UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Functional profiling of human genomic data using the protein interactome

Luz M. Garcia Alonso

Supervisor: Dr. Joaquín Dopazo

September 2015

Declaration

I hereby declare I myself carried out all the experiments described in this thesis, except where indicated in the text. The work presented here took place in the Systems Biology Department in the Centro de Investigacion Principe Felipe under the supervision of the Dr. Joaquin Dopazo. Also, I declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Luz M. Garcia Alonso

Acknowledgements

Son muchas las personas que han contribuido a la realización de esta Tesis, tanto de manera directa, aportando ideas y contribuyendo a la discusión, como de manera indirecta, por aligerar la carga y llenarla de buenos momentos. A todos ellos, muchas gracias.

A mi director de tesis, Joaquín Dopazo, por abrirme las puertas de su laboratorio y darme todos los medios e ideas necesarios para hacer esta tesis. Muchas gracias por la oportunidad brindada.

A mi tutor Joaquín Cañizares, por su buena disposición, sus consejos y su ayuda burocrática.

To the reviewers and tribunal members, for the time and expertise dedicated to evaluate this thesis.

A Jose Carbonell, Ignacio Medina y Pablo Minguez, de los que he aprendido básicamente todo. Sois parte muy importante de este trabajo. Gracias por las enseñanzas, consejos y visión crítica. A Nacho, por confiar en mi aun cuando al empezar no sabía que significaba Perl. A Pablo por sus lecciones sobre PPIs, aguantar mis avalanchas de mails y acogerme en Heidelberg. A Josete en especial, por su paciencia y temple ante mis ideas geniales de bombero.

A todos los bioinfos, por hacer que cada día en el CIPF fuera un buen día. Habéis hecho de estos 4 años de trabajo una experiencia que va mas allá del terreno profesional. Es un lujo ir a trabajar con tantas ganas y eso os lo debo a vosotros. A cada una de las supermujeres que han formado parte del pasillo de las chicas, mi segunda casa: Eva, Alicia, Rosa, Lorena, Verónica, Patricia y Marta. A Roberto y Pako, por compartir el disfrute por la absurdez. A Pablo y Ruben, por su magia para solucionarlo todo. A Quique, por sus lecciones magistrales. A Martina y Paco, por su derroche de buena energía. A Sonia, per la seva disponibilitat a ajudar en tot. A la inagotable Ana, por ser un ejemplo de

vitalidad y fuerza. A Lorena, mi compañera de sufrimiento en Boston. A Cankut, David, Marta, Davide, Cristina, Alex, Stefan, Fernando, Pablo, Joaquín, Pedro, Rafa, Mónica y Eugenia por los buenos momentos de desconexión dentro y fuera del CIPF.

Als amics de la Pobla, perquè no s'està tan agust a cap lloc com a casa. A Aida, per estar sempre ahi per a traure'm del cau i viure una nova batalla. A Sele, per tants anys d'aventures.

A los biotecs, por los buenos momentos a lo largo de estos ya 10 años. A Carla y Marta, por ese viaje post-tesis que nos marcaremos.

A la meua Mara, per ser màgica.

Als meus pares, avies i la resta de la família, per cuidar-me, consentir-me i animar-me diàriament. I a la meua germana Inma, per cuidar-los a tots.

Als meus avis, molt especialment, perquè la malaltia ací tractada els va afectar a ells també.

Y a Pepe, porque mucho del tiempo dedicado a este trabajo era tuyo.

Abstract

Our understanding of the biological mechanisms for most common human diseases is far from complete. Even with well established genetic landscapes, our capacity to make accurate phenotypical predictions or determine personalized disease risk using genetics alone is not possible for most diseases due to our lack of understanding of the mechanisms by which genetic alterations cause disease. Several suggestions have been proposed to explain this manifested lack of direct relation between genotype and phenotype, including interactions with other molecules, pleiotropy and environmental perturbations. Due to their essential role in carrying cellular functions, proteins and its interactions seem crucial to translate genomic data to phenotypic states. In this thesis I present three different and independent approaches to integrate human genomic data with prior knowledge in terms of protein-protein interactions (PPIs). The overall objective is, by making use of the interactome structure, to propose functional hypotheses that help to interpret the genetic variability observed in different human phenotypes. First I developed a methodology to extract the network component associated to any gene list ranked by any experimental parameter, as the one coming from case-control genome-wide associations studies. Second I performed a systematic analysis of human variants in the context of the protein interactome. There I study how the interactome structure can help us to explain the amount of apparently deleterious variation observed in actual populations and, therefore, give insight in its role in shaping the patterns of variability. Results are compared against somatic mutation found in Leukemia patients. Finally, I structurally resolved the protein interactome and used it to study how somatic mutations found in primary tumours distribute across the interacting interfaces and identify those with a potential role in driving oncogenesis. Although each chapter covers a different ques-

tion, all of them demonstrate the potential of the interactome in helping to interpret genomic variation observed under diverse research scenarios.

Resum

El nostre coneixement sobre els mecanismes biològics causants de la majoria de malalties humanes comuns és encara pobre. Tot i que en l'actualitat tenim mapes genètics d'alta resolució, la nostra capacitat per a fer prediccions fenotípiques certeres utilitzant únicament marcadors genètics és encara molt baixa degut a que no entenem les bases moleculars a traves de les quals les alteracions genètiques condicionen un fenotip de malaltia. Entre les principals causes d'aquesta aparent falta de relació directa entre genotip i fenotip estan la complexitat introduïda per les interacciones moleculars, els fenòmens de peleiotropia i la influencia dels factors externs. Degut al paper clau en dur a terme la majoria de funcions cel·lulars, les proteïnes i les seues interaccions han adquirit una especial atenció en la traducció de les dades genotípiques en estats fenotípics. Aquesta tesi presenta tres estartègies diferents per a la integració de dades genòmiques humanes amb la xarxa d'interaccions proteiques (interactoma). L'objectiu comú és, fent ús de l'estructura del interactoma, proposar hipòtesis funcionals que ajuden a interpretar els patrons de variabilitat genètica observats en diferents estats fenotípics. En primer lloc, es proposa una metodologia per a extraure el component de l'interactoma associat als gens rellevants en una llista ranquejada per qualsevol paràmetre experimental, com l'estadístic derivat d'estudis d'assocaició de genoma. En segon lloc, es descriu un anàlisi sistemàtic de les variants genètiques observades en humans sans en el context del interactoma. Ací s'analitza com l'estructura del interactoma pot ajudar a explicar l'aparent elevada quantitat de variants deletèries observades en els últims estudis poblacionals de sequenciació de genomes. Els resultats son comparats amb les mutacions somàtiques observades en pacients de Leucèmia. Finalment, es presenta un estudi de les mutacions somàtiques observades en tumors primaris de més de 20 tipus utilitzant una versió del

interactoma més resolutiva, que inclou l'estructura tridimensional de les proteïnes. Encara que cada estudi presentat en la tesi planteja resoldre qüestions diferents, tots ells demostren el potencial del interactoma de proteïnes en ajudar a interpretar la variació genòmica humana observada en un context tant poblacional com de malaltia.

Resumen

Nuestro conocimiento acerca de los mecanismos biológicos causantes de la mayoría de enfermedades humanas comunes es pobre aún. Incluso con mapas genéticos de alta resolución, nuestra capacidad para hacer predicciones fenotípicas certeras o determinar el riesgo de una persona a padecer una enfermedad utilizando solamente marcadores genéticos es muy baja. Entre las principales causas de esta aparente falta de relación directa entre genotipo y fenotipo están las interacciones moleculares, los fenómenos de pleiotropía y la influencia de los factores externos. Debido al papel esencial que ejercen en llevar a cabo las funciones celulares, las proteínas y sus interacciones han adquirido una atención especial en la traducción de los datos genotípicos a estados fenotípicos. En esta tesis se presentan tres estrategias diferentes para la integración de datos genómicos humanos con la red de interacciones proteicas (interactoma). El objetivo común de todas ellas es, haciendo uso de la estructura del interactoma, proponer hipótesis funcionales que ayuden a interpretar los patrones de variabilidad observados en diferentes estados fenotípicos humanos. Primero, se propone una metodología para extraer el componente del interactoma asociado a los genes relevantes en una lista ranqueada por cualquier parámetro experimental, como el estadístico derivado de los estudios de asociación genómicos. Es segundo lugar se describe un análisis sistemático de las variantes genéticas observadas en humanos sanos en el contexto del interactoma. En él se estudia cómo la estructura del interactoma puede ayudar en explicar la aparentemente elevada cantidad de variantes deléreas observadas en los últimos estudios poblacionales de secuenciación de genomas. Los resultados son comparados con las mutaciones somáticas observadas en pacientes de Leucemia. Finalmente, se presenta un estudio de las mutaciones somáticas observadas en tumores primarios utilizando una versión del interactoma que

incluye la estructura tridimensional de las proteínas. Aunque cada estudio presentado en la tesis pretende resolver preguntas diferentes, todos ellos demuestran el potencial del interactoma de proteínas en ayudar a interpretar la variación genómica humana observada en un contexto tanto evolutivo como de enfermedad.

Contents

List of Figures

List of Tables

We need to remember that whereas mathematics is the art of the perfect and physics the art of the optimal, biology, because of evolution, is only the art of the satisfactory.
Sydney Brenner

CHAPTER 1

Introduction

1.1 Systems Biology: aim and definition

Every human, as all living individuals, store the information required for developing themselves in the sequence of its genome, in the form of DNA packed into high-order chromatin. The genome encodes the information to formulate the functional molecules (RNA and proteins), which orchestrate all cellular processes that give rise to the living being. While RNA is responsible of protein-template and gene expression regulatory roles, proteins are considered the main players of final biological functions. These do not operate alone but are organized into complex circuits in a way that most cellular functions are the outcome of an intricate network of interactions perfectly coordinated. Through these circuits, proteins crosstalk back to the genome to regulate, by means of transcriptional regulators and epigenetic changes, which part of it is expressed in a particular cell type at a specific moment. It is thus the cooperativity between molecules what ultimately drives the development and behavior of living organisms.

The complexity of the cell system is evidenced among the vast amount of studies to identify the roots of most human diseases. These, although successful in identifying genes behind rare Mendelian diseases, have failed to uncover direct causalities for most complex pathologies. Contrary to Mendelian diseases where highly penetrant alterations enable to establish direct associations with the causal gene, common pathologies arise from a more complex interplay between different molecular perturbations. This yet undefined complexity is the reason why the molecular mechanisms underlying most complex diseases are still an unsolved paradigm. Hence, approaches that adequately capture the exquisite complexity of the cell are urgently needed to achieve accurate descriptions of the mechanisms driving common diseases.

Recently, Systems Biology has been proposed as a new candidate field to cope with biological complexity. Much controversy exists with respect to the aims, scope and approaches of Systems Biology, and several definitions of the field have been proposed. Here I take the definition

proposed by De Backer et al. (2010), who escapes from the simplistic dichotomy of ”reductionist molecular biology” versus ”holistic systems biology”, and considers Systems Biology as a complementary new stage in the development of Molecular Biology with common biological questions, but formulating more complex, system-wide hypotheses and extending its approaches with new system-wide (omics) experiments and mathematical models. See De Backer et al. (2010) for a deeper review of the System Biology definition and approaches.

The research described in this thesis is addressed under a Systems Biology perspective, where genomic data is integrated with models that reflect the complexity of the cells to interpret human genomic variation observed under different phenotypic scenarios.

1.2 Human protein interactome

1.2.1 Definition

Protein-protein interactions (PPIs) are physical contacts between two or more proteins driven by biochemical events and/or electrostatic forces. The protein interactome describes the full collection of PPIs that can occur in a cell and offers us an unprecedented level of detail of all the molecular circuits governing cell functions.

To consider a contact between two protein as PPI, this should be direct and specific, not just a generic touch (Bahadur et al., 2004). For example, every protein at one point is in contact with the ribosome and most with the proteasome by means of generic interactions. Since these physical contacts between proteins imply general but not specific functions, these are excluded from the PPI definition.

PPIs can be classified as stable (also called permanent) or transient, and both types can be either strong or weak (Nooren and Thornton, 2003). Stable interactions have relatively long lifetime, and mostly occur in proteins forming complexes (such as DNA damage repair complex). On the contrary, transient interactions form and break down briefly by means of combinations of non-covalent bonds such as hydrogen bonds, ionic interactions, Van der Waals forces, or hydrophobic bonds. Frequently, the conditions that enable transient physico-chemical contacts are dependent on other biochemical changes, such as the interaction with other proteins, the addition of functional groups (such as protein phosphorylation, acetylation, etc) or its localization in the cell (Perkins et al., 2010). Thus, transient PPIs are expected to control most of cellular regulatory processes through signalling pathways.

1.2.2 PPIs identification

In humans, the interactome is predicted to range between 130,000 and 600,000 PPIs, although no more than 80,000 have been already experimentally observed (Mosca et al., 2013c). Experimental identification

of PPIs is done either at large or small scale with two main technologies that produce different PPI data types. The most used techniques in measuring direct physical interactions between protein pairs and complex are, respectively, Yeast-two-hybrid (Y2H) (Ito et al., 2000, 2001) and Tandem Affinity Purification coupled to Mass Spectrometry (TAP-MS) (Berggård et al., 2007). An overview of experimental proteomic techniques applied to measure PPIs can be found in Berggard et al. (Berggård et al., 2007).

The quality of such data has been a controversial subject. Initial studies comparing results from several PPI detection methods showed little overlap between experiments, yet due to methodological bias toward the identification of particular types of proteins or interactions (e.g. Y2H system has troubles detecting proteins of greater abundance and stability) or yet to troubles in reaching the saturation point, which results in incomplete coverage (Von Mering et al., 2002). Additionally, for some systems such as Y2H, the false positive rate accounted for almost a half of the total data generated in 2002 (Von Mering et al., 2002). Despite widely questioned a decade ago, latest high-throughput PPIs screenings showed an improved accuracy of the detection methods, reducing the false positive rate and reproducibility problems (Lage, 2014; Rolland et al., 2014). Methodological refinements together with quality awareness-raising initiatives promoted by the collecting databases, with the implementation of common standards in curation practices (Mosca et al., 2013c; Orchard, 2014), have allowed reliable PPIs to be available to the whole research community. But still, due to the diversity in nature of PPIs, the transient character (Perkins et al., 2010), the conditionality (Grossmann et al., 2015) and the lack of a unique technique able to detect all of them, a combination of approaches is required to cover accurately the whole interactome.

1.2.3 Functional modules

There is a wide consensus on the fact that the biological functionality of the cell arises from the cooperative behaviour of sets of molecules and its interactions. This organization is captured in the human protein interactome which, as for other organisms, display a hierarchical topology (Han et al., 2004; Rual et al., 2005). Proteins are usually arranged forming neighborhoods of proteins highly interconnected within them but sparsely connected to the rest of the network. Such sets of highly interconnected molecules define operational entities, identifiable by network clustering algorithms, to which different elementary functions can be attributed. This lead to the concept of modularity in biological networks, applicable to the protein interactome, metabolic networks, signalling pathways and networks of gene-gene interactions (Ravasz et al., 2002). Actually, the relationship between common functionality and interaction has been successfully used to predict protein function (Sharan et al., 2007).

Modular patterns also extend to genes related to similar human pathologies. Gene products associated to a common disease also exhibit an increased tendency to interact among them, co-express in the tissue affected and display coherent functions according to GO annotations (Oti and Brunner, 2007; Taylor et al., 2009; Aerts et al., 2006). For example, genes associated to Ataxia were observed to occupy a common region in the interactome and display a shorter distance between them than random subsets and share common partners (Lim et al., 2006). This proximity property of disease genes to locate closer each other is being extensively used to prioritize new candidate disease genes in genomic-scale studies (Oti et al., 2006; Köhler et al., 2008).

Important efforts are being made to identify disease-associated modules. That is, network components whose perturbation are believe to initiate disease phenotype in humans. Despite the extensive focus on functional modules based on GO (or other unstructured labels) for the biological interpretation of different types of genomic experiments

(gene expression microarrays, large-scale genotyping), conceptualizing a function simply as a label shared by a set of genes resulted in a poor description of the cellular complexity, since ignores gene relationships. PPIs provide a more realistic representation of such relationships beyond categorical definitions (Minguez et al., 2009; Minguez and Dopazo, 2010). The use of the interactome as a theoretical scaffold that relates proteins among them allows depicting subnetworks of interacting proteins associated to features in genomic experiments. The identified subnetworks or modules can be considered a higher level, structured description of functional modules operating in the cell (Ideker et al., 2002; Mitra et al., 2013). Such subnetworks can serve as hypothesis building scaffolds that guide researchers to zoom into the molecular mechanisms involved in particular function or disease (Ideker and Lauffenburger, 2003; Mitra et al., 2013). Consequently, much attention has been focused over the last decade in methods aimed to interrogate the interactome with other molecular profile to identify these subnetworks accomplishing specific functions.

1.2.4 The interactome as a network

Taking all together, the PPIs can be modelled as an undirected graph, where nodes represent proteins and edges the interactions. The resulting network acquires a particular shape (topology), displaying a set of global-scale properties (also called emergent properties) not observable when studying the PPIs in isolation (Barabási and Albert, 1999; Goh et al., 2002; Jeong et al., 2001; Barabási, 2009).

One of the key findings from topological studies is that biological networks share patterns on its organization, being these patterns significantly different from those observed in random network models (e.g. Erdos Reny model). That is, all of them follow a common principle: the great majority of the proteins have very few connections whereas few proteins concentrate the great majority of connections (Barabási and Albert, 1999). The high-degree nodes are often called *hubs*, and are thought to

carry relevant roles in their networks. This non-random distribution of the connectives approximates a "power-law" and states that the probability that a chosen protein has k PPIs approximates: $P(k) \sim k^{-\gamma}$, where $2 < \gamma < 3$.

Another common feature observed in complex biological networks, also observed in other non-biological real networks such as the World Wide Web (WWW), is that the average distance between any two proteins in the network is small so that any node can be reached through a few edges (Giot et al., 2003). This property, called 'small-world effect', arises from the existence of highly centred proteins (proteins with high relative betweenness centrality), which connect the whole network. The direct consequence is that local perturbations in a central node could reach every other node very quickly.

1.2.5 Functional and evolutionary implications of the interactome topology

The fact that previously mentioned properties (modularity, hierarchy, degree distribution and small-world) are observed among different biological networks across different species suggest that there are evolutionary mechanisms constraining the maintenance of them. Although the reasons and mechanisms are not yet fully understood, several hypotheses have been proposed to explain the biological consequences of this acquired structure. Recent publications propose it is the consequence of how these networks grow, by expanding through *preferential-attachment process* where new nodes are linked to existing nodes with a probability proportional to the number of connections of the later (Barabási and Albert, 1999). Albeit this approach successfully explains degree distributions in other real networks, the biological mechanism that support the preferential attachment hypothesis in PPI networks is less evident. Probabilistically, is less likely that a newly introduced protein incorporates the structural characteristics to specifically interact to a wide range of proteins and become a hub. However, considering that new proteins

are formed mostly by duplications and fusions of existing genes (Pastor-Satorras et al., 2003), and not by *de novo* insertion of random sequences, probabilistic models that ignore evolutionary forces shaping molecular networks are far from provide a complete explanation.

The emergence of proteins with high connectivity and centrality suggests that these may play a key role. From an evolutionary perspective, one can hypothesize that topology displayed by biological networks have been selected in virtue of the advantages brought in terms of stability and robustness against random errors, such as genetic alterations, environment perturbations or stochastic variation (Taylor, 2013). Theoretical studies have demonstrated that biological networks are topologically robust against the removal of random nodes, but that the removal of the most connected ones (simulating an attack) drastically alters the network's topology (Albert et al., 2000). This is in clear opposition to the behaviour observed in random networks, where random errors and attacks can not be distinguished (Albert et al., 2000). Since the likelihood of removing a highly connected protein is significantly small due to its low frequency, most interactions in the network, and thus its global properties, would remain intact under random perturbations. Interestingly, studies in several species observed that most of the genes that when depleted decrease robustness, expressed as a proxy of cell survival, code for proteins that are highly connected in the interactome (Jeong et al., 2001; Hahn and Kern, 2005) and evolve more slowly than proteins with low connectivity (Kim et al., 2007). Taking together, these results suggest that genetic variation in these key proteins is constrained by natural selection. Paradoxically, the robustness conferred by these proteins can suppose a mechanism that facilitates genetic variation in the rest of the genes (Levy and Siegal, 2008). Understanding how living systems deal effectively with perturbations such as genetic variability and exploit it to evolve remains a challenging goal for evolutionary systems biology.

From the point of view of disease, there is a vast amount of research supporting a strong relationship between the topological role of a protein in the interactome and its association to human diseases. First studies

showed that the protein products of genes driving cancer tend to have higher connectivity than non-cancer proteins (Wachi et al., 2005; Jonsson and Bates, 2006). These authors proposed that altered genes in cancer are key for the proliferation of the tumor cells and, therefore, display the same topological properties that essential genes. In the same sense, Goh et al. (2007) and Feldman et al. (2008) compared essential genes to cancer and mendelian disease genes and found that topological values for centrality and connectivity of the later tend to lie in between the essential or tumorgenic genes and those not included in any previous class (that is non-disease genes). Combining these results with the relationship between centrality and essentiality, the general conclusion is that genes driving disease are not randomly located in the network. Although at the publication of the first evidences there was concern about the potential bias in PPI networks towards the more studied genes (that is disease genes), new studies in less biased interactomes reinforce initial tendency (Rolland et al., 2014). Considering disease as the alteration of the overall stability, these observations evidence the role of hub genes in safeguarding phenotypic robustness.

1.2.6 Structural details

Despite the potential of PPI networks, in combination with other molecular profiles, to identify sets of proteins associated to the phenotype under study, it proved difficult to derive biological conclusions from the discovered subnetworks. Although pointing at the underlying functional modules, these subnetworks can only be seen as descriptive scaffolds rather testable hypotheses (Ideker and Lauffenburger, 2003). The reason behind is the lack of molecular details encoded in the interactome model. That is, the interactome has been modelled abstractly (ie. an undirected graph), where proteins represent graph-theoretical nodes ignoring its structural details and the stereochemistry of the interactions.

A protein physically interacts with its partner through a region on its surface, called interface (Janin, 1995; Jones and Thornton, 1996). For

most cases, proteins interact to each other at specific domains (ordered regions) (Pawson and Nash, 2003). One example is the SH2 and SH3 domain containing proteins which binding require a phosphotyrosine residue in the context of a peptide motif. However, there is increasing evidence that intrinsically disordered regions also participate in functional binding (Uversky et al., 2005). In both cases, the binding specificity for a pair of interacting interfaces is determined by the amino acid residues in the neighborhood of the final 3D structure that each partner acquires. This is specially important in studies integrating PPIs with genomic data. Single nucleotide or short indel (insertion or deletion) variants, can introduce changes in the amino acid sequence that modify the physico-chemical properties of the protein and, consequently, can affect the interacting ability and alter its function. As confirmed experimentally, mutations altering different protein domains can not be expected to cause identical effects on protein properties and functions (Zhong et al., 2009; Wang et al., 2012). Under a scenario where the biological functions are believed to raise from the network of molecular interactions, interactions seem to be a more rational level of abstraction than genes. In fact, only for humans, Uniprot database reports more than 4000 missense mutations experimentally observed to affect the interaction ability of the protein with its partner (Consortium, 2011).

The introduction of the molecular structure onto the biological networks is still in its beginnings (Mosca et al., 2013a). The reasons behind are the lack of structural data for most of the proteins and the lack of a clear decision making process to asses the effect of a single amino acid change on its structure and function. Additionally, integrating structural information would add a new level of detail to consider when modelling the interactome and would require a more complex approach to analyze it (Ideker and Lauffenburger, 2003).

To our knowledge, only few studies reported the systematic integration of disease mutation data onto protein structures. Zhong et al investigated how distinct molecular defects in proteins lead to distinct perturbations in the interactome and, ultimately, phenotypic con-

sequences (Zhong et al., 2009). In its study, deletions and early stop gain mutations are modelled as node loss, whereas nonsynonymous mutations and small indels are modelled as edge specific perturbations. They observed that a half of the human disease mutation affect specific edges and that mutation in genes responsible for different diseases tend to be located in distinct interacting interfaces. In another study, Wang et al observed that missense mutations associated to human diseases were enriched on the interacting interfaces of proteins and that phenomenas like gene pleiotropy or locus heterogeneity can be explained by their location within its interfaces (Wang et al., 2012). Similar observations were described recently in (Mosca et al., 2015). All these studies highlight how structurally resolved interactomes can produce more detailed and useful models. More importantly, the message derived is that mutations that perturb distinct protein activities may be key in explaining heterogeneity between individuals.

1.3 Human genomics

1.3.1 Sequencing the human genome

Since the genetic theory of inheritance culminated with description of the molecular structure of DNA more than 60 years ago (Watson et al., 1953), the analysis of the genome sequence has centered most of the effort in biological research. In the seventies, the development of two major techniques, the enzymatic DNA sequencing method and the invention of the polymerase chain reaction (PCR) (Mullis et al., 1992), supposed an extraordinary revolution in human genetics and molecular diagnostic. These techniques together with the improvement in computer capabilities enabled the automation of DNA sequencing of small sequences, creating the basis for studying genetic diseases, population ancestry and evolution (Sanger et al., 1977; Hunkapiller et al., 1991; Swerdlow et al., 1990).

At the end of the eighties, the research focus moved from the gene level to the genome level with the ambitious Human Genome Project (HGP), which objective was, for the first time, to read the whole sequence of a human genome (Adams et al., 1991). The HGP initiated officially in 1990 and it took more than a decade until the publication of the first sequence of the human genome in 2001 (Lander et al., 2001; Venter et al., 2001). This draft represented the genomic road-map to guide geneticist and marked a turning point for modern biomedical research, with the birth of the genomic era. However, even with the tools available by that time, the progress towards the characterization of the genetic variability in large populations and its implication for human health was far from being accessible. In the meanwhile, DNA microarrays where used to screen relatively common polymorphisms among large cohorts, and helped to identify polymorphic risk alleles (Hindorff et al., 2009). Nevertheless, these only provide information for around one million genetic variants, which restricts the scope for research and discoveries (Manolio et al., 2009).

The scenario changed some years ago with the introduction of the

Next-Generation Sequencing (NGS) technologies (Margulies et al., 2005; Shendure et al., 2005; Bentley et al., 2008). Today, more than a decade since the first genome sequence of a human being was published, sequencing a human genome in feasible time at accessible costs has turned out to be a reality (Schuster, 2008). This unprecedented capacity has led to an explosion of collaborative projects to sequence the genomes of both patients suffering genetic diseases and healthy humans. The objective is to create a detailed catalogue of human genetic variation to help in the diagnosis of genetic diseases and personalized therapy design.

1.3.2 Genotype to phenotype: the challenge of heterogeneity

The advances in genome sequencing technologies are fueled by the basic need of understanding how genotypes embedded in the genome give rise to phenotypes. However, the endeavour of making use of genomic data to assist medical decisions has run into a wall. The publication of the genomic landscape of human populations and cancer cohorts has revealed a much larger amount of variability among individuals than expected. On the one hand, sequencing projects in human populations (Consortium et al., 2010; Fu et al., 2013) have found a vast amount of apparently deleterious variation in the genome of normal, healthy individuals (Xue et al., 2012; MacArthur et al., 2012). The question here arises as to how do individuals with such divergent genomes handle this perturbations to display stable healthy phenotypes (Waddington, 1942). On the other hand, cancer sequencing projects (McLendon et al., 2008; Hudson and Jennings, 2011; Weinstein et al., 2013) have revealed a stunning heterogeneity between tumors, both between patients with the same cancer type and within the same patient. The question in this second scenario arises as to how do divergent genomes converge to tumour initiation and progression (Vogelstein et al., 2013). Both scenarios pose a significant challenge to identify the genomic alterations that contribute to disease phenotype and impedes progress towards personalized treat-

ment. The following sections of the chapter describe these two problems in more detail.

In general, the challenge faced here is how to extract knowledge from the massive amount of genomic data we are accumulating. Since transforming genome sequence information into a final living being requires numberless cooperative actions mediated by its encoded molecules, the process would necessarily need the contextualization of this data under a model that fully integrates the structure, function and organization (i.e. complexity) of the molecular interactions that give rise the cell behaviour. However, even with a good map of the molecular circuits describing the complex machinery of the cell, the absence of a theory on how cooperative processes in cell emerge from the information coded in the genome complicates the application of sequencing data to its full potential (Brenner, 2010).

1.3.3 Deleterious variability in healthy populations

Deleterious variants are those expected to impact negatively the reproductive fitness of its carrier by affecting severely the biochemical function of protein-coding genes (Sunyaev et al., 2001). Due to its expected negative impact, deleterious variants are of tremendous interest as they are the perfect candidates to contribute to human disease. Nevertheless, population-scale sequencing studies (Consortium et al., 2010; Fu et al., 2013) have reported an unanticipated large amount of these variants in healthy individuals, contradicting this view. Although the possibility that any of the used donors eventually become ill cannot be dismissed, seem unlikely that they have suffered extensively from any genetic disease (MacArthur and Tyler-Smith, 2010; Xue et al., 2012; Nothnagel et al., 2011; MacArthur et al., 2012). This apparently deleterious variation is not restricted to coding regions but also seems to occur in other non-coding, regulatory elements, such as miRNAs (Carbonell et al., 2012), transcription factor binding sites (TFBSs) (Spivakov et al., 2012) and other genomic elements (Lappalainen et al., 2013).

The origin of this apparent excess of damaging variants has been attributed to the combination of a recent accelerated human population growth with a weak purifying selection (Keinan and Clark, 2012; Tennessen et al., 2012). Fu an colleges estimated that around the 85% of the predicted deleterious variants arose in the last 5,000 to 10,000 years and that the selection has not had enough time to clear them (Fu et al., 2013). Thus, while those alleles with strong negative character are expected to be filtered out from the population by natural selection, the selection pressure over not as strong deleterious mutations can be weak, keeping them at lower frequencies. As a consequence, human populations have increased its genetic heterogeneity and the burden of Mendelian disorders.

Conservative estimators quantify that there are no less than 250 loss-of-function (LoF) variants in coding genes per sequenced genome, 100 of them apparently related to human diseases, and around 30 in a homozygous state (MacArthur and Tyler-Smith, 2010; Xue et al., 2012; MacArthur et al., 2012). The presence of disease-associated mutations in the genome of healthy individuals raises doubts about the true pathogenicity of these variants and has led to the reconsideration of its putative causal effect. However, although some errors are expected in disease-variant associations, specifically for rare diseases where the small population size can lead to ambiguous assignments of causality, the presence of such number of known disease alleles indicates that many of these variants coexists at low frequency in the human population without apparent consequences (MacArthur et al., 2012). There is thus an urgent need to decipher the mechanisms by which specific deleterious variants can have a clear pathological effect under some conditions while in others seem apparently innocuous. This fact poses a major challenge for clinical geneticists in the identification of true disease-causing genetic mechanisms and supposes a bottleneck in the translation of genome sequence data to the clinics.

From an evolutionary point of view, this excess of functional variation evidences that individuals are able to maintain appropriate healthy

phenotypes in the presence of such amount of theoretically perturbing variants. The question is how this robustness is achieved. Different reasons have already been proposed to account for the maintenance of deleterious variants in human populations. These include severe recessive disease alleles in homozygous state; variants implicated in late onset phenotypes, that can not be efficiently screened by natural selection; reduced penetrance phenotypes which require additional genetic and/or environmental factors for expression (Raj et al., 2010); gene redundancy, where the effect of a mutation is compensated by another gene with similar function (Hoffmann, 1991); and, finally, sequencing and annotation errors (Xue et al., 2012; Nothnagel et al., 2011; MacArthur et al., 2012).

Despite the expected nonessentiality of most of the carrying genes, these variants, in combination, may still have an effect on the phenotype of the carrier. A study in yeast including genome-wide sequence and expression data concluded that robustness to random mutations is not only due to gene redundancy but to epistatic interactions between unrelated genes (Wagner, 2000). Examples supporting this mechanisms can be found in large metabolic networks in yeast, where these can compensate for complete loss of enzymatic reactions by exploiting alternative pathways (Szathmary, 1993). Another example is the buffering mechanisms via the heat-shock proteins against genetic variation observed in developmental processes in Drosophila (Queitsch et al., 2002). To what extent molecular interactions can buffer deleterious variability needs to be studied.

1.3.4 Genetic heterogeneity among cancer patients

Cancer is a heterogeneous collection of diseases sharing common traits (Weinberg, 2007), being the most relevant the ability of the cancer cells to proliferate and pass the disease phenotype to its descendants. In the oncogenesis process, a cell evolves from normal to cancerous through the accumulation of heritable alterations that shape the cell machinery to achieve a series of attributes (i.e. convergent evolution) called the

"Cancer Hallmarks" (Hanahan, 2000; Hanahan and Weinberg, 2011) by means of successive rounds of clonal proliferation, genetic diversification and clonal selection (Nowell, 1976; Stratton et al., 2009) (see Figure 1.1). Governed by the biological imperative to survive and perpetuate, in a way that resembles Darwinian evolution but on accelerated timescales, cells carrying alterations that confer an advantage are positively selected. Thus, most oncogenic alterations are somatic, acquired throughout life by means of replication errors, damage in the DNA repair machinery or even mutagen agents, such as UV-light or tobacco (Greenman et al., 2007).

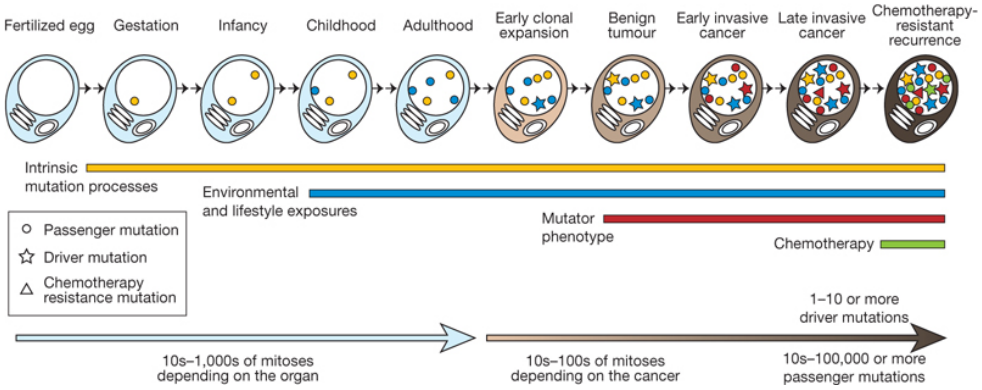

Figure 1.1: Evolutionary theory of oncogenesis. Image taken from Stratton et al. (2009). Acquisition of somatic mutations through normal cell division from the fertilized egg to a cancer cell. During the processes, several factors such as mutagens or DNA repair defects may contribute to the mutational burden. Mutations are grouped as passenger mutations, with minor or null contribution to cancer phenotype; driver mutations, which favour clonal expansion; and/or chemotherapy resistance mutations.

The emergence of NGS technologies is allowing us to decode cancer genomes at unprecedented resolution. The generated data is being collected by projects like The Cancer Genome Atlas (TCGA) (McLendon et al., 2008) and the International Cancer Genome Consortium (ICGC) (Hudson and Jennings, 2011), which contain genomic information from thousands of patients for more than 30 different tumor types. Under the assumption that accurate descriptions of the genome sequence would

provide insights into the oncogenesis process, much effort is being done in identifying every difference between normal and cancer genomes. However, the cancer genome scenario evidenced to be much more intricate than expected. Studies from different cancer types agree in three major points. First, only few genes are mutated at high frequency, revealing a "long tail" distribution (Greenman et al., 2007) (see Figure 1.2). Even in these genes mutated at high frequencies, the observed mutations are usually different and the patient carrying them often display different prognosis and treatment response. Secondly, high heterogeneity exists among different patients and within them, being not two genomes equal, even those from the same cancer type. Finally, common mutations are found in patients with tumors located in distinct tissues, with different histological patterns, which prove that identical alterations may trigger similar phenotypes in different contexts (Gerlinger et al., 2012; Vogelstein et al., 2013; Lawrence et al., 2013).

