final scores of the 42 module identification methods applied in Sub-challenge 1 for each of the six networks, as well as the overall score summarizing performance across networks (same method identifiers as in **Table 1**). Scores correspond to the number of unique trait-associated modules identified by a given method in a network (evaluated using the hold-out GWAS set at 5% FDR, see Methods). Ranks are indicated for the top ten methods. The last two rows show the performance of consensus predictions derived from the challenge submissions and randomly generated modules, respectively.

(C) Robustness of the overall ranking was evaluated by subsampling the GWAS set used for evaluation 1,000 times. For each method, the resulting distribution of ranks is shown as a boxplot. The rankings of method *K1* are substantially better than those of the remaining teams (Bayes factor < 3, see Methods).

(D) Number of trait-associated modules per network. Boxplots show the number of trait-associated modules across methods, normalized by the size of the respective network. See also **Fig. S1B**.

Figure 3

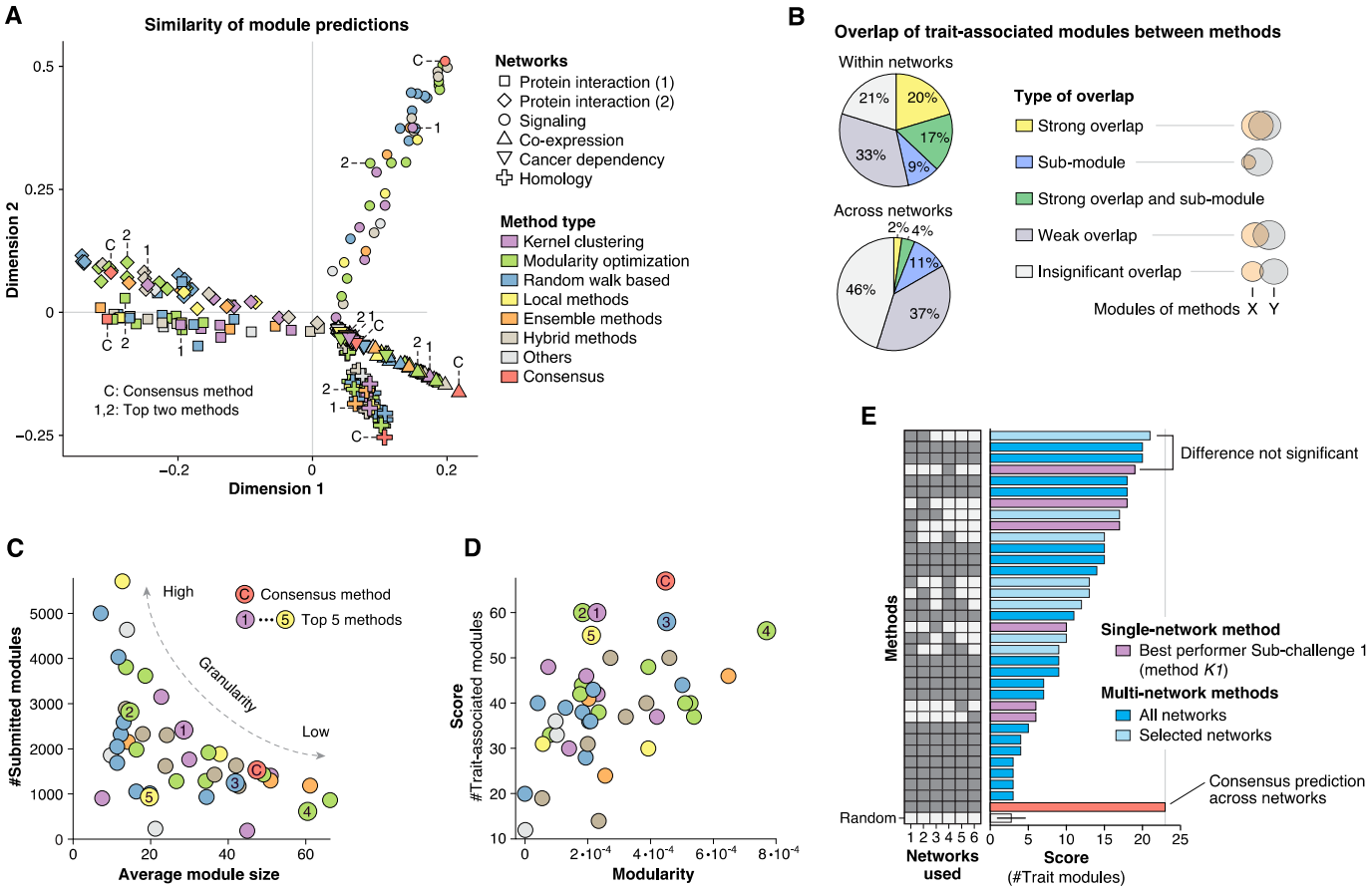


Figure 3: Complementarity of module predictions from different methods and networks.

(A) Similarity of module predictions from different methods (color) and networks (shape). The closer two points are in the plot, the more similar are the corresponding module predictions (multidimensional scaling, see Methods). Top performing methods tend to be located far from the origin (the top three methods are highlighted for each network). Top methods do not cluster close together, suggesting dissimilar modular decompositions (see also **Fig. S3A**).

(B) Comparison of GWAS trait-associated modules identified by all challenge methods. Pie-charts show the percentage of trait modules that show overlap with at least one trait module from a different method in the same network (top) and in different networks (bottom). We distinguish between strong overlap, sub-modules, weak but significant overlap, and insignificant overlap (Methods).

(C) Total number of predicted modules versus average module size for each method (same color scheme as in Panel A). There is a roughly inverse relationship between module number and size because modules had to be non-overlapping and did not have to cover all genes. The top five methods (highlighted) produced modular decompositions of varying granularity. See also **Figs. S3B-D**.

(D) Challenge score (number of trait-associated modules) versus modularity is shown for each method (same color scheme as in Panel A). Modularity is a topological quality metric for modules based on the fraction of within-module edges (Newman and Girvan, 2004). While there is modest correlation between the two metrics (r=0.45), the methods with the highest challenge score are not necessarily those with the highest modularity, presumably because the intrinsic scale of modularity is not optimal for the task considered in the challenge.

(E) Final scores of multi-network module identification methods in Sub-challenge 2 (evaluated using the hold-out GWAS set at 5% FDR, see Methods). For comparison, the overall best-performing method from Sub-challenge 1 is also shown (method *K1*, purple). Teams used different combinations of the six challenge networks for their multi-network predictions (shown on the left): the top-performing team relied exclusively on the two protein-protein interaction networks. The difference between the top single-network module predictions and the top multi-network module predictions is not significant when sub-sampling the GWASs (**Fig. S1D**). The last two rows show the performance of multi-network consensus predictions (obtained by integrating single-network submissions from Sub-challenge 1 across networks) and randomly generated module predictions, respectively.

Figure 4

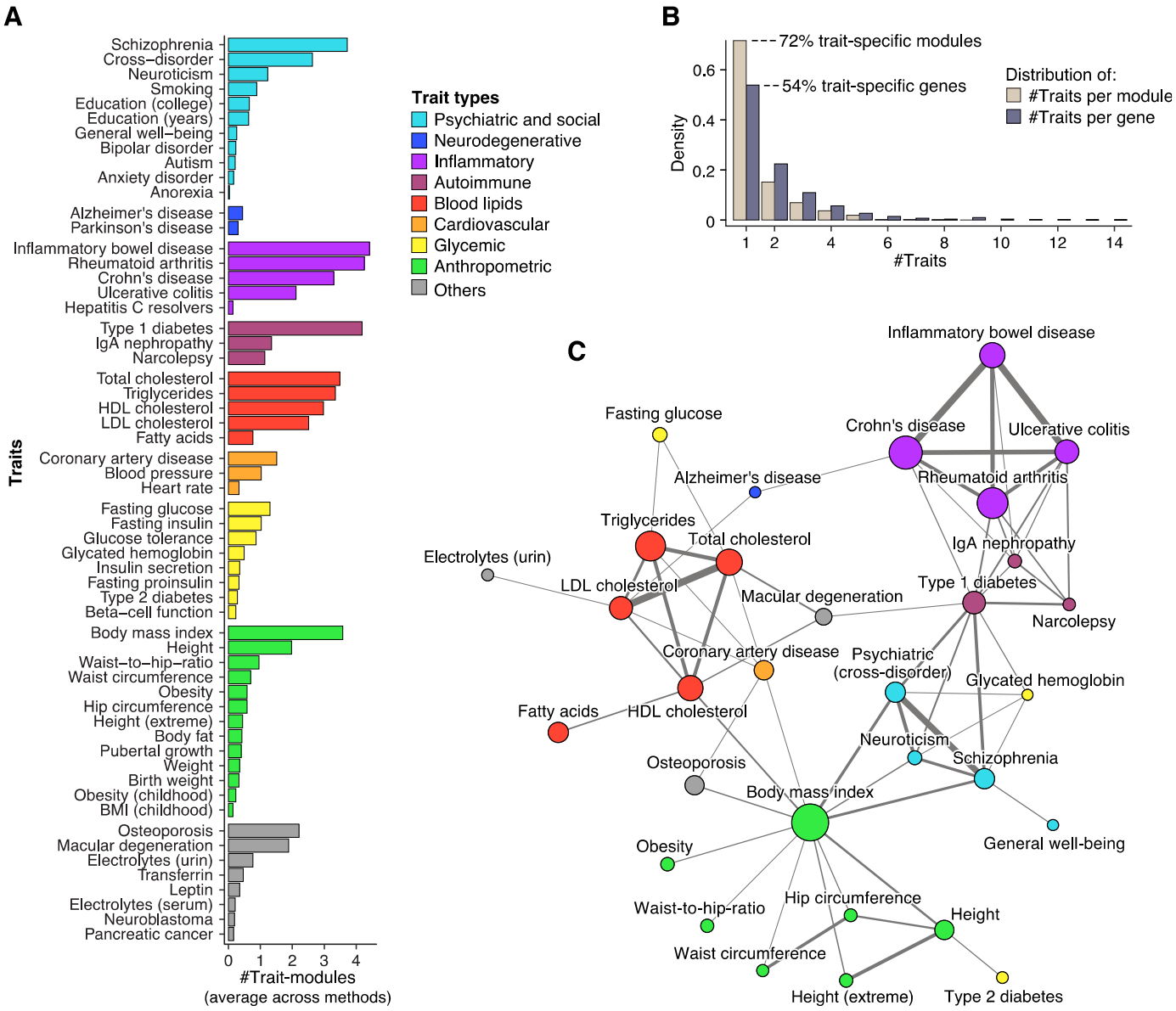


Fig. 4: Overlap between modules associated with different traits and diseases.

(A) Average number of trait-associated modules identified by challenge methods for each trait. For traits where multiple GWASs were available, results for the best-powered study are shown.

(B) Histograms showing the number of distinct traits per trait-associated module (brown) and gene (grey). 72% of trait-associated modules are specific to a single trait, while the remaining 28% are hits for multiple traits. In contrast, only 54% of trait-associated genes are specific to a single trait.

(C) Trait network showing similarity between GWAS traits based on overlap of associated modules (force-directed graph layout). Node size corresponds to the number of genes in trait-associated modules and edge width corresponds to the degree of overlap (Jaccard index; only edges for which the overlap is significant are shown, see Methods). Traits without any edges are not shown. Traits of the same type (color) tend to cluster together, indicating shared pathways.

Figure 5

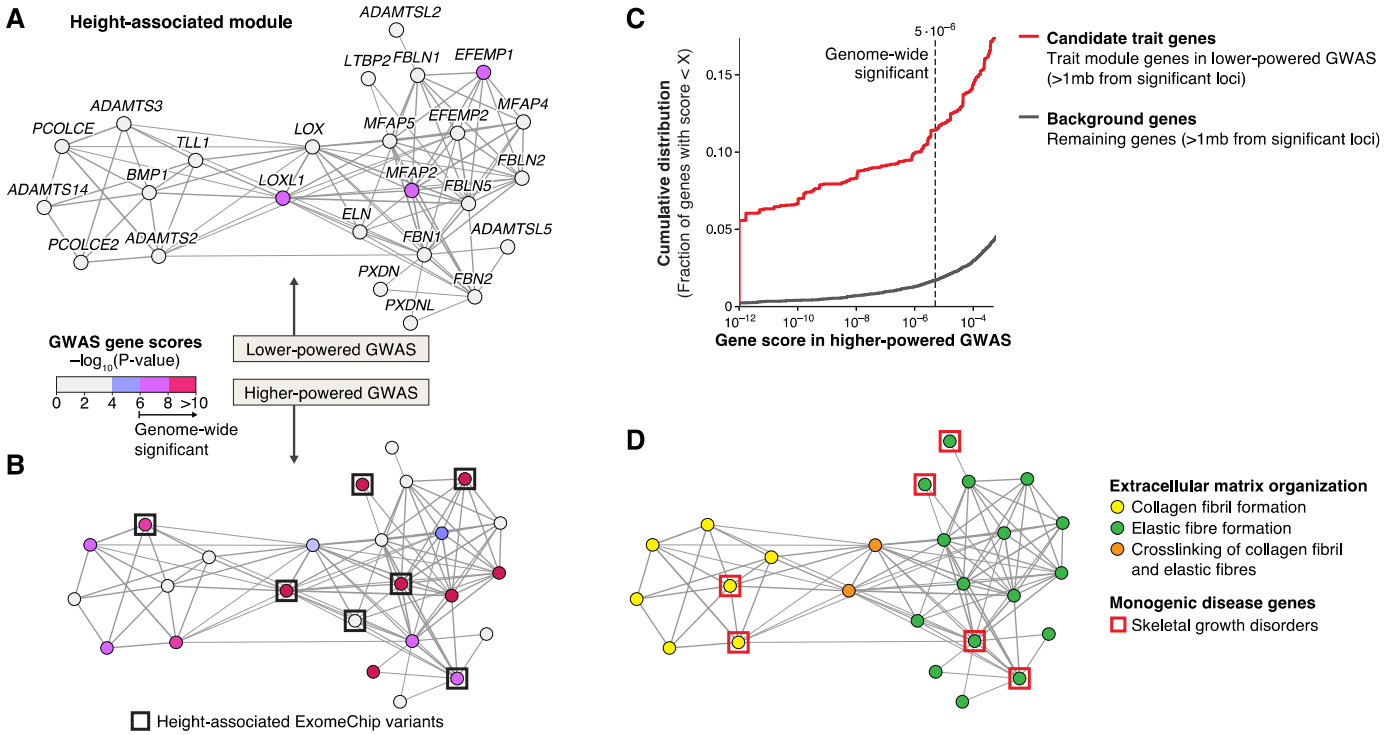


Figure 5: Support of trait-module genes in diverse datasets.

(A) Example module of the consensus method in the STRING protein interaction network (force-directed graph layout). The module shows modest association to height (*q*-value = 0.04) in the GWAS by Randall et al. (2013) (lower-powered than the GWAS shown in Panel B). Color indicates GWAS gene scores. The signal is driven by three genes from different loci with significant scores (pink), while the remaining genes (grey) are predicted to be involved in height because of their module membership.

(B) The module from **Panel A** is supported in the higher-powered GWAS (Wood et al., 2014) (*q*-value = 0.005). 45% of candidate trait genes (grey in **Panel A**) are confirmed (pink). In addition, 28% of module genes have coding variants associated to height in an independent ExomeChip study published after the challenge (Marouli et al., 2017) (black squares, enrichment *p*-value = 1.9E-6). See also **Fig. S4B**.

