Computational approaches and resources to support translational research in human diseases

Janet Piñero Gonzalez

TESI DOCTORAL UPF / ANY 2105

DIRECTORS DE LA TESI

Dra. Laura I. Furlong

DEPARTAMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Acknowledgements

First and foremost, I wish to thank my supervisor, Laura I Furlong for her support, her guidance, her patience, for the scientific discussions, and the conversations, for being understanding and kind in moments where my personal situation was delicate. I am also thankful for the excellent example she has set as a scientist, woman and mother.

I am grateful to Baldo Oliva, Joaquín Dopazo and Xavier de La Cruz, for agreeing to be on my thesis board.

I want to thank my professors at the Facultad de Biología de la Universidad de La Habana, specially Georgina Espinosa, Olimpia Farnés, Alina Forrellat, Fabiola Pazos, and Joaquin Díaz Brito.

A special mention goes to Jordi Deu-Pons, for all his help setting up onexus and the DisGeNET web platform.

I want to express my gratitude to the colleagues in my group, in special to two former members: Montserrat Cases, and Solène Grosdidier. I have shared unforgettable moments with both of them.

To James Dalton, the first person who became my friend here. That, I will not forget. The fact that we are still friends, when so much water has passed under our bridges proves my theory that true friendships are forged in times of battle.

To Sophia Derdak, for the operas, book recommendations, and pieces of pie that we shared together.

To Steve Laurie for the shared coffees, the conversations, and for his always fair and pertinent comments about science and other important topics.

To my distant but present friends. The ones that managed to follow my adventures from afar, and shared good and bad times with me. First and foremost, Ana Elsa Batista (love you!), Sahily Martinez (thanks for being there these last two years, for paying attention, it is all a person needs sometimes), Delvin Nodarse (for the phone calls, and the conversations, and the laughs), Allein Plain (for demonstrating that guys also know how to care about long-distance friends), Vivian Gonzalez (for your always smart insights, for caring, we have proved that email is still a way of actual communication). A huge thanks to them, for proving that distance is just an illusion.

I cannot thank enough a group of people that have made my stay here at PRBB more than ***awesome***, and that have made me feel like I do belong somewhere (the average Cuban emigrant of my age has serious issues with belonging). We have shared birthdays, coffees, conversations, happiness, sadness, and so many other things (what happens in Vegas, stays in Vegas), that I won't even try to list them. They are Núria Queralt, Alba Gutiérrez, María Saarela, Carina Oliver, Martina Gasull, Rodny Hernandez, Àlex Bravo, Alfons Pauner, Miguel A. Sánchez, and Gabriel González.

I also want to thank other PRBB people, with whom I have spinned and body pumped, volleyballed, hiked, and dinned and drank and even sang karaoke, and talked (meaningful talks, I do love talking). Special thanks to the youngsters from the

Regulatory Genomics Group and from the Biomedical Genomics Group (years 2011-2015).

I owe sincere gratitude to Aymée Robainas Barcia and Gustavo Rodríguez for providing so many enjoyable moments for my whole family (beach, dinners, mountains, movies). For sharing those holiday celebrations that are so difficult for me.

These acknowledgements would not be complete if I did not mention Abel Gonzalez Perez for many, many… so many things. Scientifically speaking, I am thankful for all the comments, the English corrections, for always being available to discuss any science issue, or question, or opinion, for his support at the very beginning when I was taking my first steps in this brave new world. On a personal note, I thank him for taking care of Pablo during the times I was absolutely immersed in my thesis. Thanks for believing in me more than I did, for the patience with my character, for accepting me and loving me the way I am, for making me feel secure, and safe.

Last but not least, I want to thank Pablo for being patient during these holidays, and for trying to take care of me, even when that is not his job. For showing me what's life about, for teaching me the greatest thing I'll ever learn.

"No one can whistle a symphony. It takes an orchestra to play it." The symphony has been amazing, thank you all, for letting me be part of it.

Abstract

In the last two decades, biomedical research has experienced a revolution driven by technological advances that have resulted in the dramatic growth in the volume and variety of data. The fragmented nature of this process has produced bottlenecks in the analysis and extraction of knowledge from this sea of information. To overcome this hurdle, better catalogs that integrate heterogeneous data types, offer easy access to users, and at the same time, support automatic workflows, are needed. With this in mind, we have developed DisGeNET, a discovery platform that contains information on more than 17,000 genes related to over 14,000 diseases, which makes it one of the largest repositories of its kind, and therefore a valuable resource to support bioinformatics research. We have used DisGeNET to study the global, local, and mesoscale properties of disease genes in the context of protein interaction networks. To produce an accurate analysis of the mesoscale properties of the protein interaction networks, we first compared the network partitions generated by two popular clustering algorithms, in order to assess how this choice would impact the follow-up biological analysis. Using the best performing algorithm we then explored the network properties of disease genes, and found that the analysis of the properties of disease genes as a whole is not very informative, given that groups of genes associated to different disease classifications exhibit different, and sometimes, opposite behavior. Then we evaluated the relationship between the network properties of different groups of disease genes and their tolerance to likely deleterious germline variants across human populations. Finally, we have developed a new method to study disease comorbidities, by combining DisGeNET and molecular networks data in a network medicine approach and we applied this method to the analysis of COPD, one of the most prevalent respiratory diseases, and its comorbidities.