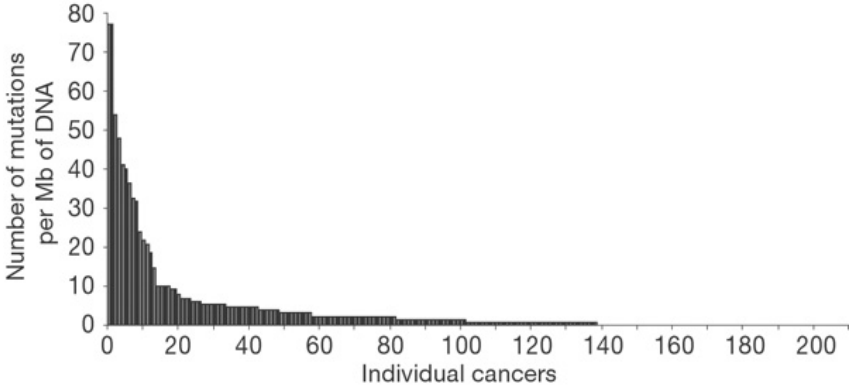

Figure 1.2: Prevalence of somatic alterations in human cancer genomes. Image taken from Greenman et al. (2007). X axis represents Mb of DNA whereas Y axis the number of somatic alterations (single base substitutions, insertions/deletions and complex mutations) in 210 individual human cancers.

The cause of this complex landscape is that tumor cells harbor a great amount of genetic alterations compared to normal cells, mostly due to genome instability or deficiencies on the DNA repair machinery (Lothe et al., 1993; Fishel et al., 1993). As a consequence, a large number of

relatively random alterations are generated (Zheng et al., 2014). Thus, most of the alterations are believed to confer no relevant advantage to tumors (called passenger alterations) whereas only alterations in few genes would contribute to the oncogenic phenotype (called driver alterations) (Greenman et al., 2007). The identification of these driver alterations is one of the most anxious objectives in oncology research (Weinstein et al., 2013; Vogelstein et al., 2013) since it can guide the diagnosis of new patients but, more importantly, point at new targets for therapeutic intervention.

Classically, methods aimed to identify cancer drivers using genomic data have focused on looking for genes which mutation frequencies are higher than expected, considering the overall mutation rate (Greenman et al., 2007). However, approaches based on the overall frequency, although accurate in identifying highly mutated genes such as TP53, fail in identifying less frequent causal mutations. To overcome this bias, several methods have been proposed. Most of them apply a gene-centric perspective in a way that the mutation rate of a gene is considered independent one from another gene, so that highly mutated genes such as TP53 do not mask the signal from the others. In this sense, proposed methods analyze whether a mutated gene displays properties similar to those displayed by known driver genes such as function, tissue expression and evolutionary and network properties (D'Antonio and Ciccarelli, 2013). Others focus on the mutation pattern along the gene sequence, such as the preferential accumulation of functional mutations rather than neutral (Gonzalez-Perez and Lopez-Bigas, 2012) or the mutation clusterization in hotspots (Gonzalez-Perez and Lopez-Bigas, 2012), in phosphorylation sites (Reimand and Bader, 2013) or protein domains (Porta-Pardo and Godzik, 2014). See Gonzalez-Perez et al., 20013 for a deep review.

While much effort is being done in identifying driver genes, the functional interpretation of its genetic alterations in a systematic way is still in its beginning. Specifically, the relevance of the protein interacting interfaces in tumorigenesis, although key in mediating cell signalling, is still poorly understood. Recently, Espinosa et al. (2014) performed

a comprehensive characterization of COSMIC mutations with the aim of identifying common protein structural properties that help to predict new driver mutations. Among other properties, they found a significant amount of missense mutations in bound-forming residues at protein binding interfaces. Their observations suggest that mutations altering PPIs may be a mechanism for signalling aberrations in cancer.

An important advantage of gene-centric approaches is that they can introduce prior functional knowledge and, ultimately, find a potential mechanistic implication for the somatic mutations. For example, they can identify specific mutated phosphorylation sites within genes that perturb signaling networks by altering protein modifications (Reimand and Bader, 2013), which in turn adds an additional source of evidence. Moreover, these methods allow to study a gene as a multifunctional factory, a more realistic assumption as different protein sites are responsible for different functions. It may be the case when an oncogenic signal can be only initiated if a specific functional site of the protein is altered, but not other sites. As observed for Mendelian diseases (Wang et al., 2012), considering the different structural and functional sites of the multifaceted genes may help to understand why patients that, although share common driver genes, display high variability in prognosis and therapy response (Andreyev et al., 2001; Pao et al., 2005; Alamo et al., 2014). Research in this direction is essential to improve personalized cancer treatment.

1.4 Thesis outline

Linking genomic variation to phenotypes is one of the most anxious objectives in Systems Biology. Due to its role in accomplishing cellular functions, proteins and its interactions provide an excellent model reflecting cellular complexity. This thesis presents different approaches to integrate human genomic data with the molecular circuits defined by PPIs. The overall objective is, by making use of the interactome, to propose functional hypotheses that help to interpret the genetic variability observed in different human phenotypes. The research is distributed into three chapters. Although each one covers a different question, all of them demonstrate the potential of the interactome in helping to interpret the genomic variation observed under diverse research scenarios. Each chapter starts by giving a short overview of the problem under study and defines the questions that remain open, followed by a definition of the individual objectives to achieve. Next, it gives a detailed description of the data and methods used. Finally, all the results are exposed and discussed.

The **second chapter** comes up with the demand for methodologies for the functional profiling of genome-scale experiments introducing PPIs data. In the past, methods for functional have genomics been developed to analyze simple, unstructured module definitions, such as GO, to account for the common functionality of a group of genes. Despite the success of methods based on GO modules, conceptualizing a function simply as a label shared by a set of genes results in a poor description of the cellular complexity. PPIs provide a useful and extensively used representation of such relationships. The use of the interactome as a theoretical scaffold that relates proteins among them may allow to depict subnetworks of interacting proteins associated to features in genomic experiments. These subnetworks provide a zoom in map of the implicated cell circuits that serves as a source for future hypotheses. Here, I propose a general methodology for the identification of interactome modules hidden in sorted lists derived from high-throughput experiments, such

as a ranking of candidate genes from Genome Wide Association Studies (GWAS) or Differentially Expressed Genes (DEGs) from transcriptomics analysis. First, a framework to collect and assemble protein interactomes is defined, with special emphasis in the quality of PPIs. Next, a research is conducted to identify a parameter that help us to identify patterns of cooperativity, in terms of PPIs, in a list of genes. Following, we propose an algorithm to identify relevant subnetworks in a ranked list of genes. Finally, the method is applied to a real case, a GWAS study in Bipolar Disorder.

With the decrease in the economic and time costs of sequencing an human genome, different research initiatives are focusing in sequencing large amount of human genomes from both healthy donors and disease patients. Opposite to previous technologies such as DNA microarrays, sequencing technologies discovered a variation in the genome that is orders of magnitude greater than the one facet previously. Among this variation, an unexpectedly high number of apparently deleterious variants have been discovered in healthy human populations, which suggests that observing the occurrence of a deleterious variant is a necessary, although not sufficient, condition for it to have a pathological effect. This fact poses a major challenge for clinical geneticists in the hunting for of true disease-causing mutations in personalized medicine. Hence, there is an urgent need for methodologies to guide in the selection of causal or susceptibility genetic alterations. We claim that a prior step would necessarily face first with the challenge of deciphering why deleterious mutations can have a pathological effect in some individuals but cause not such obvious effect in other carriers. Taking the interactome as a model that reflects a certain degree of complexity of the cell physiology, the **third chapter** investigates its role in enabling deleterious mutation burden in human populations to be compatible with normal condition. The hypothesis stated here is that the actual interactome topology could be buffering the impact of deleterious variants, thus permitting what seems to be a high mutation load. To test to which extent this hypothesis is compatible with the observed genetic variability in human

populations, I perform a systematic study of deleterious variants from 1,330 exomes of healthy humans. First, a study of the global topological properties of proteins concentrating deleterious variants is shown, with special emphasis to those associated to cancer and Mendelian diseases. Next, interactome integrity of each individual is assessed, followed with a functional an structural study to characterize the roots of the integrity maintenance. Finally, the results are compared to somatic variants from 42 Chronic Myeloid Leukemia (CLL) patients.

The third chapter gives us some hints about how drivers can disregulate signalling circuits by affecting key proteins in the network. The **fourth chapter** is motivated by the poor knowledge on how mutations affecting to PPIs are related to cancer occurrence and progression. The research presented here goes a step beyond and describes the analysis of somatic mutations from 5920 cancer patients of 33 different cancer types in the context of the three-dimensional (3D) structurally resolved interactome. This new high resolution version of the interactome has the advantage that it can provide testable mechanistic hypothesis instead of abstract graph entities. As a first step, the systematic distribution of missense mutations among interacting interfaces is studied. Next, a protein centric-approach is used to predict PPI interfaces with statistically unexpected mutation rates. Through an example, the presented study demonstrates that mutations in different interacting sites of the same gene significantly correlate with different clinical outcome, thus providing a mechanistic explanation for patient heterogeneity.

Finally, the manuscript closes with a summarized discussion of the results and an enumeration of the general conclusions achieved.

CHAPTER 2

A methodology for functional profiling using protein-protein interactions data

Part of the work presented in this chapter is adapted from the following publication:

Garcia-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158–e158

2.1 Overview and objectives

As described in the introduction, there is extended consensus on the modular organization of the cellular functions embedded in the molecular interaction networks (Ravasz et al., 2002; Han et al., 2004; Rual et al., 2005). This modular property acquires particular importance in the cases when the molecular mechanisms underlying a disease are unknown. Here, we can use the results of a high-throughput experiment, describing a particular condition, to query the interactome scaffold to identify subnetworks enriched in the most relevant molecules from the experiment. These subnetworks can serve as hypotheses building scaffolds that lead future research (Ideker and Lauffenburger, 2003; Mitra et al., 2013). Contrary to the methods using Gene Ontology (GO) or other discrete categories, biological networks do not pre-define functional modules, instead they help to build it. The advantage is, therefore, its potential to discover new functional modules instead of being limited to the known ones.

Numerous approaches have been proposed to query the interactome with other genome-wide data to seek biological modules (Ideker et al., 2002; Ideker and Lauffenburger, 2003; Mitra et al., 2013). Most of these methods have been designed to deal specifically with gene expression data and use a scoring function based on the values of differential expression (node-based methods) or co-expression (edge-based methods). Such scoring functions are applied together with different search strategies to identify the subnetwork with the highest score compatible with the observed gene expression. However, the complexity of the interactome generates an enormous search space, which makes of the labour of finding subnetworks a NP-hard problem, as requires infeasible run-times. Other simpler methods rely on the pre-selection of gene sets with imposed thresholds, which constitutes a drawback since these threshold are often arbitrary and may affect the final biological conclusions (Minguez et al., 2009). A complete revision of the state of the art of the existing methods can be found in Mitra et al. (2013).

In this chapter, I propose a new approach to identify functional interactome modules applicable to any kind of genome-wide experiment that avoids the necessity of pre-selecting genes with arbitrary thresholds. Specifically, here I focus on developing an heuristic methodology for subnetwork enrichment analysis using the protein interactome. The chapter work-flow is structured according three main objectives:

1. To build a network by collecting and curating all experimentally-derived PPIs.

2. To develop of a methodology for identifying subnetworks associated to high-throughput experiments. This step would require:

   (a) To identify a parameter that help to identify a subnetwork enrichment.

   (b) To establish a framework for the identification of subnetwork enrichment in a list of genes ranked according its relevance in a high-throughput experiment.

3. Application of the method to extract and characterize disease-associated subnetwork from:

   (a) GWAS in Bipolar disorder.

   (b) Differential expression analysis in Fanconi anemia.

2.2 Materials and methods

2.2.1 Construction of protein interactomes

There are several primary repositories for protein-protein interactions determined experimentally, which differ in the data collection protocols, scope, accuracy and annotation quality criteria (Ooi et al., 2010). For this purpose, we generated two different interactomes: one including all the physical protein-protein interactions, and a high-quality interactome based on those interactions detected with, at least, two different techniques (Von Mering et al., 2002).

First, PPIs datasets were downloaded from three main sources: MINT (Ceol et al., 2010), IntAct (Aranda et al., 2010) and BioGRID (Stark et al., 2011) downloaded on April 2011. These databases provide the data in the PS-MI format (Hermjakob et al., 2004; Samuel et al., 2007) which contains the minimum information required for reporting a molecular interaction experiment (MIMIx) (Orchard et al., 2007). The terminology included in these file follows a controlled vocabulary organized in the Molecular Interactions (MI) ontology, fact that makes the different resources to be comparable and integrable in a straightforward way.

Next, a common protocol was applied to build both interactomes. First, only PPIs between proteins in the UniProt Swiss-Prot were selected (Consortium, 2011). Secondly, self-interaction were discarded. Finally, PPIs which interaction type implies a "physically association" (MI:0407) were retained. Additionally, to build the high-confidence interactome, we applied an strict filter to remove potential artefactual interactions by picking only those interactions detected with two different experimental methods. To avoid taking PPIs determined through similar experiments (e.g. "two hybrid array" and "two hybrid pooling approach"), MI ontology terms under the "experimental interaction detection method" category were mapped to the six top terms (Minguez et al., 2009).

Finally, species-specific interactomes were build for *Arabidopsis*

thaliana, *Drosophila melanogaster*, *Escherichia coli (strain K12)*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* .

2.2.2 Modelling the interactome as a network

As mentioned in the introduction, taking all the PPIs together, these can be modelled as an undirected graph, where nodes represent proteins and each physical interaction between them defines an undirected edge (edges that do not have an assigned direction).

The resulting network acquires a particular shape, called topology. Some topological properties can help to understand the biological role of the elements in the network (Yeger-Lotem et al., 2004). Graph theory has established the basis for studying these proprieties (Luscombe et al., 2004). Although there are several topological parameters that may used to describe a network, here there is a selection of the parameters that provide a biological meaning:

- **Connection degree** (k), may be the more intuitive parameter, refers to the number of edges (PPIs) of a protein.

- **Shortest path** is a measure of the centrality between two nodes and is defined as the minimum number of edges needed to be traversed in a network to get from one node to another.

- **Betweenness centrality** $C_b(\nu)$ is a measure of the relevance of a protein within the network. Is defined as the fraction of shortest paths between protein pairs s and t that pass through the protein of interest ν. It is calculated as:

$$C_b(\nu) = \sum_{s \neq \nu \neq t \in V} \frac{\sigma_{st}(\nu)}{\sigma_{st}} \tag{2.1}$$

  where $\sigma_{st}(\nu)$ denotes the number of shortest paths which pass through a node and σ_{st} the total number of shortest paths in the

graph. **Relative betweenness centrality** $C'_b(\nu)$ is calculated as:

$$C'_b(\nu) = \frac{2 \times C_b(\nu)}{n^2 - (3n - 2)} \tag{2.2}$$

where n is the total number of proteins in the graph. Normalizing by dividing by the maximum betweenness centrality establishes comparability between graphs of different sizes.

- **Closeness centrality** $C_c(\nu)$ Another measure of centrality. Represents the distance from a given node (ν) to the remaining nodes and accounts for how close is (ν) from any other node. It is defined by the inverse of the average length of the shortest paths from (ν) to all the other nodes in the graph:

$$C_c(\nu) = \frac{n - 1}{\sum_{s \neq \nu \in V} \sigma_{s\nu}} \tag{2.3}$$

- **Clustering Coefficient** $H(\nu)$ measures the degree to which the direct neighbours of a node ν tend to be connected between them as well. It is calculated as:

$$H(\nu) = \frac{2e_n}{n_\nu(n_\nu - 1)} \tag{2.4}$$

being e_n the number of edges among the proteins connected to the protein ν, and n_ν the number of neighbours of protein ν.

Network properties studied along this thesis were calculated using *igraph* library in R 2.12.3.

2.2.3 Identification of candidate functional subnetworks

Given a list of genes of biological relevance, we define its corresponding interactome module as the subnetwork formed by the shortest paths that connect the genes from the list. This subnetwork is called the Minimal Connected Network (MCN). However, obtaining the MCN

associated to the genes relevant in a high-throughput experiment is not enough to consider this as a module of action. Instead, an evaluation of the candidate MCN is needed to determine whether there is an enrichment in terms of PPIs. Our assumption is that a list of genes representing a functional module should form a MCN which properties differ from the random expectation. Thus, we can evaluate the candidate MCN by assessing how different is from MCNs obtained with lists of randomly selected genes.

MCNs can be characterized by the local topological properties that each node acquires in the network (node-level parameters) or by global properties related to the whole network (network-level parameters). Among the node level parameters, there is the connection degree, relative betweenness and clustering coefficient, all of them described at the section 2.2.2. Among the network-level parameters we can quantify the total number of nodes, number of connections and number of components (i.e. a set of proteins connected among them and separated from the rest). Furthermore, we can also combine the global network features and test the average number of nodes or connections per component.

False Discovery rate (FDR) and power

With the purpose of finding a network property that helps us to characterize a real functional module in the interactome, different parameters were evaluated in terms of power or sensibility (true-positives rate) and false discovery rate or specificity (false-positive rate). Power is defined as the probability of declaring a MCN as significantly different from a random subnetwork when it has been obtained from a list of actual functionally-related proteins. On the other hand, the false-positive rate is defined as the probability of declaring a MCN as significant when it is obtained from lists of randomly selected proteins.

The different nature of parameters requires a distinct strategy of network evaluation:

- **Testing node-level parameters.** The comparison of two net-

works through their node-level parameters can be seen as checking if the two samples of measurements (e.g. the connectivity degree values taken by the proteins in a real MCN against those taken by the proteins in a random MCN) are different or not. Non-parametric tests as the two sample Wilcoxon test, Kolmogorov–Smirnov test and the common area under both distributions (Martínez-Camblor et al., 2008) were studied. The use of non-parametric tests is more appropriate as the normality of parameter values for the interactome proteins can not be assumed.

- **Testing global network parameters**. Global network parameters are described with a single value (e.g. number of components in a real MCN against the number of components in a random MCN) and are, consequently, easier to interpret but more difficult to compare in terms of statistic significance since no null distribution models are easily available for all measured parameters. Thus, given a distribution of values generated from random networks, a statistical p-value can be estimated as the corresponding percentile of the parameter value over that null distribution built with random MCN.

False discovery and power rate tests were performed using the *S. cerevisiae* interactome. *S. cerevisiae* PPIs seem to be reached the best coverage and is likely to be the most comprehensive description of a species-specific interactome in eukaryote.

Generation of real MCNs (functional modules)

Functional network modules described in the literature in (Roguev et al., 2008; Han et al., 2004; Shachar et al., 2008) were used as *bona fide* real subnetworks. Also KEGG pathways (Kanehisa et al., 2011) and GO terms (Ashburner et al., 2000), which are known to be rich in network component (Minguez et al., 2011; Minguez and Dopazo, 2010). Specifically, GO-defined modules among levels 6 and 12 were selected to avoid

either general or highly specific GO terms. Since we may work among a range of input list sizes, lists containing 20, 50 and 100 proteins were collected from all these sources. A total of 156 modules were analyzed.

Generation of random MCNs

In order to approximate the false discovery rate that the combination of a parameter with a particular test can produce, it was necessary to build up a collection of MCN from lists of randomly chosen proteins. To cover the broad extend of possible conditions, we generated random list containing 20, 50 and 100 genes, for both the curated and the non-curated interactome. To obtain a distribution of values for any of each conditions, 2000 random samples were generated. The values derived from the random MCN can be used as a pre-calculated confidence interval when a MCN found in a new dataset is tested.

2.2.4 Algorithm for subnetwork enrichment analysis in a ranked list

Here we aim to define an approach to query the interactome with genome-wide data to identify biological modules. Instead of taking a subselection of the genes based on a fixed threshold, the algorithm proposed starts with the complete list of gene identifiers involved in a genomic experiment, ranked according the relevance to the conditions studied (eg. differential expression statistical score when comparing two conditions). The ranking parameter is, therefore, used as a guide to scan for subnetwork enrichment through the entire ranked list of molecules. This strategy, similar to the GSEA strategy, avoids the imposition of a gene-based threshold to pre-select a limited number of genes for further network enrichment analysis. In contrast, the algorithm seeks for sets of genes connected among them and coordinately associated to high (or low) values of the ranking parameter. Since we look for a subnetwork property, there is no need in pre-selecting a fixed number of nodes based on arbitrary thresholds.

The algorithm proposed starts with a ranked list $S = (g_1, ..., g_n)$ of n molecules. The ranked list S is subdivided into a sequence of additive partitions $S_k = (g_1, ..., g_k; k \leq n)$ of size k. The proteins corresponding to any of the partitions are mapped onto the interactome scaffold and the MCN is extracted. For the identification of the MCN, the shortest paths among all the pairs of nodes in the list are calculated using *Dijkstra* algorithm (Dijkstra, 1959). Then, the parameter of interest (z_k, defined as the average nodes per component of the MCN) is calculated for each MCN. Finally, we seek for the most relevant partition (the sub-list S_best) as follows:

1. First, ordering the parameter of interest z_k according to the ranked list, all relative maxima are identified. The partitions selected S_k^{max} represent partitions incorporating a new protein capable of connecting to the proteins in the previous partitions.

2. Next, the score L_k is computed as $L_k = (z_k - 1)/(k - 1)$ for all the selected partitions S_k^{max}. The score can be seen as a balance between the increase in connected nodes and the distance to the top of the ranked list ($k = 1$). We choose the partition S_{best} and index k_{best} corresponding to the highest L_k score computed in b) form the S_k^{max} chosen in (a).

3. Finally, an empirical p-value is calculated as the proportion of random sub-lists of k_{best} molecules (which corrects the size effect) with an average of nodes per component greater than z_{kbest}.

When the MCN is build for each partition, only proteins contained in the partition are considered (that is, direct PPIs). However, we can also consider another scenario in which proteins not included in the partition are used to connect proteins contained in the partition (called external intermediates).

Additionally, the algorithm considers the option of incorporating seed genes, that is genes from which the MCN should be build. These seeds may represent genes that are of interest because they have already

been associated to the condition under study. In this case, the seed list $S_{seed} = (g_1, ..., g_m)$ of m molecules is forced to be part of the whole list, S_K, defined as $S_K = S_{seed} + S_k$. The selection procedure is the same than described above but keeping always the S_{seed} molecules within the list.

All statistical tests were performed using R software environment. The final algorithm was introduced in Babelomics (Medina et al., 2010), a web platform for the analysis of omics data with advanced functional profiling.

2.2.5 GWAS analysis in bipolar disorder

Anonymous genotype data from bipolar disorder patients was downloaded from the WTCCC (Burton et al., 2007). A total of 2000 Caucasian patients and 1500 controls, both from United Kingdom genotyped on the Affymetrix $500K$ mapping array were analysed. Basic association test from Plink toolset was used to perform the GWAS (Purcell et al., 2007), which compares allele frequencies between cases and control. The association study generated a list of SNPs ranked by the p-value, an indicator of the strength of the association. Next, SNPs mapping within or in the neighborhood of genes were retained. Finally, the list was filtered so that the SNP with the smallest p-value from each gene was retained. Thus, we obtain a list of genes ranked according the p-value of the most associated SNP. This SNP to gene mapping is identical to the performed by similar functional-based approaches such as PBA (Medina et al., 2009; Wang et al., 2010). This ranked list is used to illustrate the performance of the algorithm presented in the previous section.

2.3 Results and discussion

2.3.1 Collection and curation of protein interactomes

PPIs sources comparison

The first step of this thesis was to integrate and curate the three main PPIs sources (IntAct, MINT and BioGRID). With this purpose, PPIs datasets were downloaded and curated as described in section 2.2.1. First, proteins from different sources were referred to the corresponding uniprot identifier and, therefore, they were comparable. Tables 2.1 and 2.2 show the overlap among proteins and PPIs described in the databases. Intact showed the best overlap. Either the protein and the PPI overlap between databases was not high enough, which ratify that a methodology to combine different PPI sources as developed here is indispensable to achieve a proper coverage.

	IntAct	MINT	BioGRID
IntAct	51,112		
MINT	81,51 %	30,890	
BioGRID	73,84%	59,29%	31,720

Table 2.1: Overlap among proteins described in databases. Overlap in interactor proteins (as of January 2011) enlisted in the respective databases IntAct (Aranda et al., 2010), MINT (Ceol et al., 2010) and BioGRID (Stark et al., 2011). Each column shows the overlap as a percentage of the total protein number of the database. The diagonal shows the absolute number of proteins in each database.

	IntAct	MINT	BioGRID
IntAct	205,686		
MINT	65,12%	87,901	
BioGRID	40,95%	59,98%	128,779

Table 2.2: Overlap among PPIs described in databases. Overlap in PPIs (as of January 2011) enlisted in the respective databases IntAct (Aranda et al., 2010), MINT (Ceol et al., 2010) and BioGRID (Stark et al., 2011). Each column shows the overlap as a percentage of the total PPIs number of the database. The diagonal shows the absolute number of PPIs in each database.

PPIs curation process

Figure 2.1 describes the number of non-redundant proteins and PPIs obtained after each step of the curation process (Figure 2.1). A total of 693,079 interactions were downloaded, and 537,682 (77.6%) of them were confirmed to be physical interactions between proteins. The rest of interactions were removed when a Uniprot identifier for one of the interactors was not found (0.35% of interactors) or when a physical interaction was not experimentally proved (22.1% of interactions). In the verified physical PPI dataset, 302830 (56.3%) were predicted to be of high-confidence. The large number of PPIs that could not be considered as high-confidence PPIs by the "two different detection methods" criteria (Von Mering et al., 2002) evidences the need for an additional quality assessment.

Protein interactomes per species

As a result of the curation process, two scaffold interactomes were generated for each species: a "curated protein interactome" containing high-confidence physical PPIs detected with two different techniques, and a "non-curated protein interactome" containing all physical interactions. The following figure summarizes the number of non-redundant proteins and PPIs obtained per species (Figure 2.2):

The species with larger amount of PPIs were *S cerevisiae* and *H sapiens*, which reached almost 90,000 interactions. Focusing further in PPIs detected by at least two different experimental methods, *S cerevisiae* displayed the largest set whereas the human interactome decreased to the same size than *D melanogaster*. Thus, the proportion of PPIs detected with two different techniques among all PPIs ranges from 26.2% in humans to 37% in yeast and 55.9% in fruit fly.

The PPIs coverage per species (the number of known PPIs among the number of expected PPIs) can be evaluated by comparing with estimators of the veritable interactome size. While empirical studies estimate over 30,000 PPIs in *S cerevisiae* (Von Mering et al., 2002), the

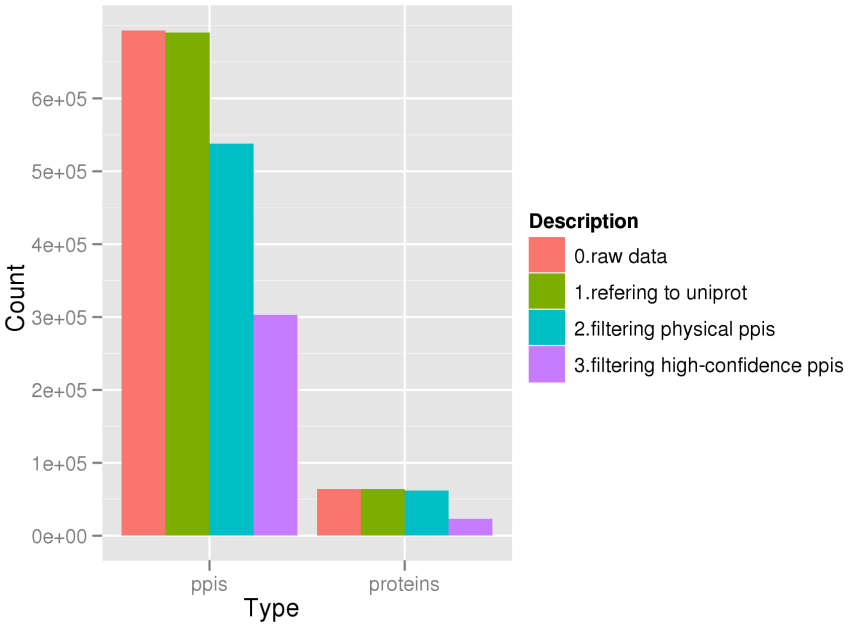

Figure 2.1: Number of non-redundant proteins and PPIs at the different points of the curation process. Bars represent the absolute number of non-redundant proteins and PPIs (first chart and second chart respectively) per curation process step.

estimated size of the human interactome ranges between 200,000 and 375,000 interactions PPIs (Bork et al., 2004; Ramani et al., 2005). Then, assuming that PPIs not detected with two different techniques are false-positives, the coverage for *S cerevisiae* interactome is 107%. Moreover, since the expected false-positive rate rages from 12% to 50% (Ito et al., 2001; Mrowka et al., 2001; Venkatesan et al., 2008) and the percentage of PPIs detected with only one technique is about 63% in *S cerevisiae*, it is possible to hypothesize that the non-curated interactome may contain true PPIs and, consequently, the *S cerevisiae* it would be greater than expected. Due to the high coverage, *S cerevisiae* interactome may be considered suitable for testing methodologies working with a PPIs scaffold.

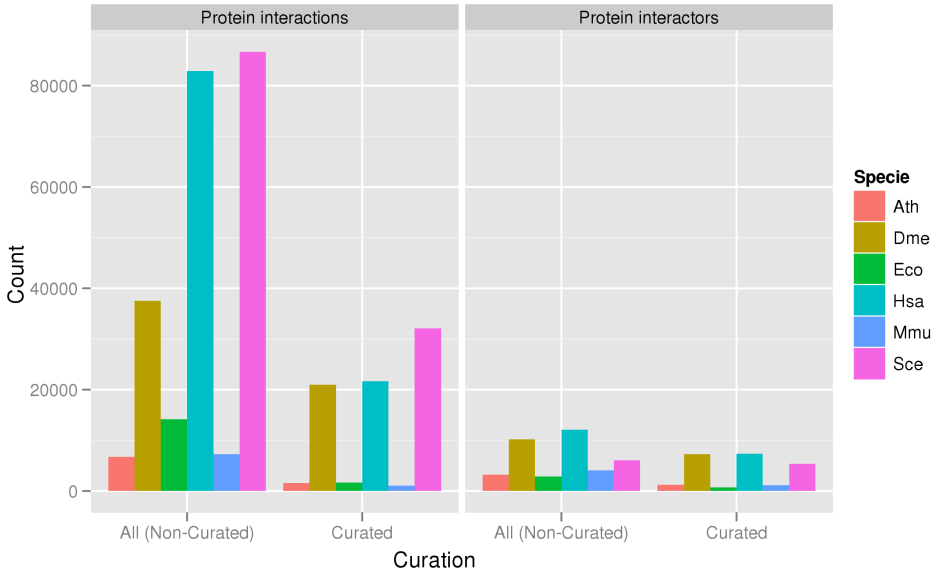

Figure 2.2: Number of non-redundant proteins and PPIs per species. Bars represent the absolute number of non-redundant proteins and PPIs (first chart and second chart respectively) per species and curation depth. *Arabidopsis thaliana* (Ath), *Drosophila melanogaster* (Dme), *Escherichia coli (strain K12)* (Eco), *Homo sapiens* (Hsa), *Mus musculus* (Mmu) and *Saccharomyces cerevisiae* (Sce) .

With respect to the human interactome, used in future analyses of this thesis, the coverage of the high-confidence PPIs elucidated among the estimated size (200,000-375,000 PPIs) is around 6-10%. Considering all the retrieved PPIs, the coverage grows to more than 22-40%. This observation indicates that much more experimental work is needed to achieve the expected number of PPIs.

2.3.2 Study of the network parameters characteristic of real functional subnetworks

Once the protein interactome (the working scaffold) has been defined and characterized, it is possible to move to the second part of this chapter: to develop a methodology for identifying an enrichment subnetwork in

high-throughput data. The first concern is to study how real functional interactome modules differ from random ones and find out a network parameter reporting this difference. The performance of the different network parameters was tested in the *S cerevisiae* interactome since is expected to be the most complete.

For this study, we have collected real functional subnetworks that were used as a test set on which the efficiency of common local and global network parameters was assessed in terms of power and false-positive rates. The collected real functional modules include KEGG pathways, subnetworks described in the literature and some GO modules. Proteins from these real modules were extracted into lists that cover different list sizes (20, 50 and 100 nodes). Each list was mapped onto the interactome and the corresponding MCN was extracted. MCN derived from real subnetworks with a clear function were compared to MCN generated from random lists of the same size. Thus, for a given number of genes N, an empirical simulated distribution can be derived by repeatedly selecting N genes randomly from the genome, then looking for the MCN that connects them and measuring the parameter of interest. Repeating this procedure 2000 times allows deriving the distribution sought.

As described in Materials and Methods section 2.2.1, we have studied two scenarios:

- (i) Subnetworks found within sets of proteins with direct connections among them

- (ii) Subnetworks found within sets of proteins with either direct connections or connected through one intermediate protein not present in the set.

The second scenario (ii) represents a common situation in large-scale genomic analysis. In many proteomic analyses, some of the proteins present in the sample are simply not detected because of the sensitivity of the technique. In the case of transcriptomic experiments, it is quite common that the noise affecting to individual probes representative of

the genes (and the corresponding gene products) makes some of them present different values of the statistic. In an ideal situation, a group of proteins that co-express and conform, for example, a complex should appear together in a differential expression experiment and should easily be detected by a conventional test that look for network enrichment. In a real situation, it is quite common that as a consequence of noise or experimental errors some proteins of the subnetwork are missing in the experiment (in spite of being actually involved in the network structure). It can also happen that some proteins (key in the definition of the network) do not change their expression across the compared conditions, thus a differential expression experiment did not report them in the result. Thus, looking for networks within a set of proteins, allowing for some connections provided by proteins not in the set, increases enormously the sensitivity of the network detection method and makes it more robust against noise. It also allows overcoming some intrinsic limitations of experimental designs based on differential expression, such as the difficulty of detecting networks in which some of the nodes do not differentially express across the conditions compared.

As stated in subsection 2.2.2, there are several parameters and statistical tests that we can use to perform subnetwork comparisons. Some of them are defined at the node level while others are global network features. The following sections show the power and false-positive rate observed for each network parameter and comparison method under different list sizes and MCN inference approaches (allowing or not external intermediates).

Node level parameters performance

In the node-level parameters test, the objective is to identify which combination of parameter (connection degree, betweenness and clustering coefficient), test (two sample Wilcoxon test, Kolmogorov–Smirnov test or check the common area under the distribution) and sampling method (choosing randomly one node from each random MCN, calcu-

lating the mean or applying a bootstrap procedure) displays the best performance in distinguishing between real and random MCNs. Figures 2.3 and 2.4 represent, respectively, the false-positive rate and the discrimination power for each tested combination.

As shown in Figure 2.3, when comparing node level parameters in MCN without intermediate nodes, the false-positive rate taking the mean reference were not suitable. Also, bootstrap sampling procedure was not sufficient for clustering coefficient parameter and common area test analyses. Random reference seemed to be the most satisfactory for both false and true positive rate performances. Regarding the test, Kolmogorov-Smirnov two sample test provided the best balance between the discrimination power and the false-positive rate for all parameters.

With respect to the performance in MCN allowing an intermediate node (see Figure 2.4), there was a marked increase in false-positive rate, with the exception of the bootstrap sampling method, which gives a marked decrease in the discrimination power. Then, no tested combination is appropriated when intermediate nodes are allowed in the MCN.