(C) Support of candidate trait genes across eight different traits for which lower- and higher-powered GWASs are available in our hold-out set. The lower-powered GWASs were used to predict candidate trait genes, i.e., genes within trait modules that do not show any signal (GWAS gene score <4) and that are located far away (>1mb) from any significant GWAS locus (cf. grey genes in **Panel A**). The plot shows the cumulative distribution of gene scores in the higher-powered GWASs for candidate trait genes (red line) and all other genes (grey line, see Methods).

(D) Functional annotation of genes in the height-associated module from **Panel A**. Genes implicated in monogenic skeletal growth disorders are highlighted (red squares, enrichment *p*-value = 7.5E-4). See also **Table S2**.

Figure 6

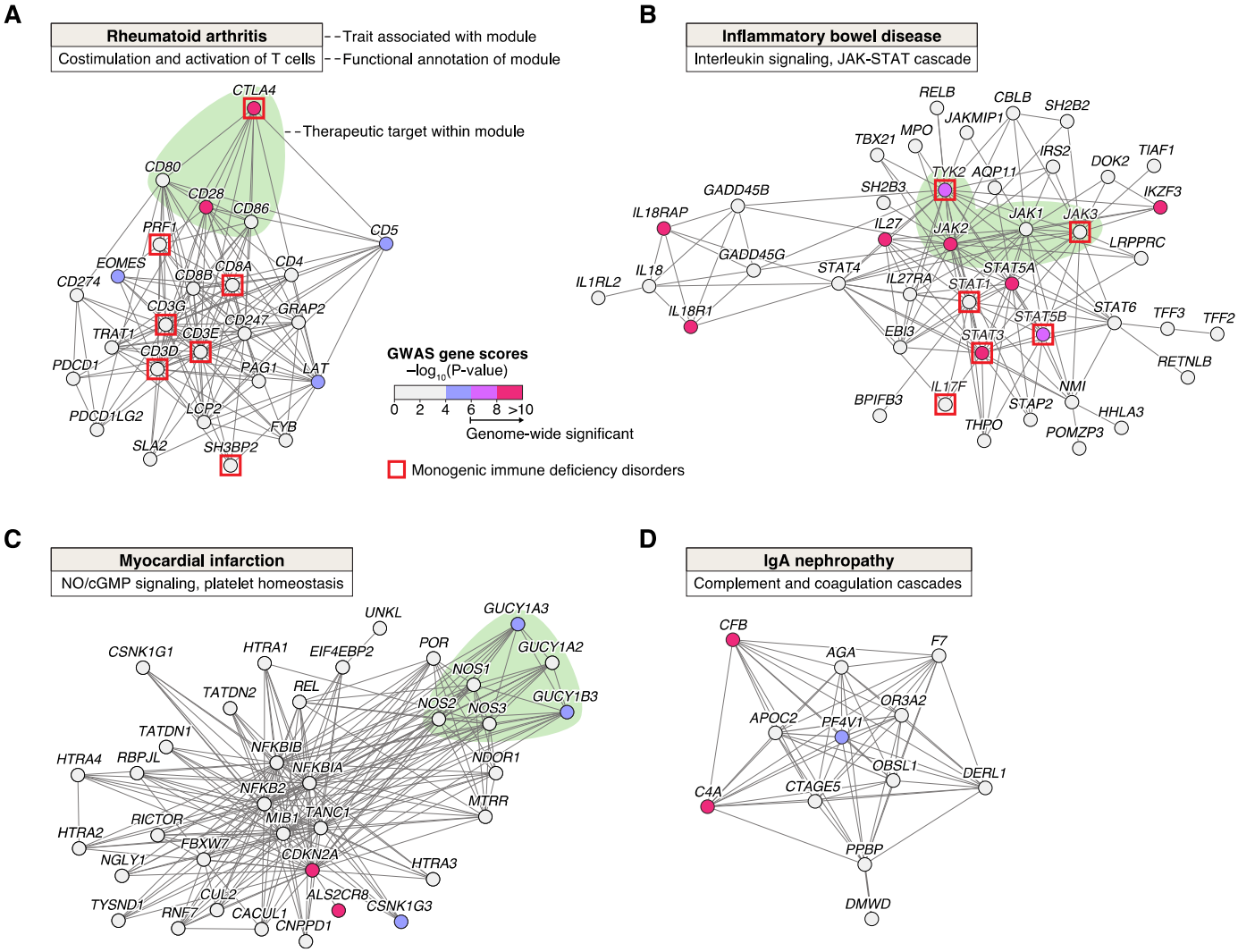


Figure 6: Example trait modules comprising therapeutically relevant pathways.

(A, B and C) Three trait-associated modules in the STRING protein interaction network identified using the consensus method (similar results were obtained for other modules and traits, **Tables S3, S4**). Node colors correspond to gene scores in the respective GWAS. For the two inflammatory disorders (A and B), red squares indicate genes causing monogenic immunodeficiency disorders (enrichment *p*-values of 4.1E-8 and 1.2E-6, respectively).

(A) Module associated with rheumatoid arthritis (*q*-value = 0.04) involved in T cell activation. A costimulatory pathway is highlighted green: T cell response is regulated by activating (*CD28*) and inhibitory (*CTLA4*) surface receptors, which bind B7 family ligands (*CD80* and *CD86*) expressed on the surface of activated antigen-presenting cells. The therapeutic agent CTLA4-Ig binds and blocks B7 ligands, thus inhibiting T cell response.

(B) A cytokine signalling module associated with inflammatory bowel disease (*q*-value = 0.0006). The module includes the four known Janus kinases (*JAK1-3* and *TYK2*, highlighted green), which are engaged by cytokine receptors to mediate activation of specific transcription factors (*STATs*). Inhibitors of JAK-STAT signaling are being tested in clinical trials for both ulcerative colitis and Crohn's disease (Neurath, 2017).

(C) Module associated with myocardial infarction (*q*-value = 0.0001). The module includes two main components of the NO/cGMP signaling pathway (highlighted green): endothelial nitric oxide synthases (*NOS1-3*), which produce the gas nitric oxide (NO) used as signal transmitter, and soluble guanylate cyclases (*GUCY1A2*, *GUCY1A3* and *GUCY1B3*), which sense NO leading to formation of cGMP. The cGMP signal inhibits platelet aggregation and leads to vascular smooth muscle cell relaxation; it is a therapeutic target for cardiovascular disease as well as erectile dysfunction (Kraehling and Sessa, 2017).

(legend continued on next page)

(D) Module associated to IgA nephropathy (IgAN; *q*-value = 0.04). The module was identified using the best-performing method (*K1*) in the InWeb protein interaction network. Besides finding complement factors that are known to play a role in the disease (*CFB* and *C4A*), the module implicates novel candidate genes such as the chemokine *Platelet Factor 4 Variant 1* (*PF4V1*) from a sub-threshold locus, and is enriched for coagulation cascade, a process known to be involved in kidney disease (Madhusudhan et al., 2016) (see also **Fig. S5**).

Methods

Network compendium

A collection of six gene and protein networks for human were provided by different groups for this challenge. The two protein-protein interaction and signaling networks are custom or new versions of existing interaction databases that were not publicly available at the time of the challenge. The remaining networks were yet unpublished at the time of the challenge. This was important to prevent participants from deanonymizing challenge networks by aligning them to the original networks. The original networks, anonymized networks and the mappings from gene symbols to anonymized IDs are available on the challenge website.

Networks were released for the challenge in anonymized form. Anonymization consisted in replacing the gene symbols with randomly assigned ID numbers. In Sub-challenge 1 each network was anonymized individually, i.e., node *k* of network A and node *k* of network B are generally not the same genes. In Sub-challenge 2 all networks were anonymized using the same mapping, i.e., node *k* of network A and node *k* of network B are the same gene. Since the networks were unpublished, it was practically impossible for participants to infer the gene identities. Participants also agreed not to attempt to infer gene identities as part of the challenge rules.

All networks are undirected and weighted, except for the signaling network, which is directed and weighted. Basic properties and similarity between the networks are shown in **Figs. 1A** and **S2E**. Below we briefly summarize each of the six networks. Detailed descriptions of networks 4, 5 and 6 are available on GeNets, a web platform for network-based analysis of genetic data (http://apps.broadinstitute.org/genets).

Network 1: STRING protein-protein interaction network

The first network was obtained from STRING, a database of known and predicted protein-protein interactions (Szklarczyk et al., 2015). STRING includes aggregated interactions from primary databases as well as computationally predicted associations. Both physical protein interactions (direct) and functional associations (indirect) are included. The challenge network corresponds to the human protein-protein interactions of STRING version 10.0, where interactions derived from text-mining were removed. Edge weights correspond to the STRING association score after removing evidence from text mining. The network was provided by Damian Szklarczyk and Christian von Mering (University of Zürich).

Network 2: InWeb protein-protein interaction network

The second network is the InWeb protein-protein interaction network (Li et al., 2017). InWeb aggregates physical protein-protein interactions from primary databases and the literature. The challenge network corresponds to InWeb version 3. Edge weights correspond to a confidence score that integrates the evidence of the interaction from different sources.

Network 3: OmniPath signaling network

The third network is the OmniPath signaling network (Türei et al., 2016). OmniPath integrates literature-curated human signaling pathways from 27 different sources, of which 20 provide causal interaction, 7 deliver undirected interactions. These data were integrated to form a directed weighted network. The edge weights correspond to a confidence score that summarizes the strength of evidence from the different sources.

Network 4: GEO co-expression network

The fourth network is a co-expression network based on Affymetrix HG-U133 Plus 2 arrays extracted from the Gene Expression Omnibus (GEO) (Barrett et al., 2011). In order to adjust for non-biological variation, data were rescaled by fitting a loess-smoothed power law curve to a collection of 80 reference genes (ten sets of ~8 genes each, representing different strata of expression) using nonlinear least squares regression within each sample. All samples were then quantile normalized together as a cohort. This approach is described fully in (Subramanian et al., 2017). After filtering out samples that did not pass quality control, a gene expression matrix of 22,268 probesets by 19,019 samples was obtained. Probes were mapped to genes by averaging and the pairwise Spearman correlation of genes across samples was computed. The matrix was thresholded to include the top 1M strongest positive correlations resulting in an undirected, weighted network. The edge weights correspond to the correlation coefficients.

Network 5: Achilles cancer co-dependency network

The fifth network is a functional gene network derived from the Project Achilles dataset v2.4.3 (Cowley et al., 2014). Project Achilles performed genome-scale loss-of-function screens in 216 cancer cell lines using massively parallel pooled shRNA screens. Cell lines were infected with a library of 54,000 shRNAs, each targeting one of 11,000 genes for RNAi knockdown (~5 shRNAs per gene). The proliferation effect of each shRNA in a given cell line could be assessed using Next Generation Sequencing. From these data, the dependency of a cell line on each gene (the

gene essentiality) was estimated using the ATARiS method. This led to a gene essentiality matrix of 11,000 genes by 216 cell lines. Pairwise correlations between genes were computed and the resulting co-dependency network was thresholded to the top 1M strongest positive correlations, analogous to how the co-expression network was constructed. Project Achilles data was kindly provided by Aviad Tsherniak and Barbara Weir (Broad Institute).

Network 6: CLIME homology-based network

The sixth network is a functional gene network based on phylogenetic relationships identified using the CLIME (clustering by inferred models of evolution) algorithm (Li et al., 2014). CLIME can be used to expand pathways (gene sets) with additional genes using an evolutionary model. Briefly, given a eukaryotic species tree and homology matrix, the input gene set is partitioned into evolutionarily conserved modules (ECMs), which are then expanded with new genes sharing the same evolutionary history. To this end, each gene is assigned a log-likelihood ratio (LLR) score based on the ECMs inferred model of evolution. CLIME was applied to 1,025 curated human gene sets from GO and KEGG using a 138 eukaryotic species tree, which resulted in 13,307 expanded ECMs. The network was constructed by adding an edge between every pair of genes that co-occurred in at least one ECM. Edge weights correspond to the mean LLR scores of the two genes.

Challenge structure

Participants were challenged to apply network module identification methods to predict functional modules (gene sets) based on network topology. Valid modules had to be non-overlapping (a given gene could be part of either zero or one module, but not multiple modules) and comprise between 3 and 100 genes. Modules did not have to cover all genes in a network. The number of modules per network was not fixed: teams could submit any number of modules for a given network (the maximum number was limited due to the fact that modules had to be non-overlapping). In Sub-challenge 1, teams were required to submit a separate set of modules for each of the six networks. In Sub-challenge 2, teams were required to submit a single set of modules by integrating information across multiple networks (it was permitted to use only a subset of the six networks).

The challenge consisted of a leaderboard phase and the final evaluation. The leaderboard phase was organized in four rounds, where teams could make repeated submissions and see their score on each network. Due to the high computational cost of scoring the module predictions on a large number of GWAS datasets (see next section), a limit for the number of

submissions per team was set in each round taking into consideration our computational resources and the number of participating teams. The total number of submissions that any given team could make over the four leaderboard rounds was thus limited to only 25 and 41 for the two sub-challenges, respectively. For the final evaluation, a single submission including method descriptions and code was required per team, which was scored on a separate set of GWASs after the challenge closed to determine the top performers.

The submission format and rules are described in detail on the challenge website (https://www.synapse.org/modulechallenge).

Challenge scoring

We have developed a novel framework to empirically assess module identification methods on molecular networks using GWAS data. In contrast to functional gene annotations and pathway databases such as GO, which sometimes originate from similar types of functional genomics data as the network modules, GWAS data are orthogonal to the networks and thus provide an independent means of validation. In order to cover diverse molecular processes, we compiled a large collection of 180 GWAS datasets from public sources. The collection was split into two sets of 76 and 104 GWASs used for the leaderboard phase and the final evaluation, respectively (**Table S1**).

Gene and module scoring using Pascal

SNP-trait association p-values from a given GWAS were integrated across genes and modules using the Pascal (pathway scoring algorithm) tool (Lamparter et al., 2016). Briefly, Pascal combines analytical and numerical solutions to efficiently compute gene and module scores from SNP p-values, while properly correcting for linkage disequilibrium (LD) correlation structure prevalent in GWAS data. To this end, LD information from a reference population is used (here, the European population of the 1000 Genomes Project was employed as we only included GWASs with predominantly European cohorts). Compared to alternative gene scoring methods that rely on Monte Carlo simulations, Pascal is about 100 times faster and more precise (Lamparter et al., 2016). The fast gene scoring is critical as it allows module genes that are in LD, and can thus not be treated independently, to be dynamically rescored. This amounts to fusing the genes of a given module that are in LD and computing a new score that takes the full LD structure of the corresponding locus into account. Finally, Pascal tests modules for enrichment in high-scoring (potentially fused) genes using a modified Fisher method, which avoids any p-value cutoffs inherent to standard binary enrichment tests. As background gene

set, the genes of the given network were used. Lastly, the resulting nominal module p-values were adjusted to control the FDR via the Benjamini-Hochberg procedure. A snapshot of the Pascal version used for the challenge is available on the challenge website.

Scoring metric

In Sub-challenge 1, the score for a given network was defined as the number of modules with significant Pascal p-values at a given FDR cutoff in at least one GWAS (called trait-associated modules). Thus, modules that were hits for multiple GWAS traits were only counted once. The overall score was defined as the sum of the scores obtained on the six networks (i.e., the total number of trait-associated modules across all networks). For the official challenge ranking a 5% FDR cutoff was defined, but performance was further reported at 10%, 2.5% and 1% FDR.

Module predictions in Sub-challenge 2 were scored using the exact same methodology and FDR cutoffs. The only difference to Sub-challenge 1 was that submissions consisted of a single set of modules (instead of one for each network) and there was thus no need to define an overall score. As background gene set, the union of all genes across the six networks was used.