Resumen

En las últimas dos décadas, la investigación biomédica ha experimentado una revolución gracias a los avances tecnológicos, que han producido un incremento dramático en la cantidad y la diversidad de datos biomédicos disponibles. Este proceso ha ocurrido de manera fragmentada, y en consecuencia los datos se encuentran almacenados en distintos repositorios, lo cual impone barreras a la hora de integrarlos, analizarlos y extraer conocimiento a partir de ellos. Para superar estas barreras, es necesario contar con recursos computacionales que integren esta información, y ofrezcan un fácil acceso a la misma, permitiendo al mismo tiempo su análisis automatizado. En respuesta a esta necesidad hemos desarrollado DisGeNET, una plataforma orientada a la exploración de las causas genéticas de las enfermedades humanas, que contiene actualmente información sobre más de 14.000 enfermedades y 17.000 genes. En esta tesis, describimos el uso de DisGeNET en combinación con métodos de redes complejas para el estudio de las propiedades de los genes asociados a enfermedades en el contexto de redes de interacción entre proteínas. Para ello, evaluamos previamente en qué medida la utilización de distintos algoritmos de reconocimiento de comunidades en redes afecta a los resultados de los análisis e influencia su interpretación biológica. A continuación, caracterizamos las propiedades de redes de los genes asociados a enfermedades como conjunto y también en sub-

grupos, empleando diferentes criterios de clasificaciones de las enfermedades. Posteriormente, evaluamos cómo estas propiedades de redes están relacionadas con la tolerancia a mutaciones posiblemente deletéreas en grupos de genes asociados a distintas clases de enfermedades, mediante el análisis de datos generados por las nuevas tecnologías de secuenciación. Finalmente, desarrollamos una nueva metodología para explorar los mecanismos moleculares de la comorbilidad, basada en la combinación de datos obtenidos de DisGeNET, con redes de interacción de proteínas y la aplicamos al estudio de las comorbilidades de la enfermedad pulmonar obstructiva crónica.

Preface

The main objective of the work reported in this thesis is the development of computational tools and network biology approaches to contribute to the understanding of the molecular mechanisms underpinning human diseases. My work is therefore framed within efforts to bridge the gap between genotype and phenotype, one of the most pressing among the current challenges of biomedical research. Although current technological advances have fostered the production of large quantities of different types of biological data on a daily basis, we are still far from understanding how alterations of the genetic information end up producing a disease phenotype. These technological advances have expanded the catalog of available omics data, that now includes transcriptomics, proteomics, metabolomics, methylomics, lipidomics, and importantly, genomics data. Genomics data are being produced at a cost and speed that has outpaced our capacity for meaningful interpretation. We are witnessing an unprecedented capacity of description of a patient from a genetic and molecular point of view, and the challenge is to translate this knowledge into clinical actions in the treatment and, even better, prevention of diseases. In order for this to happen, we need appropriate standards to homogeneously annotate and exchange the data, integrative databases to overcome the current trend to store information in "discipline silos", computational tools and new analytic strategies and methods to organize, explore, and translate the data into knowledge.

Table of Contents

Index of figures

1. INTRODUCTION

To present the background of this work, I first review our current knowledge on the molecular genetics of human diseases (Section 1.1 of the Introduction), and I provide a historical perspective of the development of the experimental approaches that have produced that knowledge. I then introduce the systems and nomenclatures currently used to classify and annotate human diseases (Section 1.2). These two first sections introduce the third one, which reviews the computational resources that contain our knowledge about the genetic basis of diseases. The final section of the Introduction (1.4) is dedicated to the use of system biology approaches that in the past two decades have exploited the wealth of data on biological networks produced by high-throughput experimental techniques to advance the research on the molecular mechanisms of human diseases.

1.1 The genetic basis of diseases

Diseases are perturbations to the structure or function of parts of the body that cause alterations of the normal homeostatic processes. They are usually characterized by specific signs, symptoms, and biochemical and laboratory findings. Diseases usually affect specific organs and tissues, but they can also manifest at the system level. They can be caused by external agents, trauma, or genetic anomalies.

Mendel is considered the father of modern genetics and his research in the 1860's on the transmission of several traits in the garden green peas laid the grounds for our current understanding of the molecular basis of the inheritance of phenotypes[1]. The first genetic disorder described was an inborn error of metabolism, alcaptonuria, found by Sir Archibald Garrod in 1902 (Garrod 1902).

Genetic diseases are caused by alterations in the DNA sequence of the germline, or certain somatic cells of an organism. These abnormalities may involve changes in the DNA sequence of a single –or several– genes, or changes in the number or structure of the chromosomes. These changes might be inherited from the parents, but also caused by *de novo* mutations, or somatic mutations. Diseases are usually classified according to their underlying genetic architecture as monogenic diseases, oligogenic diseases, mitochondrial diseases, chromosomal abnormalities, and complex diseases (Figure 1).