Results are in accordance with the high diversity in subnetworks shapes. Subnetworks representing a signalling pathway are expected to be little connected and exhibit an elongated shape whereas subnetworks representing protein complexes are dense connected and display a ball shape (Johnson and Hummer, 2011). Since betweenness, clustering coefficient and connection degree are highly shape dependent, they seem not suitable for MCNs discrimination.

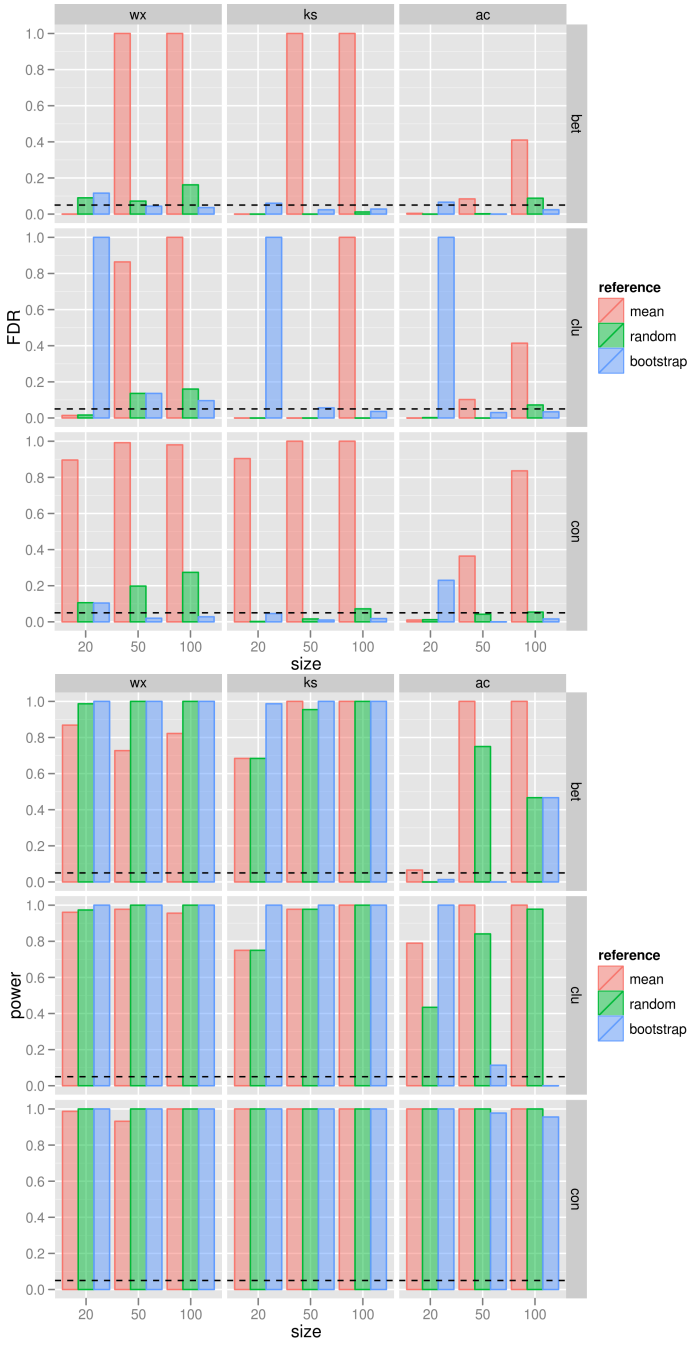

Figure 2.3: Comparative analysis of the discriminatory power of node-level parameters with not intermediates. The X axis accounts for the MCN size whereas Y axis for the false discovery rate (FDR) and true-positive rate (power). Arrangement of charts in rows and columns correspond to topological parameter (con: Connection degree; bet: Betweenness; clu: Clustering coefficient) and test (Wx: Wilcoxon two sample test, ks: Kolmogorov-Smirnov two sample test; ac: Common area test). Colour correspond to the reference taken for comparison.

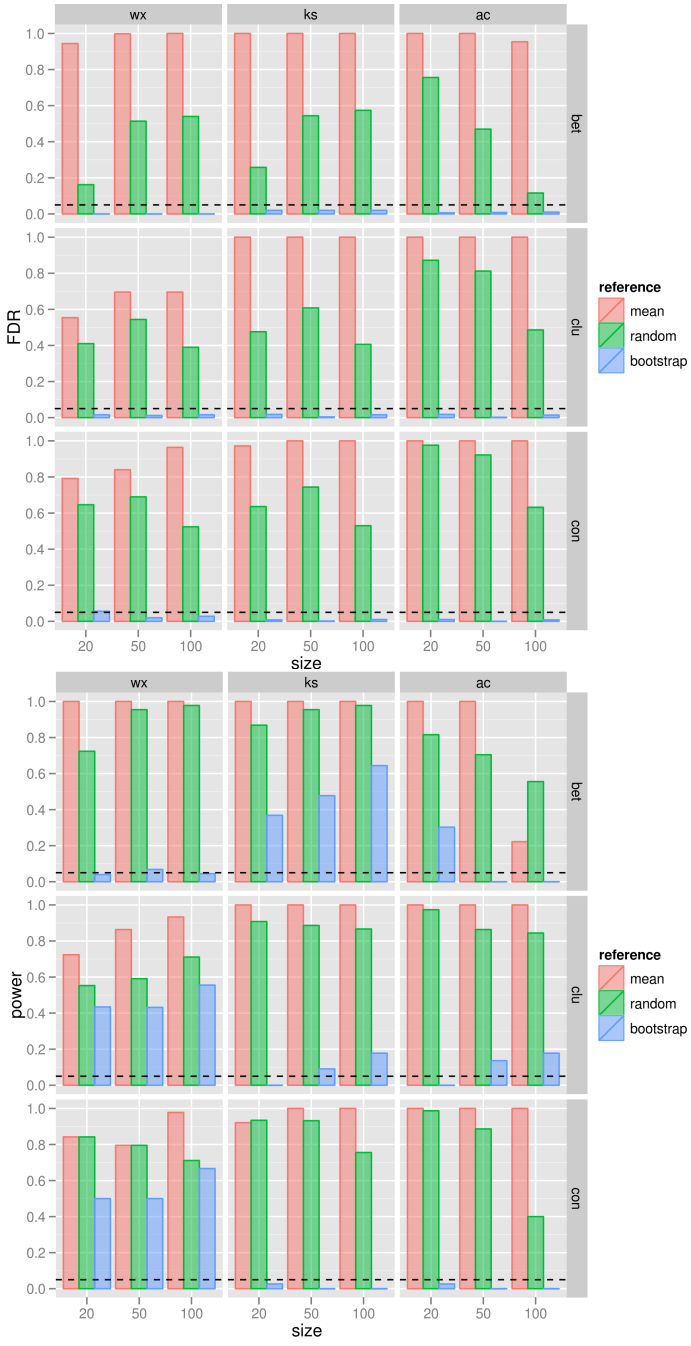

Figure 2.4: Comparative analysis of the discriminatory power of node-level parameters considering an intermediate. The X axis accounts for the MCN size whereas Y axis for the false discovery rate (FDR) and true-positive rate (power). Arrangement of charts in rows and columns correspond to topological parameter (con: Connection degree; bet: Betweenness; clu: Clustering coefficient) and test (Wx: Wilcoxon two sample test, ks: Kolmogorov-Smirnov two sample test; ac: Common area test). Colour correspond to the reference taken for comparison.

Global network parameters performance

In the evaluation of global network features evaluation we dealt with only one number (e.g. number of components in the subnetwork under study) and not a vector of values (distribution of topological values acquired by the proteins in the network under study). The concern was to determine whether the global number of connections, nodes, components and average number of nodes or connections per component was greater than random. It is possible to estimate a p-value simply by accounting for the frequency of such feature in the reference set.

Figure 2.5 shows that the most sensitive among the global network parameters is the average number of nodes per component. This feature also demonstrates to be robust to the inclusion of intermediate nodes. Thus, the average number of nodes per component was selected to define a subnetwork enrichment. Biologically, the best performance of the average number of nodes per component proves that the only constraint in real subnetworks is that subnetwork members should aggregate to a connected component, independently of the subnetwork shape and edge density.

2.3.3 NetworkMiner: a tool for subnetwork enrichment analysis in a ranked list

Defined the parameter discriminating between real and random MCNs, it is possible to introduce a method to search for subnetwork enrichment in high throughput data. High-throughput experiments end with a list of molecules ranked according its corresponding measurement. In principle, the ranking values are supposed to be derived from a genomic experiment and must have, consequently, a biological meaning. For example, it can be the value of a t-test statistics derived from a differential expression experiment, thus accounting for the higher level of expression in one of the conditions compared; it can also be a p-value in a GWAS, thus accounting for the association strength of each of the genes with a disease, etc. Obviously, this methodology is not

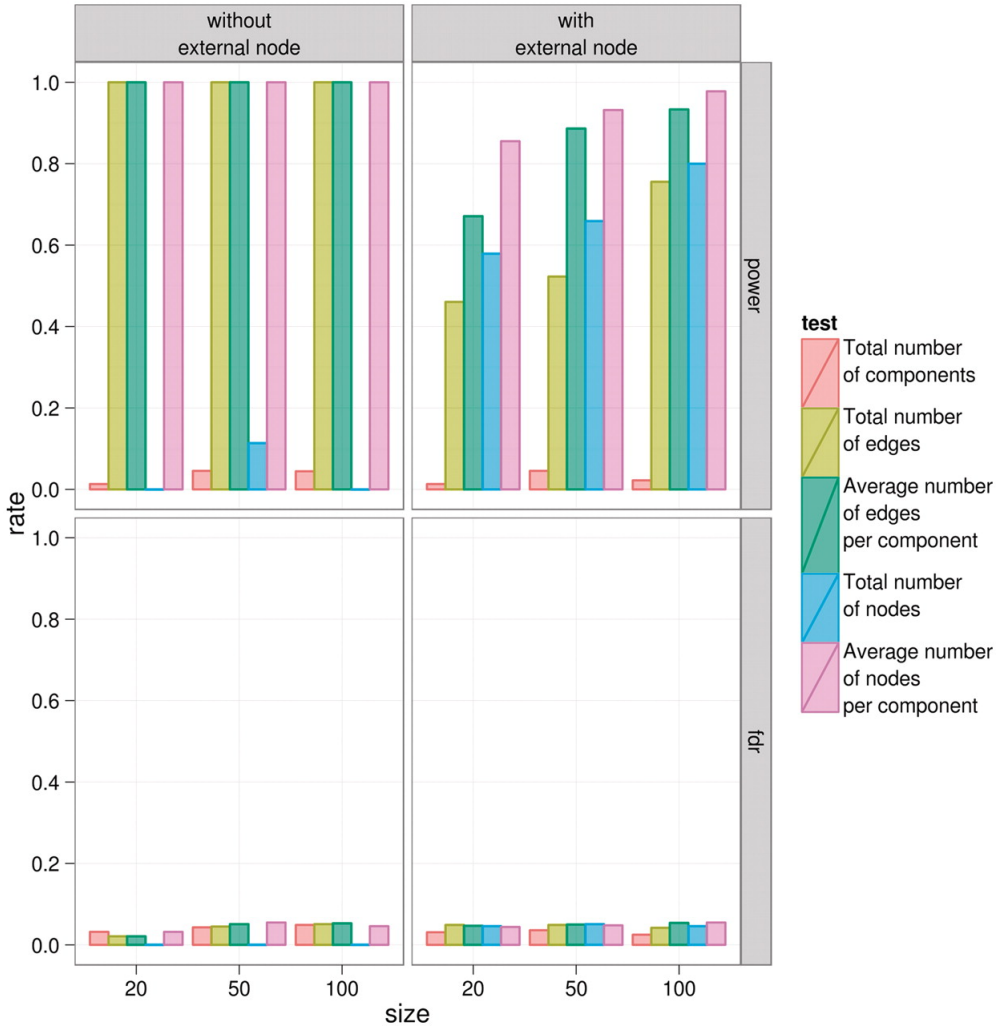

Figure 2.5: Comparative analysis of the discriminatory power of different network-level parameters. The x-axis accounts for the MCN size. Arrangement of charts in rows and columns corresponds to intermediate node inclusion and false/true-positive rates (FDR: false-positive rate; power: true-positive rate). Color corresponds to the feature tested. Image taken from (Garcia-Alonso et al., 2012).

restricted to genotyping or differential gene expression and other ranking values representing the results of other types of experiments are also possible. Then, the interpretation must be done accordingly to

the biological property that this particular ranking value is representing. This ranking parameter, which is informative of the molecule relevance for the phenotype under study, is used as a guide to scan for subnetwork enrichment through the entire list. This ensures us to avoid a threshold imposition which may affect the final biological conclusions of the analyses. The full algorithm is described in the 2.2.4 section at the Materials and Methods part. The algorithm was introduced in Babelomics, a web platform for the analysis of omics data (Figure) and it is fully documented in `https://github.com/babelomics/babelomics/wiki/Gene-Set-Network-Enrichment-(Network-Miner)`.

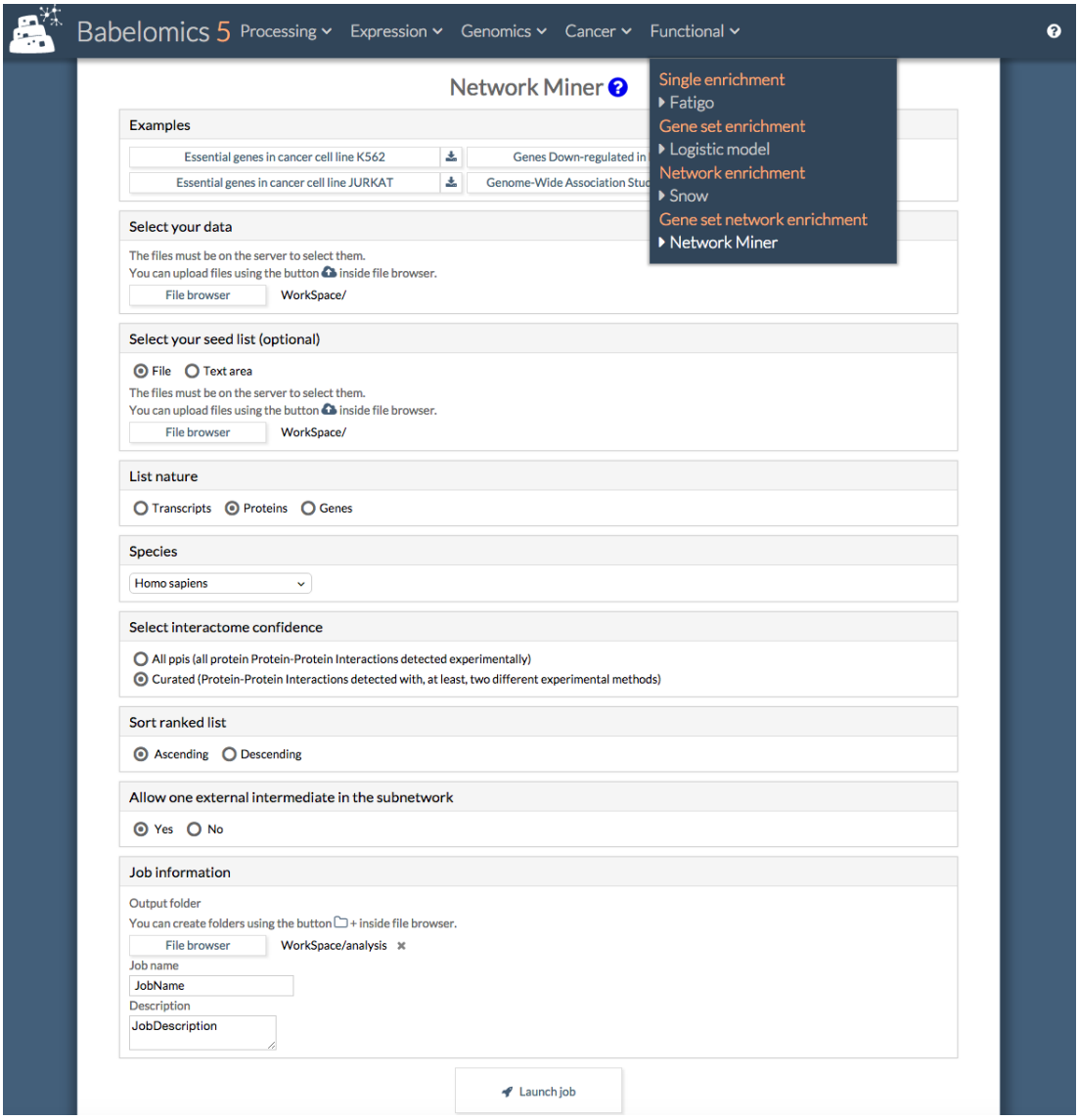

Figure 2.6: Snapshot of Network-Miner tool in Babelomics. Documentation on how to use the tool can be found in the following url `https://github.com/babelomics/babelomics/wiki/Gene-Set-Network-Enrichment-(Network-Miner)`.

2.3.4 Applications

GWAS analysis in Bipolar disorder

To illustrate the the potential of the method, we applied it to GWAS data from a bipolar disorder study (Burton et al., 2007). A total of 2000 Caucasian UK patients of bipolar disorders and 1500 controls genotyped on the Affymetrix $500K$ mapping array were studied. After performing a GWAS analysis, we obtained a list of genes ranked by the smallest p-value among its SNPs. This list is used by NetworkMiner, which looks for the significant subnetworks associated to the lowest p-value of the association test, i.e. subnetworks associated to the bipolar disorder. The network analysis was performed allowing an intermediate node in the MCN. Since several genes have already been associated to bipolar disorder, we included these as seed genes for the NetworkMiner algorithm.

Figure 2.7 shows the MCN significantly associated to bipolar disorder ($P \leq 0.05$). This MCN contains two separated components, which includes 10 genes highly associated to the disease and 11 additional genes connected to them. Three of the genes already known to be associated to the disease, FXYD6 (Choudhury et al., 2007), INO1 (Shamir et al., 2007) and (Willmroth et al., 2007), belong to the network found. The network is enriched in genes belonging to bipolar disorder related GO biological processes: learning, cognition, nervous system development ($P = 0.0364$) and, nerve growth factor receptor signaling pathway, this last one marginally significant ($P = 0.0561$).

This example is the typical case where clear genetic associations are not found mainly because its heritability depends on multiple genes of small effect size (Plomin et al., 2009). None of these small effect genes will obtain a significant value in a gene-based test, but all of them will have simultaneously a low p-value and consequently will be closer to the top side of the ranked list. If these genes are part of an interacting network, then network analysis methodologies will discover them as collectively associated to the disease through their connections.

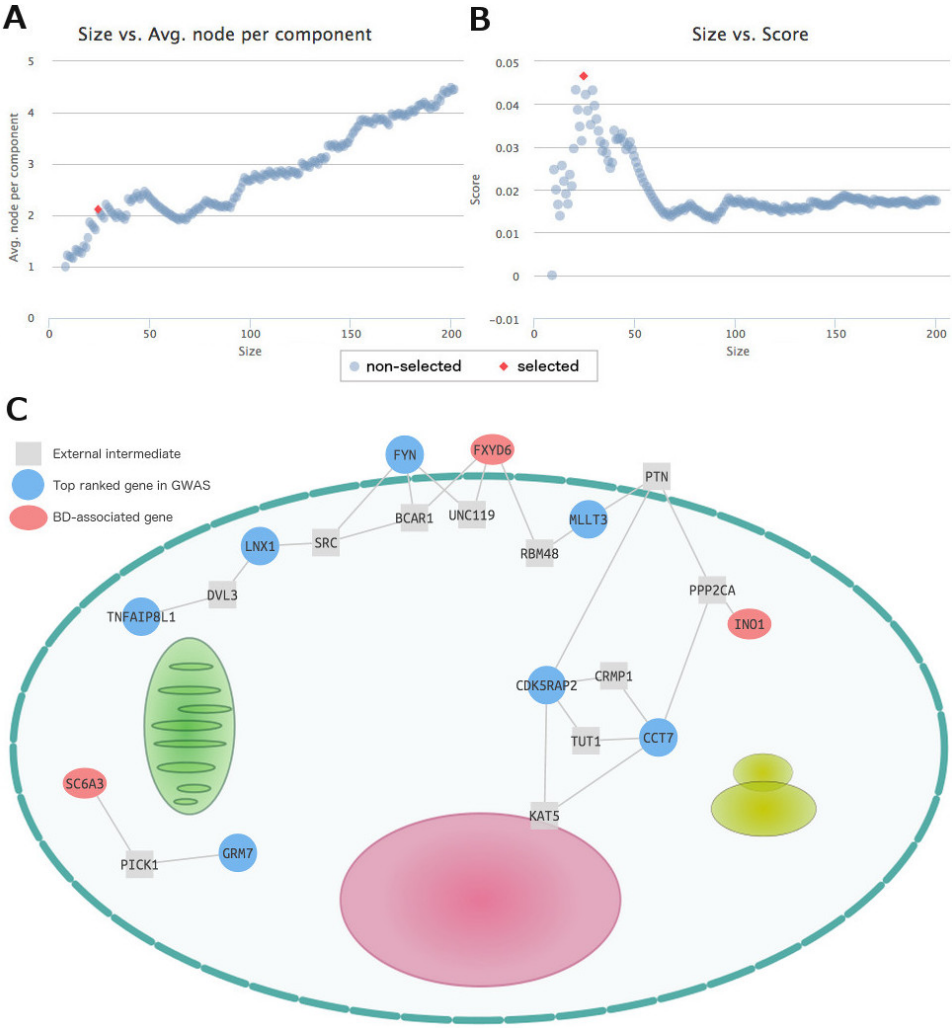

Figure 2.7: Subnetwork found among the genes most associated to bipolar disorder in a GWAS. Average nodes per component (**A**) and score (**B**) as a function of the sublist size. Significant subnetwork (**C**). Selected sublist is labelled with a red dot in the plots. Subcellular location of the genes is displayed by the cell layout used by the NetworkMiner software, based on GO cellular component.

CHAPTER 3

Role of the interactome in the maintenance of deleterious variability in human populations

Part of the work presented in this chapter was published in:

Garcia-Alonso, L., Jiménez-Almazán, J., Carbonell-Caballero, J., Vela-Boza, A., Santoyo-López, J., Antiñolo, G., and Dopazo, J. (2014). The role of the interactome in the maintenance of deleterious variability in human populations. *Molecular systems biology*, 10(9)

3.1 Overview and objectives

Deleterious genetic mutations are those predicted to affect gene function and decrease the fitness of the carrier. Contrary to expectations, genome sequencing studies of healthy human populations revealed a high prevalence of these type of mutations, with a relevant number in a homozygous state. This amount of deleterious variants poses a major challenge for personalised medicine: to distinguish between true disease-causing variants from the large background of deleterious variants (but no pathogenic) present in healthy human genomes (Xue et al., 2012). Mechanistic understanding of why deleterious mutations in some genes can have a pathological effect but cause not obvious ill effect when affect other genes remains still elusive, which supposes a major bottleneck for the application of genome sequence data in the clinical praxis.

Several mechanisms have been proposed to explain such tolerance to deleterious variants: these can have a recessive effect that requires the mutation to be homozygous to produce a disease phenotype; the disease condition might have a reduced penetrance in a way that requires additional factors for its expression; or the symptoms appear in older ages (later onset phenotypes) and gene redundancy (Xue et al., 2012; Nothnagel et al., 2011; MacArthur et al., 2012). Although valid to explain part of the buffering, fail in providing an explanation to why different individuals carrying the same deleterious mutation may display very different phenotypes. This evidences that there are additional potential sources that can confer phenotypic robustness to the carries. Here we hypothesise that the way the system is organized may provide an additional buffering mechanism of internal perturbations.

The concept of robustness in biology is gaining much attention (Masel and Siegal, 2009). Most studies about the role of biological networks in phenotypic robustness have focused on networks of genetic (that is epistatic) interactions (Wagner, 2000). Less attention has received the network of protein-protein interactions (PPIs). Theoretical studies showed how the actual topology could provide a mechanisms to buffer

random loss of its nodes (simulating random perturbation), but it was very sensitive to the removal of highly connected and central proteins (simulating an attack) (Albert et al., 2000). However our knowledge on the possible contribution of the protein interactome in buffering disadvantageous mutations has not been studied yet at the population scale.

In this chapter we aim to decipher the role of the interactome in enabling deleterious mutation load in human populations to be compatible with normal condition. Given the potential role of biological networks in assuring the robustness of cell systems against mutations, our hypothesis is that the actual interactome topology could be buffering the impact of deleterious variants, thus permitting what seems to be a high mutation load. In order to check the extent to which this hypothesis is compatible with recent observations on human variability, I accomplished the following objectives:

1. Analysis of the coding sequences (exomes) of 1,330 healthy individuals together with 41 individuals with chronic lymphocytic leukaemia to extract either germline and somatic variants with a potential impact in the protein coding genes.

2. Description of the deleterious variability found in the newly sequenced Spanish population.

3. Identification of potentially deleterious variants among individuals and its validation.

4. Population-based analysis of the global topological properties of the interactome proteins carrying potentially deleterious variants.

5. Characterization of the changes on the interactome structure of each individual caused by homozygous deleterious variants and quantification of the differences between real and simulated populations.

6. Study of the distribution of deleterious variants among the modular structure and functions of the interactome.

7. Comparison between the distribution of germline and somatic variants among the modular structure and functions of the interactome.

3.2 Materials and methods

3.2.1 Human healthy individuals exome sequencing data

The Spanish population (MGP)

Exome sequence data from 252 human individuals was retrieved from the Medical Genome Project (`http://www.medicalgenomeproject.com`). These individuals are healthy humans, with the absence of current known disease or genetic condition in the family history. However, diseases appearing at older ages cannot be completely ruled out. Since the samples were sequenced in the context of the Medical Genome Project, this population was called MGP. Samples were obtained in accordance with the approved protocols of the respective institutional review boards for the protection of human subjects. The study conformed to the tenets of the declaration of Helsinki. Data was retrieved in VCF format, which contained all the variants found in the population, without any quality filter.

The 1000 Genomes Project populations (1KGP)

Together with the MGP population, 13 other human populations we used in this study. These include: Asian populations CHB Han Chinese in Beijing, China (97 donors), CHS Han Chinese South (100 donors) and JPT Japanese in Tokyo, Japan (89 donors); American populations MXL Mexican Ancestry in Los Angeles, CA (66 donors), PUR Puerto Rican in Puerto Rico (55 donors) and CLM Colombian in Medellin, Colombia (60 donors); African populations YRI Yoruba in Ibadan, Nigeria (88 donors), LWK Luhya in Webuye, Kenya (97 donors) and ASW African Ancestry in Southwest USA (61 donors); and European populations TSI from Tuscany in Italia (98 donors), FIN Finnish from Finland (93 donors), GBR British from England and Scotland (89 donors), CEU which are Utah residents (CEPH collection) with Northern and Western European ancestry (85 donors).

The exome sequences of the human populations described above were downloaded from the 1000 Genomes Project (Siva, 2008) web page (`http://www.1000genomes.org/`) in multi-sample VCF format (February 2012 release), containing high quality variants. The total number of individuals studied in all the populations is of 1078.

3.2.2 Cancer donors

Paired samples of Chronic Lymphocytic Leukemia (CLL)

41 paired tumor and normal exome samples of Chronic Lymphocytic Leukemia patients not mutated for IGHV (Quesada et al., 2012) were considered for this analysis. The exome data was downloaded from the EGA repository (ID: EGAD00001000044) in fasta format and processed as specified in the following section.

3.2.3 Analysis of exome sequencing data

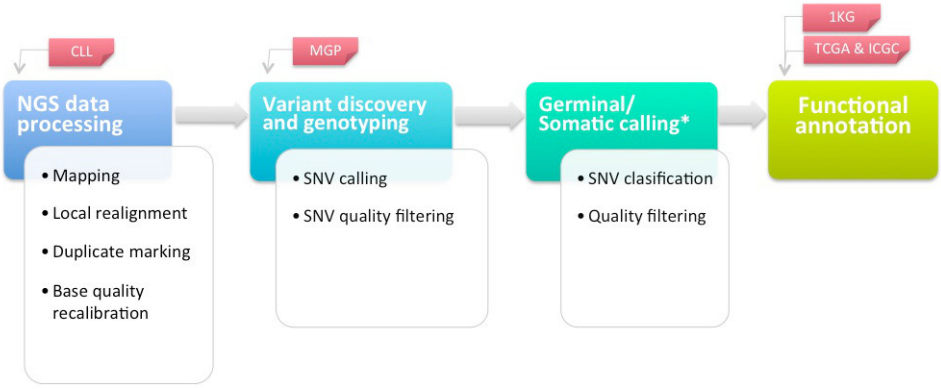

Figure 3.1: Framework for variant discovery and genotyping from NGS data. *Step only applicable to paired tumor and normal samples

From sequences to variant calls

For each dataset, the retrieved data was at different steps of the WES analysis framework. CLL data were raw sequences in fastq format, MGP were unfiltered variants in VFC format and 1KGP were filtered variants in VCF format too. Figure describes the pipeline used to process the samples and the step at which each dataset was integrated. In brief, sequence reads were aligned to the reference human genome build GRCh37 (hg19) by using the Burrows-Wheeler Alignment tool (Li and Durbin, 2009). Reads correctly mapped were further filtered with SAMtools (Li et al., 2009), which was also used for sorting and indexing mapping files. Only high quality sequence reads mapping to the reference human genome in unique locations were used for variant calling. The Genome Analysis Toolkit (GATK) (McKenna et al., 2010) was used to realign the reads around known indels and for base quality score recalibration. Identification of single nucleotide variants and indels was performed using GATK. The SNV calls were re-examined and standard hard filtering parameters were applied to remove possible artifacts(DePristo et al., 2011) considering: total read depth, the number of individuals with coverage at the site, the fraction of variant reads in each heterozygote, the ratio of forward and reverse strand reads for reads carrying reference and variant alleles, and the average position of variant alleles along a read. The somatic variant calling was carried out with the specialized software Mutect (Cibulskis et al., 2013).

Variant functional annotation

Once defined the variants for all the individuals mentioned above, its functional consequence was assessed with VARIANT software (Medina et al., 2012) and selected those affecting either the protein sequence or the mRNA transcription/translation.

3.2.4 Selection of variants with functional impact

Deleterious variants in healthy individuals

A key step is to identify those variants that have an effect on the molecular function of the gene products. First, as highly frequent variants are not expected to be damaging variants, we removed those variants with a frequency higher than 90% in the population (the reference allele is observed at a low frequency in that population). Variants located in intronic, upstream, downstream or intergenic regions, as well as variants with synonymous or unknown functional consequence were filtered out. Only nonsynonymous, stop loss, stop gain and splicing disrupting variants were considered.

Should be noted here that we are seeking for deleterious but no pathological variants. The definitions of deleterious and pathological variants proposed by MacArthur et al. (2014) were taken here. Deleterious variants are those that reduce the reproductive fitness of carriers, and would be targeret by purifying natural selection. In contrast, pathogenic mutations contribute mechanistically to disease.

There are several methods to compute the deleteriousness of a variant. Since there is no unique method that outperforms the rest, the criteria followed here combines complementary methods to achieve more reliable predictions (Thusberg et al., 2011). Specifically, the putative impact and deleteriousness of these variants was computed as a combination of: SIFT (Kumar et al., 2009), Polyphen (Adzhubei et al., 2010) damage scores and phastCons (Siepel et al., 2005) conservation score. Since the conservation score is the only parameter applicable to any type of position, we have used it as a primary filter. Thus stop loss, stop gain and splicing disrupting variants with phastCons conservation score higher than 200 are selected as deleterious. In the case of nonsynonymous variants, a SIFT score lower than 0.05 or a Polyphen score higher than 0.95 are also required to consider them as deleterious.

Computational validation of deleterious variants

To asses the reliability of the predictions, we performed an in silico structural analysis of the impact of the predicted mutations in the proteins. Only proteins structurally solved in the PDB (Bernstein et al., 1978) were used here for validation. Protein sequences were downloaded from UniProt database (Consortium, 2011) and were used to build three-dimensional models using the RaptorX program (Källberg et al., 2012). The program performs a template-based protein structure modelling, applying single- and multiple-template threading methods. The three-dimensional model was used to predict the effect that single point mutations has over the stability of protein, using the SDM software (Worth et al., 2011). SDM calculates a stability score that accounts for the free energy difference between the wild-type protein and the corresponding mutated protein.

Additionally, we used some sequence-based features, such as changes in the charge and the polarity of the protein to further assess the severity of the impact produced by the change. Changes in charge and polarity were defined exclusively on the basis of the type of residue substitution. Changes in polarity and charge were based uniquely on the residue changed. Polarity changes were measured in a hydrophobicity scale of 0 (LIFWCMVY), 1 (PATGS) or 2 (HQRKNED) (Mirkovic et al., 2004). Changes in the total protein charge were estimated on the basis of the charges of the residues: positive (RK), negative(ED) or non-charged (LIFWCMVYPATGSHQN).

Finally, we included SNAP predictions (Bromberg and Rost, 2007), since it was observed to reach the best sensitivity compared to other methods (Thusberg et al., 2011). Either SDM and SNAP were not applied systematically but only for validation purpose due to it is not optimized for high-throughput analyses.

3.2.5 Construction of a global human protein interaction network

Protein binding interactions

Section 2.2.1 already introduced the sources of PPIs used and the curation approach applied to build high confidence PPIs. The achieved interactome was modelled as described in the section 2.2.2. In the same sense, all network topological parameters used in this chapter follow the definition proposed in section 2.2.2. Network properties studied along this thesis were calculated using *igraph* library in R 2.15.1.

Determining the tissue specificity of human protein interactions

To determine which protein interactions can occur in a particular cell or tissue type of the human body, we used global gene expression data (Lukk et al., 2010). Our assumption is that if two proteins are not expressed in a tissue then the interaction cannot occur and, therefore, is removed from the tissue-specific interactome. Human gene expression data was downloaded from the online resource Gene Expression Atlas (http://www.ebi.ac.uk/gxa/array/U133A). The experiment contains consistently normalized human gene expression data matrix of 5372 samples integrated from 206 public experiments of a HG-U133A array platform gathered from ArrayExpress and GEO websites. Lukk et al. (2010) annotated these samples by adding biological variables (like cell line, disease state, organism part, developmental stage) that are not present in the original publication. By using this annotation, we selected the samples classified as "normal" for the "4 meta group" variable. Thus, a total of 1065 experiments representing 76 normal tissues where used in this analysis.

The following criteria was used to define if a gene is considered as present or absent in a tissue:

- Probe Calling at Sample Level: The MAS5 Detection Calling method from the affy R package was used to analyse the expression data.

The MAS5 Detection Calling method is based on the Wilcoxon signed rank-based gene expression presence/absence detection algorithm and identifies whether a particular transcript is present or absent.

- Transcript Calling at Sample Level: Probes mapping different transcripts were removed. A transcript was considered present in a sample if more than 50% of the mapping probes were present.

- Transcript Calling at Tissue Level: Human and mouse samples were assigned to a tissue by Lukk et al. (2010). Tissues with less than 2 samples were removed. A transcript was considered present in a tissue if it is expressed in at least one sample of the tissue.

3.2.6 Deciphering the effect of the deleterious variants on the interactome

Robustness of the interactome structure to homozygous deleterious variants of healthy individuals

The objective is to quantify the global damage that the deleterious variants cause on the interactome. To achieve this, individual interactomes were built by removing those nodes affected by homozygous deleterious variants from the network and the impact that such subtraction of nodes has on the interactome structure was studied. In particular, the impact over the interactome is assessed by measuring the following network properties: i) separation into isolated components, via the total number of components or the size of the giant component; ii) Connectivity loss: via the total number of remaining edges and iii) Increase of path lengths, by measuring the network diameter (largest shortest path) or the average path length.