Robustness analysis of challenge ranking

To gain a sense of the robustness of the ranking with respect to the GWAS data, we subsampled the set of 104 GWASs used for the final evaluation (called the “test set”) by drawing 76 GWASs (same number of GWASs as in the leaderboard set; note that we have to do subsampling rather than resampling of GWASs because the scoring counts the number of modules that are associated to at least one GWAS, i.e., including the same GWASs multiple times does not affect the score). We applied this approach to create 1,000 subsamples of the test set. The methods were then scored on each subsample.

The performance of every method *m* was compared to the highest-scoring method across the subsamples by the paired Bayes factor K_m. That is, the method with the highest overall score in the test set (all 104 GWASs) was defined as reference (i.e., method *K1* in Sub-challenge 1). The score *S(m, k)* of method *m* in subsample *k* was thus compared with the score *S(ref, k)* of the reference method in the same subsample *k*. The Bayes factor K_m is defined as the number of times the reference method outperforms method *m*, divided by the number of times method *m* outperforms or ties the reference method over all subsamples. Methods with $K_m < 3$ were considered a tie with the reference method (i.e., method *m* outperforms the reference in more than 1 out of 4 subsamples).

Module identification methods

Here we provide an overview of module identification approaches applied in the two sub-challenges, including a detailed description of the top-performing method. Full descriptions and code of all methods are available on the challenge website (https://www.synapse.org/modulechallenge).

Overview of module identification methods in Sub-challenge 1

Based on descriptions provided by participants, module identification methods were classified into different categories (**Fig. 2A**). Categories and corresponding module identification methods are summarized in **Table 1**. In the following, we first give an overview of the different categories and top-performing methods, and then describe common pre- and post-processing steps used by these methods:

- **Kernel clustering**. Instead of working directly on the networks themselves, these methods cluster a kernel matrix, where each entry (*i*, *j*) of that matrix represents the closeness of nodes *i* and *j* in the network according to the particular similarity function, or kernel that was applied. Some of the kernels that were applied are well-known for community detection, such as the exponential diffusion kernel based on the graph Laplacian (Kondor and Lafferty, 2002) employed by method *K6*. Others, such as the LINE embedding algorithm (Tang et al., 2015) employed by method *K3* and the kernel based on the inverse of the weighted diffusion state distance (Cao et al., 2013, 2014) employed by method *K1*, were more novel. Method *K1* was the best-performing method of the challenge and is described in detail below.

- **Modularity optimization.** This method category was, along with random-walk-based methods (see below), the most popular type of method contributed by the community. Modularity optimization methods use search algorithms to find a partition of the network that maximizes the modularity *Q* (commonly defined as the fraction of within-module edges minus the expected fraction of such edges in a random network with the same node degrees) (Newman and Girvan, 2004). The most popular algorithm was Louvain community detection (Blondel et al., 2008). At least eight teams employed this algorithm in some form as either their main method or one of several methods. The top team of the category (method *M1*), which ranked second overall, first sparsified networks by removing low confidence edges. A mixture of several established community detection algorithms was then employed in order to search for a partition that optimized

modularity. Importantly, these algorithms were extended with an additional resistance parameter that penalized merging of communities (Arenas et al., 2008); increasing the resistance parameter thus led to partitions with a larger number of communities. Communities above the size limit (100 nodes) were subdivided recursively by reapplying the same community detection algorithms to the corresponding subnetworks (see below).

- **Random-walk-based methods.** These methods take inspiration from random walks or diffusion processes over the network. Several teams used the established Walktrap (Pons and Latapy, 2005) and Infomap (Rosvall et al., 2009) algorithms. The top team of this category (method *R1*) used a sophisticated random-walk method based on multi-level Markov clustering (Satuluri et al., 2010). The method modifies basic Markov Clustering in two ways. First, a hierarchical view of the graph is considered by successively coarsening neighborhoods into fewer supernodes. The clustering is first run on the coarsened graph, enabling the detection of communities at varying scales. Second, a balance parameter is introduced that adjusts for nodes to preferentially join smaller communities, thus leading to more balanced community sizes. Similar to method *M1* described above, networks were first sparsified and communities above the size limit were recursively subdivided. While we did not include kernel methods in the "random walk" category, several of the successful kernel clustering methods used random-walk-based measures within their kernel functions.

- **Local methods**. Only three teams used local community detection methods, including agglomerative clustering and seed set expansion approaches. The top team of this category (method *L1*) first converted the adjacency matrix into a topology overlap matrix (Ravasz et al., 2002), which measures the similarity of nodes by their topological overlap based on the number of neighbor they have in common. The team then used the SPICi algorithm (Jiang and Singh, 2010), which iteratively adds adjacent genes to cluster seeds such as to improve their local density.

- **Hybrid methods**. Seven teams employed hybrid methods that leveraged clusterings produced by several of the different main approaches listed above. These teams applied more than one community detection method to each network in order to get larger and more diverse sets of predicted modules. The most common methods applied were Louvain (Blondel et al., 2008) hierarchical clustering, and Infomap (Rosvall et al., 2009). Two different strategies were used to select a final set of modules for submission: (1) choose a single method for each network according to performance in the leaderboard

round, and (2) select modules from all applied methods according to a topological quality score such as the modularity or conductance (Fortunato and Hric, 2016).

- **Ensemble methods.** Much like hybrid methods, ensemble methods leverage clusterings obtained from multiple community detection methods (or multiple stochastic runs of a single method). However, instead of selecting individual modules according to a quality score, ensemble methods merge alternative clusterings to obtain potentially more robust consensus predictions (Lancichinetti and Fortunato, 2012). Our method to derive consensus module predictions from team submissions is an example of an ensemble approach (described in detail below).

Besides the choice of the community detection algorithm, there are other steps that critically affected performance, including pre-processing of the network data, setting of method parameters, and post-processing of predicted modules. We describe successful approaches employed by challenge participants to address these issues below (pre- and post-processing steps of challenge methods are also summarized in **Table 1**):

- **Pre-processing**. Data pre-processing often plays a key role in the analysis of noisy data, such as biological network data. Most networks in the challenge were densely connected, including many edges of low weight that are likely noisy. Some of the top teams (e.g., *M1*, *R1*, *L1*) benefitted from sparsifying these networks by discarding weak edges before applying their community detection methods. An added benefit of sparsification is that it typically reduces computation time. Few teams also normalized the edge weights of a given network to make them either normally distributed or fall in the range between zero and one. Not all methods required pre-processing of networks, for example the top performing method (*K1*) was applied to the original networks without any sparsification or normalization steps.

- **Parameter setting**. Most community detection methods have parameters that need to be specified, typically to control the resolution of the clustering (the number and size of modules). While some methods have parameters that explicitly set the number of modules (e.g., the top-performing method *K1*), other methods have parameters that indirectly control the resolution (e.g., the resistance parameter of the runner-up method *M1*). Teams used the leaderboard phase to optimize the parameters of their method. Note that teams could make at most 25 submissions during the leaderboard phase, which limited the parameter space that could be explored in particular for methods with multiple parameters. While there were also methods that had no parameters to set (e.g.,

the classic Louvain algorithm), these methods have an intrinsic resolution that may not always be optimal for a given network and target application.

- **Post-processing**. Depending on the target application, the output of community detection methods may need to be post-processed. In biological networks, most methods typically lead to highly imbalanced module sizes. That is, some modules may be very small (e.g., just one or two genes), while others are extremely large (e.g., thousands of genes). Both extremes are generally not useful to gain biological insights at the pathway level. In the challenge, module sizes were thus required to be between 3 and 100 genes. Since current community detection methods generally do not allow such constraints on module size to be specified, teams used different post-processing steps to deal with modules outside of this range. A successful strategy employed by teams to break down large modules was to recursively apply their method to each of these modules. Alternatively, all modules of invalid size were merged and the community detection method was re-applied to the corresponding subnetwork. Finally, modules with less than three genes were often discarded (i.e., the corresponding genes were not included in any of the submitted modules). Some teams also discarded larger modules that were deemed low quality according to a topological metric, although this strategy was generally not beneficial.

Top-performing team method

The top-performing team developed a kernel clustering approach (method *K1*) based on a distance measure called Diffusion State Distance (DSD) (Cao et al., 2013, 2014), which they further improved for this challenge (Crawford et al., in preparation). DSD produces a more informative notion of proximity than the typical shortest path metric, which measures distance between pairs of nodes by the number of hops on the shortest path that joins them in the network. More formally, consider the undirected network $G(V, E)$ on the node set $V = \{v_1, v_2, v_3, \ldots, v_n\}$ with $|V| = n$. $He^t(v_x, v_y)$ is defined as the expected number of times that a random walk (visiting neighboring nodes in proportion to their edge weights) starting at node v_x and proceeding for some fixed t steps will visit node v_y (the walk includes the starting point, i.e., 0th step). Taking a global view, we define the n-dimensional vector $He^t(v_x)$ whose ith entry is the $He^t(v_x, v_i)$ value to network node v_i. Then the DSD^t distance between two nodes v_x and v_y is defined as the $L1$ norm of the difference of their He^t vectors, i.e.

$$DSD^t(v_x, v_y) = ||He^t(v_x) - He^t(v_y)||_1.$$

It can be shown that DSD is a metric and converges as $t \to \infty$, allowing DSD to be defined independently from the value *t* (Cao et al., 2013). The converged DSD matrix can be computed tractably, with an eigenvalue computation, as

$$DSD(v_x, v_y) = ||(1_x - 1_y)(I - D^{-1}A + W)^{-1}||_1,$$

where D is the diagonal degree matrix, A is the adjacency matrix, and W is the matrix where each row is a copy of π, the degrees of each of the nodes, normalized by the sum of all the vertex degrees (in the unweighted case; weighted edges can be normalized proportional to their weight), and 1_x and 1_y are the vectors that are zero everywhere except at position x and y, respectively. The converged DSD matrix was approximated using algebraic multigrid techniques (Crawford et al., in preparation). Note that for the signaling network, edge directions were kept and low-weight back edges were added so that the network was strongly connected; i.e. if there was a directed edge from $v_x$to v_y, an edge from v_y to v_x of weight equal to 1/100 of the lowest edge weight in the network was added.

A spectral clustering algorithm (Ng et al., 2001) was used to cluster the DSD matrix of a given network. Note that the spectral clustering algorithm operates on a similarity matrix (i.e., entries that are most alike have higher values in the matrix). However, the DSD matrix is a distance matrix (i.e., similar entries have low DSD values). The radial basis function kernel presents a standard way to convert the DSD matrix to a similarity matrix; it maps low distances to high similarity scores and vice-versa. Since the spectral clustering algorithm employed uses k-means as the underlying clustering mechanism, it takes a parameter k specifying the number of cluster centers. The leaderboard rounds were utilized to measure the performance of different k. Also note that spectral clustering produces clusters of size less than 3, and clusters of size more than 100. Whenever a cluster of size less than 3 was produced, those vertices were not included in any cluster for that network. Whenever a cluster of size more than 100 was produced, spectral clustering was called recursively to split that cluster into two subclusters (i.e., k=2) until all clusters were of size < 100.

The top-performing team also used a different algorithm to search for dense bipartite subgraph module structure in half of the challenge networks. However, a post-facto analysis of their results showed that this step contributed few modules and the score would have been similar with this additional procedure omitted (Crawford et al., in preparation).

Overview of module identification methods in Sub-challenge 2

In Sub-challenge 2, few teams employed dedicated multi-network community detection methods (De Domenico et al., 2015; Didier et al., 2015). The majority of teams first built an integrated network by merging either all six or a subset of the challenge networks, and then applied single-network methods (typically the same method as in Sub-challenge 1) to modularize the integrated network. For example, the team with highest score in Sub-challenge 2 merged the two protein interaction networks and then applied the Louvain algorithm to identify modules in the integrated network. The top performing team from Sub-challenge 1 also performed competitively in Sub-challenge 2. They applied their single-network method (*K1*) to an integrated network consisting of the union of all edges from the two protein interaction networks and the coexpression network.

Similar to Sub-challenge 1, teams used the leaderboard phase to set parameters of their methods. However, besides the parameters of the community detection method, there were additional choices to be made, whether to use all or only a subset of the six networks and how to integrate them.

Consensus module predictions

We developed an ensemble approach to derive consensus modules from a given set of team submissions (see **Fig. S2A** for a schematic overview). In Sub-challenge 1, a consensus matrix C^n was defined for each network n, where each element c_{ij} corresponds to the fraction of teams that put gene i and j together in the same module in this network. That is, c_{ij} equals one if all teams clustered gene i and j together, and c_{ij} equals zero if none of the teams clustered the two genes together. The top-performing module identification method (*K1*) was used to cluster the consensus matrix (i.e., the consensus matrix was considered a weighted adjacency matrix defining a functional gene network, which was clustered using the top module identification method of the challenge). Method *K1* has only one parameter to set, which is the number of cluster centers used by the spectral clustering algorithm (see previous section). This parameter was set to the median number of modules submitted by the considered teams for the given network. The consensus module predictions described in the main text were derived from the submissions of the top 50% teams (i.e., 21 teams) with the highest overall score on the leaderboard GWAS set. (Results for different cutoffs regarding the percentage of teams included are reported in **Fig. S2C**.)

Multi-network consensus modules were obtained by integrating team submissions from Sub-challenge 1 across all six networks using the same approach (see **Fig. S2B**). The same set of teams was considered (i.e., top 50% on the leaderboard GWAS set). First, a multi-network consensus matrix was obtained by taking the mean of the six network-specific consensus matrices C^n. The multi-network consensus matrix was then clustered using method *K1* as described above, where the number of cluster centers was set to the median number of modules submitted by the considered teams across all networks.

Two additional, more sophisticated approaches to construct consensus matrices C^n were tested: (1) normalization of the contribution of each module by the module size led to similar results as the basic approach described above, and (2) unsupervised estimation of module prediction accuracy using the Spectral Meta Learner ensemble method (Parisi et al., 2014) did not perform well in this context (**Fig. S2D**).

Similarity of module predictions

To define a similarity metric between module predictions from different methods, we represented module predictions as vectors. Namely, the set of modules predicted by method m in network k was represented as a prediction vector P_{mk} of length $N_k(N_k-1)/2$, where N_k is the number of genes in the network. Each element of this vector corresponds to a pair of genes and equals 1 if the two genes are in the same module and 0 otherwise. Accordingly, for any two module predictions (method m_1 applied to network k_1, and method m_2 applied to network k_2), we calculated the distance as follows:

$$D(m_1k_1,\, m_2k_2) \;=\; 1 - \frac{<P_{m_1k_1}, P_{m_2k_2}>}{||P_{m_1k_1}||_2 \; ||P_{m_2k_2}||_2}, \qquad (1)$$

where $<.,.>$ is the Euclidean inner product, $||.||_2$ is the Euclidean norm, and D is the (symmetric) distance matrix between the 252 module predictions submitted in Sub-challenge 1 (i.e., 42 methods applied to each of six networks). The distance matrix D was used as input to the Multidimensional Scaling (MDS) analysis for dimensionality reduction in **Fig. 3A**.

Similarity between method predictions across networks was calculated in the same way. To this end, the prediction vectors P_{mk} of method m for the six networks ($k = 1,2,\ldots,6$) were concatenated, forming a single vector P_m that represents the module predictions of that method for all six networks. A corresponding distance matrix between the 42 methods was computed

using the same approach as described above (Equation 1) and used as input for hierarchical clustering in **Fig. S3A**.