Next, the extent of the damage produced by the deleterious variants on the interactomes of real individuals was studied. To evaluate so, the network properties of real individual interactomes against simulated interactomes was compared. Simulated individuals were built by

removing the same number of affected nodes selected randomly. In these simulated interactomes, the probability of a protein being affected is identical for any protein in the network. Such simulated interactomes represent the expectation of random damage on the interactome for a given number of affected proteins. We performed these comparisons at population level. Thus, for each population, 1000 interactomes with a number of affected proteins randomly sampled among the values observed in the population are generated. The average values of network properties of real and simulated interactomes are compared by means of a non-parametric Mann-Whitney test. We conducted another simulation in which proteins were removed not randomly as before but rather with a probability proportional to the observed mutation frequencies in the 1KGP population. In this scenario, the resulting simulated individuals will have deleterious variants only in proteins that are affected in normal individuals, but in random combinations that not necessarily exist in real healthy individuals. The comparison of the observed values of interactome network properties in real individuals with respect to the corresponding distribution of values obtained from the simulated population of interactomes will confirm whether the variants carried by normal population occur in the less damaging positions among all the possible locations or not.

Distribution of deleterious variants among the modular structure of the interactome

Here the interactome was divided into communities or modules by using the Walktrap algorithm (Pons and Latapy, 2006). This algorithm finds densely connected neighbourhoods, also called network communities or modules, within a graph via random walks under the assumption that short random walks are "trapped" within highly interconnected network regions. A second community detection algorithm, called Infomap (Rosvall and Bergstrom, 2008), was used to validate the results. Both algorithms were carried out using the freely available igraph R

package (`http://cran.r-project.org/web/packages/igraph/`), keeping the authors default parameters. In our study, we used only those communities composed of at least 5 proteins.

Once defined the interactome communities, we studied the distribution of the proteins containing deleterious variants. Here, for every individual, we calculated the proportion of affected proteins per module. To determine how the observed distributions deviate from to the random expectations we carried out a permutation test in which the affected proteins are distributed randomly across the interactome. Again, the probability of a protein being affected in the permutations is the same for any protein in the interactome. Then, empirical random distributions of affected proteins are obtained for each module separately by running 1000 simulations for each individual. We define the value of relative damage for each module as the percentile of the empirical random distribution corresponding to the observed proportion of affected proteins in the module. Relative damage values are rescaled between 0 (no proteins affected at all in this module) to 1 (the maximum number possible of proteins affected in this module).

Functional profiling of interactome modules

To identify the biological processes affected or protected across communities, GO enrichment test of the clusters found was carried out using the FatiGO (Al-Shahrour et al., 2004) algorithm, as implemented in the Babelomics platform (Medina et al., 2010).

3.3 Results and discussion

3.3.1 Variability in human healthy populations

Variability distribution in the Spanish population

Table 3.1 summarizes the variability in the exonic regions of the newly sequenced Spanish population. Almost one third of the variants found were private of the Spanish population, not described in dbSNP (Sherry et al., 2001), 1KGP (Consortium et al., 2010) or NHLBI Exome Sequencing Project (Fu et al., 2013). This proportion of discovery is similar to what previous observations in other sequencing projects (Fu et al., 2013). The average number of variants per individual in the coding regions of the genome analysed was about 19,000. The results document the presence of a considerable amount of potentially deleterious variation in the Spanish population. As observed in other large-scale genomic projects (Xue et al., 2012; MacArthur et al., 2012), there is an average of 1200 potentially deleterious variants per individual, of which 352 are strongly predicted as deleterious.

Figure 3.2 depicts the extent of the variability captured by the analysed Spanish population. The total number of new variants present only in Spanish population grows linearly with the number of analysed individuals and seems to be far from reaching a plateau. However, when

SNV type	Total	Avg.	Avg.*	Total (local)	Avg. (local)	Avg.* (local)
All	171406	18880.1	6906	63343	836.7	59.4
Singletons	54214	202	59.4	54214	202	59.4
Nonsynonymous	97589	9193.7	3335.5	40564	538.6	41
Synonymous	73011	9734	3596.5	21857	287.2	18
Stopgain	1852	95.8	22	1060	15.9	0.4
Stoploss	178	29.4	12	71	0.6	0.1
Splicing	4217	417.2	154.8	1842	25.1	2
Potentially deleterious	32736	1163.8	211.2	17314	141.8	3.3
Deleterious	12639	352.6	51.4	7136	51	0.3

Table 3.1: Variants summary in MGP population. Table describes the amount of total and the average per individual variants observed in MGP population classified according types. *: homozygous variants; local: MGP-specific variants (not appearing in other populations).

new variants are decomposed into rare variants (singletons) and polymorphic variants (those shared by several individuals) it is apparent that the main contribution to the private Spanish variability comes from rare variants, while polymorphic variants reached a plateau soon. Therefore, most of the polymorphisms within coding regions unique to the Spanish population have been discovered in this work and seem to be restricted to about 10,000 positions. Moreover, around one third of the variants found in the Spanish population are homozygous. This proportion decreases to a level of 7% if only Spanish-specific variants are considered. The pattern of distribution of homozygous and heterozygotes is coherent with a scenario in which most of the variants are in Hardy-Weinberg equilibrium (Stern, 1943). At low allelic frequencies of the alternative allele, heterozygotes are prevalent, while the situation is the opposite at high allelic frequencies, where many alternative alleles have been fixed in the population.

In summary, MGP population displays an excess of low-frequency nonsynonymous coding variants, most of them as heterozygotes, which agree with the observations from other populations (Coventry et al., 2010; Keinan and Clark, 2012; Li et al., 2010; Nelson et al., 2012; Tennessen et al., 2012; Marth et al., 2011; Casals et al., 2013).

Variants in proteins of the interactome among different populations

Figure 3.3 shows the average number of variants per individual in the proteins that define the interactome used in this study. As has been previously described for the complete set of human proteins in several reports of genomic variability, African populations show higher variability (over 8000 variants) than the rest of the populations (about 6500 variants), including the CLL genomes (Figure 3.3A). The average number of potentially deleterious variants (Figure 3.3B) follows a similar pattern to the total number of variants. African populations undergo more mutational load than the rest of the populations. The same pattern

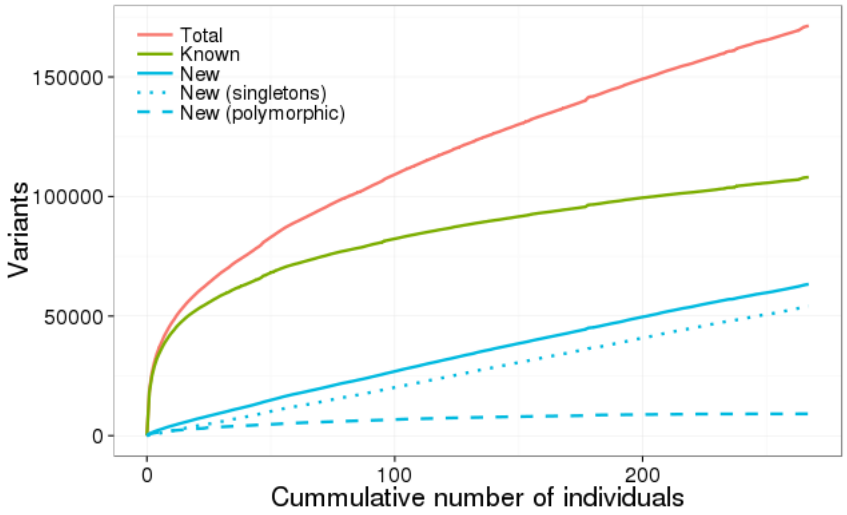

Figure 3.2: Accumulative number of new variants contributed by individuals. The red line represents the total variants, the green line represents the number of already known variants, the blue line represents the number of new variants (not present in 1KGP and dbSNP). New variants are decomposed into polymorphic (present in more than one individual of the MGP population) represented by the blue dashed line, and rare variants (present in only one MGP individual), represented by the dotted line.

is observed for the number of proteins affected by deleterious variants heterozygous state (Figure 3.3C). As expected, the Spanish population sequenced here presented a level of variation similar to that observed in non-African populations. However, this pattern is inverted when proteins with deleterious variants homozygous are analysed (Figure 3.3D). This observation is compatible with the history of the populations, with an older African population which has accumulated more variability but has filtered out deleterious variants homozygous whereas the rest of the populations underwent a relatively recent bottleneck which is reflected in a lower level of variability and a higher level of homozygosity (Lohmueller et al., 2008). This genetic fingerprint is still observable in the proteins which make up the interactome.

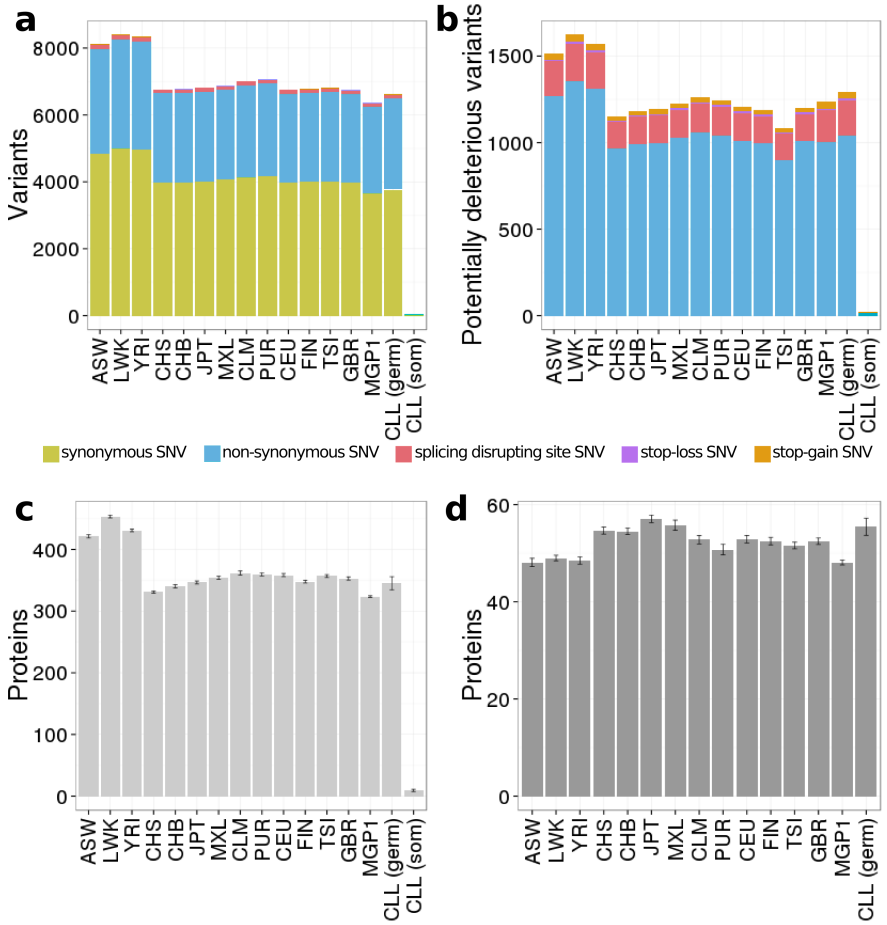

Figure 3.3: Summary of variants found in the proteins which configure the human interactome in all the populations analyzed. (A) Number of variants found; (B) Number of potentially deleterious variants; (C) Number of proteins carrying at least one deleterious variant in one of their alleles (mutation load); (D) Number of proteins carrying deleterious variants in both alleles (homozygous mutation load). Image adapted from (Garcia-Alonso et al., 2014).

In silico validation of the deleterious variants

Among potentially deleterious variants, we consider those with a clear deleterious effect, such as stop loss, stop gain and splicing disrupting conserved variants. In addition any conserved nonsynonymous variant with a SIFT score lower than 0.05 or a Polyphen score higher than 0.95 was considered deleterious, as recommended in the original publications (Kumar et al., 2009; Ramensky et al., 2002). Since the application of both scores sometimes results in contradictory predictions (Hicks et al., 2011) an in silico study was performed on a subset of 20 randomly chosen variants (8 predicted to be tolerant, 5 somatic predicted as damaging from CLL and 7 predicted as damaging from non-disease populations). Table 3.2 shows the relationship between the predictions derived from SIFT and Polyphen and the structural features calculated for the subset of selected variants. In general, a good agreement between predicted deleterious effect and unfavourable changes in the sequence and structure properties can be observed. Figure 3.4 depicts an example of this agreement.

Pred	Gene	Mut	SIFT	PPh	Cons	PDB	Polarity	Charge	SNAP	SDM	ΔΔG	Solvent accessibility
T	SPG7	T503A	0,37	0,001	609	2QZ4 (A)	1 → 1	0 → 0	N (53%)	N	0,39	63.8% (A) → 55.9% (PA)
T	NFRK	R171Q	0,36	0,224	0	3u21 (A)	2 → 2	+ → **0**	NN (58%)	N	0,23	101.5% (A) → 111.1% (A)
T	CA4	K267R	NA	0,314	0	1ZNC (A)	2 → 2	+ → +	N (92%)	S	1,2	26.8% (PA) → 33.8% (PA)
T	PDIA3	A7S	0,1	0	0	3F8U (A)	1 → 1	0 → 0	N (92%)	SD	-0,56	110.5% (A) → 103.6% (A)
T	RBBP	K215R	0,28	0	691	2XU7 (A)	2 → 2	+ → +	N (85%)	SS	0,92	41.4% (PA) → 57.3% (PA)
T	RET	L56M	0,46	0,003	278	2X2U (A)	0 → 0	0 → 0	N (89%)	SD	-0,69	3.6% (B) → 4% (B)
T	LXN	R48K	0,95	0	497	2BO9 (D)	2 → 2	+ → +	N (92%)	SD	-0,64	53.7% (PA) → 66.5% (PA)
T	BACE	P459L	0,62	0,397	0	3ZKM (A)	**1 → 0**	0 → 0	N (69%)	SS	0,81	33.8% (PA) → 48% (PA)
C	SF3B1	Y623C	0	0,999	731	2FHO (A)	0 → 0	0 → 0	NN (68%)	HS *	2,21	33.6% (PA) → 45.5% (PA)
C	SF3B1	T663I	0	0,998	671	2FHO (A)	**1 → 0**	0 → 0	NN (58%)	HS *	2,58	6.8% (B) → 8.6% (B)
C	SF3B1	K700E	0	0,999	685	2FHO (A)	2 → 2	+ → **-**	NN (58%)	SS	0,6	47% (PA) → 38.2% (PA)
C	PABP	E372G	0	0,86	745	4F02 (A)	**2 → 1**	**-** → **0**	NN (82%)	HD *	-2,8	72.2% (A) → 58% (PA)
C	PABP	R374C	0	0,998	745	4F02 (A)	**2 → 0**	+ → **0**	NN (78%)	N	-0,2	59.3% (A) → 52% (PA)
D	SDCB	E114G	0	0,234	670	1N99 (A)	**2 → 1**	**-** → **0**	N (53%)	D	-1,79	62.8% (A) → 72.4% (A)
D	RP2	R251G	0,05	0,998	584	3BH6 (B)	**2 → 1**	+ → **0**	NN (58%)	HD *	-3,25	29.4% (PA) → 66.9% (PA)
D	AGT	T207M	NA	0,991	340	2WXW (A)	**1 → 0**	0 → 0	NN (82%)	S	1,41	5.2% (B) → 2.6% (B)
D	ATF4	E322D	0,01	0,998	608	1CI6 (A)	2 → 2	- → -	N (78%)	D	-1,75	77.4% (PA) → 86.1% (A)
D	GRB7	T287M	0,03	0,895	445	4K81 (A)	**1 → 0**	0 → 0	NN (70%)	S	1,35	47.9% (PA) → 55% (PA)
D	MAVS	C79F	0	0,999	304	3J6C (A)	0 → 0	0 → 0	NN (82%)	HD *	-2,03	10.8% (B) → 8.8% (B)
D	BAIAI	T124I	0,05	0,042	525	2KXC (A)	**1 → 0**	0 → 0	N (53%)	S	1,94	44.9% (PA) → 59.5% (PA)

Table 3.2: Computational validation of variant deleteriousness prediction. Unfavourable structural properties of the mutations are highlighted in bold and underlined, and partially unfavourable properties are highlighted in bold. Columns contain: 1) prediction (T, tolerated; C, cancer and deleterious; D, deleterious), 2) gene name, 3) mutation, 4) SIFT score, 5) Polyphen score, 6) PhasCons conservation score, 7) PDB identifier of the structure template and the chain, 8-9) change in the protein polarity and charge based on changes in the charge of the changed residue (see 'Materials and Methods'), 10) SNAP prediction (N: neutral and NN: non-neutral, with the accuracy in brackets), 11) SMD prediction (N: neutral, S: stabilizing, SS: slightly stabilizing, D: destabilizing, SD: slightly destabilizing and HD: highly stabilizing and * causing protein malfunction), 12) SMD pseudo ΔΔG, and 13) the change in percentage of solvent accessibility (A: accessible, PA: partly accessible and B: buried). Table adapted from (Garcia-Alonso et al., 2014).

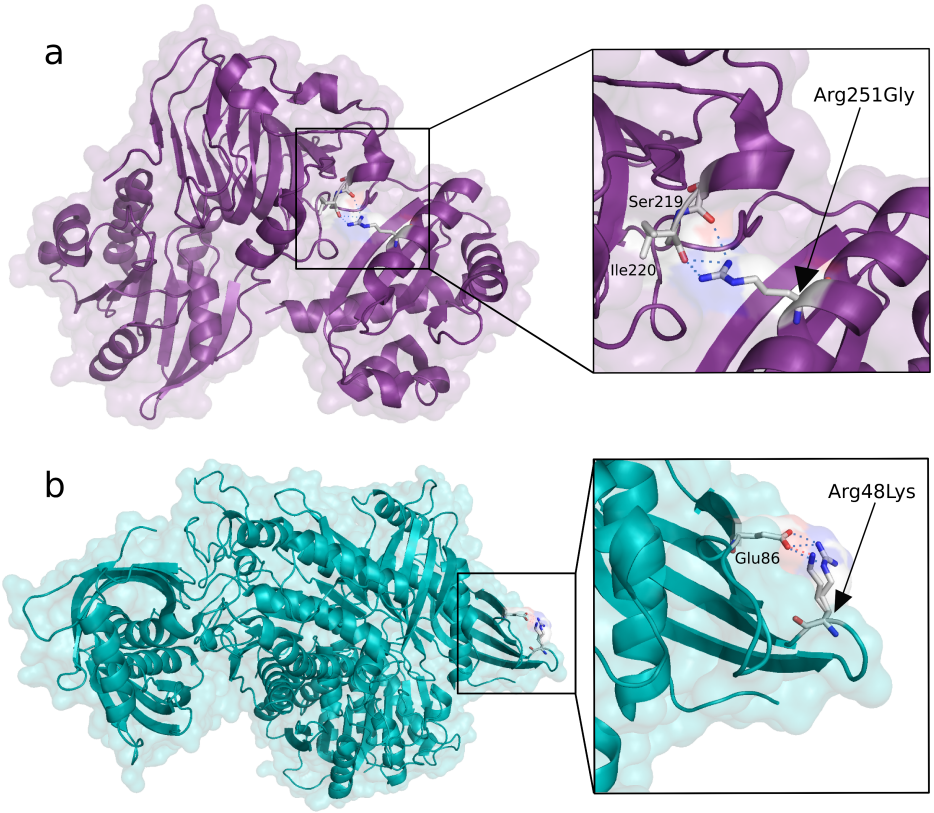

Figure 3.4: Molecular model of the human RP2 (a) and LXN (b) proteins and detailed view of the altered amino acids (Arg251Gly and Arg48Lys, respectively). A) The amino acid change Arg251Gly in the RP2 protein was predicted as deleterious according to SIFT and PolyPhen thresholds. The original residue (Arg251) of α-helix forms a hydrogen bond with the Ser219 and Ile220, however the new residue is highly destabilizing. Specifically, the new residue (Gly) is uncharged, more hydrophobic and smaller than the original, which causes that the positive charge will be lost and the amino acid will not be in the correct position, hampering the establishment of the original hydrogen bond. B) The amino acid change Arg48Lys in the LXN protein was classified as tolerant according to the criteria used. The new amino acid, whose substitution was predicted as tolerant by SIFT and Polyphen, does not cause a significant change in protein stability, maintaining the same charge and polarity as the wild-type residue. Image taken from (Garcia-Alonso et al., 2014).

3.3.2 Topological role of proteins carrying deleterious variants

We analysed the occurrence of deleterious mutations in proteins with different network properties in the interactome. Figure 3.5 shows the number of interactions corresponding to the proteins affected by a deleterious variant either in both alleles (homozygous) or in only one allele (heterozygous) or not affected by any deleterious variant, in at least one individual. It also shows the number of interactions observed in proteins with deleterious somatic mutations in CLL, proteins corresponding to monogenic diseases and the subset of somatic mutations in CLL corresponding to cancer driver genes (Vogelstein et al., 2013). The number of interactions in proteins with both alleles affected by a deleterious variant in healthy individuals was significantly lower than the number of interactions observed either in proteins with only one allele affected (FDR-adjusted Mann-Whitney U test $P = 5.44 \text{ x } 10^{-4}$) or in unaffected proteins ($P = 5.22 \text{ x } 10^{-5}$). Proteins carrying only one allele affected by a deleterious variant showed a slightly lower number of interactions than unaffected proteins, although the difference is not significant in this case, probably because they have no pathogenic effect in either case.

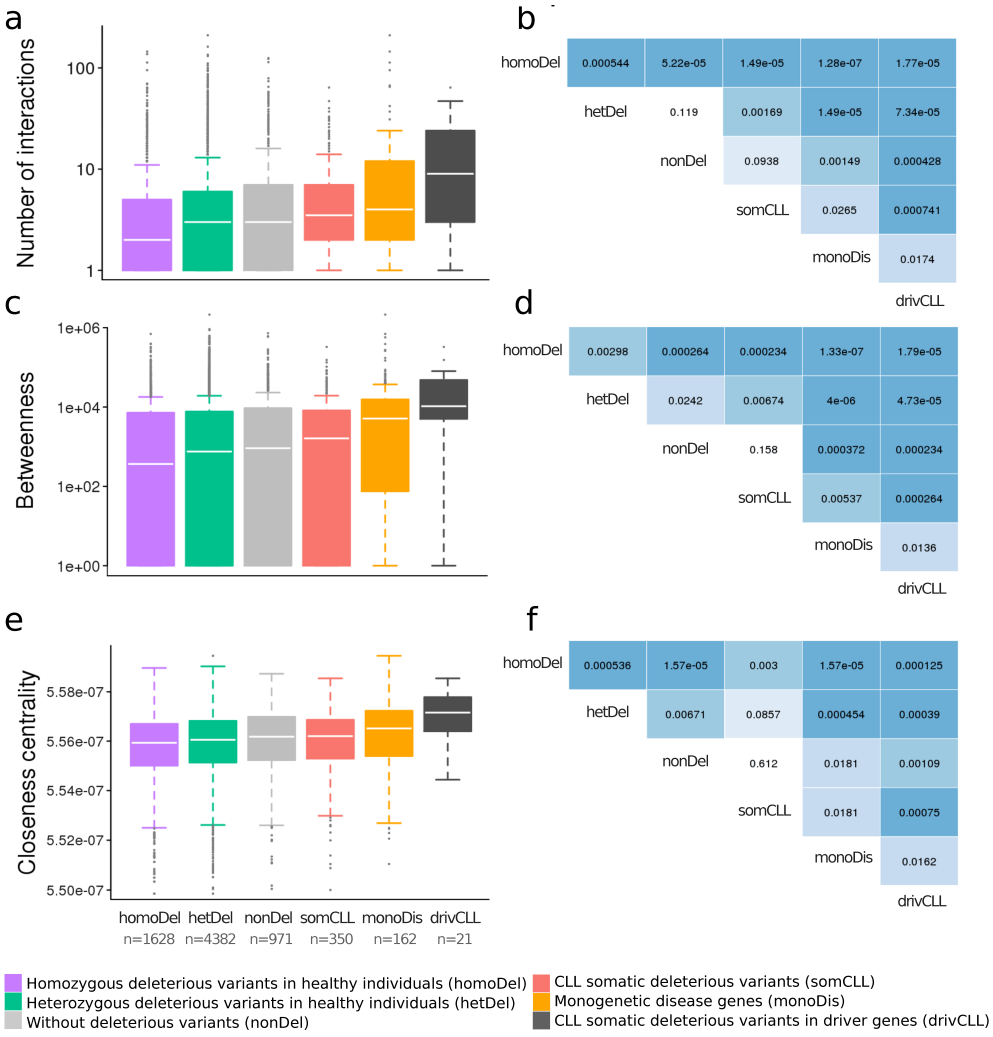

Figure 3.5: Connection degree, betweenness and closeness centrality of proteins affected by deleterious variants. (a): From left to right: Number of interactions in proteins affected by deleterious variants in both alleles (homozygous), in only one allele (heterozygous), not affected by any deleterious variant, proteins affected (homozygous or heterozygous) in a pathological condition (somatic variants in CLL), proteins affected by monogenic diseases and the subset of somatic variants in CLL which occur in cancer driver proteins (Vogelstein et al., 2013). (b): Significance of the comparisons tested by the rank sum (Mann-Whitney U test) with FDR multiple testing adjustments. (c): Betweenness in the same groups of proteins as in a. (d): Significance of the comparisons tested as in b. (e): Closeness centrality in the same groups of proteins as in a. (f): Significance of the comparisons tested as in b. Image adapted from (Garcia-Alonso et al., 2014).

In a scenario of mutational disease represented by all the CLL proteins carrying somatic mutations (driver and passenger variants), the number of interactions in affected proteins was significantly higher than in healthy homozygous ($P = 1.49$ x 10^{-5}) and the healthy heterozygote ($P = 0.00169$) scenarios, as expected. The proteins affected by monogenic diseases displayed a significantly higher number of connections than the CLL proteins carrying somatic mutations ($P = 0.0265$) (and obviously more than the deleterious homozygous and heterozygous and unaffected proteins in healthy individuals, see Figure 3.5b). However, if only cancer driver proteins carrying somatic deleterious mutations in CLL are considered, the number of connections was significantly higher than any other subset of proteins analysed, including monogenic disease proteins (see Figure 3.5b). The analysis studying the relationship between the same sets of genes and other properties such as betweenness (Figure 3.5c and d) and closeness centrality (Figure 3.5e and f) was repeated, obtaining a similar trend. The results demonstrate a clear relationship between the degree of pathogenicity of the scenario and the connectivity of the proteins affected.

Figure 3.6 depicts how the number of connections, the closeness centrality and the betweenness present a weak, but significant negative correlation (Spearman's rank correlation coefficient $\rho = -0.0661$, $P = 1.34$ x 10^{-7}, $\rho = -0.0536$, $P = 1.93$ x 10^{-5} and $\rho = -0.0534$, $P = 2.05$ x 10^{-5}, respectively) with the frequency of occurrence of deleterious variants in the population (both homozygous and heterozygous). This trend, although negative as well, is not significant in the case of homozygous, probably due to the lower sample size. On the contrary, in the pathological condition represented by CLL, the network properties number of connections ($\rho = 0.152$, $P = 0.0116$), betweenness ($\rho = 0,118$, $P = 0.051$) and closeness centrality ($\rho = 0.128$, $P = 0.0335$) are positively correlated with the recurrence of the mutation across patients.

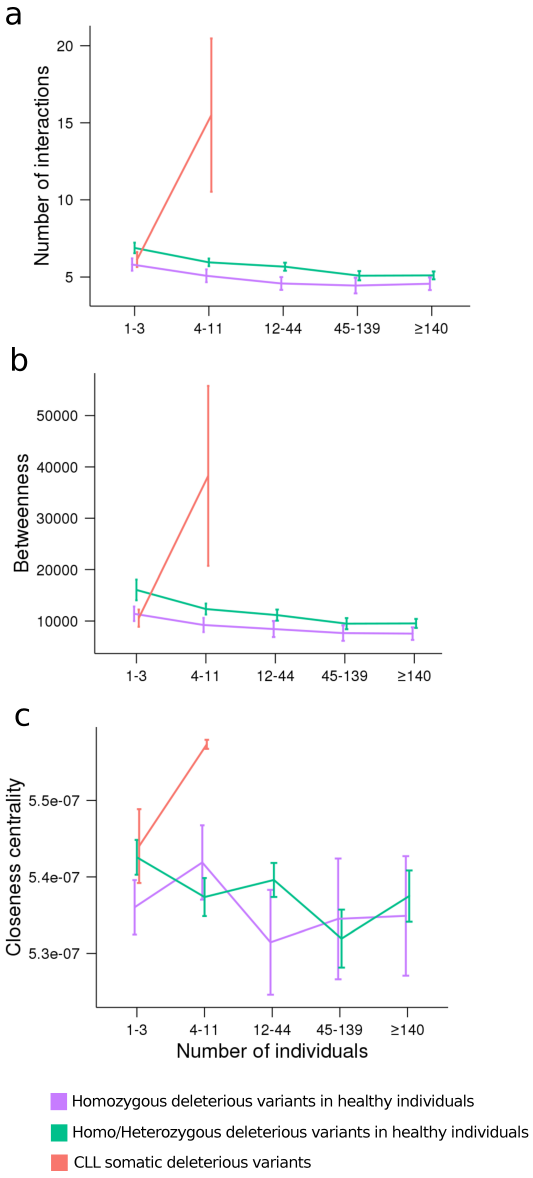

Figure 3.6: Mean connectivity (a), betweenness (b) and closeness centrality (c) for proteins undergoing deleterious variants. The blue line represent deleterious variants in both alleles (homozygous), and the green line deleterious variants in at least one allele (homozygous+heterozygous), grouped according to the number of individuals in normal populations (1KGP genomes and Spanish populations) in which they were observed. The red line represents CLL somatic heterozygous deleterious variants observed in growing number of individuals (within the sample of patients). The plots include 1SD bars. Image adapted from (Garcia-Alonso et al., 2014).

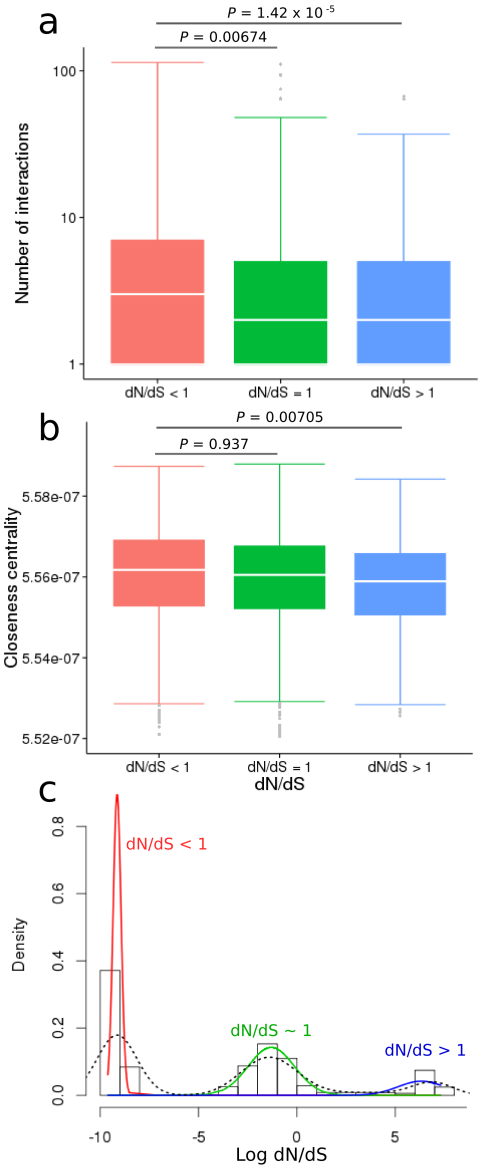

Figure 3.7: Boxplots displaying the trends of proteins under positive selection (with a ratio of nonsynonymous (dN) to synonymous (dS) mutations greater than 1), neutral selection (with dN/dS approximately equal to 1) and negative selection (dN/dS¡1), obtained as in (Serra et al., 2011), with respect to different network properties: A) Number of interactions and B) Closeness centrality. The significance level of the mean differences is given as the Mann Whitney U rank sum P value. C) Density plots for the Log dN/dS values using the normalmixEM procedure, from the R "mixtools" package, with default parameters. Image taken from (Garcia-Alonso et al., 2014).

From the point of view of disease, a high degree of connectivity between proteins mutated in the same disease state has been reported (Wachi et al., 2005; Jonsson and Bates, 2006; Goh et al., 2007; Feldman et al., 2008). First studies showed that the protein products of genes driving cancer tend to have higher degree than non-cancer proteins (Wachi et al., 2005; Jonsson and Bates, 2006), suggesting that altered genes in cancer are key for the proliferation of the tumor cells and, therefore, display the same topological properties that essential genes. More global studies (Goh et al., 2007; Feldman et al., 2008) compared essential genes, to cancer and monogenic diseases and found that topological values for centrality and connectivity of the later tend to lie in between the essential or tumorgenic genes and those not included in any previous class (that is non-disease genes). The general conclusion is that genes driving disease are not randomly located in the network.

Previous evolutionary studies documented a preferential occurrence of adaptive events at the periphery of the human protein interaction network (Fraser et al., 2002; Kim et al., 2007). It was confirmed that the distribution of selective pressures, measured as the ratio of non-synonymous to synonymous variants, across the network properties used here (number of interactions, betweenness and closeness centrality) was consistent with what was previously observed: proteins under positive selection tend to be placed in the periphery of the network while proteins under negative selection tend to be in the internal regions (see Figure 3.7).

3.3.3 Robustness of the interactome structure to homozygous deleterious variants of healthy individuals

The effect that the specific combination of deleterious variants carried by any healthy individual has on the interactome was studied. Since variants which produce a loss of function were considered, the recessive (and most plausible) scenario was tested. This was achieved by removing proteins from the interactome when they were affected by deleterious

variants in both alleles (homozygous for the alternative allele). Then, the impact that this subtraction had over the interactome structure was calculated (see Material and Methods, subsection "Selection of deleterious variants" for details). This impact is inferred by measuring the changes in several global network properties such as the number of connections, the average length of shortest paths and the number of components. These parameters account for the interconnectedness and integrity of the interactome (Albert et al., 2000). The values obtained for these parameters in the 1KGP and MGP populations correspond to interactomes of healthy individuals.

In order to understand the basis of the robustness of the interactome against the deleterious variants carried by normal individuals the normal interactomes were compared with simulated interactomes in which the same number of damaged proteins was randomly removed (see Methods). The comparison between the real and simulated interactomes resulted in significant differences between them in the network parameters measured. Real normal populations (1KGP, Spanish population and CLL germinal line) always have more connections than simulated individuals (compare real populations bar to simulated populations with uniform probability bar in Figure 3.8a). Moreover, these connections preserved in real individuals are organized in a way which maintains a significantly lower average length of shortest paths (same comparison in Figure 3.8b), a distinctive feature of biological networks, and avoids disconnection from the giant component (same comparison in Figure 3.8c). In other words, real individuals have significantly more structured and less affected interactomes than simulated individuals for the same number of removed (damaged) proteins. The results were highly significant for the 1KGP population and still significant but with higher p-values for the MPG and CLL populations, due to the smaller sample sizes (see Figure 3.8).

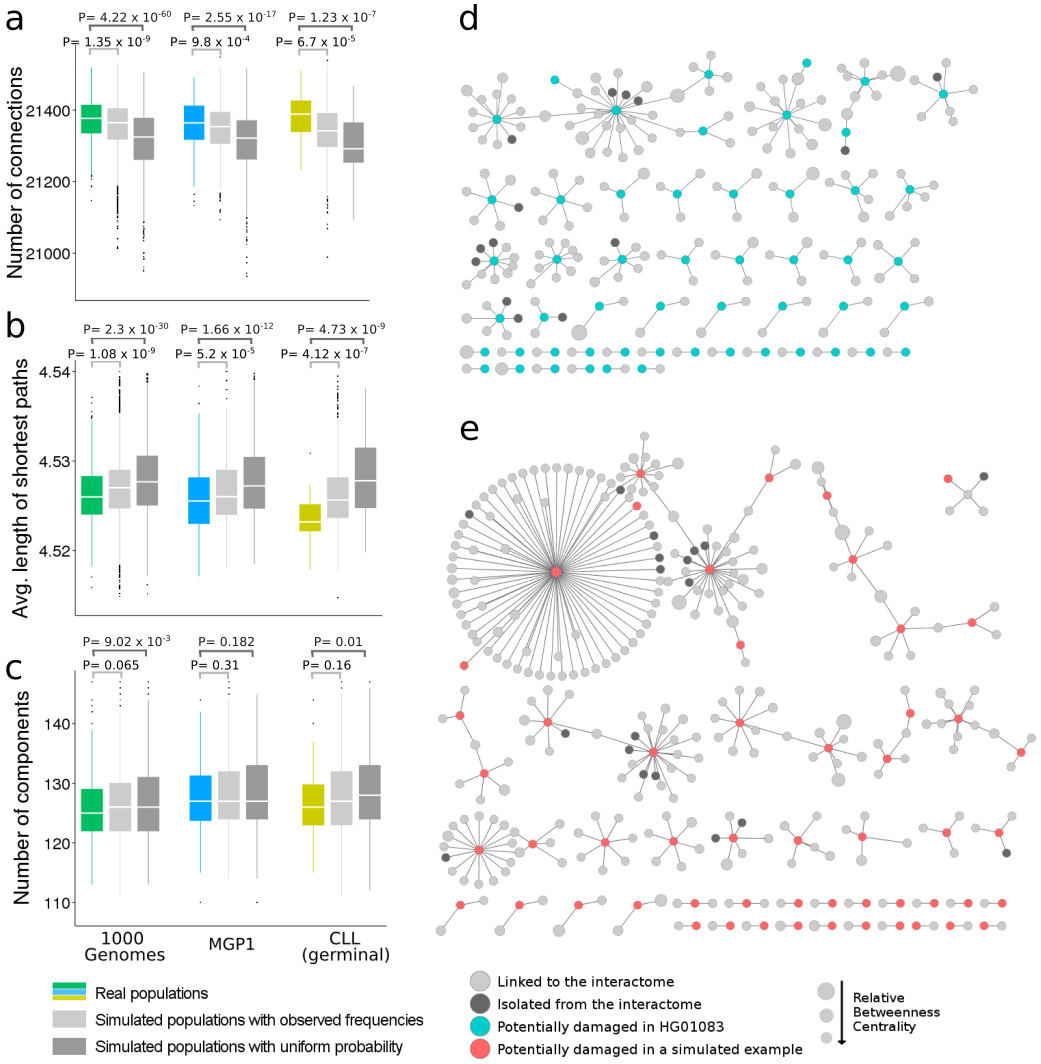

Figure 3.8: Impact of potentially deleterious variants on the interactome of real and simulated individuals. Comparison of the interactome damage between real and random individuals after removing proteins containing homozygous deleterious variants. The comparison was performed using 1KGP populations (green box), Spanish population MPG1 (blue box) and the germinal variants of the CLL patients (yellow box) and contrasting their distributions with the corresponding simulated distribution (grey boxes). Two different scenarios are simulated: Simulated populations with uniform probability and simulated populations with observed frequencies. The effects on the global network topology were defined by: (a) the number of connections in the remaining interactome, (b) the average length of the shortest paths and (c) the total number of isolated components. Visual illustration of the network components lost after removing nodes corresponding to damaged proteins in (d) a real individual from 1KGP (HG01083 of the PUR population) and (e) a simulated individual with the same number of damaged proteins. Image taken from (Garcia-Alonso et al., 2014).

This simulation demonstrates that healthy individuals carry deleterious variants in a specific set of proteins whose deletion minimizes the impact on the interactome structure. However, it is not clear whether this low impact is due to the actual individual proteins observed in the population or whether it occurs because proteins with deleterious variants are only tolerated in specific combinations which minimize the damage to the interactome structure. To address this question another simulation was conducted in which deleterious mutations were assigned to proteins according to their observed mutation frequencies in healthy individuals (1KGP and MGP populations). Unlike the previous simulation, the simulated individuals only carried deleterious variants in proteins which are affected in normal individuals, but in random combinations which do not necessarily exist in real healthy individuals.

Although not as remarkable as in the previous simulation, the difference between real and simulated values was also significant. Again, real normal populations had significantly more connections than simulated individuals (compare real populations bar to simulated populations with observed frequencies bar in Figure 3.8a), connections which result in a network with shorter shortest paths between components (see how average lengths of shortest pathways change across real and simulated populations in Figure 3.8b) and have a tendency to display fewer isolated components (same comparison in Figure 3.8c). The p-values were higher, and in some cases non-significant (number of components for MGP and CLL germinal populations, probably due to their small sizes), as the effect of removing the acceptable combination of damaged proteins is not as strong as the effect of removing random proteins. The results obtained as a whole suggest that only a limited number of variants in specific combinations are tolerated by the interactome for compatibility with a healthy condition.

An example visually illustrates the type of connections lost in the simulation with random occurrences of deleterious variants when compared to the type of connections lost in the case of observed deleterious variations. Figure 3.8d depicts an example of sub-networks disconnected

from the interactome of a normal individual from the 1KGP, because both alleles of the gene coding the connecting protein had deleterious variants. Figure 3.8e shows an example taken from a simulated individual. It is clear that while interactomes of real individuals are slightly trimmed off by the deleterious variants they carry, the interactomes of simulated individuals undergo more serious damage having larger disconnected portions.

3.3.4 Robustness validation in tissue-specific interactomes

As commented before, the interactome used here represents the collection of almost all high-quality PPIs that are known in human. However it is quite unlikely that all these proteins are present at the same time, and therefore the interaction can occur. Concerned by this fact, we built more realistic interactomes. Specifically, we used tissue-specific gene expression arrays obtained from normal samples to compute the likelihood of a transcript to be expressed in that tissue. Using the gene expression as a proxy of the protein to be expressed, we built tissue-specific interactomes by removing all the proteins, and so the corresponding interactions, that are likely not expressed. With this new versions of the interactomes, we repeated the analysis described in the previous section. Figures 3.9 and 3.10 represent the p-values from the comparison between real population and simulated populations with uniform probability and with observed frequencies, respectively. In both figures, rows represent the tissue-specific interactomes whereas the columns the network parameters evaluated: average shortest paths, number of proteins that remain in the giant component, number of isolated components and number of edges. As a control for the simulation, the total number of nodes remaining in the networks is also shown in the figures, showing no difference between real and simulated populations. From the figures we can observe that the results obtained in the previous section hold with the tissue-specific interactomes, although the strength of the differences is smaller for the comparison between real populations to simulated populations with prior probabilities.

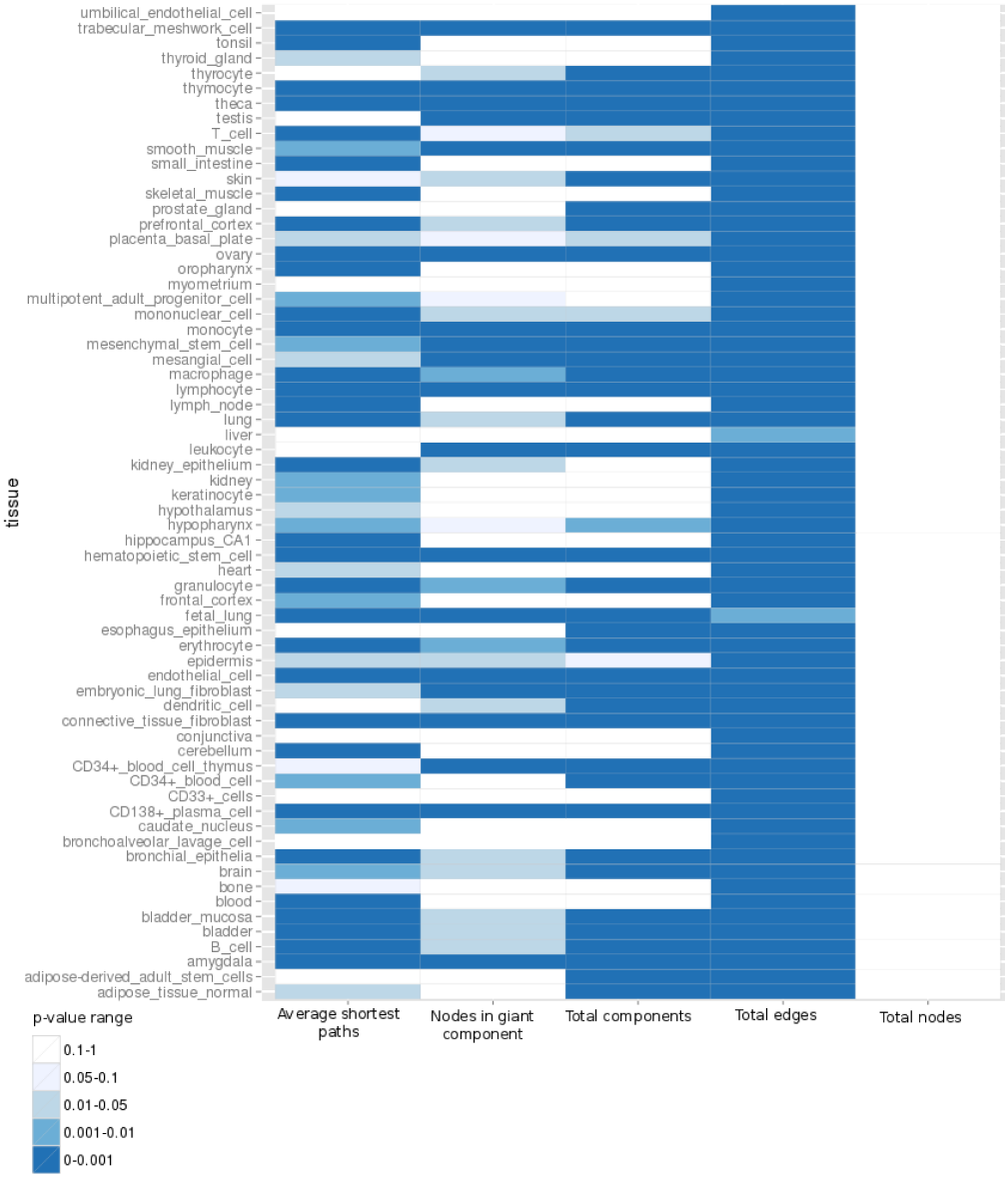

Figure 3.9: Heatmap of the impact of deleterious variants in tissues compared to simulated populations with uniform probability. Rows represent the interactomes from different tissues whereas that columns the network parameters evaluated. The color code represents the strength of the p-value from the comparison.

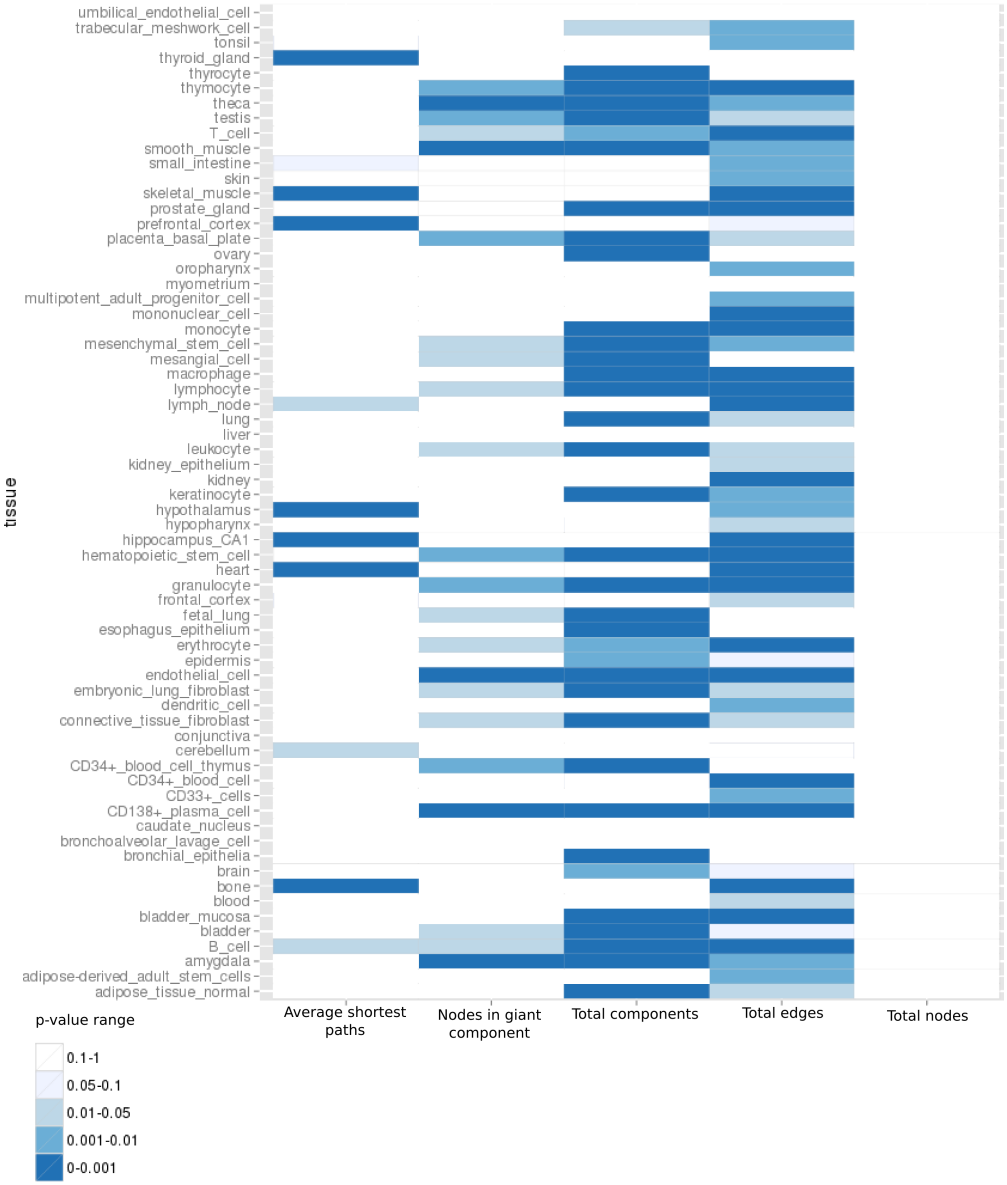

Figure 3.10: Heatmap of the impact of deleterious variants in tissues compared to simulated populations with observed frequencies. Rows represent the interactomes from different tissues whereas that columns the network parameters evaluated. The color code represents the strength of the p-value from the comparison.

3.3.5 Distribution of deleterious variants among the interactome modules

In order to understand the reasons why such specific combinations of deleterious variants cause both minimal disruption to the interactome and have no associated pathological effects, their location within the network of protein interactions was studied. Firstly a summarized representation of the interactome was derived by detecting neighbourhoods of densely connected sub-graphs which define communities, or modules of highly interacting proteins (Pons and Latapy, 2005; Rosvall and Bergstrom, 2008). These modules can be considered functional entities which enable the biological interpretation of the results. Then, the distribution of genes carrying alleles affected by deleterious variants across the modules was studied in individuals from the Spanish population and the 1KGP populations. The pattern of distribution of affected modules across populations is defined by conventional hierarchical clustering using the Euclidean distances between them. The clustering obtained was quite coherent with the geographical origins and history of the analysed populations (Figure 3.11). The Spanish population is located close to the rest of the European populations as well as to Latin Americans populations, with whom they share common ancestors. The deleterious germinal variants found in CLL patients are located close to the Spanish population, probably because it is mainly composed of Spanish CLL patients. On the contrary, the distribution of mutations of somatic deleterious mutations of CLL (Figure 3.11) follows a pattern inverse to the rest of the normal populations. This anomalous distribution clusters this sample outside of any human population.

The same clustering methodology was applied to group the modules. The analysis resulted in the definition of five main clusters. The two clusters at the bottom are composed of highly affected modules, enriched in proteins with deleterious variants. The central cluster is composed of protected modules, with a lower proportion of proteins with deleterious variants than expected by chance. And the two upper clusters correspond

to an intermediate situation.

The distribution of cell functionalities across the modules is depicted in Figure 3.11. The cluster containing protected (and often central) modules is enriched in GO terms related to essential cellular functions, such as gene expression, translation, protein targeting, and chromatin organization. Conversely, the most external clusters contain cell functionalities acquired later in evolution, mainly related to signalling immune response and cell communication (central and part of the upper clusters in Figure 3.11).

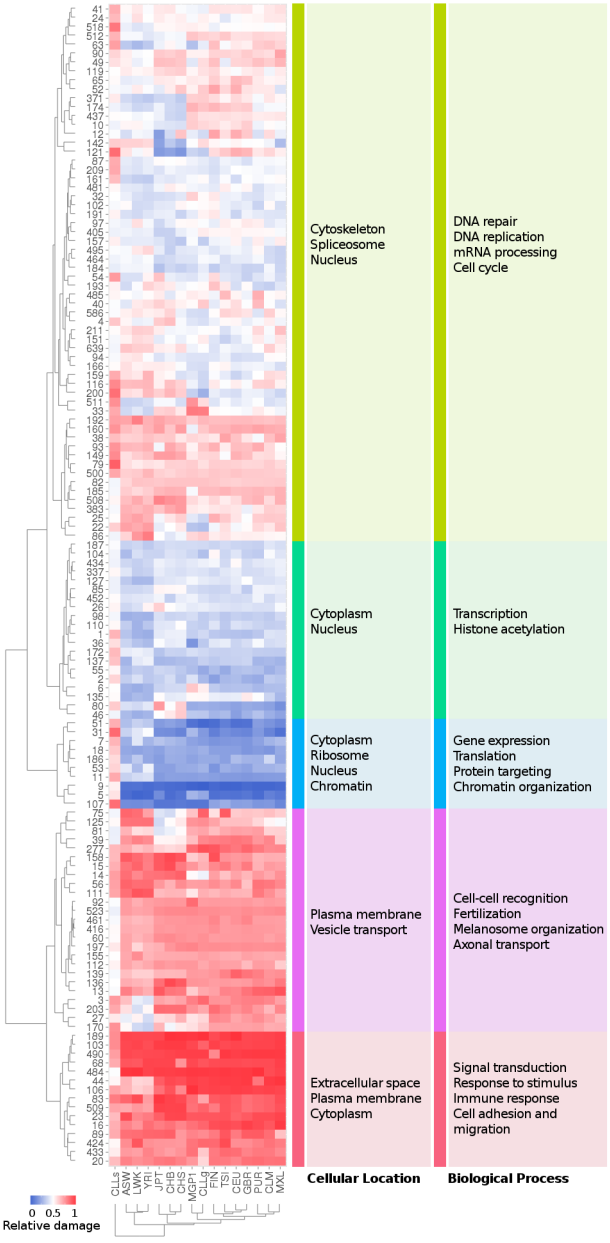

Figure 3.11: Heatmap of interactome modules defined by the Walktrap clustering algorithm (rows) and the 1KGP, the Spanish MGP and germinal and somatic CLL patients populations (columns). The color code represents the relative damage of the module, which accounts for the deviation in the proportion of affected proteins in the module from the random expectation distribution. The color code represents the relative damage value, which ranges between 0 (blue: no proteins affected at all in this quartile) to 1 (red: the maximum possible number of proteins affected in this quartile). On the right of the figure the main GO terms significantly enriched in any of the five main row clusters defined by conventional hierarchical clustering using the Euclidean distances are displayed. Image taken from (Garcia-Alonso et al., 2014).

3.3.6 Comparison of germinal to somatic cancer-specific mutations

Finally, we focused on comparing the distribution of deleterious variants in genes across the different communities in both the germinal line (which would represent a normal genome) and somatic mutations in the cancer samples (corresponding to a pathological condition) of CLL patients. The germinal line of CLL patients presents a pattern of distribution of variants indistinguishable from normal individuals (Figure 3.11). Modules located at the periphery of the interactome are considerably enriched in affected proteins in healthy individuals, while the opposite tendency is observed in internal modules (as portrayed in Figure 3.12A). The extent of this trend is confirmed by the significant negative correlation (Spearman correlation test $P \leq 0.001$) of a measure which accounts for the centrality of a module in the interactome (closeness centrality) with the normalized proportion of affected proteins with respect to the random expectation (relative damage of the module) (Figure 3.12B). However, the pattern of somatic mutations in CLL is completely different to any other population and is actually inverted to the pattern observed in normal individuals. Figure 3.12C documents the inverse trend of distribution of mutations when represented on the interactome of modules. As opposed to the case of normal populations, deleterious somatic mutations in CLL are over-represented in internal modules of the interactome (Figure 3.12D). The significance of this trend is confirmed by the significant positive correlation (Spearman correlation, $P \leq 0.01$) existent between a measure of the module centrality within the interactome (closeness centrality) and the proportion of affected proteins with respect to the random expectation (relative damage of the module) (Figure 3.12D). The opposite trends observed both in normal populations and in somatic mutations of CLL patients have been confirmed using different interactomes and different algorithms for defining modules within them (See Table 3.3).

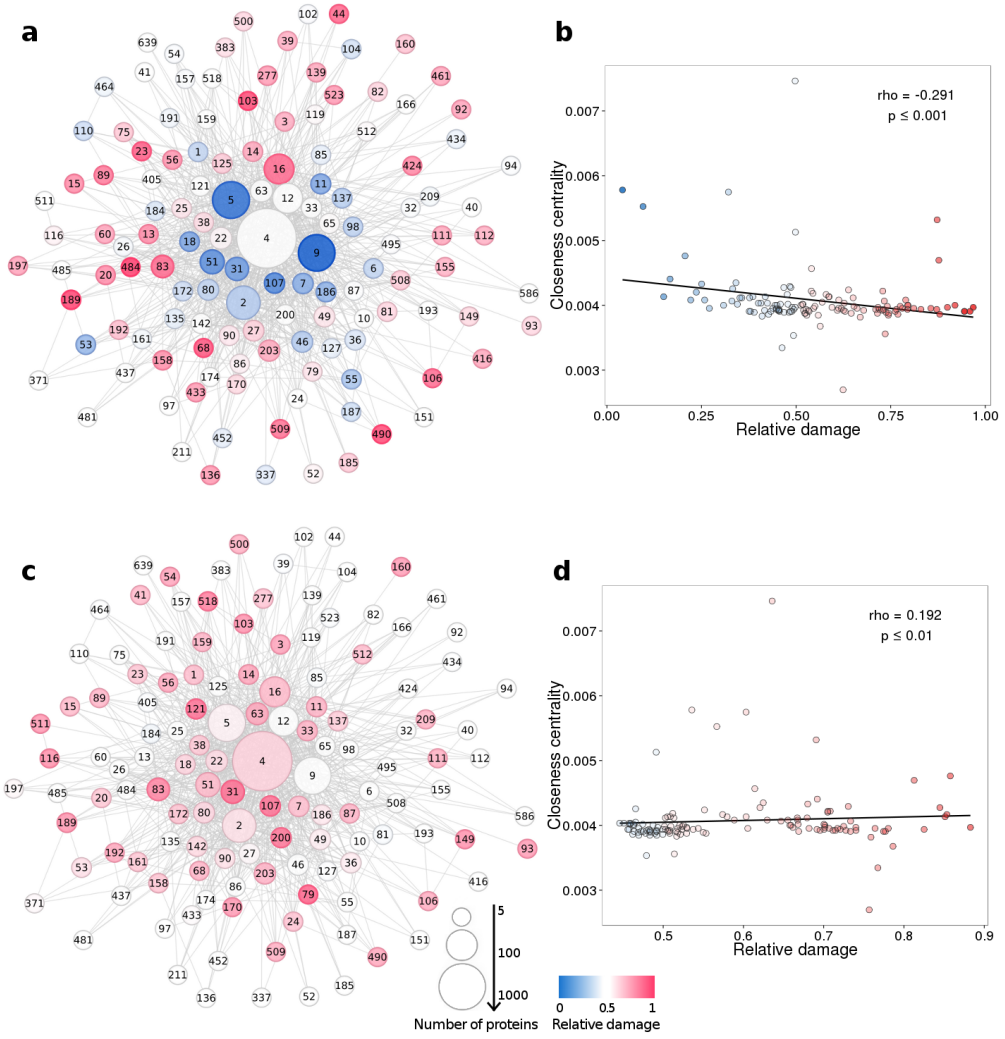

Figure 3.12: Relationship between proteins carrying deleterious variants and the module centrality. a) Distribution of proteins with deleterious variants in human populations and the germinal line of the CLL patients across the interactome of modules (defined by the Walktrap clustering algorithm). Two modules are connected if there is at least one interaction between one of their proteins. The numbers in the nodes are module identifiers. The size of the node is proportional to the number of proteins which it contains. (b) Module closeness centrality as a function of the relative damage of the module. (c) Distribution of proteins with somatic deleterious variants in CLL exomes across the interactome of modules. (d) Module closeness centrality as a function of the relative damage of the module for the somatic CLL exomes. The color code represents the relative damage value, which ranges between 0 (blue: no proteins affected at all in this quartile) to 1 (red: the maximum possible number of proteins affected in this quartile). Image taken from (Garcia-Alonso et al., 2014).

In order to check whether this observation was reflecting the centrality of individual proteins or whether it was accounting for the centrality of the modules, the data was reanalysed using as a grouping variable only the centrality of each protein within the network. Here, the interactome was divided into 4 regions according to the closeness centrality distribution quartiles and then, the distribution of damaged proteins among the four regions was calculated for each individual. The result obtained was the same: the peripheral regions of the interactome accumulated more proteins affected by deleterious mutations than expected by chance, whereas the internal region displayed a remarkable reduction ($P = 3.96$ x 10^{-6} Mann-Whitney U test) in affected proteins (See Figure 3.13). Thus, the burden of deleterious variability observed in a protein seems to be related to the centrality of the protein.

Sample	Interactome	Detection algorithm	Rho	P-value
1KGP, MGP and germinal CLL	Curated	Walktrap	-0.292	≤ 0.001
1KGP, MGP and germinal CLL	Curated	Infomap	-0.159	≤ 0.001
1KGP, MGP and germinal CLL	Non-Curated	Walktrap	-0.13	0.28
1KGP, MGP and germinal CLL	Non-Curated	Infomap	-0.11	≤ 0.01
1KGP, MGP and germinal CLL	STRING	Walktrap	-0.186	≤ 0.01
1KGP, MGP and germinal CLL	STRING	Infomap	-0.205	≤ 0.01
Somatic variants CLL	Curated	Walktrap	0.192	≤ 0.01
Somatic variants CLL	Curated	Infomap	0.176	≤ 0.001
Somatic variants CLL	Non-Curated	Walktrap	0.321	≤ 0.01
Somatic variants CLL	Non-Curated	Infomap	0.211	≤ 0.01
Somatic variants CLL	STRING	Walktrap	0.28	≤ 0.001
Somatic variants CLL	STRING	Infomap	0.322	≤ 0.001

Table 3.3: Validation of the relationship between the module centrality and damage. Different network module detection algorithms (Infomap and Walktrap) and three protein interactomes were used. Table adapted from (Garcia-Alonso et al., 2014).

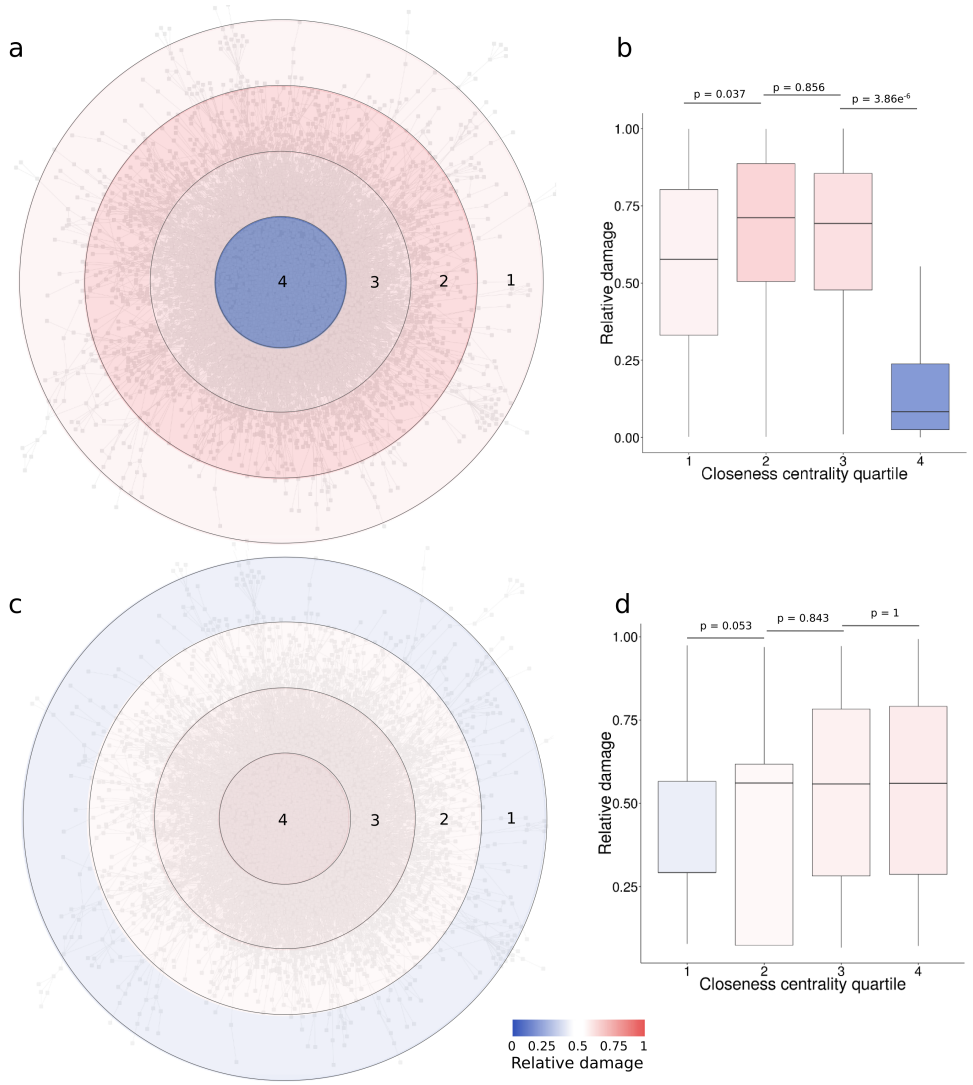

Figure 3.13: Relationship between the distribution of the deleterious variants and the interactome centrality in normal populations and CLL patients. A) Distribution of proteins with deleterious variants in healthy human populations and the germinal-line CLL exomes across the interactome. B) The corresponding boxplots representing the distribution of relative damage in each individual in any of the quartiles. C) Distribution of proteins with deleterious variants in the somatic CLL exomes, representative of a pathological condition, across the interactome. D) Boxplots representing the distribution of relative damage of each individual in any of the quartiles for the somatic CLL exomes. The interactome was divided into four sectors of proteins according its closeness centrality (from 1, peripheral to 4, central). The color code represents the relative damage value, as in Figure 3.12. Image taken from (Garcia-Alonso et al., 2014).

Summarizing, our results strongly suggest that the pathogenic role of deleterious mutations is highly correlated with the impact on the interactome integrity caused by the combined apparently deleterious variants of the affected proteins, which is also related to the location of such proteins within the interactome. That is, affected proteins in healthy individuals are concentrated in peripheral modules, avoiding internal modules. However, the most important factor which sheds light on the mechanisms by which the interactome can bear a large number of proteins with deleterious mutations is related to the way in which affected proteins are specifically combined in healthy individuals. Affected proteins in healthy individuals tend to occur in combinations which preserve short path lengths. When the same proteins occur in random combinations, the length of the shortest paths significantly increases. Thus, the structural constraints imposed by the preservation of shortest paths may contribute to the relative higher tolerance for deleterious mutations observed in the periphery of the interactome. In the periphery, combinations of affected proteins that preserve shortest path lengths are easier to find than in internal regions of the interactome. This property could only be observed by means of an individualized analysis of the healthy subjects.

CHAPTER 4

Deciphering mutational oncogenic signatures on structurally resolved protein interacting interfaces

Part of the work presented in this chapter was published as preprint in:

Garcia-Alonso, L. and Dopazo, J. (2015). Mutational oncogenic signatures on structurally resolved protein interacting interfaces. *bioRxiv*, page 016204

4.1 Overview and objectives

Aberrant protein activities caused by mutations altering PPIs have been associated to cancer initiation and progression. A well known example is the Y42C mutation in BRCA2 that inhibits its interaction with the protein A, essential for DNA repair, replication and recombination, leading to the accumulation of DNA damage (Wong et al., 2003). From a more global point of view, studies applying a network centric perspective in cancer (Goh et al., 2007; Feldman et al., 2008), in agreement with our previous observation that somatic mutations from CLL patients tend to fall in central modules of the interactome, suggest that the impairment of protein interactions at the core of the interactome can be a common mechanism in tumor evolution. In fact, the products of cancer driver genes participate in a greater number of interactions (network hubs) (Jonsson and Bates, 2006) and are more likely to influence a larger number of biological processes (pleiotropy) (Yu et al., 2008). These observations, although pointing at global properties of genes involved in cancer, can only be seen as descriptive rather than propose testable hypotheses. The reason behind is the lack of molecular details in the way the interactome is modeled (ie. an undirected graph), where proteins represent graph-theoretical nodes ignoring its structural details. However, not all the mutations from the same protein have the same consequence but its impact depends on the stereochemical nature of the change and, ultimately, on its location within the carrier protein, which evidences the importance of integrating the structural properties in approaches aimed to decipher the effect of cancer mutations in protein coding genes.

We hypothesize that somatic mutations are more likely to confer a functional change and, therefore, to be selected if they alter a molecular interaction, specially if they occur in proteins that govern essential biological processes. This thesis continues with investigating the role of the protein interacting interfaces in the tumorgenic process through a systematic analyses of somatic mutations affecting structurally resolved protein-protein interaction interfaces. The specific objectives in

this chapter are:

1. To collect somatic point mutations from cancer patients form the TCGA and ICGC repositories and annotate the consequences of the change at protein level.

2. To reconstruct the three-dimensional (3D) structure of the PPIs form the human interactome.

3. To study the distribution of the cancer missense mutations (amino acid–changing) among the structural region types of the interactome proteins.

4. To identify protein interacting interfaces accumulating unexpected amounts of somatic variants.

5. To characterize the clinical implications of the cancer mutations in the enriched interacting interfaces.

6. To describe the topological properties of the enriched interacting interfaces in the context of the protein interactome.

7. To build a molecular map enclosing the mutational oncogenic signatures found on the structurally resolved protein interactome.

4.2 Materials and methods

4.2.1 TCGA and ICGC cancer datasets

Two main data sources were used to retrieve somatic mutations from different cancers: ICGC (release 15.1) and TCGA (curated dataset available at Synapse ref. *syn1729383*, (Kandoth et al., 2013)). The reason why we decided to use *syn1729383* instead of processed data downloded directly from TCGA data portal is that Kandoth and colleges made an effort to standardize mutation data from the different cancer types. Briefly, they reprocessed the data 1) to eliminate known, recurrent false positives and germline single nucleotide polymorphisms (SNP) present in the dbSNP database; 2) to eliminate low quality calls; and 3) to transfer all variant coordinates to GRCh37 and reannotate them using the Gencode human transcript annotation imported from Ensembl release 69 (for more details of the standardization process, see Synapse documentation `https://www.synapse.org/#!Synapse:syn1729383`).