Overlap between trait-associated modules

Three different metrics were considered to quantify the overlap between trait-associated modules from different methods and networks. The first metric was the Jaccard index, which is defined as the size of the intersection divided by the size of the union of two modules (gene sets) A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

The Jaccard index measures how similar two modules are, but does allow the detection of sub-modules. For example, consider a module A of size 10 that is a submodule of a module B of size 100. In this case, even though 100% of genes of the first module are comprised in the second module, the Jaccard index is rather low (0.1). To capture sub-modules, we thus considered in addition the percentage of genes of the first module that are comprised in the second module:

$$S(A,B) = \frac{|A \cap B|}{|A|}.$$

Lastly, we also evaluated the significance of the overlap. To this end, we computed the p-value p_{AB} for the overlap between the two modules using the hypergeometric distribution. P-values were adjusted using Bonferroni correction given the number of module pairs tested.

Based on these three metrics, we categorized the type of overlap that a given trait-module A had with another trait-module B as:

(1) *strong overlap* if $J(A,B) \geq 0.5$ and $p_{AB} < 0.05$;

(2) *submodule* if $J(A,B) < 0.5$ and $S(A,B) - J(A,B) \geq 0.5$ and $p_{AB} < 0.05$;

(3) *partial overlap* if $J(A,B) < 0.5$ and $S(A,B) - J(A,B) < 0.5$ and $p_{AB} < 0.05$;

(4) *insignificant overlap* if $p_{AB} \geq 0.05$.

This categorization was used to get a sense of the type of overlap between trait modules from all methods (see **Fig. 3B**).

Trait similarity network

We defined a network level similarity between GWAS traits based on overlap between trait-associated modules. To this end, we only considered the most relevant networks for our collection of GWAS traits, i.e., the two protein interaction, the signaling and the co-expression network (see **Fig. 2D**). For a given network, the set of "trait-module genes" G_T was obtained for

every trait T by taking the union of the modules associated with that trait across all challenge methods. (If different GWASs were available for the same trait type (see **Table S1**), the union of all corresponding trait-associated modules was taken). The overlap between every pair of trait-module gene sets G_{T_1} and G_{T_2} was evaluated using the Jaccard index $J(G_{T_1}, G_{T_2})$ and the hypergeometric p-value $p_{T_1T_2}$ as described in the previous section. P-values were adjusted using Bonferroni correction. For the visualization as a trait-trait network in **Fig. 4C**, an edge between traits T_1 and T_2 was added if the overlap was significant ($p_{T_1T_2} < 0.05$) in at least three out of the four considered networks, and node sizes and edge weights were set proportional to the average number of trait-module genes and the average Jaccard index across the four networks, respectively.

Evaluation of candidate trait genes

Trait-associated modules comprise many genes that show only borderline or no signal in the corresponding GWAS (called "candidate trait genes"). To assess whether modules correctly prioritized candidate trait genes, we considered eight traits for which older (lower-powered) and more recent (higher-powered) GWAS datasets were available in our test set (**Fig. S4A**). This allowed us to evaluate how well trait-associated modules and candidate trait genes predicted using the lower-powered GWAS datasets were supported in the higher-powered GWAS datasets.

We only considered candidate trait genes that were predicted solely because of their membership in a trait-associated module, i.e., that did not show any signal in the lower-powered GWAS as defined by: (i) a high gene p-value (p > 1E-4, i.e., two orders of magnitude above the genome-wide significance threshold of 1E-6) and (ii) genomic location of more than one megabase away from the nearest significant locus of the corresponding GWAS. Gene p-values were computed using Pascal as described above (see "Gene and module scoring using the Pascal tool"). Finally, the Pascal p-value of all candidate trait genes was evaluated for the higher-powered GWAS. Since there is a genome-wide tendency for p-values to become more significant in higher-powered GWAS data (Boyle et al., 2017), Pascal p-values were also evaluated for a background gene set (all genes that meet the two conditions (*i*, *ii*) but do not belong to trait-associated modules of the lower-powered GWAS). **Fig. 5C** shows the cumulative distribution of Pascal p-values for the candidate trait genes as well as the background genes.

Functional enrichment analysis

In order to test network modules for enrichment in known gene functions and pathways, we considered diverse annotation and pathway databases. GO annotations for biological process, cellular component, and molecular functions were downloaded from the GO website (http://geneontology.org, accessed on January 20, 2017). Curated pathways (KEGG, Reactome, and BioCarta) were obtained from MSigDB version 5.2 (http://software.broadinstitute.org/gsea). We also created a collection of gene sets reflecting mouse mutant phenotypes, as defined by the Mammalian Phenotype Ontology (Blake et al., 2017). We started with data files HMD_HumanPhenotype.rpt and MGI_GenePheno.rpt, downloaded from the Mouse Genome Informatics database (http://www.informatics.jax.org) on February 21, 2016. The first file contains human-mouse orthology data and some phenotypic information; we then integrated more phenotypic data from the second file, removing the two normal phenotypes MP:0002169 ("no abnormal phenotype detected") and MP:0002873 ("normal phenotype"). For each remaining phenotype, we then built a list of all genes having at least one mutant strain exhibiting that phenotype, which we considered as a functional gene set.

Annotations from curated databases are known to be biased towards certain classes of genes. For example, some genes have been much more heavily studied than others and thus tend to have more annotations assigned to them. This and other biases lead to an uneven distribution of the number of annotations per genes (annotation bias). On the other hand, the gene sets (modules) tested for enrichment in these databases typically also exhibit bias for certain classes of genes (selection bias) (Glass and Girvan, 2014; Young et al., 2010). Standard methods for GO enrichment analysis use the hypergeometric distribution (i.e., Fisher's exact test), the underlying assumption being that, under the null hypothesis, each gene is equally likely to be included in the gene set (module). Due to selection bias, this is typically not the case in practice, leading to inflation of p-values (Glass and Girvan, 2014; Young et al., 2010). Following Young et al. (2010), we thus used the Wallenius non-central hypergeometric distribution to account for biased sampling. Corresponding enrichment p-values were computed for all network modules and annotation terms (pathways). The genes of the given network were used as a background gene set. For each network, module identification method, and annotation database, the $M \times T$ nominal p-values of the M modules and T annotation terms (pathways) were adjusted using Bonferroni correction.

Data and software availability

Challenge data, results, and code are available from the challenge website (https://synapse.org/modulechallenge). This includes:

- Official challenge rules;
- Gene scores for the compendium of 180 GWASs used in the challenge plus 5 additional GWASs obtained after the challenge (GWAS SNP p-values are available upon request);
- The molecular network collection (anonymized and deanonymized versions);
- Module identification method descriptions and code provided by teams;
- The final module predictions of all teams for both sub-challenges;
- Consensus module predictions for both sub-challenges;
- Method scores at varying FDR cutoffs;
- Individual module scores for all GWASs;
- Enriched functional annotations for all modules (GO, mouse mutant phenotypes, and diverse pathway databases);
- A snapshot of the PASCAL tool and scoring scripts.

The latest version of PASCAL and the source code is also available from the PASCAL website (https://www2.unil.ch/cbg/index.php?title=Pascal) and GitHub (https://github.com/dlampart/Pascal).

Supplementary Figures and Tables

Figure S1

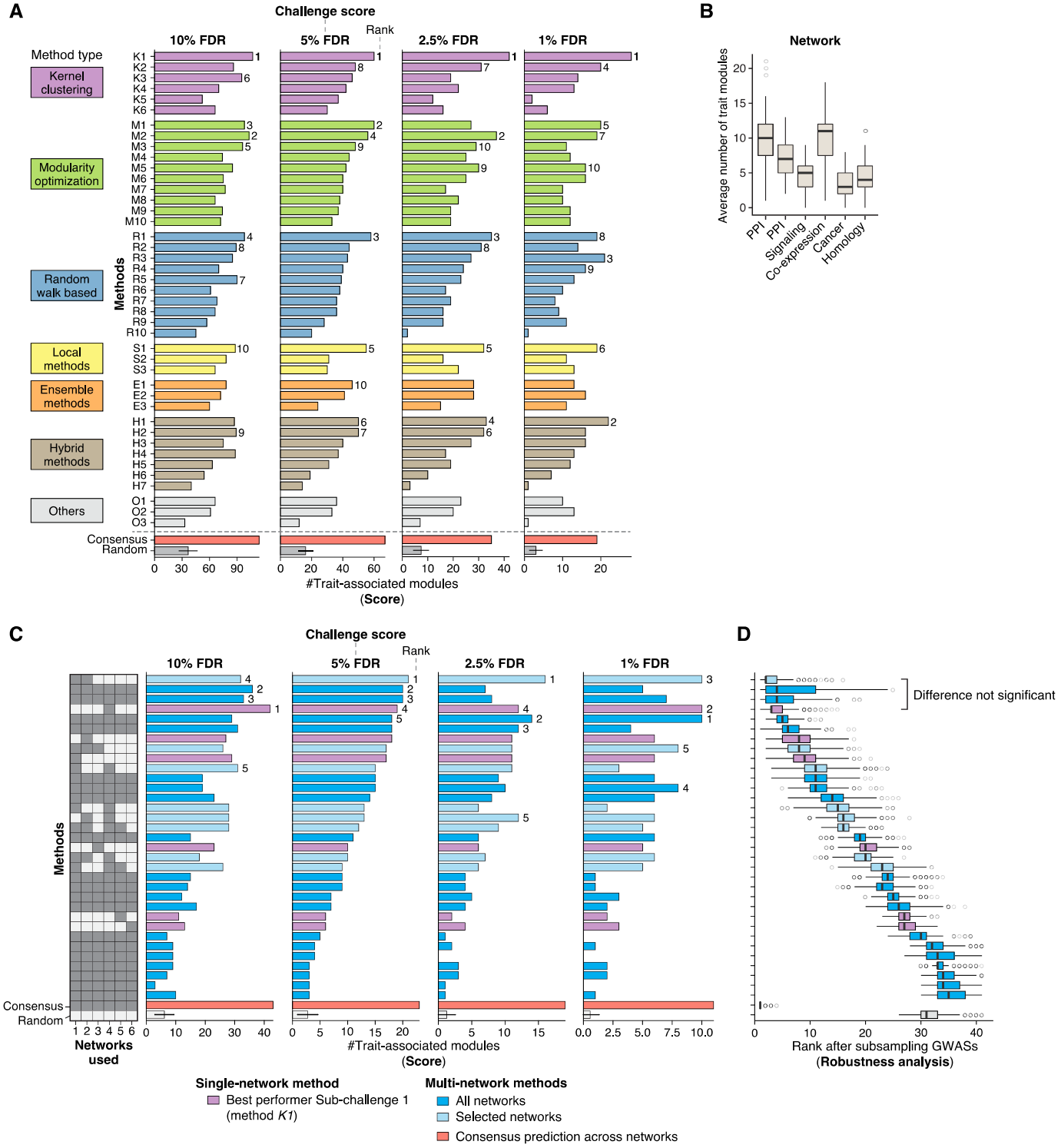


Figure S1. Assessment of Module Identification Methods, Related to Figures 2 and 3.

(legend on next page)

Figure S1. Assessment of Module Identification Methods, Related to Figures 2 and 3.

(A) Overall scores of the 42 module identification methods applied in Sub-challenge 1 at four different FDR cutoffs (10%, 5%, 2.5%, and 1% FDR). For explanation see legend of **Fig. 2B**, which shows the scores at 5% FDR (the predefined cutoff used for the challenge ranking). The top-performing method (*K1*) ranks first at all four cutoffs. The consensus prediction achieves the top score at 10% and 5% FDR, but not at the more stringent cutoffs.

(B) Average number of trait-associated modules across all methods for each of the six networks. The most trait modules are found in the two protein-protein interaction (PPI) and the co-expression networks. Related to **Fig. 2D**, which shows the average number of trait modules relative to network size.

(C) Final scores of multi-network module identification methods in Sub-challenge 2 at four different FDR cutoffs (10%, 5%, 2.5%, and 1% FDR). For explanation see legend of **Fig. 3E**, which shows the scores at 5% FDR (the predefined cutoff used for the challenge ranking). Ranks are indicated for the top five teams (ties are broken according to robustness analysis described in **Panel D**). The multi-network consensus prediction (red) achieves the top score at each FDR cutoff.

(D) Robustness of the overall ranking in Sub-challenge 2 was evaluated by subsampling the GWAS set used for evaluation 1,000 times. For each method, the resulting distribution of ranks is shown as a boxplot (using the 5% FDR cutoff for scoring). Related to **Fig. 2C**, which shows the same analysis for Sub-challenge 1. The difference between the top single-network module prediction and the top multi-network module predictions is not significant when sub-sampling the GWASs (Bayes factor < 3).

Figure S2

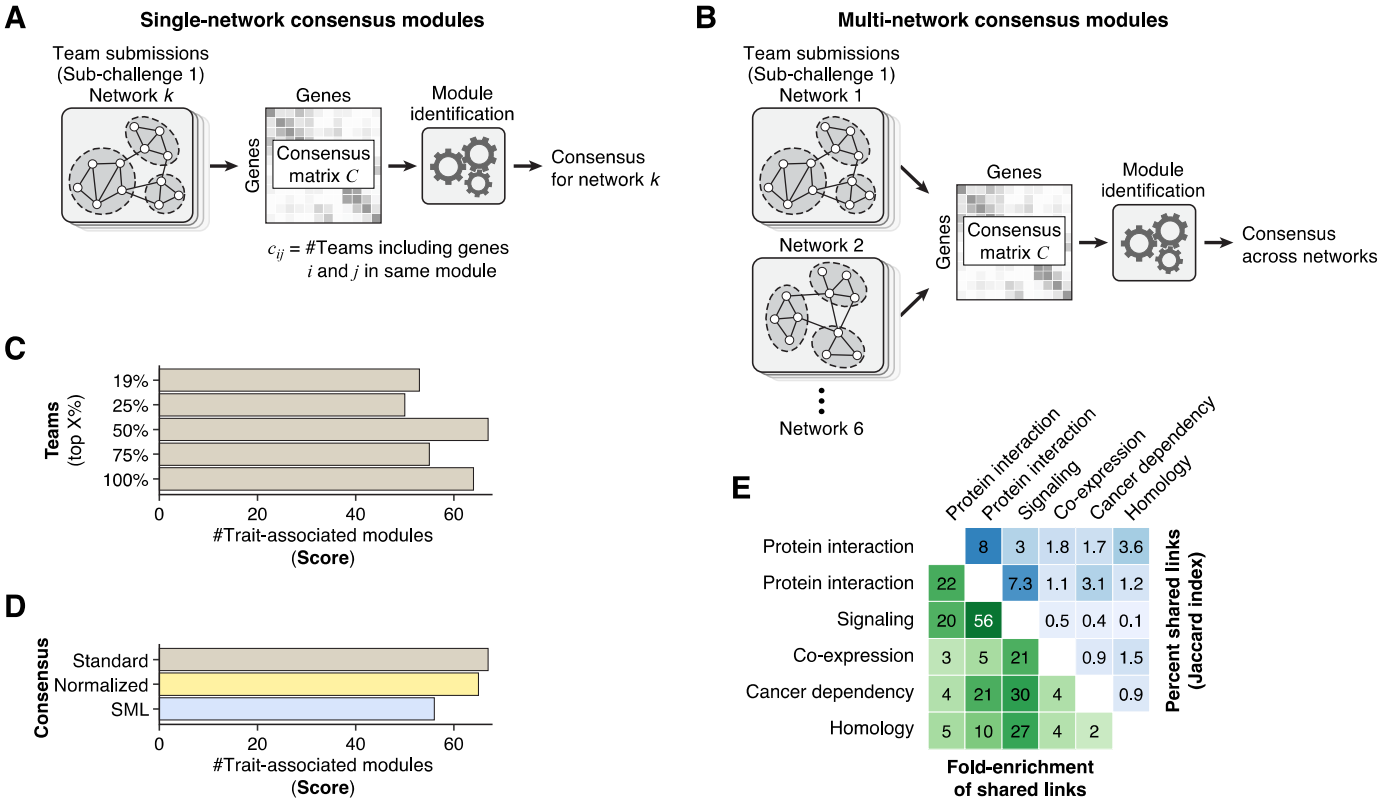


Figure S2. Consensus Module Predictions, Related to Figures 2 and 3.