When merging data from ICGC and TCGA, several consideration should be taken. The following list enumerates them and describes how we solved each point:

1. **Sample type.** Not all the samples are primary tumors but there are also cell lines, tissue from metastases and peripheral blood. Here, only samples from primary tumors are selected.

2. **TCGA sample redundancy.** TCGA uses different barcodes for the same sample (see barcode info), if they were sequenced several times or by several pipelines. To avoid treating their variants as observed recurrence across patients, we relabelled the samples by removing the vial letter (reducing the sample IDs to the form "TCGA-XX-XXXX-XX"), so that we only have one sample ID per sample, and merged its calls.

3. **Donor redundancy.** ICGC has collected some cancers from TCGA, so they will be redundant when we merge both datasets. To avoid

duplicity, we merged the data from the same donor if the sample identifier matches.

4. **Donor with different samples.** To avoid treating their variants as observed recurrence across patients, donors with more than one sample are removed.

5. **Different annotation pipelines.** Mutations have been processed and annotated using different pipelines and data sources, which may influence the consequences assigned to each one. In order to make the data comparable, we retrieved only genome coordinates and the reference and alternative alleles and reannotated them using VARIANT software (see next section).

6. **Merged COADREAD.** As TCGA pancancer tumor-types COAD (Colon) and READ (Rectum) are treated as the same tumor type "COADREAD" (Colorectal), we did the same with the ICGC tumours COAD-US and READ-US.

Moreover, apart from the previous steps, we discarded those donors with no exonic somatic variants in any protein of the interactome. We discarded also the THCA-SA cancer type from ICGC due to the abnormal amount of natural germinal variants in the processed file (germinal variants extracted from the 1KGP and Spanish MGP individuals). Finally, donors for which the number of mutations deviate by three times the standard deviation were excluded.

4.2.2 Analysis of exome sequencing data

Variant functional annotation

As commented in the previous section, we merged the processed datasets from the ICGC and TCGA/Synapse and extracted the genome coordinates and the reference and alternative alleles generating a bed file. Next, each variant was mapped to the corresponding transcript/protein and the functional consequence was computed by VARIANT (Medina

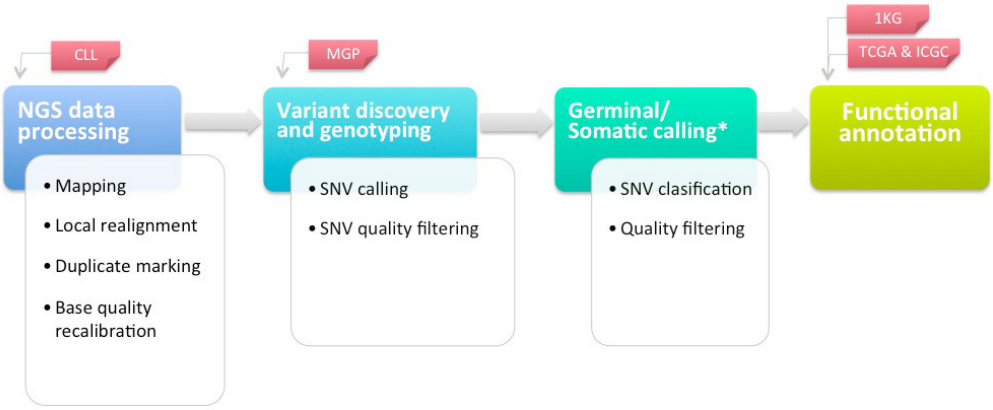

Figure 4.1: Framework for variant discovery and genotyping from NGS data. Data extracted from TCGA and ICGC are processed variant calls included at the 4th step of our pipeline (Functional annotation of variant calls).

et al., 2012) software. Only point mutations were selected for further analysis.

Detection of mutations under positive selection among cancer patients

OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012) and OncodriveCLUST (Tamborero et al., 2013a) were used to search for variants under positive selection across the cohort of tumor samples (pan-can analysis). OncodriveFM identifies genes with a bias towards accumulation of mutations with high functional impact whereas OncodriveCLUST searches for genes with significantly clustered mutations among the gene sequence. OncodriveFM and OncodriveCLUST were run on the merged mutation dataset. (Gonzalez-Perez et al., 2013).

4.2.3 Construction of the structurally resolved human protein interactome

Defining the interacting interfaces of each PPI

The protein interactome used in the previous chapter (section 2.2.1 for more details) was resolved structurally by predicting the protein interaction interfaces of each PPI. For this purpose, folloging the homology modeling approach proposed by Wang and colleges (Wang et al., 2012), co-crystal structures are used as a gold-standard evidence to resolve the structural details of PPIs. For each protein in the interactome, protein domain definition proposed in Pfam and the protein sequence-Pfam mappings were retrieved Pfam database (Finn et al., 2013). Next, pairs of interacting Pfams that were observed to physically interact in at least one high-resolution three-dimensional co-crystal structure in the Protein Data Bank (Bernstein et al., 1978) were collected from 3did (Mosca et al., 2013b) (release of February 2014) and iPfam (Mosca et al., 2013b) (release of September 2013). Finally, each PPI was interrogated to contain a Pfam-Pfam interaction data. When two proteins were shown experimentally to interact and also contain interacting Pfam domains, the Pfam domains were predicted to be responsible for the interaction and were consider them as the interaction interfaces.

Prediction of ordered/disordered protein sequences

Protein ordered and disordered sequence regions were estimated with the DISOPRED-V2 software using default parameters (Ward et al., 2004). The input protein sequences (fasta files) were downloaded from the UniProt database. Evolutionary constraint PhasCons of positions falling in ordered and disordered regions was compared with the non-parametric Wilcoxon test.

4.2.4 Statistical analysis of the mutations distribution among protein region types

The distribution of the pancancer missense mutations with respect to protein regions was evaluated with a permutation test by comparing the observed mutation count (number of mutations across all donors) in ordered protein sequences, interacting Pfam domains and non-interacting Pfam domains, to a null distribution estimated using a permutation approach. Specifically, the permutation consisted of randomly reassigning mutations to protein sequence positions using all proteins from the structurally resolved interactome, so that the total number of mutations in the interactome is always the same as in the observed case. The p-value is calculated as the frequency of the observed value in the null distribution.

4.2.5 Identification of significantly mutated protein interacting interfaces

Identification of PPI interfaces enriched in somatic mutations was assessed PPIprotein-centric approach. Specifically, statistical significance of mutations in each PPI interface was estimated with one tailed binomial test using overall protein mutation ratio as background. Here, given the observed number of missense mutations on the a given interacting domain (X) out of the total number of missense mutations in the protein (N) and the ratio of residues from the protein occupying the domain (R), a one-tailed binomial model was used to compute the probability of observing equal to or greater than X mutations in the domain when the null hypothesis is true (mutations distribute equally inside and outside the interacting domain). We consider that FDR P-values ≤ 0.05 indicate that the ratio of observed mutations in the interacting domain is significantly greater than the ratio of mutation along the ordered protein sequence. Interfaces with less than 5 mutations were discarded.

In order to avoid selecting hyper-mutated sequence regions (Liu et al., 2013), we performed a second test over the synonymous mutations, which are rather not expected to have a functional implication at

the protein level. Interacting interfaces enriched in synonymous mutations where discarded from our predictions since we cannot distinguish whether such bins are just a consequence of hyper-mutation phenomena or, instead, are accumulated due to the conferred selective advantage.

Finally, GOLGA4, RYR2, RYR1 and KRT2 were also excluded from the list as they have been proposed as likely false positives from the methods identifying drivers.

4.2.6 Survival analysis

Survival data was downloaded from TCGA and ICGC data portals. Estimation of overall survival for each patient group was calculated using the Kaplan-Meier method implemented in the *survival* package in R. This method uses the clinical variables *donor age at last followup*, *donor age at diagnosis* and *donor vital status* to quantify the proportion of patients still surviving after a given period of time after its diagnosis. The survival comparison was analyzed using the log-rank test.

4.2.7 Identification of the Minimal Connected 3D Network mutated in cancer

The interaction network between predicted interaction protein interfaces was created using the SNOW method (Minguez et al., 2009). SNOW detects the largest Minimal Connected Network (MCN) linking all the input proteins and tests if network interconnectivity is significantly greater than the corresponding random expectations. To construct the interaction network, we used the structurally resolved interactome allowing the incorporation of one external connecting protein. SNOW algorithm was rewritten so that we add only external proteins to participate in the MCN if they directly interact with, at least, two enriched interfaces. An empirical distribution of the random expectation of the *average number of nodes per component* parameter for a network of N components was obtained by repeatedly sampling random sets of N protein interfaces from the complete interactome. Then, the real value of the

parameter from the MCN between the interfaces of interest is obtained and contrasted with respect to their corresponding random expectations. Finally, the network was visualized using the CellMaps web visualization tool `http://cellmaps.babelomics.org/`.

4.3 Results and discussion

4.3.1 Cancer donors

After merging the somatic exonic data using the criteria described in the section 4.2.1 and filtering donors with no somatic variants in the proteins of the interactome, we collected a total of 5920 cancer donors from 33 different cancer types (Table 4.1). Figure 4.2 shows the number of donors per cancer type and the project source, being the BRCA the cancer with more donors analyzed. Figure 4.3 shows that pediatric and liquid tumours contain far fewer mutations whereas melanomas and lung tumors harbor the highest frequencies. Mutational rates are consistent with previous observations (Vogelstein et al., 2013; Kandoth et al., 2013).

Source	Number of donors	Number of cancer types
TCGA	1807	17
ICGC	1764	20
Shared	2349	9
Total	**5920**	**33**

Table 4.1: Total number of donors and cancer types retrieved from TCGA and ICGC.

4.3.2 Construction of a three dimensional structurally resolved protein interactome

One of the key starting points of this chapter was to resolve the three dimensional structure of the PPIs from the interactome. We followed the homology modeling approach proposed by Wang and colleges (Wang et al., 2012). Specifically, we considered all binary interactions from our curated interactome (see section 2.2.1 for more details) that contain a Pfam domain pair interacting in at least one co-crystal structure in the PDB and, therefore, the specific interactor interfaces of each participant are known. That is, we use co-crystal structures as a gold-

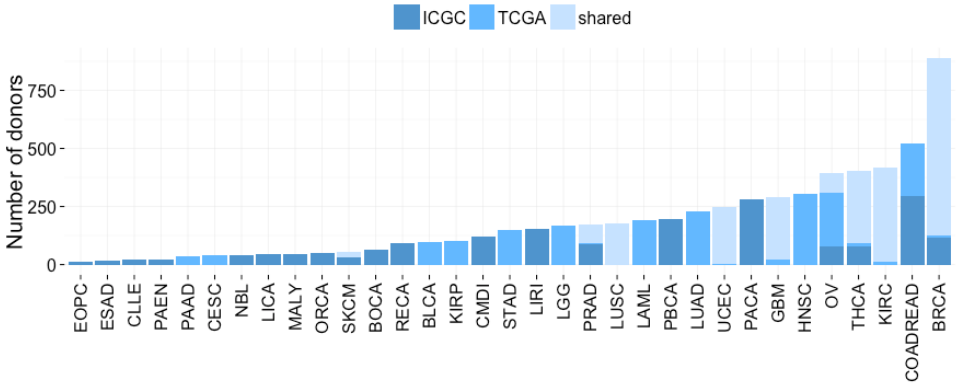

Figure 4.2: Number of donors per cancer type and source. Each bar represents the total count and the color the source from where the donor was obtained.

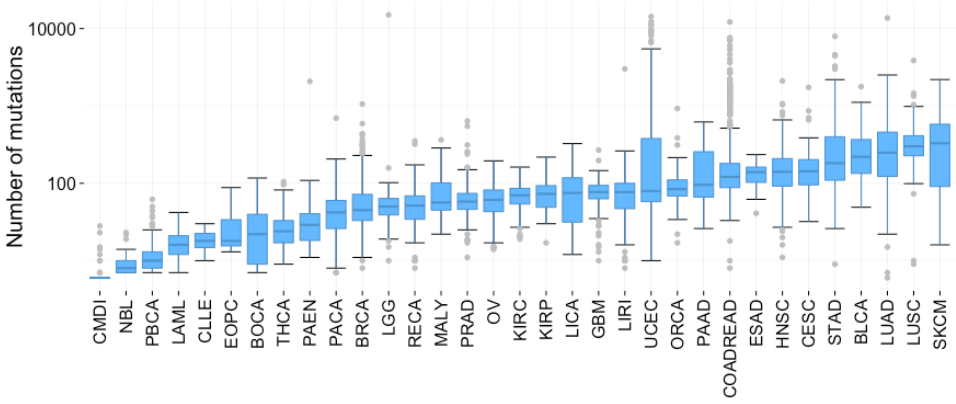

Figure 4.3: Distribution of mutation frequencies for each tumour type.

standard proof that these interactions do occur. The obtained 3D interactome consisted of a total of 7580 unique interacting domains in 13160 PPI between 4996 proteins. This new version of the interactome provides a high-resolution and accurate description of the molecular interactions and is a valuable resource for interpreting the massive amount of genomic data generated from thousands of patients.

4.3.3 Relevance of the protein interacting interfaces in cancer

With the data in hand, this chapter continues with exploring the distribution of the somatic mutation among the protein interacting Pfams. In general, Pfam domains have strong overlap with ordered regions. Since the disordered/unstructured regions of proteins are less restricted spatially and, without some exceptions, evolve more rapidly than ordered/structured regions (Brown et al., 2002), we expect that more somatic mutations would occur in these regions. This unequal evolutionary constraint, observable in the proteins of the interactome ($P \leq 0.001$, Wilcoxon test, Figure 4.4 A), can bias the results in such a way that the enrichment of mutations in the interacting Pfam domains would be underestimated. To overcome this bias, we calculated the probability estimate of each residue in the sequence being disordered and splitted the protein sequence in ordered/disordered regions (Figure 4.4 B).

We retrieved total of 176316 missense somatic mutations mapping 4846 proteins in the interactome. 69563 (39.29%) of these mutations were located within the interacting Pfam interfaces, 16822 (9.5%) in other Pfam domains, 41949 (23.69%) in other ordered regions and 48732 (27.52%) in disordered regions (Figure 4.4 C). These mutations were tested for patterns of positive selection using both oncoDriveFM and oncoDriveCLUST methods, obtaining a total of 22543 variants under positive selection in the tumorigenic process: 10270 (45.56%) in interacting Pfams, 2134 (9.47%) in other Pfams, 4607 (20.44%) in other ordered regions and 5532 (24.54%) in disordered regions (Figure 4.4 D).

Next, we aimed to investigate whether cancer somatic mutations were differently distributed among the distinct protein regions. First, we studied the preferential location of missense mutations for either ordered or disordered regions by comparing the observed number of mutations in the ordered regions for all interactome proteins to the expected distribution, obtained by permuting the variants among the whole protein sequences (Figure 4.5 A). Results show an enrichment for the somatic

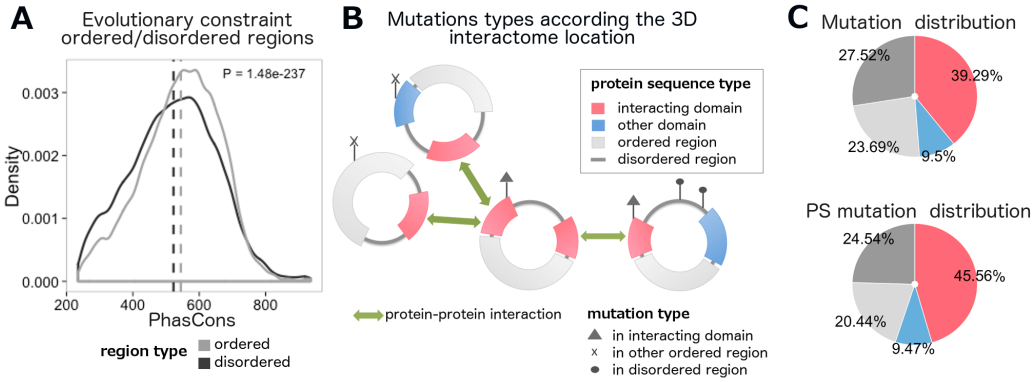

Figure 4.4: Mapping of cancer missense mutations in the structurally resolved protein interfaces. A Evolutionary constraint distribution for somatic missense mutations in ordered (light grey) and disordered (dark grey) protein sequences. Significance p-value for the comparison is displayed on the top corner (Wilcoxon test). **B** Classification of cancer missense mutations according its location in the structurally resolved proteins of the interactome. **C** and **D** proportions of, respectively, all and positively selected (PS) cancer missense mutations among interacting Pfam domains, non-interacting Pfam domains, other ordered and disordered protein sequences.

mutations in ordered regions ($P \leq 0.001$, Permutation test). Second, we focused on studying if there was a tendency for the cancer mutations to be located in the interacting domains. Here we also detect a significant overrepresentation of cancer mutations ($P \leq 0.001$, Permutation test), compared against the random expectations given the mutation frequency among ordered regions (Figure 4.5 B). When splitting the mutations according whether they were predicted to be under positive selection, we observe the same patter for both group of mutations ($P \leq 0.001$ both, Permutation test). Finally, focusing on the non-interacting domains (Figure 4.5 C), all somatic mutations are significantly underrepresented in the non-interacting domains ($P \leq 0.001$, Permutation test) whereas no significance was found for mutations under positive selection ($P = 0.16$, Permutation test).

Focusing on the interacting interfaces, we classify them according whether they occur in proteins that occupy the periphery of the

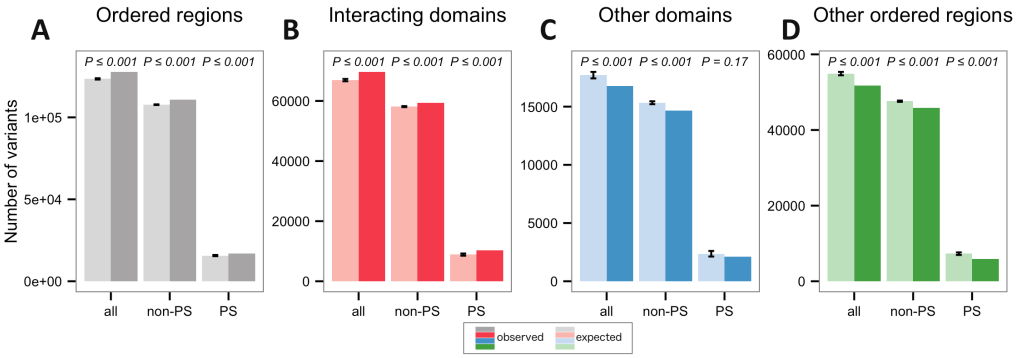

Figure 4.5: Distribution of cancer missense mutations among the structurally resolved protein region types. The bars represent the total number of missense mutations observed (dark color) and expected (light color) in ordered protein sequences (**A**), interacting Pfam domains (**B**), non-interacting Pfam domains (**C**) and other ordered regions (**D**). P-values are shown on top of bars (Permutation test). For the study of the mutation frequency in the ordered protein sequences (A), the distribution of the expected values was obtained by permuting mutations across the whole protein sequence (considering both ordered and disordered regions). For the study of the mutation frequency in the PPI interfaces, non-interacting domains and other ordered regions (B, C and D), the distribution of the expected values was obtained by permuting mutations across the ordered protein sequence. Expected error bars represent standard errors of the non-paramvalue from permutations. All: all missense mutations; PS: mutations under positive selection in pancancer dataset according oncoDriveFM and oncoDriveCLUST methods.

interactome (defined as the proteins that fall in the first quartile of the distribution for the closeness centrality parameter, Q1), in proteins with an intermediate centrality (in the second and third quartiles, Q2-Q3) or central proteins (in the fourth quartile, Q4) and repeated the same test performed inf Figure 4.5-B for each protein group. The results display a more precise and opposite signal between the mutations under positive selection and the rest of mutations (Figure 4.6). The mutations under positive selection significantly concentrate in the interacting interfaces of the central proteins ($P \leq 0.001$, Permutation test), confirming that the impairment of the central binding interactions is a positively selected network hallmark in cancer development. In contrast, the rest of the mutations display a tendency toward the network periphery, behav-

ing similarly to the distribution of the germinal variants from healthy individuals.

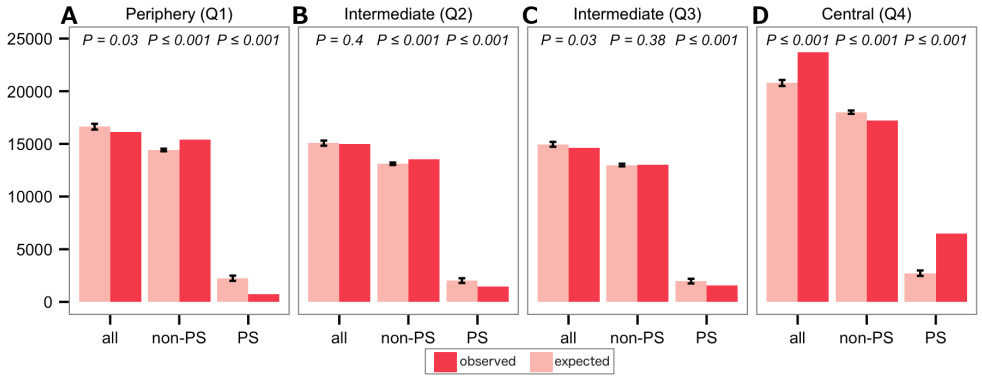

Figure 4.6: Distribution of missense mutations among structurally resolved PPI interfaces according to the network centrality. The bars represent the total number of missense mutations observed (dark color) and expected (light color) in PPI interfaces for proteins located in the periphery of the interactome (defined as the proteins that fall in the first quartile of the distribution of the closeness centrality parameter, Q1), in intermediate regions (in the second and third quartiles, Q2-Q3) or occupying central positions (in the fourth quartile, Q4). P-values are shown on top of bars (Permutation test). The distribution of the expected values was obtained by permuting mutations across the ordered sequences of the interactome proteins. Expected error bars represent standard errors of the mean value from permutations. All: all missense mutations; PS: mutations under positive selection in pancancer dataset according oncoDriveFM and oncoDriveCLUST methods.

4.3.4 Identifying significantly mutated protein interacting interfaces

Motivated by the enrichment of cancer mutations in the binding interfaces of the interactome and its potential functional implications for cancer development, we aimed to search for proteins with a bias in their mutation rates towards its interacting interfaces and, therefore, are likely to contribute to tumor evolution. An expected effect of oncogenic mutations in protein binding interfaces is an alteration in the interaction with its partners. Thus, focusing on each specific interacting domain, we performed a protein-centric mutation enrichment analysis with one tailed

binomial test (see section 4.2.5). We consider that FDR P-values ≤ 0.05 indicate that the ratio of observed mutations in the interacting domain is significantly greater than the ratio of mutation along the ordered protein sequence. Interfaces with less than 5 mutations were discarded.

As the mutation rates are not homogeneous across the genome (Liu et al., 2013) and our results could be biased towards hyper-mutated sequence regions, we performed a second test over the synonymous mutations, which are not expected to have a functional implication at the protein level. We assume that those interacting interfaces that are also enriched in synonymous mutations follow the baseline distribution of somatic mutations and were discarded since we cannot be attribute a direct functional implication.

Systematic analysis of the somatic mutations from each cancer type reveals 83 (FDR $P \leq 0.05$) proteins concentrating its somatic mutations on the interacting interfaces (Figure 4.7). Further analysis across the merged pan-cancer dataset leads to the identification of 161 additional proteins (Figure 4.8), which sums up a total of 252 significantly enriched interacting interfaces in 248 proteins. These predicted interfaces encode potential molecular mechanism for a total of 4308 missense mutations.

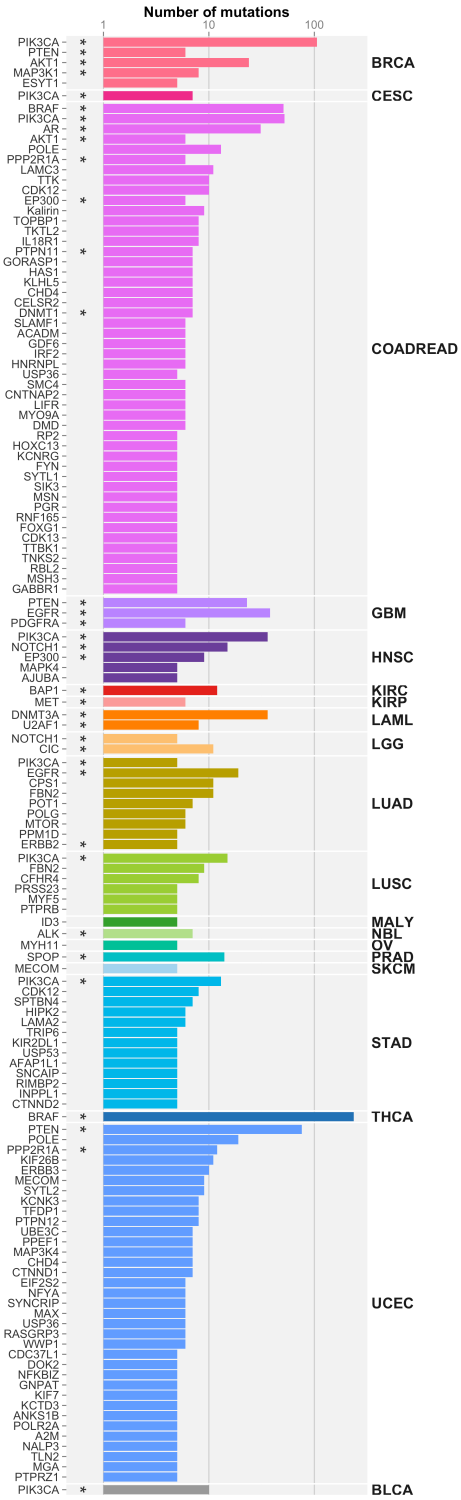

Figure 4.7: Proteins with significantly mutated interacting interfaces per cancer type. Barplot shows the number of missense mutations for each enriched Pfam-protein colored based on the cancer type. Proteins are ordered according to the decreasing frequency of mutations in the interacting interface. $*$: Cancer Driver list proposed by Vogelstein et al. (2013).

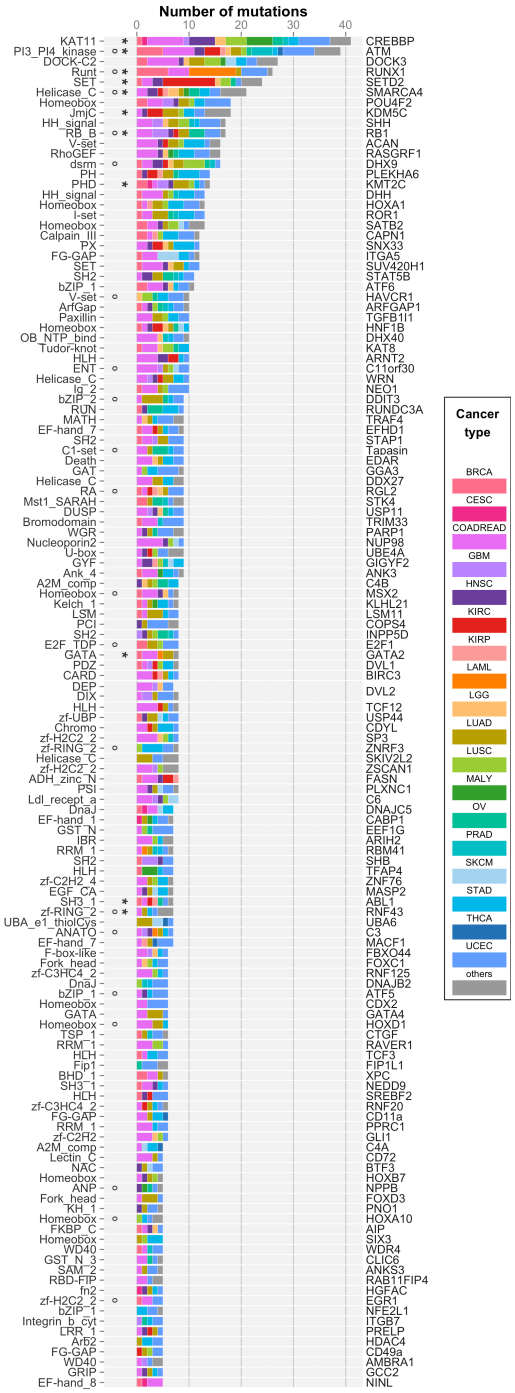

Figure 4.8: Significantly mutated protein interacting interfaces in pancancer analysis. Barplot shows the number of missense mutations for each enriched Pfam-protein colored based on the cancer type. Proteins are ordered according to the decreasing frequency of mutations in the interacting interface. *o*: driver genes identified in oncoDriveFM and oncoDriveCLUST pancancer analysis. ∗: Cancer Driver list proposed by Vogelstein et al. (2013).

Most known cancer genes accumulate mutations on the interacting interfaces

Among the identified proteins, 32 (13%) are in the Cancer Driver list proposed by Vogelstein et al. (2013) ($P = 4.25 \text{ x } 10^{-18}$, Fisher's exact test; genes labelled with * in Figure 4.8). Also, over 38.80% of the mutations in the identified interacting interfaces are considered under positive selection as defined by the pancancer analysis using oncoDriveFM and oncoDriveCLUST predictors ($P \leq 0.001$, Fisher's exact test), whereas the rest 61.19% of mutations would imply new mechanisms.

Functional profiling of the enriched proteins using FatiGO tool (Al-Shahrour et al., 2004) from Babelomics (Medina et al., 2010) reveals an overrepresentation (compared to the whole interactome members) of processes and pathways that are hallmarks of cancer (Hanahan, 2000; Hanahan and Weinberg, 2011), such as regulation of gene expression, regulation of the cell cycle and apoptosis, chromatin modification, protein processing, tyrosine kinase signalling pathways and KEGG pathways in cancer (FDR $P \leq 0.05$, Fisher's exact test). This observation demonstrates that the presented strategy is useful to propose new candidate drivers.

Focusing on the individual predicted interfaces, for example, our results highlight the pleckstrin homology (PH) kinase domain in AKT1 (Figure 4.9 A), a member of the AKT protein, which links several key processes including metabolism, proliferation, cell survival, growth and angiogenesis. PH domain-kinase domain interactions are necessary in maintaining AKT in an inactive state through autoinhibitory interactions and mutations in the PH-kinase interface constitutively active AKT, which aberrant activity leads to cellular transformation (Parikh et al., 2012). These AKT1 mutants are not effectively inhibited by allosteric AKT (which are being investigated in preclinical and clinical testing) highlighting the AKT1 mutational status has important implications for the choice of treatment in the clinic (Calleja et al., 2009; Wu et al., 2010).

Another example is the meprin and traf homology (MATH) do-

main in SPOP protein (Figure 4.9 B), an ubiquitin ligase that promotes the ubiquitin-mediated degradation of the proteins binding to its substrate recognition domain. SPOP substrates are proteins implicated in transcriptional regulation of genes involved in essential cellular functions. Some examples among its substrates are NCOA2 and NCOA3, master activators of several transcription factors such as AR, a well known cancer driver gene (Li et al., 2011). Loss of function SPOP mutations at the MATH domain hampers the interaction with its substrates, another mechanism by which activation may occur in human cancers by reverting the attenuation of AR transcriptional activity (Geng et al., 2013).

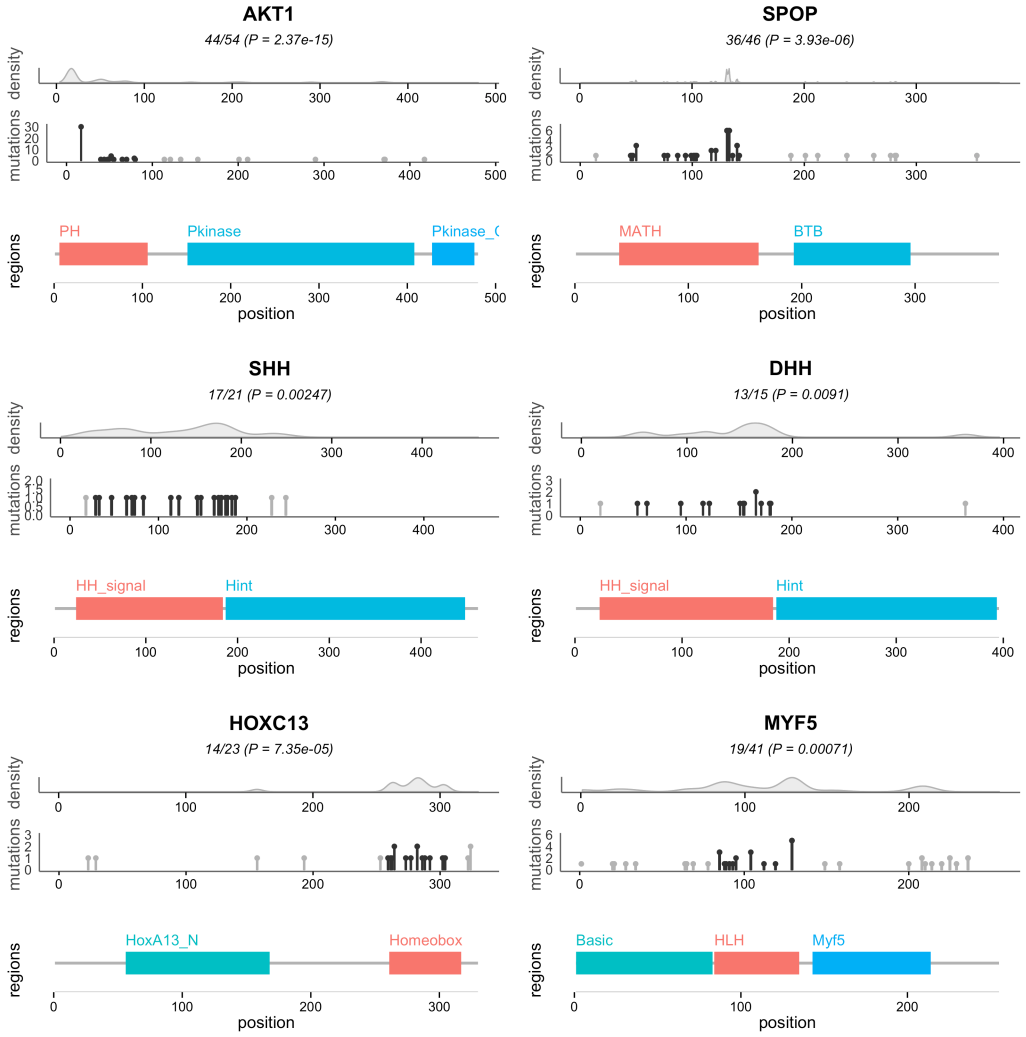

Figure 4.9: Examples of proteins with an enriched interacting domain. From top to bottom: (2) mutation density among ordered protein sequence, (2) total mutation counts for the predicted PPI interface (black) and other protein regions (gray); and (3) location of Pfam domains among the protein sequence. The predicted PPI interface is colored in red.

Novel candidate genes accumulating mutations on the interacting interfaces

Our test highlights novel candidates such as SHH and DHH, members of the Hedgehog (Hh) pathway, or MYF5 and HOXC13, both transcription factors (Figure 4.9 C-F). SHH and DHH proteins are of special interest since they are critical regulators for tissue differentiation and, in adulthood, are involved in the maintenance of homeostasis. Aberrant Hh pathway activity has been implicated in a broad variety of tumors and has been hypothesised to play an important role in the formation and maintenance of cancer stem cells (CSCs). Although it has become clear that aberrant activity of Hh pathway either by point mutations of the downstream proteins (PTCH1, SMO or SUFU) or ligand over-expression leads oncogenic signalling (Ruch and Kim, 2013), to our knowledge, few mutations affecting directly the sequence of the SHH and DHH have been proposed as a tumorigenic mechanism (Oro et al., 1997). The cause may be the relatively low frequency of its mutations when each cancer type is studied separately. This highlights that integrated pan-cancer approach can be crucial in the detection of new cancer mechanisms. Particularly, the altered domain is responsible for the HH signal and directly binds to HHIP, which regulates Hh signalling negatively (Chuang and McMahon, 1999). Therefore we propose that the impairment of the binding of SHH and DHH with its inhibitor could be a hypothetical mechanism for cancer, although further studies should be conducted to corroborate its implications.

Approach comparison

Opposite to the conventional methods based on mutation frequency that search for highly recurrently mutated genes, our strategy studies each gene separately. As observed in other "gene-centric" (Gonzalez-Perez and Lopez-Bigas, 2012; Tamborero et al., 2013a; Reimand and Bader, 2013), our approach is able to detect individual interacting interfaces whose mutational rate is low but unexpected given the protein-wide

number of mutations. In fact, our results include products from the long tail of genes with low frequency mutations (ex. SHH and DHH, members of the Hedgehog (Hh) pathway, or MYF5 and HOXC13, both transcription factors). After this approach was designed and applied to the cancer data, we learned about the publication of a similar study by Porta-Pardo and Godzik (2014) called e-Driver. As the approach we propose here, e-Driver studies the accumulation of cancer mutations on pre-defined gene functional regions using one tailed binomial tests, which supports the use of this test for such objective. However, they do not correct for the accumulation of synonymous variants.

Yet, our approach has a major limitation: it requires structural information on PPIs, which is not available from more than a half of the known interactome. Moreover, since the method specifically search for deviations from even distribution, genes for which mutations are homogeneously distributed across the sequence, such as tumor suppressors, cannot be detected. Also, genes that accumulate mutations in other functional regions but not their interfaces will not be detected. Thus, our method is complementary to other methods and the capture of all the players in oncogenesis would require the combination of the different strategies (Tamborero et al., 2013b).

4.3.5 Topological properties of enriched interfaces

Our work would be incomplete without an analysis of the topological properties of the affected interfaces. The basic questions to assess are: Do the affected interfaces have a preference to occupy central positions? Do they involve more interactions than the non-affected interfaces? Are they more likely to interact with one another, or are they spread around the interactome? To address these questions, we mapped the enriched domains onto our structurally resolved protein interactome and examined its topological properties. As expected, 33.66% of the proteins with predicted PPI interfaces are located in central positions (Q4, $P = 0.0138$, Fisher's exact test). Next, interested in deciphering if the

predicted interfaces were concentrating more interactions, we observed that enriched interfaces participate in more interactions than the interfaces of non-enriched proteins ($P = 0.0333$, Wilcoxon test). Interestingly, no difference was found compared to non-enriched interacting interfaces of the same predicted proteins ($P = 0.479$, Wilcoxon test), which suggests that the hub role of cancer genes seems a property relative to the whole protein more than to the enriched interface (Figure 4.10 B).

Biological processes involved in oncogenesis converge in common regulators that modulate crosstalk between them (Bustin, 1998). To study whether this property is reproduced in our predicted interfaces, we look for predicted interfaces that interact directly and observed a total of 37 interactions in which both interfaces in the PPI are enriched in mutations. To evaluate if the predicted interfaces have a higher tendency to interact with one another, we permuted 1000 times the interactor labels of the network while preserving the total connectivity of each protein. Results show that in only 0.3% of the random cases we obtain a value equal or greater than the observed ($P = 0.003$, Permutation test, Figure 4.10 C). This analysis was repeated using the non-enriched interfaces of the same predicted proteins and observed 25 direct interactions, and no significant difference from the random expectation was detected ($P = 0.73$, Permutation test, Figure 4.10 D).

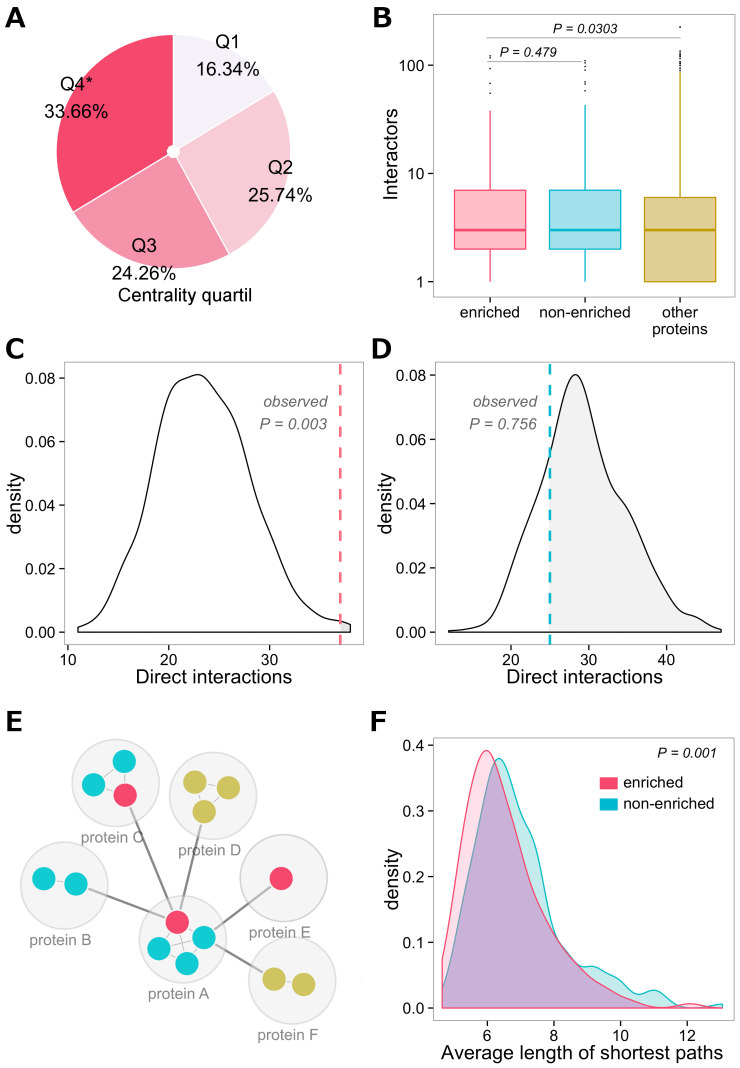

Figure 4.10: Topological properties of affected interfaces. (A) Proportion of proteins with enriched interfaces in each centrality quartile. **(B)** Interaction degree comparison between enriched interfaces, non-enriched interfaces in the same predicted proteins and interfaces in non-predicted proteins. **(C)** Expected distribution and observed value (dashed red line) for the number of direct partners between the enriched interfaces. **(D)** Expected distribution and observed value (dashed blue line) for the number of direct partners between the non-enriched interfaces in the same predicted proteins. **(E)** Graph model of the structurally resolved protein interactome. **(F)** Average length distribution of the shortest paths within enriched interfaces (red) and non-enriched interfaces from the same proteins.

Motivated by these results, we remodeled the interactome into an undirected graph by an alternative approach in which nodes represent protein interfaces (instead o the whole protein) and edges the interactions either between or within proteins (Figure 4.10 E). This new version of the interactome incorporates a new level of complexity that discriminates between different edges from the same protein when interact through different interfaces. In order to study the crosstalk between enriched and non-enriched PPI interfaces, all occurring in the predicted proteins, we computed the distances within each set of interfaces by means of the average length of the shortest path. We found that the distribution of shortest network distances is skewed towards shorter paths for the enriched interfaces, which can be a consequence of this preferential interaction affinity of the enriched interfaces to interact with each other as compared with non-enriched (Figure 4.10 F). The relevance of this observation is that it points out at the function-centric organization of the interactome, where different proteins can cause similar clinical disorders when they affect regions regulating common functions.

4.3.6 The 3D cancer interactome: new insights into the cancer hallmarks

To extract a rationale map of the proteins with driver interacting interfaces, we calculated the MCN allowing one intermediate using the SNOW tool (Minguez et al., 2009). The identified subnetwork involves 535 interactions between 293 proteins (153 are external intermediates added to connect the network, and 15 of them are known-driver genes) (Figure 4.11). As commented before, the subnetwork contains 39 direct interactions between predicted interfaces of 37 proteins (MCN, $P = 0.019$, SNOW Permutation test), that is, interactions in which both PPI interfaces are enriched in somatic mutations (bold edges in Figure 4.11 B). Over the 56.12% of the donors have at least one somatic mutation in the interfaces of the MCN PPIs.

Several well-known cancer processes and pathways are embedded in

the predicted subnetwork, such as the PIK3-AKT, the ErbB, Jak-STAT, AKT/mTOR signalling pathway, PTEN dependent cell cycle arrest and apoptosis, the DNA-dependent binding transcriptional regulators, etc, which validates our approach. Again, when performing a functional enrichment over both affected proteins and direct partners, we observe an overrepresentation (compared to the whole interactome members) of processes and pathways that are hallmarks of cancer, such as transcription factors, chromatin modificators, phosphorylation processes, regulators of cell development and death, and even proteins implicated in the vascular development and angiogenesis (FDR $P \leq 0.05$, Fisher's exact test).

However, this subnetwork provides also novel hypotheses for regulation of transcription processes potentially involved in tumor progression. One example is the network component formed by the members of basic helix-loop-helix (bHLH) family: ID3, MAX, MGA, TFAP4, TCF12, TCF3, MYF5 (Figure 4.12 A). All these proteins are members of a well-known group of transcriptional regulators of genes involved in the regulation of gene expression and cell fate. They form homo/heterodimers by the non-covalent interaction between the bHLH domains, which is required for an efficient DNA binding. Interestingly, three recent exome sequencing studies of Burkitt's lymphoma patients provided convincing support for the idea that ID genes may function as tumor-suppressors. Concretely, mutations affecting TCF3 or its negative regulator ID3 are found in the 70% of sporadic Burkitt's lymphoma, blocking the interaction between TCF3 and ID3 breaking the negative regulatory loop created by ID3 (Figure 4.12 B) (Schmitz et al., 2012; Richter et al., 2012). Moreover, other studies have described a crosstalk between the HH and WNT/β-catenin pathways, which cooperate inducing expression of some bHLH proteins in cancer (Javelaud et al., 2011). In this sense, HH signal activates bHLH family expression during development and inappropriate activation of bHLH signaling in individual cells may contribute to tumor initiation, as observed for Rhabdomyosarcoma (Gerber et al., 2006).

As for the HH pathway members, much of the previous work has

focused on studying the gene expression patterns of the bHLH family in cancer but, to our knowledge, mutations in the subnetwork formed by the bHLH TFs have not been directly implicated yet. As described before, only few works have found an association between ID3, TCF12 and TCF3 proteins and tumor initiation. Our results suggest that mutations in this network component may cause aberrant expression of genes involved in proliferation/cell fate determination by affecting binding events between these TFs and, thus, contribute to the progression of the malignant phenotype (Figure 4.12 B).

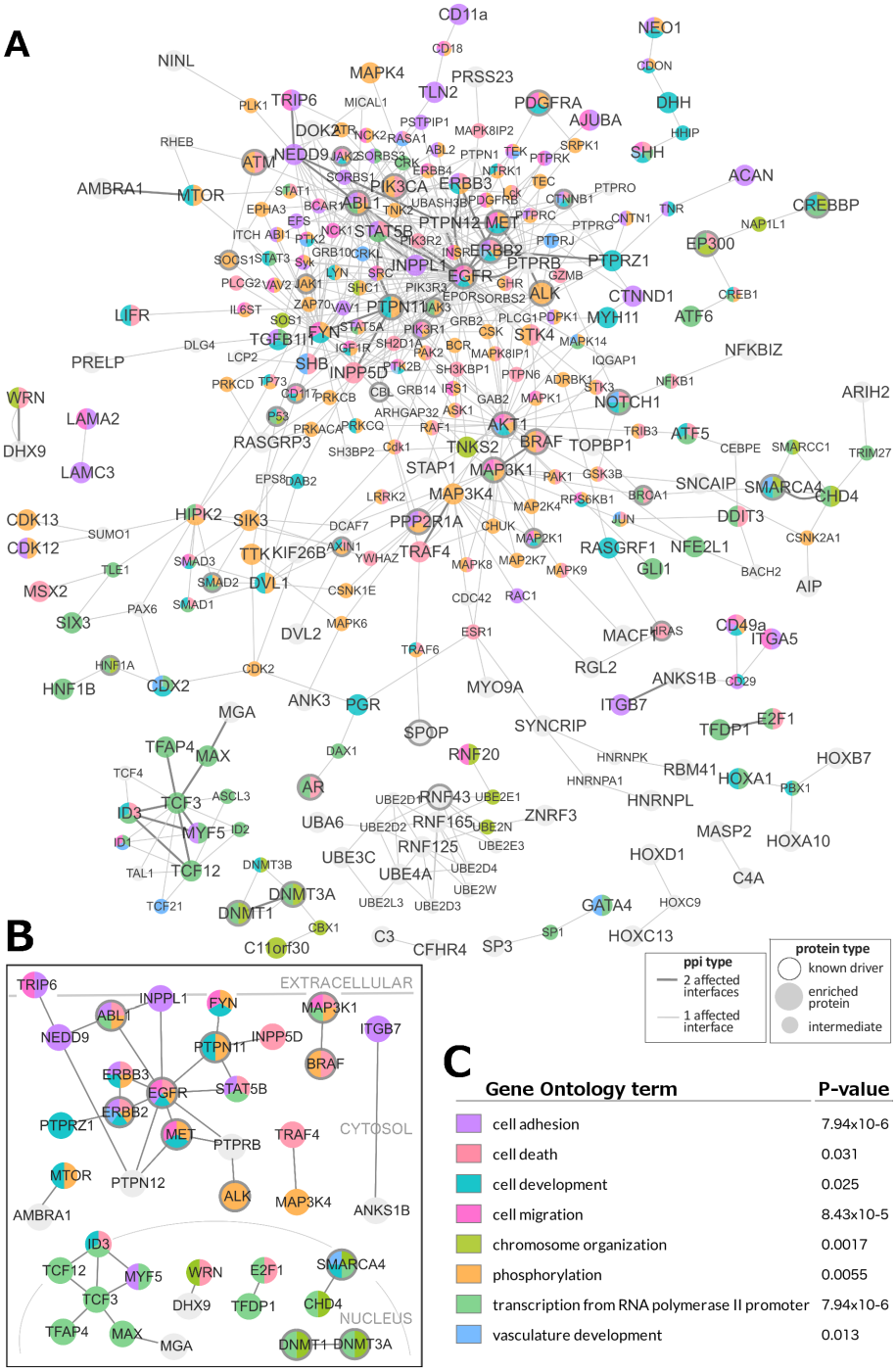

Figure 4.11: 3D subnetwork of enriched interacting domains. **(A)** MCN between the enriched interfaces allowing one intermediate. **(B)** MCN between the enriched interfaces without intermediates (only direct interactions). **(C)** Main GO Biological Processes enriched in the MCN.

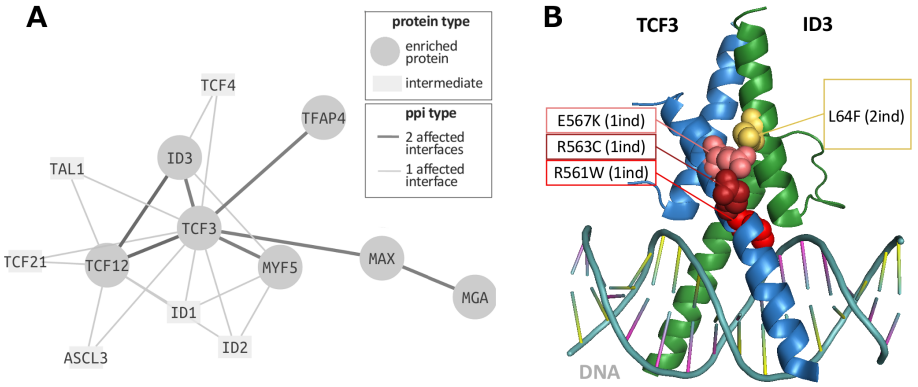

Figure 4.12: Cancer mutations in interacting domains in bHLH family of TFs. Cancer mutations in interacting domains in bHLH family of TFs. (A) Interacting interfaces of bHLH enriched in cancer mutations. Dark grey nodes indicate proteins carrying the enriched interfaces. Dark grey edges indicate interactions where both interfaces are enriched **(B)** Mutations from TCF3 and ID3 on the 3D structure of bHLH domain predicted to impact the dimer interacting interface (Model extracted from PDB ref. 2LFH)

.

4.3.7 Clinical relevance of mutations affecting interacting interfaces: survival implications

To explore the prognosis significance of the oncogenic candidate interacting interfaces, we downloaded survival clinical data from Breast Cancer (BRCA) patients. Breast cancer is the most commonly diagnosed cancer among women in Europe, causing 131,200 deaths in 2012 (Ferlay et al., 2013). Our aim at this step is to study the role of the mutations located on the interacting interfaces with the tumor evolution. Our hypothesis is that if hampering an interaction is a relevant mechanism for the tumor progression, then patients for which driver mutations are located at a specific interface (and mutations that considerably change the overall protein structure) would display a similar disease evolution. Thus, for each significant interacting interface in our analysis, we split the patients into two groups: one containing mutations in the interacting interface and the other with mutations in the same protein but outside the interacting interface.

Results show that mutations in the PI3Ka domain (exon 9) of PIK3CA were strongly associated with increased survival compared to the rest of mutations (Figure 4.13 A). However, a bibliography research reveals that the prognostic value of PIK3CA mutations at different regions in breast cancer remains controversial. Whereas some studies have reported that the presence of H1047R (PI3-PI4 kinase, exon 20) mutation was strongly associated with the absence of lymph node metastasis and better prognosis (Barbareschi et al., 2007; Kalinsky et al., 2009), our results and three other studies report that exon 9 mutations are associated with increased survival in breast cancer (Lai et al., 2008; Mangone et al., 2012; Arsenic et al., 2014) and lung adenocarcinoma (Zhang et al., 2013). Splitting patients into fourth groups (with mutations in the predicted PPI interface, H1047R high frequent mutation, mutations in the rest of PIK3CA and no mutations in PIK3CA) also reveals a difference in survival between donors (Figure 4.13 B-C), being the patients with mutation in the predicted interacting interface those with better pro-

gnosis. The difference is maintain when the information of the direct interactors is included, although due to the low recurrence of the latest, an independent validation of the mutations in PI3KR1 and PI3KR3 can not be assessed (Figure 4.13 D). These mutations were previously investigated in Glioblastoma in primary studies from the TCGA, where it was proposed that mutations on the PIK3a domain might prevent inhibitory contact with PI3KR1 and PI3KR3, causing constitutive PI3K activity (McLendon et al., 2008) (Figure 4.13 E). Recently, Hao et al. (2014) observed that PI3Ka mutant proteins, but not H1047R mutant proteins, interact with IRS1 and that the disruption of the interaction between PIK3a domain mutant and IRS1 inhibits tumor cell proliferation. Additional studies on larger and more homogeneous series of patients are necessary to verify the real significance of the association and the mechanisms behind. Nevertheless, results evidence that mutations in different PIK3CA regions may play different roles in tumor evolution.

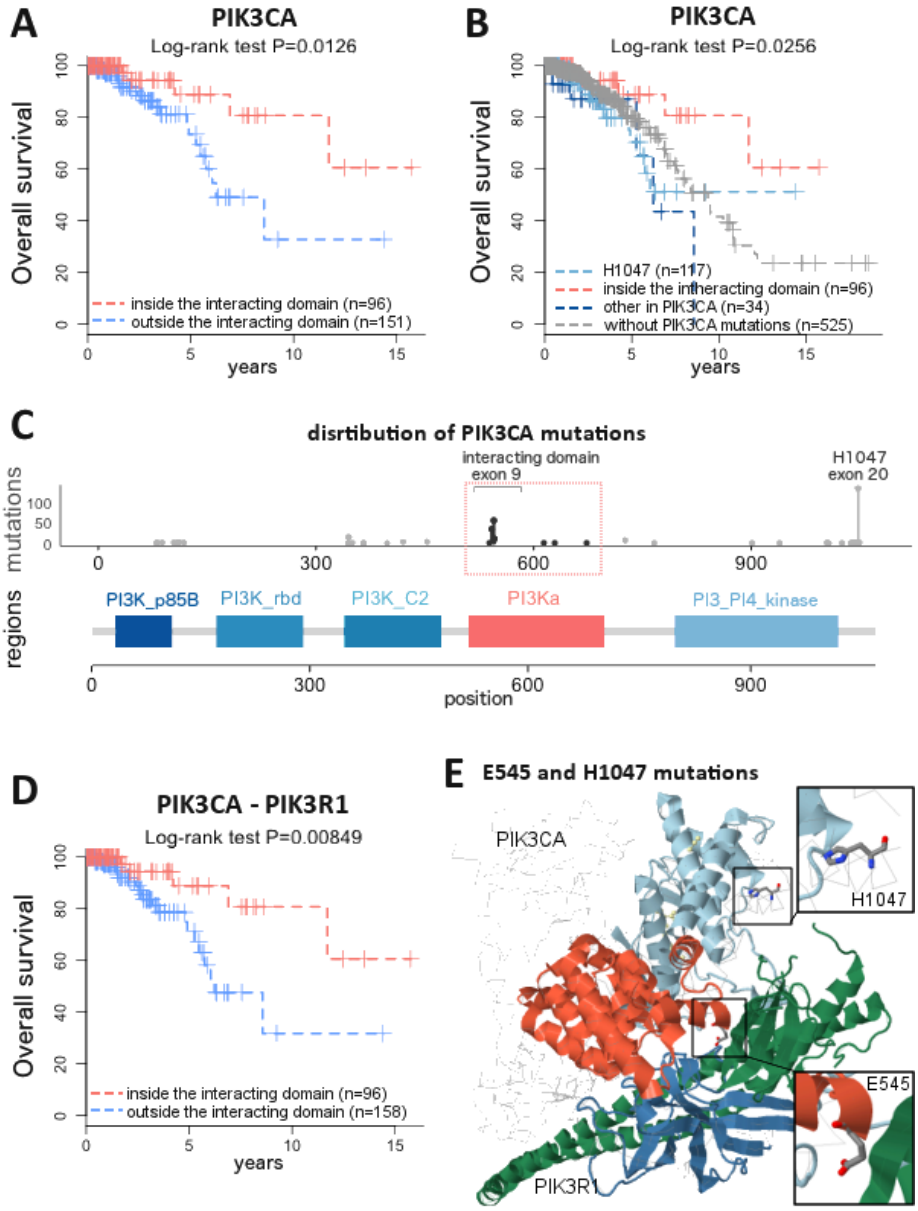

Figure 4.13: Survival analysis on PIK3CA mutations in BRCA patients. **(A)** Kaplan-Meier survival plots for patient with mutations inside (red) or outside (blue) the PIK3CA interacting domain. **(B)** Kaplan-Meier survival plots for patient with mutations inside the PIK3CA interacting domain (red), H1047 mutation (light blue), other PIK3CA mutations (dark blue) or wild type PIK3CA (gray). **(C)** Location of mutations among PIK3CA sequence domains **(D)** Kaplan-Meier survival plots for patient with mutations inside (red) or outside (blue) the PIK3CA-PIK3R1 interacting domains **(E)** Mutations location in the 3D structure of the interaction between PIK3CA (domains in orange and light blue) and PIK3R1 (domains in green and dark blue). Model extracted from PDB ref. 4JPS

.

Taken all together, our results suggest that qualitative changes in protein interactions could explain heterogeneity between cancer patients better than considering isolated genes, ignoring the structural details that mediate its communication with other molecules.

CHAPTER 5

Summarizing discussion

The so-called *post-genomic* era has brought a technical revolution in biological science that supposed an unprecedented ability to generate genetic data. Still, the challenge faced by both *pre-* and *post-genomic* is the same: to bridge the link between genotype and phenotype. Since transforming genome information into a final organisms requires numberless cooperative actions mediated by its encoded molecules, sequencing the genome and describing the extracted data represents only the beginning (Brenner, 2000). The road-map to readily asses how changes in the genome sequence impact phenotype necessarily needs the contextualization of this data under a model that fully integrates the structure, function and organization (i.e. complexity) of the molecular interactions that give rise to the cell physiology. Due to their essential role in carrying cellular functions, proteins and its interactions provide a level of abstraction that fits these needs. Protein interactome reflects a high degree of this cellular complexity and represents an intermediate layer between genotype and phenotype. Following this reasoning, this thesis proposes the interactome as a theoretical scaffold for the interpretation of genomic data.

Along the manuscript, different approaches to integrate human genomic data with PPIs are described. The overall objective is, by making use of the interactome, to propose functional hypotheses that guide the interpretation of genomic variability under different phenotype conditions. Although each one covers a different question, all of them demonstrate the potential of the interactome in helping to interpret the genomic variation observed under diverse research scenarios.

5.1 Assembling the human protein interactome

Systematic collection of PPIs for global studies requires a strict protocol that ensures the identification of confident PPIs. Due to the its diversity in nature of and the lack of a unique technique to detect all PPI types, to cover accurately the whole interactome requires to turn to a combination of approaches. Here, PPIs were retrieved from three different databases that collect data derived from a wide range of experimental studies and make it easily accessible and integrable through standard formats. Following the advice derived from Von Mering et al. (2002) to reduce the false positives PPIs, only those determined by at least two methods were used. The two methods criteria is a strict filter that decreases the PPIs coverage but, since some experimental detection methods have been quantified to sweep along 50% of false positives, we took this criteria in favour of accuracy.

Despite carefully curated, the use of prior knowledge always entails a risk of obtaining results biased towards well-studied biological processes (Das and Yu, 2012). Thus, the information related to understudied proteins could be underestimated in comparison with that of the well-studied proteins. It might be argued that this effect could, for example, enhance the differences in network parameters between disease genes and other classes. Concerned by this fact, databases integrating large-scale screenings were specifically selected as sources of PPIs, avoiding knowledge-based sources. Due to its large-scale nature, these screenings are expected to suffer less from this bias towards well-studied proteins.

5.2 NetworkMiner: searching for gene alterations that collectively, though PPIs subnetworks, associate with physiological states

Once assembled the protein interactome, the next step of this thesis focused in developing a method . The development of the method required an exhaustive analysis to identify the combination of parameter and test that best distinguishes between a PPI subnetwork carrying a coherent function from a random one. The results point to the average number of nodes per component of the MCN as the most sensitive parameter to discover real networks and distinguish them from random networks. The biological meaning of this parameter displaying the best discriminative power proves that the only constraint in functional PPIs subnetworks is that its nodes aggregate to a connected component, independently of the subnetwork shape (i.e. cascade-like in signalling pathways or hairball-like for protein complexes).

The final tool, called NetworkMiner, searches for genes at the top of the ranked list that would, collectively, contribute to the ranking experimental parameter. Thus, the approach looks for patterns of cooperativity, which is expected to be the behaviour displayed by genes associated to complex diseases. Although the method is used to identify the interactome module associated to those genes carrying SNPs with high discriminative power between Bipolar Disorder patients and controls, the input list of genes can be ranked by any criterion. Consequently, it is applicable to a large variety of experimental or theoretical scenarios. This is the main difference with respect to previously published methods, which focus only on gene expression (Ideker et al., 2002; Sohler et al., 2004).

In a similar way to the GSEA methods (Subramanian et al., 2005), NetworkMiner avoids pre-selecting genes with arbitrary cut-off. In contrast, this is defined according the tendency of the top genes to form significantly connected modules. With this tool, we extend the variety of tools for functional profiling of genomics data introducing PPIs.

Contrary to methods using functional labels such as GO terms, PPIs do not pre-define functional modules, instead they help to define it from experimental data, thus offering a great potential to discover new functional modules instead of being limited to the known ones. NetworkMiner is freely available at the Babelomics web platform (`http://babelomics.bioinfo.cipf.es/`), and is integrated with other methodologies such as FatiGO and SNOW, providing all together a powerful framework for the functional profiling of genomic data.

5.3 The interactome as a buffer of deleterious variability

In the past years, vast amounts of data are being generated with the purpose of identifying the mechanistic roots for most human diseases. The observation of a non-negligible amount deleterious variants in healthy individuals, among which there are a few with pathological associated effects, is challenging the search for disease-causing variants. The second chapter of the thesis is motivated by the need to characterize the load of genomic variation in human populations and propose a rationale for their tolerance. The endeavour requires first to understand how humans deal with such amount of deleterious variation and, within this, how pathological variants can contribute to a disease phenotype in some individuals but apparently be innocuous in others. In this part of the thesis we propose that the cell system organization may provide an additional buffering mechanism of internal perturbations. This idea is not new and has been long studied in other organisms (Albert et al., 2000; Hahn and Kern, 2005).

Taking advantage of the availability of a wealth of genomes of healthy individuals, the exome variants of more than 1330 healthy humans were analysed in the context of the protein interactome. Our results provide a rationale for the tolerance to potentially deleterious variants based on protein network topology. First, the individualized observations made in healthy subjects and CLL patients, completed with the analysis of proteins mutated diseases, strongly suggest that the pathogenic role of deleterious mutations is highly correlated with the impact on the interactome integrity caused by the combined deleterious of the affected proteins, which is also related to the location of such proteins within the interactome. In this senses, affected proteins in healthy individuals are concentrated in peripheral modules, avoiding internal modules. However, the most important factor which sheds light on the mechanisms by which the interactome can bear a large number of proteins with deleterious

mutations is related to the way in which affected proteins are specifically combined in healthy individuals. Affected proteins in healthy individuals tend to occur in combinations which preserve short path lengths. When the same proteins occur in random combinations, the length of the shortest paths significantly increases. The same results are observed in tissue-specific interactomes. Thus, the structural constraints imposed by the preservation of interconnectivity, and therefore central proteins, seem to restrict genetic deleterious variation to the periphery of the interactome. This is an expected finding since at the periphery, combinations of affected proteins that preserve shortest path lengths are easier to find than in internal regions of the interactome. Since periphery is in closer contact with the environment (proteins from this regions are enriched in plasma membrane and extracellular proteins), the interactome shape may have evolve to enlarge its ability to adapt under changing conditions. Consequently, the system remains robust to perturbations but at the same time, allocates variability and allows evolutionary diversification with low immediate impact in the system organization (Levy and Siegal, 2008). Hence, evolutionary results evidence that proteins under positive selection tend to be placed at the periphery of the interactome, whereas proteins under negative selection tend to have a central location in the interactome (Kim et al., 2007).

The patterns observed in human populations were compared to those displayed by somatic variants found in CLL primary tumour. Interestingly, somatic variability form cancer patients showed a distribution completely opposite to germline variants. Whereas germline variation reflects evolutionary variability and constraints during population evolution, recurrent somatic variants in cancer patients reflect tumour evolution. Cancer driver genes affect to very fundamental processes, such as cell division, control of gene expression or DNA repair, which leads these genes to display properties similar to essential ones (Wachi et al., 2005; Jonsson and Bates, 2006). This result agrees previous observation made in studies comparing network properties of disease genes (Goh et al., 2007; Feldman et al., 2008). It may be interpreted that, to

allow a tumour to evolve, advantageous mutations with high immediate impact in the signalling circuits are needed so that the cell can acquire notable selective advantages (Nowell, 1976; Stratton et al., 2009). Since high impactability seems a property restricted to the core of the network, this may explain the opposite accumulation tendency between germline and somatic variation.

Taken all the results together, the general conclusion of this chapter is that the deleterious character of a variant obviously depends on the damage it causes to the protein, but ultimately, it is a system's property that critically depends on the specific combination of affected proteins and its relative location within the interactome.

5.4 Oncogenic signatures on PPI interfaces

Motivated by the fact that cancer associated genes tend to occupy central roles in protein networks, the last chapter focuses on studying somatic variability distribution among PPIs. For this purpose, we structurally solved the interactome in order to identify the specific protein regions responsible for the interactions (interfaces). This new version of the interactome is more resolutive and allows to study proteins as multifunctional factories instead of monolithic graph entities (Zhong et al., 2009; Wang et al., 2012). First, the exploration of the distribution of missense mutations among the different regions of the interactors showed a clear pattern: these tend to concentrate in ordered regions of the proteins and, within these, they have a clear preference for the interfaces. Moreover, when the centrality of the protein is considered, we observed a clear tendency of the somatic mutations to occupy central positions. Finally, when mutations are divided into two groups, one including the mutations under positive selection and other the rest, we observe a more specific and opposite pattern between them. Mutations under positive selection concentrate in central regions and avoid the periphery, evidencing that the impairment of the central interactions is a positively selected network hallmark in cancer development. In contrast, the rest of the mutations display a tendency toward the network periphery, behaving similarly to the distribution of the germinal deleterious variants from control individuals, as defined in the previous chapter.

Also we propose a new approach to identify interfaces whose mutation may be advantageous for cancer development. The approach follows the philosophy of other gene-centric approaches and does not only focuses on the frequency of its mutations but in the way they are distributed along the protein (Gonzalez-Perez and Lopez-Bigas, 2012; Reimand and Bader, 2013; Porta-Pardo and Godzik, 2014). By studying the PPI interfaces, the approach identified more than 250 interacting interfaces candidate to drive cancer. Several sources of evidence support the potential role of the identified mutations: first, they accumulate in a specific

region instead of being homogeneously distributed among the protein; second, these regions are conserved domains, which functions have been evolutionary maintained; and third, these regions mediate PPIs, which are crucial in mediating the transmission of the biological signals. Thus, these interacting interfaces can be seen as mechanistic hypotheses candidates to explain the molecular basis behind the genetic-cancer associations for most of the known cancer genes. This fact is the most significant difference from frequency-based methods, which does not propose direct functional insight into the identified genes.

The power of the approach is evidenced by the ability of detecting low frequency mutations, which direct role in oncogenesis is sustained by independent published studies. Examples of these mutations are those falling in the HH domain of SHH and DHH (Oro et al., 1997) and in the bHLH domain of the bHLH family of transcription factors (Schmitz et al., 2012; Richter et al., 2012). Moreover, the approach identified interacting interfaces which mutational state relates to patients prognosis, being able to explain survival differences in patients of the same tumor type and with the same mutated gene (Lai et al., 2008; Mangone et al., 2012; Arsenic et al., 2014).

Briefly, the research presented here offers a novel approach for interpreting cancer genomes and provides a new source for hypotheses for most cancer driver genes. The results evidence that consider genes as multifunctional effectors instead of homogeneous black boxes can shed light on both clinical and genetic heterogeneity. In a scenario where is becoming evident the heterogeneity between cancer patients and the complexity of predicting prognosis outcome, zooming into the molecular consequences of genomic variants and contextualizing them in the network of molecular interactions would constitute a step forward towards personalized medicine. Due to the structural information on PPIs is still limited, a major caveat of the method, we expect our results to represent only the tip of the iceberg.

5.5 Future directions

5.5.1 The interactome model

As stated before, the use of prior knowledge has a major limitation: to bias the results towards well studied processes. To overcome the lack of coverage of PPIs is an urgently needed task. In this sense, at least two challenges should be addressed to go beyond this bottleneck. The first one should focus in improving the quality assessment and curation process of PPIs. Recent approaches are considering additional sources of evidence that the number of supporting detection methods, such as orthology, interacting conserved domains, co-expression and common functionality, etc (Schaefer et al., 2012), which are expected to reach a better balance between accuracy and coverage. The second task should reinforce the efforts in the experimental determination of PPIs. During the writing of this manuscript, Rolland et al. (2014) published a new PPIs screening that resulted in an interactome about 30% larger, more homogeneous and less biased than the one available from small-scale studies. Global study of disease genes revealed significant connectivity for cancer gene products, providing unbiased evidence that supports the results obtained in this thesis. Both, systematic PPIs screenings and improvements in the curation processes will soon provide a high quality impartial interactome. Although newest unbiased interactome topologies seem to agree with former ones, the analysis presented here should be repeated to asses the robustness of the observations as the interactome reaches more coverage.