(A) Schematic of the approach used to generate single-network consensus module predictions for Sub-challenge 1. For each network, module predictions from the top 50% of teams were integrated in a consensus matrix *C*, where each element c_{ij} gives the fraction of teams that clustered gene *i* and *j* together in the same module in the given network (performance as the percentage of considered teams is varied is shown in **Panel C**). The overall score from the leaderboard round was used to select the top 50% of teams, i.e., the same set of teams was used for each network. The consensus matrix of each network was then clustered using the top-performing module identification method of the challenge (method *K1*; see Methods).

(B) The approach used to generate multi-network consensus module predictions for Sub-challenge 2 was exactly the same as for single-network predictions, except that team submissions from all networks were integrated in the consensus matrix *C*. In other words, as input we still used the single-network predictions of the top 50% of teams from Sub-challenge 1, but instead of forming a consensus matrix for each network, a single cross-network consensus matrix was formed. This cross-network consensus matrix is then clustered using method *K1* as described above (see Methods).

(C) Scores of the single-network consensus predictions as the percentage of integrated teams is varied. We considered the top 25%, 50%, 75% and 100% of teams, as well as the top eight (19%) teams (these are the teams that ranked 2nd, or tied with the team that ranked 2nd, at any of the considered FDR cutoffs).

(D) Performance of different methods to construct the consensus matrix *C*. In addition to the basic approach described above (*Standard*), two more sophisticated approaches to construct the consensus matrix were evaluated (*Normalized* and *SML*). In each case, the same set of team submissions were integrated (top 50%) and method *K1* was applied to cluster the resulting consensus matrix.

The first alternative (*Normalized*) is similar to the basic method but further assumes that appearing together in a smaller cluster is stronger evidence that a pair of genes is associated than appearing together in a larger cluster.

(legend continued on next page)

Thus, each cluster's contribution to the consensus matrix was normalized by the size of the cluster. Furthermore, we normalized the *ij*-entry of the consensus matrix by the number of methods that assigned gene *i* to a cluster, thus taking the presence of background genes into account. We found that the consensus still achieved the top score with these normalizations, but there was no improvement compared to the basic approach.

The second method is a very different approach called Spectral Meta Learner (SML) (Parisi et al., 2014). SML is an unsupervised ensemble method designed for two-class classification problems. Briefly, it takes a matrix of predictions, P, where each row corresponds to different samples being classified and the columns correspond to different methods. Accordingly, each matrix element P_{ij} is the class (0 or 1) assigned to sample i by method j. Under the assumption of conditional independence of methods given class labels, SML can estimate the balanced accuracy of each classifier in a totally unsupervised manner using only the prediction matrix P. The algorithm then uses this information to construct an ensemble classifier in which the contribution of each classifier is proportional to its estimated performance (balanced accuracy). The module identification problem is an unsupervised problem by its nature and we applied the SML algorithm as a new way for constructing consensus modules. For each method m and network k, we created a vector of prediction P_{mk}, of size N_{G_k} by N_{G_k}, where N_{G_k} is the number genes in network as follows:

$$P_{mk}(i,j) = 1, \ \ if \ method \ m \ puts \ genes \ i \ and \ j \ in \ the \ same \ module \quad (1)$$
$$P_{mk}(i,j) = 0, \ otherwise.$$

For each network, we constructed the prediction matrix P with each column P_m defined as above. We then provided this matrix as input to the SML algorithm. The SML algorithm outputs a consensus matrix, which assigns a weight between each pair of genes. We found that SML did not perform well in the context of this challenge, likely because the underlying assumption of SML is that top-performing methods converge to similar predictions, which was not the case here (see **Figs. 3** and **S3**).

(E) Pairwise similarity of networks. The upper triangle of the matrix shows the percent of shared links (the Jaccard index multiplied by 100) and the lower triangle shows the fold-enrichment of shared links compared to the expected number of shared links at random. The two protein-protein interaction networks are the two most similar networks with 8% of shared edges.

Figure S3

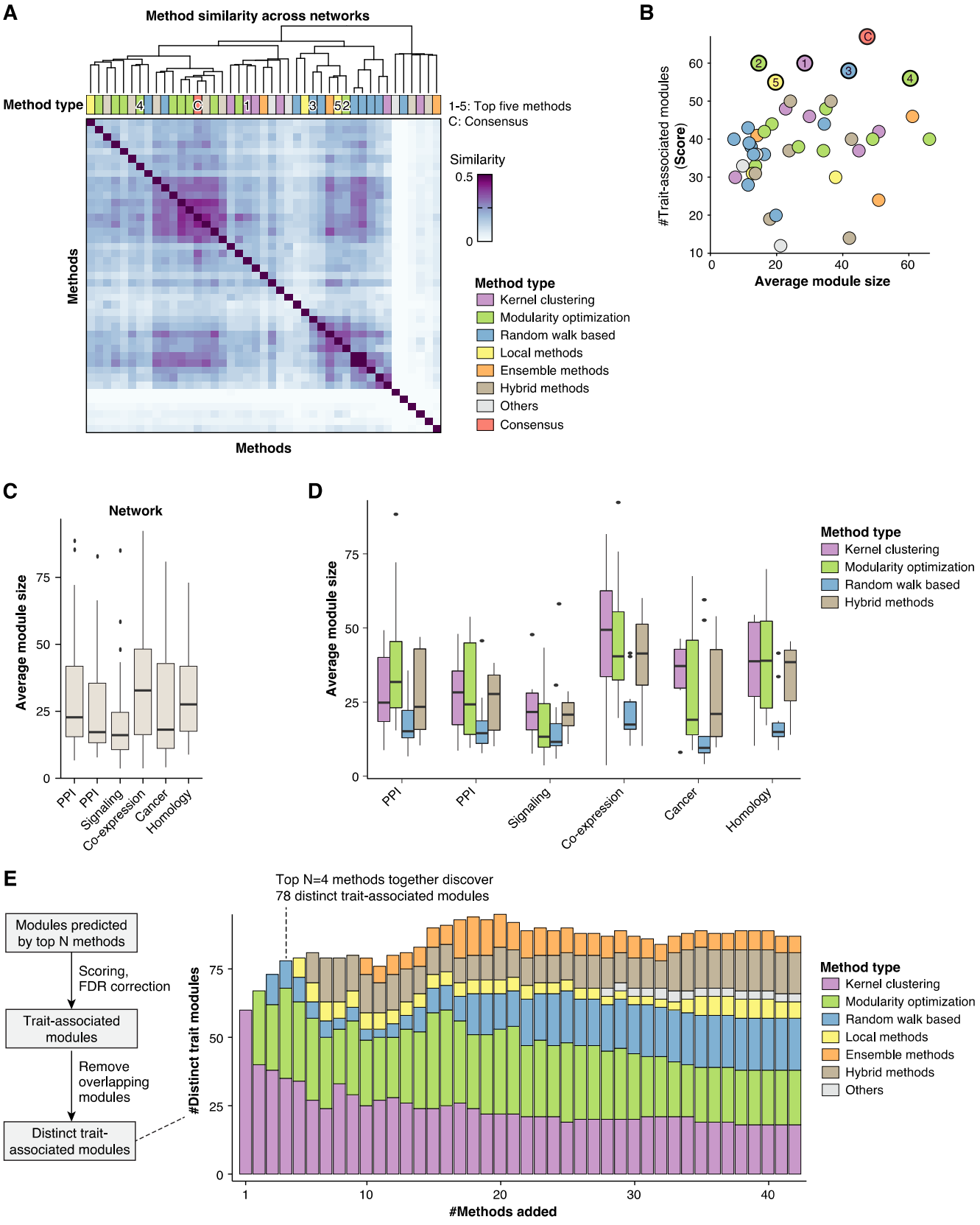


Figure S3. Complementarity of Module Identification Methods, Related to Figure 3. *(legend on next page)*

Figure S3. Complementarity of Module Identification Methods, Related to Figure 3.

All panels show results for single-network module identification methods (Sub-challenge 1).

(A) Pairwise similarity of module predictions from different methods, averaged over all networks. Similarity was computed based on whether the same genes were clustered together by the two methods (see Methods). The resulting similarity matrix was hierarchically clustered using Ward's method. The top row shows the method type. The top five methods (1-5) and the consensus (C) are highlighted. The top methods did not converge to similar module predictions (they are not all grouped together in the hierarchical clustering). Related to **Fig. 3**, which shows similarity of module predictions from individual networks.

(B) Average module size versus score for each method. The x-axis shows the average module size of a given method across the six networks. The y-axis shows the overall score of the method. Top teams (highlighted) produced modules of varying size, i.e., they did not converge to a similar module size during the leaderboard round. Methods that generated very small modules (average size < 10) were not among the top performers.

(C) Comparison of module sizes between networks. For each network, the boxplot shows the distribution of average module sizes of the 42 challenge methods. On average, modules were smallest in the signaling network and largest in the co-expression network.

(D) Comparison of module sizes between method types. For each network, boxplots show the distribution of average module sizes for kernel clustering, modularity optimization, random-walk-based, and hybrid methods (the remaining categories are not shown because they comprise only three methods each). Note that teams tuned the resolution (average module size) of their method during the leaderboard round. The variation in module size between different method categories and networks suggests that the optimal resolution is method- and network-specific. For example, teams using random-walk-based methods tended to choose a higher resolution (smaller average module size) than teams using kernel clustering or modularity optimization methods.

(E) Number of distinct trait-associated modules recovered by the top *K* methods. Given the top *K* methods, we considered the set including all modules predicted by these methods and scored them with the same pipeline as used for the individual methods in the challenge. We then evaluated how many "distinct" trait-associated modules were recovered by these methods. Distinct modules were defined as modules that do not show any significant overlap among each other. Overlap between pairs of modules was evaluated using the hypergeometric distribution and called significant at 5% FDR (Benjamini-Hochberg adjusted p-value < 0.05). From the set of trait-associated modules discovered by the top *K* methods, we thus derived the subset of distinct trait-associated modules (when several modules overlapped significantly, only the module with the most significant GWAS p-value was retained). Although the resulting scores (number of distinct trait-associated modules) cannot be directly compared with the challenge scores (because module predictions had to be strictly non-overlapping in the challenge), it is instructive to see how many distinct trait modules can be recovered when applying multiple methods. The stacked bars (colors) further show how many of the distinct trait modules are contributed by each method category. The number of distinct trait modules is not monotonically increasing as more methods are added because the larger sets of modules also increase the multiple testing burden of the GWAS scoring. The top four methods together discover 78 distinct trait-associated modules. Relatively little is gained by adding a higher number of methods.

Figure S4

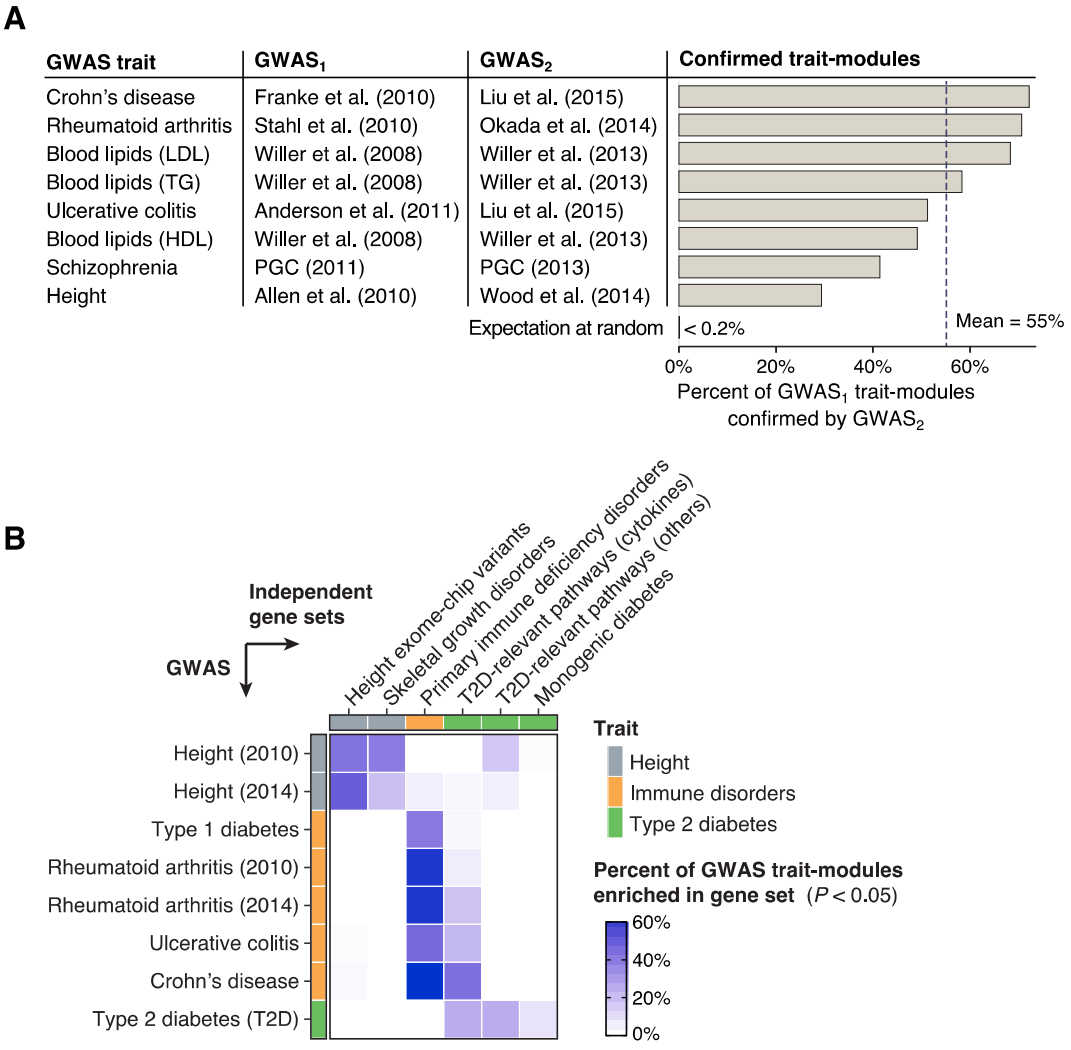


Figure S4. Support of Trait Modules in Diverse Datasets, Related to Figures 5 and 6.