There are other limitation intrinsic to the cellular model used that should be mentioned. First, although PPIs represent to some extent the cellular complexity, there are many other molecular layers that mediate the conversion of the genetic information into final phenotypic states. Examples are protein-DNA and protein-RNA interactions in chromatin remodelling and the regulation of the gene-expression, or protein-small molecules in metabolism regulation. Second, this work, as for the major-

ity of publications on protein networks, has studied the interactome as a single static entity, merging all known PPIs. However, cells are highly dynamic systems that continuously integrate and respond to molecular and environmental perturbations (Altelaar et al., 2013). In this sense, some approaches integrate static interactomes with dynamic changes in gene expression to infer case specific networks (Ideker and Lauffenburger, 2003; Taylor et al., 2009; Holme and Saramäki, 2012). Third, we should keep in mind that although two proteins with the ability to interact are expressed in the cell at the same time, this does not imply that the interaction is going to occur. For some PPIs, the physico-chemical contacts are conditional and depend on additional biochemical changes, such as post-translational modifications, interaction with other molecules or location (Perkins et al., 2010; Grossmann et al., 2015). Finally, annotating each interaction with its direction and sign would enable to take into account the directionality of the regulation and make the results more mechanistic. However, the more complex the cell model is, the closer to reality the results are, but the more difficult the modeling task becomes (Ideker and Lauffenburger, 2003). Since detailed data and modeling approaches are limited, we are obligated to choose the correct balance that best helps to test our hypotheses.

5.5.2 Modelling mutation consequences

Another issue that needs to be improved is the annotation of the functional implications of the mutations at the molecular level. Whereas the effects of deletions, insertions, and stop gain/loss are self-evident, consequences of single missense variants are rather more difficult to predict. Important efforts are being made to develop predictors of mutation deleteriousness, damage and/or functional impact, such as the SIFT and PolyPhen tools used in this thesis (Thusberg et al., 2011). Most of these tools rely on general properties, such as the sequence conservation, structural features and amino acid attributes (Riera et al., 2014). All these properties are combined into global scores (Kumar et al., 2009;

Adzhubei et al., 2010; Bromberg and Rost, 2007). Although extremely useful in assessing potential functional impact for the prioritization of disease mutation candidates, these do not provide insight about the specific molecular effect. Different mutations can have very different effects in the protein functions, such as affect the catalytic site, lead to structural instability or aggregation, alter binding of ligands, reconfigure regulatory regions, affect protein post-translational modifications, cellular localization or being neutral. In addition to protein effects, variants can also modify mRNA stability, processing or translational regulation or even the affinity of the transcription factors to DNA targets (Thusberg and Vihinen, 2009). Functionally, this implies that different mutation in the same gene may have different molecular consequences.

In a recent study, Powis and colleagues observed that cell cycle signaling differs between the different forms of mutant KRAS (Ihle et al., 2012). This heterogeneous behavior of KRAS mutations evidences that personalized medicine needs to take into account the specific mutations expressed by the tumor. To date, the majority of methods to study specific mechanisms focus on the changes in protein stability after mutations (Schymkowitz et al., 2005; Worth et al., 2011), whereas the methods to predict other specific effects in a systematic way are limited. To our knowledge, there only are available Mechismo, which proposes a framework to detect changes in interaction affinities between PPIs, protein-DNA and other chemical interactors (Betts et al., 2014); and MIMP, which predicts mutations that specifically alter kinase-binding sites in proteins (Wagih et al., 2015).

Deciphering the implications of the mutations in the molecule function will have a great impact in the drug development. For example, a mutation predicted as damaging can lead to a constitutive activation whereas another damaging in the same protein may abolish an interface required to transfer the molecular information to a partner. In each scenario, the mode of action required for a drug to revert the alteration would be different. It is likely that in the near future more initiatives such as these mentioned will arise. Analysis of genetic variability un-

der these frameworks is going to provide a more explanatory insight into its pathogenic implications and will likely be key in the future use of genomic data for personalized treatments.

CHAPTER 6

General conclusions

In this thesis, I proposed to integrate the interactome, as prior knowledge of the cell system, in the analysis of genomic data to better understand the functional implications of genetic variability in protein-coding genes in either human health and disease. This final chapter enumerates the conclusions extracted form the exposed results:

1. The diverse nature of the protein-protein interactions (PPIs) and the different biochemical basis of the detection methods requires the integration of different sources of PPIs through a strict curation pipeline to achieve a model of the human interactome comprehensive and accurate.

2. Proteins involved in common biological processes tend to form interactome modules, being the aggregation in common network components a distinctive property between functionally related and unrelated protein sets. This property can be explored in network enrichment analysis of high-throughput data to identify the subnetwork relevant for the condition under study. The proposed tool, NetworkMiner, makes use of this property and represents a simple but powerful approach able to find the PPI subnetwork component associated to extreme values of a list of genes ranked by any experimental criterion.

3. Proteins affected by deleterious variants in healthy individuals are concentrated in peripheral modules (carrying functions related to the cell periphery), avoiding internal modules (dedicated to essential functions), and are found in combinations that preserve properties related to the interactome integrity. From an evolutionary perspective, proteins under positive selection tend to be placed at the periphery of the interactome, whereas proteins under negative selection tend to have a central location in the interactome. Taken together, the actual interactome structure may have a role in maintaining deleterious variability in human populations by allocating

and, therefore, allowing genetic variation while sustaining healthy phenotypes and conferring evolutionary robustness to genetic perturbations.

4. Deleterious variants found in healthy individuals concentrate in proteins which display opposite topological properties to the proteins affected by common human disease, being this difference larger when considering cancer driver gene products. Thus, the deleterious character of a variant seems not only to rely on the damage it causes to the protein, but ultimately, it is a system's property that depends on its impact on the overall system organization.

5. Several global trends confirm the role of protein binding sites in cancer: somatic variants are more frequent in binding interfaces than in other ordered regions and tend to affect highly connected and central proteins. A significant number of well-known cancer driver genes concentrate its mutations at the interacting interfaces, suggesting that the altercation of these interfaces can be a mechanism of action for these mutations. As shown through an example, mutation location within the same protein can influence patient outcome such as survival. Thus, perturbation caused by cancer mutations in protein interactions may have a crucial role in the course of the disease and be an important factor in explaining the heterogeneity between cancer patients.

6. Globally, the thesis offers novel approaches for interpreting human genomes by zooming into the consequences of genomic variants and contextualizing them within a model of the cell.

Publications

- Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Aleman, A., Rueda, A., Vela-Boza, A., López-Domingo, J., Florido, J., Arce, P., Ruiz-Ferrer, M., Méndez-Vidal, C., Arnold, T., Spleiss, O., Alvarez-Tejado, M., Navarro, A., Bhattacharya, S., Borrego, S., Santoyo-López, J., and Antiñolo, G. (2015). 267 spanish exome sequences reveal specific distribution patterns of disease-related variants and highlight the importance of local variant frequencies. *Molecular biology and evolution*, (Under Revision)

- Porta-Pardo*, E., Garcia-Alonso*, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLOS computational biology*, (Accepted for publication)

- Garcia-Alonso, L. and Dopazo, J. (2015). Mutational oncogenic signatures on structurally resolved protein interacting interfaces. *bioRxiv*, page 016204

- Minguez, P., Letunic, I., Parca, L., Garcia-Alonso, L., Dopazo, J., Huerta-Cepas, J., and Bork, P. (2014). Ptmcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic acids research*, page gku1081

- González-del Pozo, M., Méndez-Vidal, C., Santoyo-Lopez, J., Vela-Boza, A., Bravo-Gil, N., Rueda, A., Garcia-Alonso, L., Vázquez-Marouschek, C., Dopazo, J., Borrego, S., et al. (2014). Deciphering intrafamilial phenotypic variability by exome sequencing in a bardet–biedl family. *Molecular genetics & genomic medicine*, 2(2):124–133

- Garcia-Alonso, L., Jiménez-Almazán, J., Carbonell-Caballero, J., Vela-Boza, A., Santoyo-López, J., Antiñolo, G., and Dopazo, J. (2014). The role of the interactome in the maintenance of deleterious variability in human populations. *Molecular systems biology*,

10(9)
This article has been highlighted in Nature Reviews Genetics (Burgess, 2014).

- Fernández, R. M., Bleda, M., Luzón-Toro, B., Garcia-Alonso, L., Arnold, S., Sribudiani, Y., Besmond, C., Lantieri, F., Doan, B., Ceccherini, I., et al. (2013). Pathways systematically associated to hirschsprung's disease. *Orphanet journal of rare diseases*, 8(1):187

- Garcia-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158–e158

- Bleda, M., Tarraga, J., de Maria, A., Salavert, F., Garcia-Alonso, L., Celma, M., Martin, A., Dopazo, J., and Medina, I. (2012). Cellbase, a comprehensive collection of restful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic acids research*, page gks575

- Fernández, R. M., Bleda, M., Núñez-Torres, R., Medina, I., Luzón-Toro, B., Garcia-Alonso, L., Torroglosa, A., Marbà, M., Enguix-Riego, M. V., Montaner, D., et al. (2012). Four new loci associations discovered by pathway-based and network analyses of the genome-wide variability profile of hirschsprung's disease. *Orphanet J Rare Dis*, 7:103

Bibliography

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., et al. (1991). Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656. 14

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249. 63, 146

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., et al. (2006). Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544. 7

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580. 68, 117

Alamo, P., Gallardo, A., Di Nicolantonio, F., Pavón, M. A., Casanova, I., Trias, M., Mangues, M. A., Lopez-Pousa, A., Villaverde, A., Vázquez, E., et al. (2014). Higher metastatic efficiency of kras g12v than kras g13d in a colorectal cancer model. *The FASEB Journal*, pages fj–14. 22

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382. 10, 58, 82, 139

Altelaar, A. M., Munoz, J., and Heck, A. J. (2013). Next-generation

proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48. 145

Andreyev, H., Norman, A., Cunningham, D., Oates, J., Dix, B., Iacopetta, B., Young, J., Walsh, T., Ward, R., Hawkins, N., et al. (2001). Kirsten ras mutations in patients with colorectal cancer: the 'rascal ii'study. *British journal of cancer*, 85(5):692. 22

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. *Nucleic acids research*, 38(suppl 1):D525. 31, 39

Arsenic, R., Lehmann, A., Budczies, J., Koch, I., Prinzler, J., Kleine-Tebbe, A., Schewe, C., Loibl, S., Dietel, M., and Denkert, C. (2014). Analysis of pik3ca mutations in breast cancer subtypes. *Applied Immunohistochemistry & Molecular Morphology*, 22(1):50–56. 129, 143

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29. 35

Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *Journal of molecular biology*, 336(4):943–955. 5

Barabási, A. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939):412. 8

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509. 8, 9

Barbareschi, M., Buttitta, F., Felicioni, L., Cotrupi, S., Barassi, F., Del Grammastro, M., Ferro, A., Dalla Palma, P., Galligioni, E., and Marchetti, A. (2007). Different prognostic roles of mutations in the

helical and kinase domains of the pik3ca gene in breast carcinomas. *Clinical Cancer Research*, 13(20):6064–6069. 129

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59. 15

Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16):2833–2842. 6

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1978). The protein data bank: a computer-based archival file for macromolecular structures. *Archives of biochemistry and biophysics*, 185(2):584–591. 64, 104

Betts, M. J., Lu, Q., Jiang, Y., Drusko, A., Wichmann, O., Utz, M., Valtierra-Gutiérrez, I. A., Schlesner, M., Jaeger, N., Jones, D. T., et al. (2014). Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic acids research*, page gku1094. 146

Bleda, M., Tarraga, J., de Maria, A., Salavert, F., Garcia-Alonso, L., Celma, M., Martin, A., Dopazo, J., and Medina, I. (2012). Cellbase, a comprehensive collection of restful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic acids research*, page gks575.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Current opinion in structural biology*, 14(3):292–299. 41

Brenner, S. (2000). The end of the beginning. *Science*, 287(5461):2173–2174. 135

Brenner, S. (2010). Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):207–212. 16

Bromberg, Y. and Rost, B. (2007). Snap: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research*, 35(11):3823–3835. 64, 146

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., and Keith Dunker, A. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of molecular evolution*, 55(1):104–110. 110

Burgess, D. J. (2014). Human genetics: Dissecting deleterious mutation complexity. *Nature Reviews Genetics*. 156

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678. 38, 52

Bustin, S. A. (1998). Crosstalk among cancer signalling pathways. *Molecular Medicine Today*, 4(12):511. 122

Calleja, V., Laguerre, M., Parker, P. J., and Larijani, B. (2009). Role of a novel ph-kinase domain interface in pkb/akt regulation: structural mechanism for allosteric inhibition. *PLoS biology*, 7(1):e1000017. 117

Carbonell, J., Alloza, E., Arce, P., Borrego, S., Santoyo, J., Ruiz-Ferrer, M., Medina, I., Jiménez-Almazán, J., Méndez-Vidal, C., González-del Pozo, M., et al. (2012). A map of human microrna variation uncovers unexpectedly high levels of variability. *Genome Med*, 4(8):62–62. 16

Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.-C., Gbeha, E., Hamdan, F. F., Girard, S., et al.

(2013). Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS genetics*, 9(9):e1003815. 70

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic acids research*, 38(suppl 1):D532. 31, 39

Choudhury, K., McQuillin, A., Puri, V., Pimm, J., Datta, S., Thirumalai, S., Krasucki, R., Lawrence, J., Bass, N. J., Quested, D., et al. (2007). A genetic association study of chromosome 11q22-24 in two different samples implicates the¡ i¿ fxyd6¡/i¿ gene, encoding phosphohippolin, in susceptibility to schizophrenia. *The American Journal of Human Genetics*, 80(4):664–672. 52

Chuang, P.-T. and McMahon, A. P. (1999). Vertebrate hedgehog signalling modulated by induction of a hedgehog-binding protein. *Nature*, 397(6720):617–621. 120

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219. 62

Consortium, . G. P. et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073. 15, 16, 69

Consortium, T. U. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219. 12, 31, 64

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., et al.

(2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications*, 1:131. 70

Das, J. and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1):92. 136

De Backer, P., De Waele, D., and Van Speybroeck, L. (2010). Ins and outs of systems biology vis-à-vis molecular biology: continuation or clear cut? *Acta biotheoretica*, 58(1):15–49. 4

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498. 62

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271. 37

Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Aleman, A., Rueda, A., Vela-Boza, A., López-Domingo, J., Florido, J., Arce, P., Ruiz-Ferrer, M., Méndez-Vidal, C., Arnold, T., Spleiss, O., Alvarez-Tejado, M., Navarro, A., Bhattacharya, S., Borrego, S., Santoyo-López, J., and Antiñolo, G. (2015). 267 spanish exome sequences reveal specific distribution patterns of disease-related variants and highlight the importance of local variant frequencies. *Molecular biology and evolution*, (Under Revision).

D'Antonio, M. and Ciccarelli, F. D. (2013). Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol*, 14(5):R52. 21

Espinosa, O., Mitsopoulos, K., Hakas, J., Pearl, F., and Zvelebil, M. (2014). Deriving a mutation index of carcinogenicity using protein structure and protein interfaces. *PloS one*, 9(1). 21

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences*, 105(11):4323–4328. 11, 81, 99, 140

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European journal of cancer*, 49(6):1374–1403. 129

Fernández, R. M., Bleda, M., Luzón-Toro, B., Garcia-Alonso, L., Arnold, S., Sribudiani, Y., Besmond, C., Lantieri, F., Doan, B., Ceccherini, I., et al. (2013). Pathways systematically associated to hirschsprung's disease. *Orphanet journal of rare diseases*, 8(1):187.

Fernández, R. M., Bleda, M., Núñez-Torres, R., Medina, I., Luzón-Toro, B., Garcia-Alonso, L., Torroglosa, A., Marbà, M., Enguix-Riego, M. V., Montaner, D., et al. (2012). Four new loci associations discovered by pathway-based and network analyses of the genome-wide variability profile of hirschsprung's disease. *Orphanet J Rare Dis*, 7:103.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2013). Pfam: the protein families database. *Nucleic acids research*, page gkt1223. 104

Fishel, R., Lescoe, M. K., Rao, M., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog msh2 and its association with hereditary nonpolyposis colon cancer. *Cell*, 75(5):1027–1038. 20

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752. 81

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., et al.

(2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220. 15, 16, 17, 69

Garcia-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158–e158. 49

Garcia-Alonso, L., Jiménez-Almazán, J., Carbonell-Caballero, J., Vela-Boza, A., Santoyo-López, J., Antiñolo, G., and Dopazo, J. (2014). The role of the interactome in the maintenance of deleterious variability in human populations. *Molecular systems biology*, 10(9). 72, 74, 75, 77, 79, 80, 83, 91, 93, 94, 95

Geng, C., He, B., Xu, L., Barbieri, C. E., Eedunuri, V. K., Chew, S. A., Zimmermann, M., Bond, R., Shou, J., Li, C., et al. (2013). Prostate cancer-associated mutations in speckle-type poz protein (spop) regulate steroid receptor coactivator 3 protein turnover. *Proceedings of the National Academy of Sciences*, 110(17):6997–7002. 118

Gerber, A., Wilson, C., Li, Y., and Chuang, P. (2006). The hedgehog regulated oncogenes gli1 and gli2 block myoblast differentiation by inhibiting myod-mediated transcriptional activation. *Oncogene*, 26(8):1122–1136. 125

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892. 20

Giot, L., Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y., Ooi, C., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1726. 9

Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabási, A. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690. 11, 81, 99, 140

Goh, K., Oh, E., Jeong, H., Kahng, B., and Kim, D. (2002). Classification of scale-free networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12583. 8

González-del Pozo, M., Méndez-Vidal, C., Santoyo-Lopez, J., Vela-Boza, A., Bravo-Gil, N., Rueda, A., Garcia-Alonso, L., Vázquez-Marouschek, C., Dopazo, J., Borrego, S., et al. (2014). Deciphering intrafamilial phenotypic variability by exome sequencing in a bardet–biedl family. *Molecular genetics & genomic medicine*, 2(2):124–133.

Gonzalez-Perez, A. and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic acids research*, page gks743. 21, 103, 120, 142

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). Intogen-mutations identifies cancer drivers across tumor types. *Nature methods*. 103

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158. 19, 20, 21

Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., and Stelzl, U. (2015). Phospho-tyrosine dependent protein–protein interaction network. *Molecular systems biology*, 11(3). 6, 145

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–806. 10, 139

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., et al. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93. 7, 29, 35

Hanahan, D. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70. 19, 117

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674. 19, 117

Hao, Y., Zhao, S., and Wang, Z. (2014). Targeting the protein–protein interaction between irs1 and mutant p110α for cancer therapy. *Toxicologic pathology*, 42(1):140–147. 130

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., Von Mering, C., et al. (2004). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nature biotechnology*, 22(2):177–183. 31

Hicks, S., Wheeler, D. A., Plon, S. E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human mutation*, 32(6):661–668. 73

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367. 14

Hoffmann, F. M. (1991). Drosophila abl and genetic redundancy in signal transduction. *Trends in genetics*, 7(11):351–355. 18

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125. 145

Hudson, T. J. and Jennings, J. L. (2011). International cancer genome consortium (icgc). *Cancer Research*, 71(8 Supplement):3935. 15, 19

Hunkapiller, T., Kaiser, R., Koop, B., and Hood, L. (1991). Large-scale and automated dna sequence determination. *Science*, 254(5028):59–67. 14

Ideker, T. and Lauffenburger, D. (2003). Building with a scaffold: emerging strategies for high-to low-level cellular modeling. *TRENDS in Biotechnology*, 21(6):255–262. 8, 11, 12, 29, 145

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233. 8, 29, 137

Ihle, N. T., Byers, L. A., Kim, E. S., Saintigny, P., Lee, J. J., Blumenschein, G. R., Tsao, A., Liu, S., Larsen, J. E., Wang, J., et al. (2012). Effect of kras oncogene substitutions on protein behavior: implications for signaling and clinical outcome. *Journal of the National Cancer Institute*, 104(3):228–239. 146

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574. 6, 41

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000). Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences*, 97(3):1143–1147. 6

Janin, J. (1995). Elusive affinities. *Proteins: Structure, Function, and Bioinformatics*, 21(1):30–39. 11

Javelaud, D., Alexaki, V.-I., Pierrat, M.-J., Hoek, K. S., Dennler, S., Van Kempen, L., Bertolotto, C., Ballotti, R., Saule, S., Delmas, V., et al. (2011). Gli2 and m-mitf transcription factors control exclusive gene expression programs and inversely regulate invasion in human melanoma cells. *Pigment cell & melanoma research*, 24(5):932–943. 125

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42. 8, 10

Johnson, M. and Hummer, G. (2011). Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proceedings of the National Academy of Sciences*, 108(2):603–608. 45

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20. 11

Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297. 11, 81, 99, 140

Kalinsky, K., Jacks, L. M., Heguy, A., Patil, S., Drobnjak, M., Bhanot, U. K., Hedvat, C. V., Traina, T. A., Solit, D., Gerald, W., et al. (2009). Pik3ca mutation associates with improved outcome in breast cancer. *Clinical Cancer Research*, 15(16):5049–5059. 129

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the raptorx web server. *Nature protocols*, 7(8):1511–1522. 64

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339. 101, 108

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, page gkr988. 35

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *science*, 336(6082):740–743. 17, 70

Kim, P. M., Korbel, J. O., and Gerstein, M. B. (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences*, 104(51):20274–20279. 10, 81, 140

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958. 7

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081. 63, 73, 145

Lage, K. (2014). Protein-protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. 6

Lai, Y.-L., Mau, B.-L., Cheng, W.-H., Chen, H.-M., Chiu, H.-H., and Tzen, C.-Y. (2008). Pik3ca exon 20 mutation is independently associated with a poor prognosis in breast cancer patients. *Annals of surgical oncology*, 15(4):1064–1069. 129, 143

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 14

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T.,

Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511. 16

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218. 20

Levy, S. F. and Siegal, M. L. (2008). Network hubs buffer environmental variation in saccharomyces cerevisiae. *PLoS biology*, 6(11):e264. 10, 140

Li, C., Ao, J., Fu, J., Lee, D.-F., Xu, J., Lonard, D., and O'Malley, B. W. (2011). Tumor-suppressor role for the spop ubiquitin ligase in signal-dependent proteolysis of the oncogenic co-activator src-3/aib1. *Oncogene*, 30(42):4350–4364. 118

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760. 62

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079. 62

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., et al. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics*, 42(11):969–972. 70

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., et al. (2006). A protein–protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125(4):801–814. 7

Liu, L., De, S., and Michor, F. (2013). Dna replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature communications*, 4:1502. 105, 114

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., et al. (2008). Proportionally more deleterious genetic variation in european than in african populations. *Nature*, 451(7181):994–997. 71

Lothe, R. A., Peltomäki, P., Meling, G. I., Aaltonen, L. A., Nyström-Lahti, M., Pylkkänen, L., Heimdal, K., Andersen, T. I., Møller, P., Rognum, T. O., et al. (1993). Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. *Cancer Research*, 53(24):5849–5852. 20

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010). A global map of human gene expression. *Nature biotechnology*, 28(4):322–324. 65, 66

Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312. 32

MacArthur, D., Manolio, T., Dimmock, D., Rehm, H., Shendure, J., Abecasis, G., Adams, D., Altman, R., Antonarakis, S., Ashley, E., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476. 63

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828. 15, 16, 17, 18, 57, 69

MacArthur, D. G. and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Human molecular genetics*, 19(R2):R125–R130. 16, 17

Mangone, F. R., Bobrovnitchaia, I. G., Salaorni, S., Manuli, E., and Nagai, M. A. (2012). Pik3ca exon 20 mutations are associated with poor prognosis in breast cancer patients. *Clinics*, 67(11):1285–1290. 129, 143

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753. 14

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380. 15

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al. (2011). The functional spectrum of low-frequency coding variation. *Genome biology*, 12(9):R84. 70

Martínez-Camblor, P., Uña-Álvarez, J. D., and Corral, N. (2008). k-sample test based on the common area of kernel density estimators. *Journal of Statistical Planning and Inference*, 138(12):4006 – 4020. 35

Masel, J. and Siegal, M. L. (2009). Robustness: mechanisms and consequences. *Trends in Genetics*, 25(9):395–403. 57

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303. 62

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068. 15, 19, 130

Medina, I., Carbonell, J., Pulido, L., Madeira, S., Goetz, S., Conesa, A., Tárraga, J., Pascual-Montano, A., Nogales-Cadenas, R., Santoyo, J., et al. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic acids research*, 38(suppl 2):W210. 38, 68, 117

Medina, I., De Maria, A., Bleda, M., Salavert, F., Alonso, R., Gonzalez, C. Y., and Dopazo, J. (2012). Variant: Command line, web service and web interface for fast and accurate functional characterization of variants found by next-generation sequencing. *Nucleic acids research*, 40(W1):W54–W58. 62, 102

Medina, I., Montaner, D., Bonifaci, N., Pujana, M. A., Carbonell, J., Tarraga, J., Al-Shahrour, F., and Dopazo, J. (2009). Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic acids research*, 37(suppl 2):W340–W344. 38

Minguez, P. and Dopazo, J. (2010). Functional genomics and networks: new approaches in the extraction of complex gene modules. 8, 35

Minguez, P., Dopazo, J., and Falciani, F. (2011). Assessing the Biological Significance of Gene Expression Signatures and Co-Expression Modules by Studying Their Network Properties. *PloS one*, 6(3):e17474. 35

Minguez, P., Gotz, S., Montaner, D., Al-Shahrour, F., and Dopazo, J. (2009). Snow, a web-based tool for the statistical analysis of protein–

protein interaction networks. *Nucleic Acids Research*, 37(suppl 2):W109. 8, 29, 31, 106, 124

Minguez, P., Letunic, I., Parca, L., Garcia-Alonso, L., Dopazo, J., Huerta-Cepas, J., and Bork, P. (2014). Ptmcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic acids research*, page gku1081.

Mirkovic, N., Marti-Renom, M. A., Weber, B. L., Sali, A., and Monteiro, A. N. (2004). Structure-based assessment of missense mutations in human brca1 implications for breast and ovarian cancer predisposition. *Cancer research*, 64(11):3790–3797. 64

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732. 8, 29

Mosca, R., Céol, A., and Aloy, P. (2013a). Interactome3d: adding structural details to protein networks. *Nature methods*, 10(1):47–53. 12

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2013b). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, page gkt887. 104

Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. (2013c). Towards a detailed atlas of protein–protein interactions. *Current opinion in structural biology*, 23(6):929–940. 5, 6

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., and Aloy, P. (2015). dsysmap: exploring the edgetic role of disease mutations. *Nature methods*, 12(3):167–168. 13

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? *Genome research*, 11(12):1971–1973. 41

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1992). Specific enzymatic amplification of dna in vitro: the polymerase chain reaction. *Biotechnology Series*, pages 17–17. 14

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104. 70

Nooren, I. and Thornton, J. M. (2003). Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492. 5

Nothnagel, M., Herrmann, A., Wolf, A., Schreiber, S., Platzer, M., Siebert, R., Krawczak, M., and Hampe, J. (2011). Technology-specific error signatures in the 1000 genomes project data. *Human genetics*, 130(4):505–516. 16, 18, 57

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28. 19, 141

Ooi, H., Schneider, G., Chan, Y., Lim, T., Eisenhaber, B., and Eisenhaber, F. (2010). Databases of Protein-Protein Interactions and Complexes. *Methods in Molecular Biology*, 609:145–159. 31

Orchard, S. (2014). Data standardization and sharing—the work of the hupo-psi. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(1):82–87. 6

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., et al. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature biotechnology*, 25(8):894–898. 31

Oro, A. E., Higgins, K. M., Hu, Z., Bonifas, J. M., Epstein, E. H., and Scott, M. P. (1997). Basal cell carcinomas in mice overexpressing sonic hedgehog. *Science*, 276(5313):817–821. 120, 143

Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clinical genetics*, 71(1):1–11. 7

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. *Journal of medical genetics*, 43(8):691–698. 7

Pao, W., Miller, V. A., Politi, K. A., Riely, G. J., Somwar, R., Zakowski, M. F., Kris, M. G., and Varmus, H. (2005). Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the egfr kinase domain. *PLoS medicine*, 2(3):e73. 22

Parikh, C., Janakiraman, V., Wu, W.-I., Foo, C. K., Kljavin, N. M., Chaudhuri, S., Stawiski, E., Lee, B., Lin, J., Li, H., et al. (2012). Disruption of ph–kinase domain interactions leads to oncogenic activation of akt in human cancers. *Proceedings of the National Academy of Sciences*, 109(47):19368–19373. 117

Pastor-Satorras, R., Smith, E., and Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theoretical biology*, 222(2):199–210. 10

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *science*, 300(5618):445–452. 12

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243. 5, 6, 145

Plomin, R., Haworth, C. M., and Davis, O. S. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878. 52

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pages 284–293. Springer. 89

Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218. 67

Porta-Pardo, E. and Godzik, A. (2014). e-driver: a novel method to identify protein regions driving cancer. *Bioinformatics*, page btu499. 21, 121, 142

Porta-Pardo*, E., Garcia-Alonso*, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLOS computational biology*, (Accepted for publication).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575. 38

Queitsch, C., Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889):618–624. 18

Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Beà, S., Pinyol, M., Martínez-Trillos, A., et al. (2012). Exome sequencing identifies recurrent mutations of the splicing factor sf3b1 gene in chronic lymphocytic leukemia. *Nature genetics*, 44(1):47–52. 61

Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918. 18

Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome biology*, 6(5):R40. 41

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900. 73

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(5586):1551–1555. 7, 29

Reimand, J. and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology*, 9(1). 21, 22, 120, 142

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S. H., et al. (2012). Recurrent mutation of the id3 gene in burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nature genetics*, 44(12):1316–1320. 125, 143

Riera, C., Lois, S., and de la Cruz, X. (2014). Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):249–268. 145

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S., Qu, H., Shales, M., Park, H., Hayles, J., et al. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322(5900):405. 35

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226. 6, 11, 144

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123. 67, 89

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178. 7, 29

Ruch, J. M. and Kim, E. J. (2013). Hedgehog signaling pathway and cancer therapeutics: progress to date. *Drugs*, 73(7):613–623. 120

Samuel, K., Sandra, O., Luisa, M., Bruno, A., Antony, Q., Nisha, V., Gary, B., Ioannis, X., Jerome, W., David, S., et al. (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5(9):44–52. 31

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467. 14

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS one*, 7(2):e31826. 144

Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A. L., Hodson, D. J., Buras, E., et al. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418):116–120. 125, 143

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18. 15

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The foldx web server: an online force field. *Nucleic acids research*, 33(suppl 2):W382–W388. 146

Serra, F., Arbiza, L., Dopazo, J., and Dopazo, H. (2011). Natural selection on functional modules, a genome-wide analysis. *PLoS computational biology*, 7(3):e1001093. 80

Shachar, R., Ungar, L., Kupiec, M., Ruppin, E., and Sharan, R. (2008). A systems-level approach to mapping the telomere length maintenance gene circuitry. *Molecular Systems Biology*, 4(1):172. 35

Shamir, A., Shaltiel, G., Mark, S., Bersudsky, Y., Belmaker, R. H., and Agam, G. (2007). Human mip synthase splice variants in bipolar disorder. *Bipolar disorders*, 9(7):766–771. 52

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology*, 3(1). 7

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732. 15

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311. 69

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050. 63

Siva, N. (2008). 1000 genomes project. *Nature biotechnology*, 26(3):256–256. 61

Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521. 137

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E., and Birney, E. (2012). Analysis of variation at transcription factor binding sites in drosophila and humans. *Genome Biol*, 13(9):R49. 16

Stark, C., Breitkreutz, B., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, 39(suppl 1):D698. 31, 39

Stern, C. (1943). The hardy-weinberg law. *Science*, 97(2510):137–138. 70

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724. 19, 141

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550. 137

Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles. *Human molecular genetics*, 10(6):591–597. 16

Swerdlow, H., Wu, S., Harke, H., and Dovichi, N. J. (1990). Capillary gel electrophoresis for dna sequencing: laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography A*, 516(1):61–67. 14

Szathmary, E. (1993). Do deleterious mutations act synergistically? metabolic control theory provides a partial answer. *Genetics*, 133(1):127–132. 18

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013a). Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244. 103, 120

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., et al. (2013b). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3. 121

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic

modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204. 7, 145

Taylor, N. R. (2013). Small world network strategies for studying protein structures and binding. *Computational and structural biotechnology journal*, 5. 10

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69. 17, 70

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation*, 32(4):358–368. 63, 64, 145

Thusberg, J. and Vihinen, M. (2009). Pathogenic or not? and if so, then how? studying the effects of missense mutations using bioinformatics methods. *Human mutation*, 30(5):703–714. 146

Garcia-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158–e158.

Garcia-Alonso, L. and Dopazo, J. (2015). Mutational oncogenic signatures on structurally resolved protein interacting interfaces. *bioRxiv*, page 016204.

Garcia-Alonso, L., Jiménez-Almazán, J., Carbonell-Caballero, J., Vela-Boza, A., Santoyo-López, J., Antiñolo, G., and Dopazo, J. (2014). The role of the interactome in the maintenance of deleterious variability in human populations. *Molecular systems biology*, 10(9).

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18(5):343–384. 12

Venkatesan, K., Rual, J., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K., et al. (2008). An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90. 41

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351. 14

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558. 15, 20, 21, 76, 77, 108, 115, 116, 117

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403. 6, 31, 40, 136

Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–4208. 11, 81, 140

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, 150(3811):563–565. 15

Wagih, O., Reimand, J., and Bader, G. D. (2015). Mimp: predicting the impact of mutations on kinase-substrate phosphorylation. *Nature methods*. 146

Wagner, A. (2000). Robustness against mutations in genetic networks of yeast. *Nature genetics*, 24(4):355–361. 18, 57

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854. 38

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, 30(2):159–164. 12, 13, 22, 104, 108, 142

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The disopred server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139. 104

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738. 14

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120. 15, 21

Willmroth, F., Drieling, T., Lamla, U., Marcushen, M., Wark, H.-J., and Van Calker, D. (2007). Sodium-myo-inositol co-transporter (smit-1) mrna is increased in neutrophils of patients with bipolar 1 disorder and down-regulated under treatment with mood stabilizers. *The International Journal of Neuropsychopharmacology*, 10(01):63–71. 52

Wong, J. M., Ionescu, D., and Ingles, C. J. (2003). Interaction between brca2 and replication protein a is compromised by a cancer-predisposing mutation in brca2. *Oncogene*, 22(1):28–33. 99

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). Sdm—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research*, 39(suppl 2):W215–W222. 64, 146

Wu, W.-I., Voegtli, W. C., Sturgis, H. L., Dizon, F. P., Vigers, G. P., and Brandhuber, B. J. (2010). Crystal structure of human akt1 with an allosteric inhibitor reveals a new mode of kinase inhibition. *PLoS One*, 5(9):e12913. 117

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., et al. (2012). Deleterious-and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *The American Journal of Human Genetics*, 91(6):1022–1032. 15, 16, 17, 18, 57, 69

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110. 99

Zhang, L., Shi, L., Zhao, X., Wang, Y., and Yue, W. (2013). Pik3ca gene mutation associated with poor prognosis of lung adenocarcinoma. *OncoTargets and therapy*, 6:497. 129

Zheng, C. L., Wang, N. J., Chung, J., Moslehi, H., Sanborn, J. Z., Hur, J. S., Collisson, E. A., Vemula, S. S., Naujokas, A., Chiotti, K. E., et al. (2014). Transcription restores dna repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell reports*, 9(4):1228–1234. 21

Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., et al. (2009). Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, 5(1). 12, 13, 142

There is nothing more practical
than a good theory.
Paul Dirac