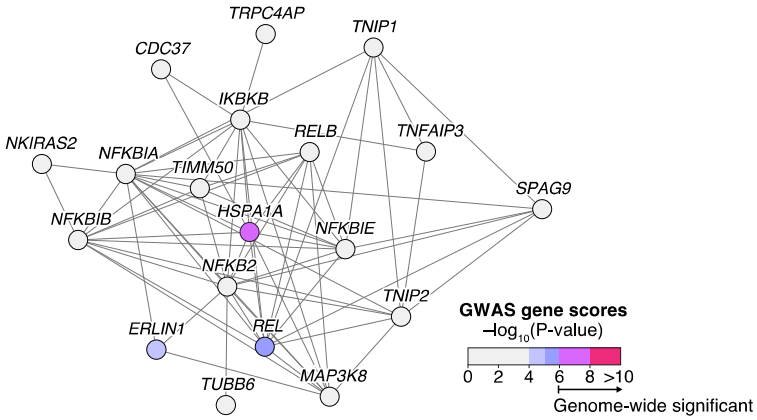
(A) Pairs of older (lower-powered) and more recent (higher-powered) GWASs used for the evaluation of module-based gene prioritization in **Fig. 5C**. The first column gives the trait and the second and third columns indicate the approximate cohort sizes of the respective GWASs. The bar plot shows the percentage of trait-associated modules from the first GWAS that are also trait-associated modules in the second GWAS. At the bottom, the expected percentage of confirmed modules at random is shown (i.e., assuming the trait-associated modules in the second GWAS were randomly selected from the set of predicted modules).

(B) Enrichment of trait-associated modules in six curated gene sets from three recent studies. The first two gene sets were taken from Marouli et al., (2017) and correspond to genes comprising height-associated ExomeChip variants and genes known to be involved in skeletal growth disorders, respectively. The third gene set was taken from de Lange et al., (2017) and corresponds to genes causing monogenic immunodeficiency disorders. Lastly, three gene sets relevant for type 2 diabetes (T2D) were taken from Fuchsberger et al. (2016) and correspond to genes in literature-curated pathways that are believed to be linked to T2D (we distinguished between genes in cytokine signalling pathways and other pathways) and genes causing monogenic diabetes. We then considered corresponding GWAS traits in our hold-out set, namely height, all immune-related disorders, and T2D. We then tested all modules associated with these GWAS traits for enrichment in these six external gene sets. Enrichment was tested using the hypergeometric distribution and p-values were adjusted to control FDR using the Benjamini-Hochberg method. The heatmap shows for each GWAS (row) the fraction of trait-associated modules that significantly overlap with a given gene set (column). It can be seen that modules associated with a given trait predominantly overlap the external gene sets that are expected to be relevant for that trait.

Figure S5

A

Module 283: IgA nephropathy, rheumatoid arthritis
NF-kappa B signaling



GO biological process	P-value
i-kappab kinase/nf-kappab signaling	1.4E-10
regulation of innate immune response	2.27E-08
positive regulation of nf-kappab transcription factor activity	3.44E-08
innate immune response-activating signal transduction	1.41E-06
stress-activated mapk cascade	1.44E-06
stress-activated protein kinase signaling cascade	1.62E-06
activation of innate immune response	1.65E-06
positive regulation of sequence-specific dna binding tf activity	1.73E-06
regulation of tumor necrosis factor-mediated signaling	2.92E-06
pattern recognition receptor signaling pathway	3.83E-06

Reactome pathways	P-value
rip mediated nfkb activation via dai	3.03E-08
traf6 mediated nfkb activation	3.94E-08
tak1 activates nfkb by phosphorylation and activation of ikks	7.65E-08
rig i mda5 mediated induction of ifn alpha beta pathways	1.4E-07
activation of nf kappab in b cells	6.95E-07
downstream signaling events of b cell receptor bcr	7.49E-06
il1 signaling	1.33E-05
signaling by the b cell receptor bcr	2.84E-05
nfkb and map kinases activation mediated by tlr4 signaling	3.67E-05
traf6 mediated induction of nfkb and map kinases upon tlr7 8 or 9 activation	3.99E-05

Mouse mutant phenotypes	P-value
decreased interleukin-6 secretion	7.14E-06
decreased b cell proliferation	2.86E-05
abnormal cytokine secretion	3.4E-05
small peyer's patches	3.92E-05
increased granulocyte number	4.48E-05
abnormal humoral immune response	5.66E-05
decreased igg1 level	6.25E-05
abnormal tumor necrosis factor secretion	9.45E-05
increased igm level	9.5E-05
increased b cell proliferation	0.00013

B

Module 203: IgA nephropathy
Complement and coagulation cascades

Visualized in **Fig. 6D**

GO biological process	P-value
regulation of complement activation	4.62E-07
regulation of humoral immune response	1.42E-06
regulation of acute inflammatory response	3.43E-06
complement activation	3.91E-06
regulation of protein processing	5.04E-06
positive regulation of leukocyte chemotaxis	4.07E-05
positive regulation of apoptotic cell clearance	5.11E-05
regulation of apoptotic cell clearance	5.25E-05
complement activation, alternative pathway	1E-04
positive regulation of leukocyte migration	1E-04

Reactome pathways	P-value
regulation of complement cascade	6.14E-06
complement cascade	8.49E-05
formation of fibrin clot clotting cascade	0.00158
initial triggering of complement	0.00187
innate immune system	0.00486
olfactory signaling pathway	0.0148
gamma carboxylation transport and amino terminal cleavage of proteins	0.0155
common pathway	0.0215
bmal1 clock npas2 activates circadian expression	0.0277
ptm gamma carboxylation hypusine formation and arylsulfatase activation	0.0318

Mouse mutant phenotypes	P-value
abnormal blood coagulation	3.2E-05
glomerulonephritis	0.00014
decreased vascular permeability	0.00015
increased spleen germinal center number	0.00029
decreased spleen germinal center size	0.0003
increased anti-single stranded dna antibody level	0.00043
abnormal spleen germinal center morphology	0.00044
decreased spleen germinal center number	0.00049
cortical renal glomerulopathies	0.00054
abnormal circulating protein level	0.00089

Figure S5. Modules Associated with IgA Nephropathy, Related to Fig. 6D.

(legend on next page)

Figure S5. Modules Associated with IgA Nephropathy, Related to Fig. 6D.

The top ten enriched GO biological processes, Reactome pathways and mouse mutant phenotypes are shown for two IgA nephropathy (IgAN) associated modules. P-values were computed using the non-central hypergeometric distribution, see Methods.

(A) IgAN-associated module identified using the consensus method in the InWeb protein-protein interaction network. The module comprises immune-related NF-κB signaling pathways. Enriched mouse mutant phenotypes for module gene homologs include perturbed immunoglobulin levels (IgM and IgG1). The module implicates in particular the NF-κB subunit *REL* as a candidate gene. The *REL* locus does not reach genome-wide significance in current GWASs for IgAN but is known to be associated with other immune disorders such as rheumatoid arthritis.

(B) Enriched annotations for the IgAN-associated module shown in **Fig. 6D**. The module comprises complement and coagulation cascades. The top two enriched mouse mutant phenotypes are precisely “abnormal blood coagulation” and “glomerulonephritis”. See main text for discussion.

Table S1

Provided as Excel file (Table_S1.xlsx)

Table S1: Collection of GWAS Datasets used for the Challenge.

The table lists the GWAS datasets used for the module scoring. The first column indicates whether the GWAS was used during the "leaderboard" or "final" evaluation phase. The five GWAS listed in the end ("extra") were not used for the scoring as they were added to the collection after the challenge. The PASCAL gene scores for all GWAS are available for download from the challenge website (file names are given in the last column). The original GWAS SNP summary statistics can be downloaded individually from the indicated sources or we can share the complete collection upon request.

Table S2

Module 175: Height
Extracellular structure organization

Mouse mutant phenotypes	P-value
abnormal cutaneous elastic fiber morphology	9.70E-04
emphysema	7.90E-03
loose skin	2.20E-02

Reactome pathways	P-value
collagen formation	9.70E-07
extracellular matrix organization	1.50E-05

GO biological process	P-value
extracellular matrix organization	0.00E+00
extracellular structure organization	0.00E+00
embryonic eye morphogenesis	6.10E-04
post-embryonic eye morphogenesis	6.20E-03
post-embryonic aaaaaorgan morphogenesis	1.40E-02
extracellular matrix disassembly	1.80E-02
collagen fibril organization	2.00E-02
supramolecular fiber organization	2.70E-02

Module 111: Inflammatory bowel disease
Interleukin signaling, JAK-STAT cascade

Mouse mutant phenotypes	P-value
decreased interferon-gamma secretion	0.00E+00
abnormal t-helper 1 physiology	5.70E-03
increased susceptibility to experimental autoimmune encephalomyelitis	1.20E-02

Reactome pathways	P-value
il 6 signaling	4.80E-05
growth hormone receptor signaling	1.80E-04
signaling by ils	2.80E-04
signaling by scf kit	1.50E-03
cytokine signaling in immune system	8.40E-03
il 7 signaling	1.10E-02
signaling by fgfr1 fusion mutants	3.70E-02

GO biological process	P-value
adaptive immune response	0.00E+00
cytokine-mediated signaling pathway	1.20E-09
cellular response to cytokine stimulus	1.30E-07
response to cytokine	2.50E-07
regulation of lymphocyte activation	9.50E-07
regulation of cell activation	3.40E-06
jak-stat cascade	5.70E-06
positive regulation of cell activation	3.30E-04
jak-stat cascade involved in growth hormone signaling pathway	1.00E-02

Module 161: Rheumatoid arthritis
Costimulation and activation of T cells

Mouse mutant phenotypes	P-value
decreased cd8-positive, alpha-beta t cell number	3.00E-05
increased interferon-gamma secretion	1.80E-03
decreased t cell proliferation	2.00E-03
increased double-negative t cell number	3.80E-03
abnormal t cell differentiation	4.10E-03
decreased double-positive t cell number	5.20E-03
abnormal t cell physiology	7.00E-03
abnormal regulatory t cell physiology	8.00E-03
abnormal double-negative t cell morphology	1.40E-02
abnormal cd8-positive, alpha beta t cell morphology	3.30E-02

Reactome pathways	P-value
generation of second messenger molecules	0.00E+00
tcr signaling	0.00E+00
pd1 signaling	0.00E+00
costimulation by the cd28 family	0.00E+00
phosphorylation of cd3 and tcr zeta chains	4.60E-07
translocation of zap 70 to immunological synapse	1.30E-05
immunoregulatory interactions	3.30E-05
downstream tcr signaling	4.90E-05

GO biological process	P-value
immune response-activating cell surface receptor signaling	0.00E+00
immune response-activating signal transduction	0.00E+00
immune response-regulating signaling pathway	0.00E+00
immune response-regulating cell surface receptor signaling	0.00E+00
regulation of cell-cell adhesion	0.00E+00
positive regulation of cell-cell adhesion	0.00E+00
regulation of cell adhesion	0.00E+00
lymphocyte costimulation	0.00E+00
t cell costimulation	0.00E+00
t cell activation	0.00E+00
positive regulation of cell adhesion	0.00E+00
positive regulation of immune response	0.00E+00
antigen receptor-mediated signaling pathway	0.00E+00
t cell receptor signaling pathway	0.00E+00
regulation of t cell activation	0.00E+00
regulation of cell activation	0.00E+00
positive regulation of cell activation	0.00E+00
positive regulation of t cell activation	0.00E+00
regulation of lymphocyte activation	0.00E+00
adaptive immune response	4.80E-05
positive regulation of interleukin-2 biosynthetic process	1.30E-04
regulation of interleukin-2 biosynthetic process	1.00E-03
positive regulation of t cell proliferation	1.60E-03
regulation of t cell proliferation	2.30E-03
positive regulation of interleukin-2 production	6.80E-03
positive regulation of lymphocyte proliferation	7.20E-03
positive regulation of mononuclear cell proliferation	7.40E-03

Module 283: Myocardial infarction
Cyclic GMP signaling, platelet homeostasis

Mouse mutant phenotypes	P-value
-	-

Reactome pathways	P-value
nitric oxide stimulates guanylate cyclase	2.4E-06
platelet homeostasis	1.3E-02

GO biological process	P-value
positive regulation of cgmp biosynthetic process	2.1E-03
nitric oxide mediated signal transduction	2.6E-03
positive regulation of cgmp metabolic process	3.5E-03
regulation of cgmp biosynthetic process	4.4E-03
regulation of cgmp metabolic process	1.6E-02

Table S2: Functional Enrichment for Example Modules, Related to Figs. 5 and 6.

Enrichment p-values for mouse mutant phenotypes, Reactome pathways and GO biological processes are shown for four example modules discussed in the main text (**Figs. 5 and 6**). P-values were computed using the non-central hypergeometric distribution and adjusted using the Bonferroni method (Methods). Results for the remaining trait-associated modules from the consensus method in the STRING protein interaction network are shown in **Table S4**. Functional enrichment analysis for additional pathway databases and modules from all methods and networks are available on the challenge website.

Table S3

Modules

ID	Trait type	GWAS trait	Module function / pathways	References
13	Psychiatric	Psychiatric (cross-disorder)	Neural development, chemical synaptic transmission	Table S4
21	Cardiovascular	Cardiovascular disease	Lipid and steroid hormone metabolism	Table S4
24	Blood lipids	Total cholesterol	Cholesterol homeostasis	Table S4
36	Inflammatory	Rheumatoid arthritis	Interferon signaling, antigen processing	Table S4
65	Blood lipids	LDL cholesterol	Protein ubiquitination	Table S4
72	Psychiatric	Neuroticism	Mitotic cell cycle	Table S4
77	Psychiatric, anthropometric	Schizophrenia, BMI	Chromatin organization, epigenetic gene silencing	Table S4
80	Glycemic	Fasting glucose	Glycerophospholipid and fatty acid metabolism	Table S4
109	Psychiatric, anthropometric	Schizophrenia, BMI	Nucleosome organization	Table S4
111	Inflammatory	Inflammatory bowel disease	Interleukin signaling, JAK-STAT cascade	Fig. 6B, Table S2
126	Anthropometric	Overweight	Activation of MAPK and JNK cascades	Table S4
161	Inflammatory	Rheumatoid arthritis	Costimulation and activation of T cells	Fig. 6A, Table S2
175	Anthropometric	Height	Extracellular structure organization	Fig. 5, Table S2
248	Anthropometric	Overweight	Lipid and insulin signaling, B cell activation	Table S4
262	Anthropometric	Height	Pre-NOTCH transcription and translation	Table S4
270	Inflammatory	Ulcerative colitis	MHC class II antigen presentation	Table S4
271	Inflammatory	Inflammatory bowel disease	Interleukin signaling, regulation of JAK-STAT cascade	Table S4
283	Cardiovascular	Myocardial infarction	Cyclic GMP signaling, platelet homeostasis	Fig. 6C, Table S2
294	Anthropometric	Body mass index (women)	Phospholipid metabolism and transport	Table S4
323	Inflammatory	Crohn's disease	TNF-mediated signaling	Table S4
326	Blood lipids	LDL cholesterol	Plasma lipoprotein particle clearance	Table S4

Module comprises:
- Core trait-specific pathways
- Other pathways

Table S3: Overview of Consensus Trait-modules in the STRING Network, Related to Fig. 6.

Overview of all 21 trait-associated modules discovered by the consensus method in the STRING protein-protein interaction network. The first three columns give the module ID, the trait type, and the specific GWAS trait that the module is associated to. We tested all modules for enrichment in GO annotation, mouse mutant phenotypes, and other pathway databases using the noncentral hypergeometric test (Methods). The putative function of each module based on this enrichment analysis is summarized in the fourth column (see **Figs. 5**, **6** and **Tables S2**, **S4** for details). Two thirds of the modules have functions that correspond to core pathways underlying the respective traits, while the remaining modules correspond either to generic pathways that play a role in diverse traits or to pathways without an established connection to the considered trait or disease. Only pathways with a well-established link to the trait were considered core pathways. Generic pathways, such as cell-cycle-related or epigenetic pathways, were not considered core pathways because they are relevant for many traits and tissues, making them more difficult to target therapeutically. For example, modules 77 and 109 are both associated with schizophrenia and comprise pathways related to epigenetic gene silencing and nucleosome organization, respectively. Although there is evidence that epigenetic mechanisms may play a role in schizophrenia, we considered this to be a generic pathway.

Table S4

Provided as Excel file (Table_S4.xlsx)

Table S4: Functional Enrichment of Consensus Trait Modules.

For each of the 21 consensus trait-modules shown in **Table S3**, all categories with a Bonferroni-corrected P-value below 0.05 are listed (Methods). Only results for mouse mutant phenotypes, Reactome pathways and GO biological process annotations are included for brevity. Full results including all tested pathway databases and all challenge modules are available on the challenge website.

C A multiplex seed-expanding approach to describe the landscape of premature aging diseases: supplementary materials

Supplementary Materials

Supplementary Tables

Table S-1: Premature aging phenotypes retrieved from HPO.

HPO Term Identifier	Term Name
HP:0007495	Prematurely aged appearance
HP:0002216	Premature graying of hair
HP:0004771	Premature graying of body hair
HP:0005328	Progeroid facial appearance
HP:0008509	Aged leonine appearance
HP:0100678	Premature skin wrinkling
HP:0000607	Periorbital wrinkles
HP:0007392	Excessive wrinkled skin
HP:0007407	Excessive skin wrinkling on dorsum of hands and fingers
HP:0007605	Excessive wrinkling of palmar skin
HP:0007517	Palmoplantar cutis laxa
HP:0007414	Neonatal wrinkled skin of hands and feet
HP:0009762	Facial wrinkling

Table S-2: *ORPHANET* diseases (Rath et al., 2012) in which at least one PA aging phenotype has been described.

Disease ID	Disease Name	Causative Genes
ORPHA:100	Ataxia-telangiectasia	ATM
ORPHA:101028	Transaldolase deficiency	TALDO1
ORPHA:1262	Böök syndrome	
ORPHA:1297	Branchio-oculo-facial syndrome	TFAP2A
ORPHA:1299	Branchioskeletogenital syndrome	CDH11
ORPHA:1318	Campomelia, Cumming type	
ORPHA:1340	Cardiofaciocutaneous syndrome	BRAF; MAP2K1; KRAS; MAP2K2
ORPHA:137608	Segmental outgrowth-lipomatosis-arteriovenous	PTEN
ORPHA:1387	Cataract-intellectual disability-hypogonadism syndrome	RAB3GAP1; RAB3GAP2
ORPHA:156156	Lipoatrophy with diabetes, leukomelanodermic papules	
ORPHA:163746	Peripheral demyelinating neuropathy-central dysmyelinating	SOX10
ORPHA:1775	Dyskeratosis congenita	NOP10; PARN; NHP2; USB1; TERC; TINF2; RTEL1; TERT; DKC1; CTC1; WRAP53
ORPHA:1807	Focal facial dermal dysplasia type III	TWIST2
ORPHA:1860	Thanatophoric dysplasia type 1	FGFR3
ORPHA:1901	Ehlers-Danlos syndrome, dermatosparaxis type	ADAMTS2
ORPHA:191	Cockayne syndrome	ADAMTS2
ORPHA:1979	Lipodystrophy due to peptidic growth factors deficiency	
ORPHA:2067	GAPO syndrome	ANTXR1
ORPHA:2078	Geroderma osteodysplastica	PYCR1; GORAB
ORPHA:209	Cutis laxa	
ORPHA:220295	Xeroderma pigmentosum-Cockayne syndrome complex	ERCC2; ERCC3; ERCC4; ERCC5
ORPHA:221008	Rothmund-Thomson syndrome type 1	
ORPHA:221016	Rothmund-Thomson syndrome type 2	RECQL4
ORPHA:228240	Elastoderma	
ORPHA:2500	Acrogeria	COL3A1
ORPHA:2617	Microcephalic primordial dwarfism, Montreal type	
ORPHA:263534	Acral peeling skin syndrome	CSTA; TGM5
ORPHA:2658	Lenz-Majewski hyperostotic dwarfism	PTDSS1
ORPHA:280365	Autosomal semi-dominant severe lipodystrophic laminopathy	LMNA
ORPHA:2834	Wrinkly skin syndrome	ATP6V0A2
ORPHA:286	Ehlers-Danlos syndrome, vascular type	COL3A1; COL5A1
ORPHA:2909	Rothmund-Thomson syndrome	
ORPHA:2959	Progeria-short stature-pigmented nevi syndrome	
ORPHA:2962	De Barsy syndrome	
ORPHA:2963	Progeroid syndrome, Petty type	SLC25A24
ORPHA:2976	Pseudoleprechaunism syndrome, Patterson type	
ORPHA:3051	Intellectual disability-sparse hair-brachydactyly syndrome	SMARCA2
ORPHA:3163	SHORT syndrome	PIK3R1
ORPHA:3322	Hoyeraal-Hreidarsson syndrome	ACD; PARN; TINF2; RTEL1; TERT; DKC1
ORPHA:3342	Arterial tortuosity syndrome	SLC2A10
ORPHA:33445	Neuroectodermal melanolysosomal disease	MYO5A
ORPHA:3437	Vogt-Koyanagi-Harada disease	FAS; PTPN22
ORPHA:3440	Waardenburg syndrome	
ORPHA:3455	Wiedemann-Rautenstrauch syndrome	
ORPHA:357074	Autosomal recessive cutis laxa type 2, classic type	ATP6V1E1; ATP6V0A2; ATP6V1A
ORPHA:363618	LMNA-related cardiocutaneous progeria syndrome	LMNA
ORPHA:381	Griscelli disease	
ORPHA:412057	Autosomal recessive cerebellar ataxia due to STUB1 deficiency	STUB1
ORPHA:435628	Keppen-Lubinsky syndrome	KCNJ6
ORPHA:500	Noonan syndrome with multiple lentigines	BRAF;PTPN11; RAF1
ORPHA:50811	Lipodystrophy-intellectual disability-deafness syndrome	
ORPHA:633	Laron syndrome	GHR
ORPHA:66633	Sensorineural hearing loss-early graying-essential tremor syndrome	

Disease ID	Disease Name	Causative Genes
ORPHA:740	Hutchinson-Gilford progeria syndrome	LMNA; ZMPSTE24
ORPHA:75496	Ehlers-Danlos syndrome, progeroid type	B4GALT7; B3GALT6
ORPHA:758	Pseudoxanthoma elasticum	ABCC6; ENPP1
ORPHA:769	Rabson-Mendenhall syndrome	INSR
ORPHA:79086	Acquired generalized lipodystrophy	
ORPHA:79087	Acquired partial lipodystrophy	LMNB2
ORPHA:79397	Epidermolysis bullosa simplex with mottled pigmentation	KRT14; KRT5
ORPHA:79474	Atypical Werner syndrome	LMNA
ORPHA:79476	Griscelli disease type 1	MYO5A
ORPHA:79477	Griscelli disease type 2	RAB27A
ORPHA:808	Seckel syndrome	ATR; CEP152; TRAIP; CENPE; CENPJ; PCNT; RBBP8; PLK4; ATRIP
ORPHA:870	Down syndrome	
ORPHA:894	Waardenburg syndrome type 1	PAX3
ORPHA:895	Waardenburg syndrome type 2	EDNRB; SOX10; KITLG; MITF; SNAI2
ORPHA:897	Waardenburg-Shah syndrome	EDN3; EDNRB; SOX10
ORPHA:90153	Mandibuloacral dysplasia with type A lipodystrophy	LMNA
ORPHA:90154	Mandibuloacral dysplasia with type B lipodystrophy	ZMPSTE24
ORPHA:902	Werner syndrome	WRN
ORPHA:90348	Autosomal dominant cutis laxa	FBLN5; ELN; ALDH18A1
ORPHA:90349	Autosomal recessive cutis laxa type 1	FBLN5; EFEMP2
ORPHA:90350	Autosomal recessive cutis laxa type 2	
ORPHA:904	Williams syndrome	TBL2; LIMK1; GTF2IRD1; CLIP2; ELN; BAZ1B; GTF2I; RFC2
ORPHA:920	Ablepharon macrostomia syndrome	TWIST2
ORPHA:99876	Ehlers-Danlos syndrome type 7B	COL1A2

Table S-3: Networks used in this study, number of nodes, number of edges and network densities.

Network	**Number of nodes**	**Number of edges**	**Density**
PPI	12 803	68 530	8.36×10^{-4}
Pathways	10 010	227 310	4.54×10^{-3}
Co-expression	10 131	922 424	1.80×10^{-2}
Complexes	8 650	91 502	2.44×10^{-3}
Disease similarity	3 193	12 941	2.54×10^{-3}

Supplementary Figures

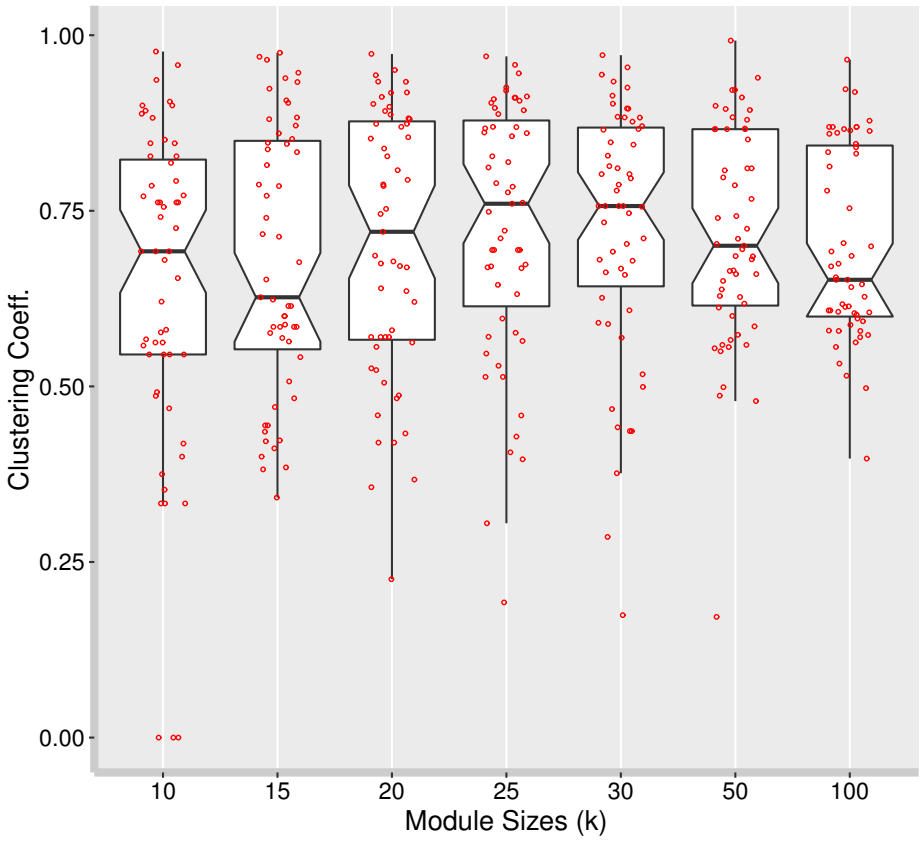


Figure S-1: Clustering coefficients for all the modules of size k. Modules are identified from repRWR-M on multiplex network for all PA diseases associated to at least one causative gene.

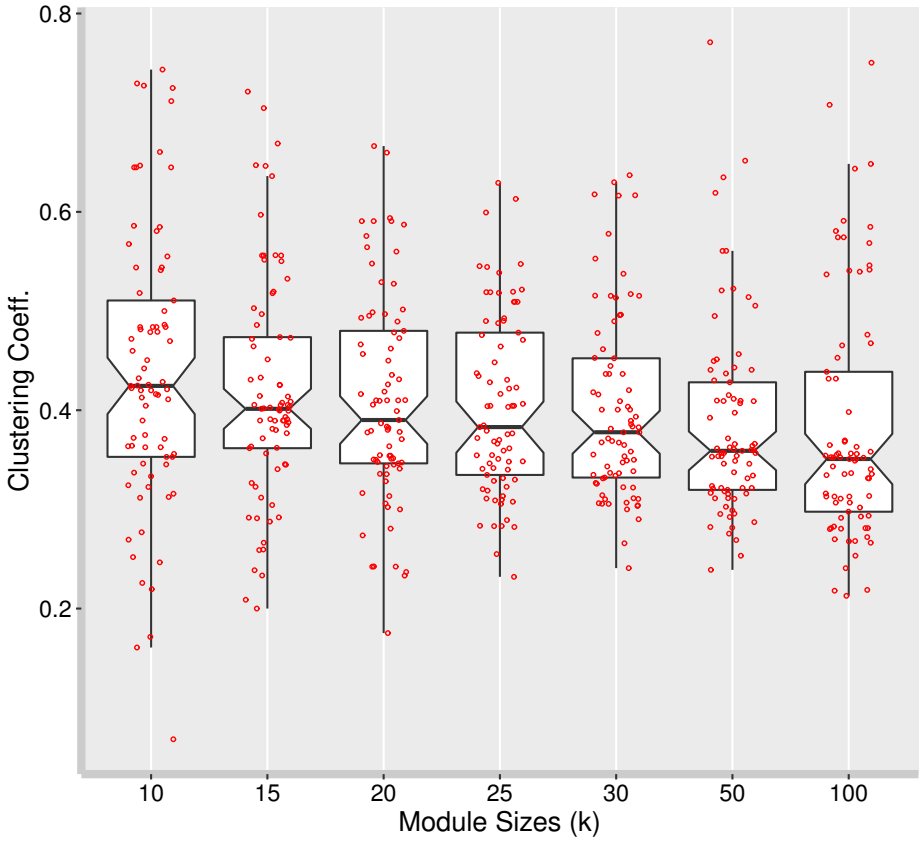


Figure S-2: Clustering coefficients for all the modules of genes, one module per PA disease, of different sizes generated by Repeated RWR-MH.

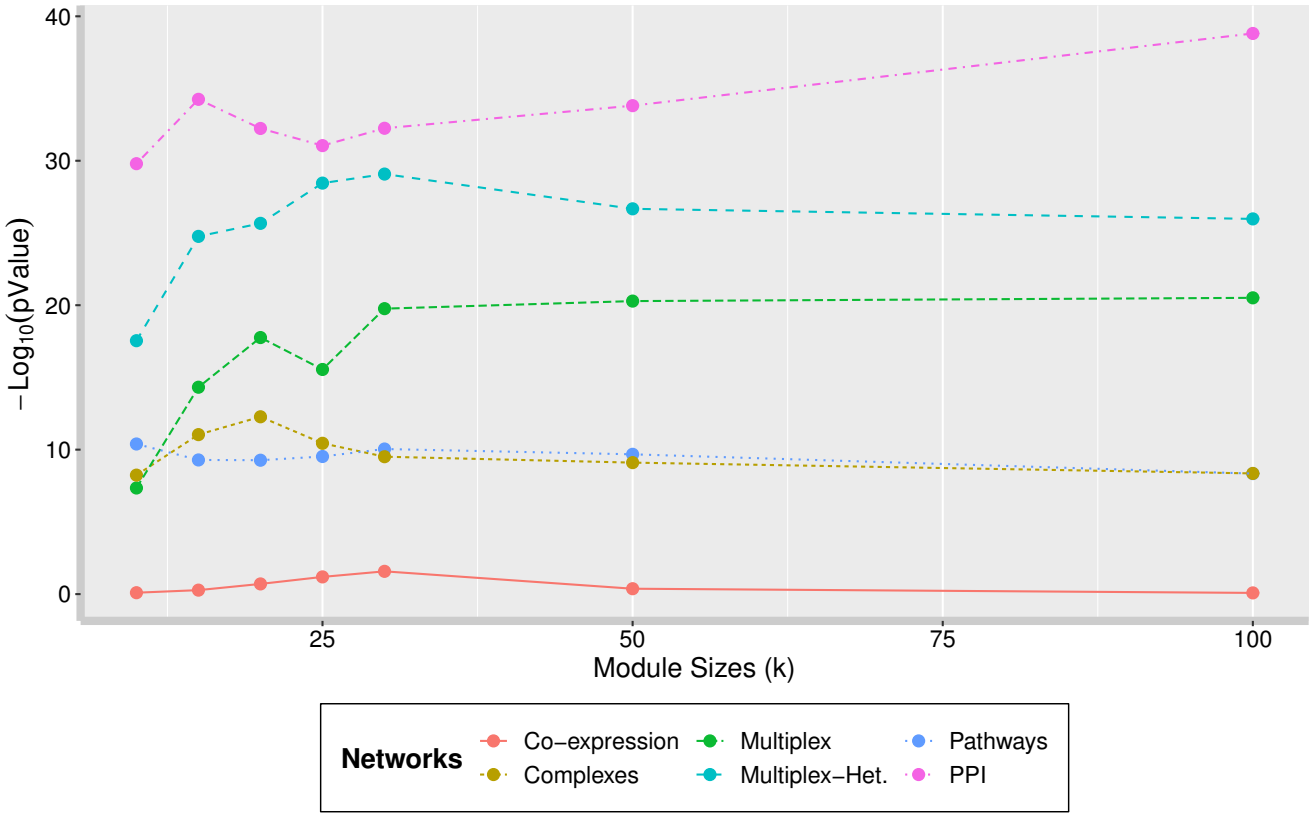


Figure S-3: Fisher's test results for the different sizes of modules and networks considered in our study. We considered the number of matching genes among our genes modules and aging-related genes from *GenAge* database without preliminar filtering.

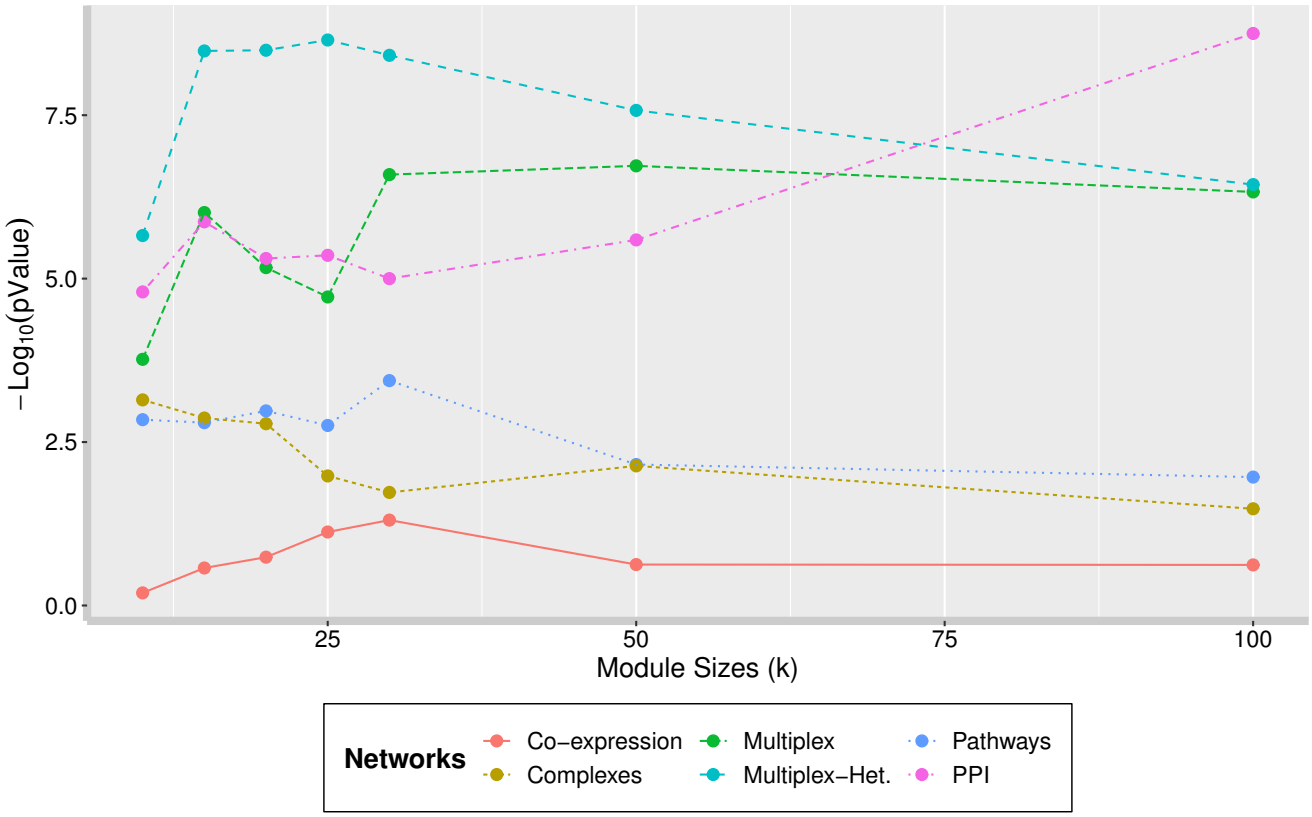


Figure S-4: Fisher's test results for the different sizes of modules and networks considered in our study. We considered the number of matching genes among our genes modules and aging-related genes from *GenAge* database after a criterion filtering.

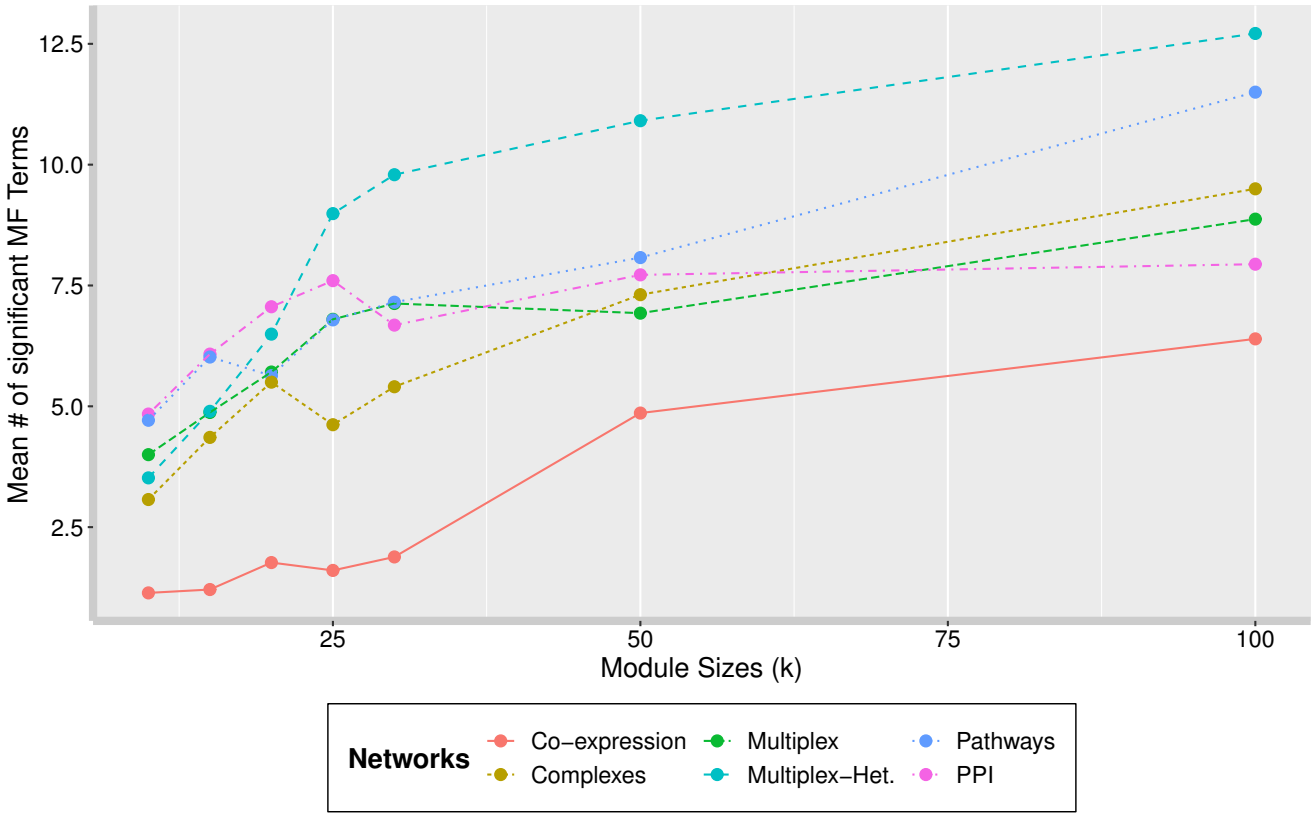


Figure S-5: The mean number of significant GO-MF annotated terms retrieved in the PA disease modules of different sizes when the repeated RWR algorithm is applied on the different networks under study.

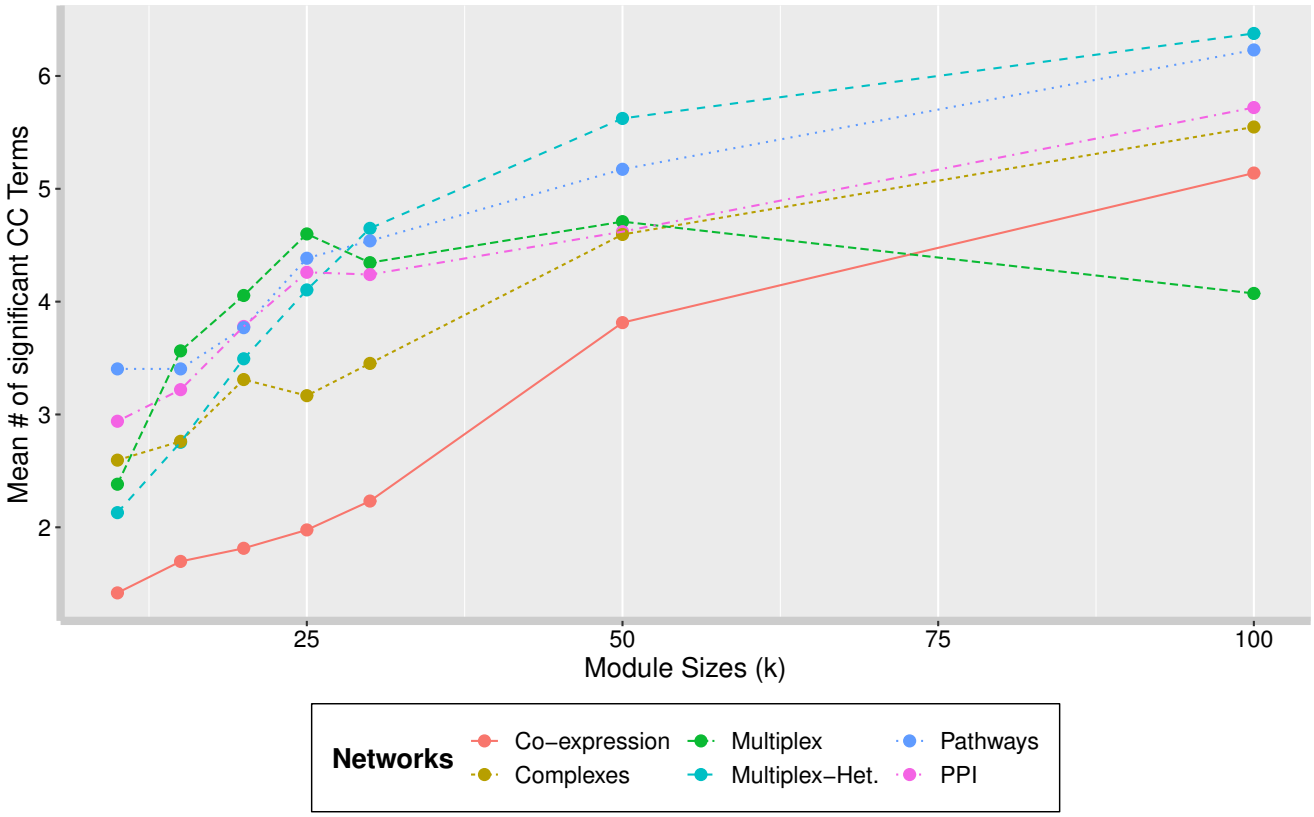


Figure S-6: The mean number of significant GO-CC annotated terms retrieved in the PA disease modules of different sizes when the repeated RWR algorithm is applied on the different networks under study.

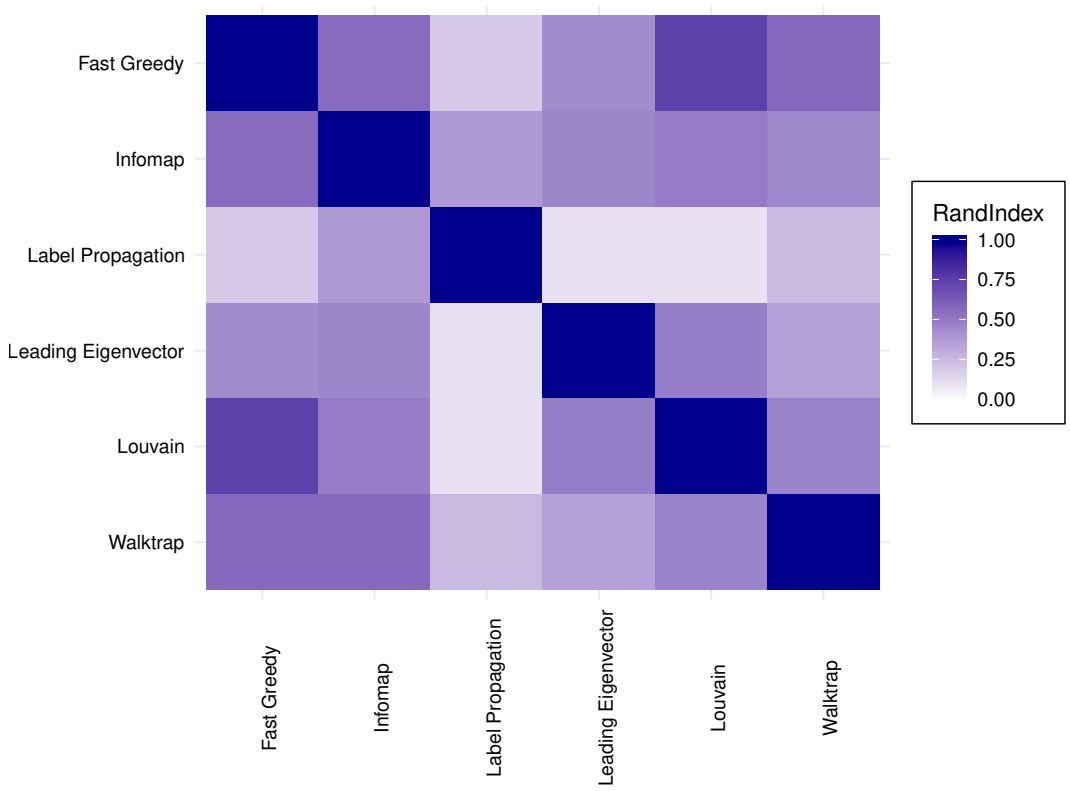


Figure S-7: Adjusted Rand Indexes between the partitions given by the different algorithms used to cluster the network of PA diseases.