1.1.1. Single gene disorders

Also called monogenic diseases, they are largely determined by the action, or lack of action, of germline[2] mutations at individual loci. The mutations maybe present on one, or in both gene alleles. These disorders are often called Mendelian diseases, because their inheritance follows a similar pattern than Mendel's traits in peas. Even when many of them are classified as rare diseases (they affect less than 1 person in 200,000 (United States definition) or 1 person in 2,000 (European Union definition), single-gene disorders as a group are responsible for a significant proportion of morbidity and mortality worldwide. More than 90% of them manifest before puberty,

[1] Phenotype: the collection of observable traits of an organism, including its morphology, its physiology at the level of the cell, the organ, and the body, and even molecular-level traits, such as gene expression profiles (Nachtomy, Shavit, and Yakhini 2007).

[2] Germline mutations: genetic alterations that are present in the cells that are destined to develop into gametes (germinal or germline alterations)

and only 1% of the cases present after the end of the reproductive period (Nussbaum, McInnes, and Willard 2007). Some examples of single gene diseases are cystic fibrosis caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene; Wilson's disease, caused by mutations in the ATPase, Cu++ transporting, beta polypeptide (ATP7B) gene; and several alterations in globin genes, which cause hemoglobinopathies such as thalassaemias, and sickle cell anemia. Even when protein-coding genes constitute less than 1.5% of the human genomes (Lander et al. 2001), the majority of Mendelian phenotypes described up-to-date are caused by genetic alterations that perturb the function, localization, or amount of proteins (Chong et al. 2015). Many of the causing mutations are exonic or splice-site mutations that change the amino acid sequence of the affected gene (Majewski et al. 2011).

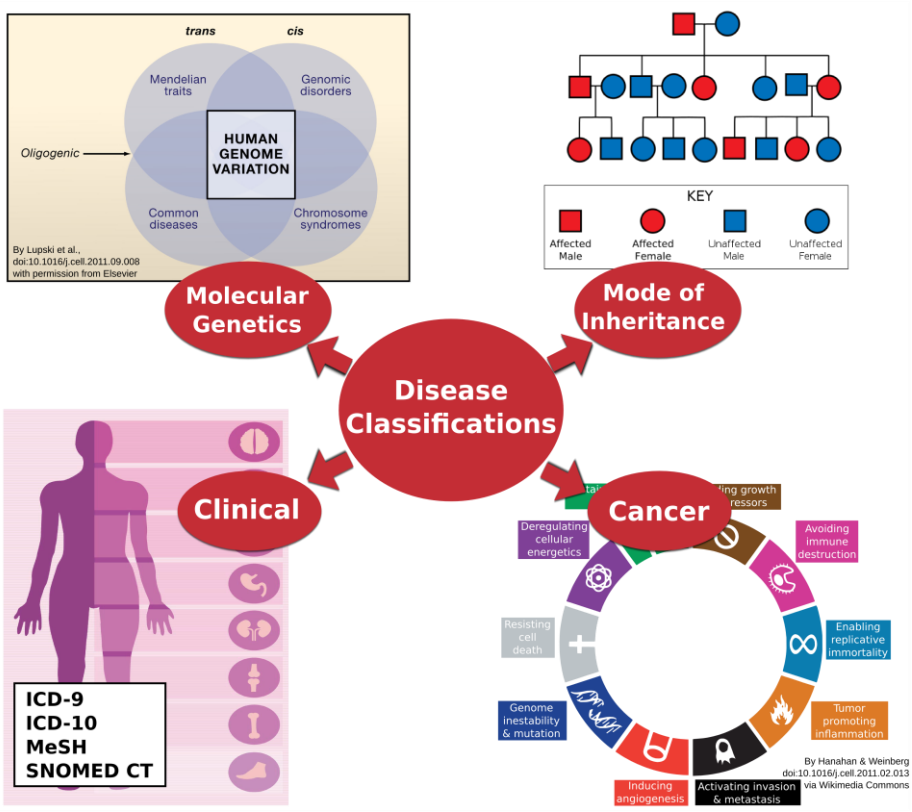

Figure 1: Classification of diseases according to different criteria. Diseases are classified according to their underlying molecular genetics, their mode of inheritance, their clinical manifestations, etc.

The pattern of inheritance of Mendelian traits depends on whether the phenotype is observed upon mutation of one (heterozygous) or both (homozygous) alleles of the causing gene. Diseases that manifest upon heterozygous changes are said to be dominant, while those that require changes in both alleles of the gene to manifest are considered recessive. If the causing gene is located in an automosome, the disease is called autosomal; if it is located within a sex chromosome it is said to be sex-linked.

Autosomal recessive diseases occur when the two alleles of a gene of an individual are mutated. Most of these mutations cause a reduction or complete loss of

the gene function. Classical examples, such as Galactosemia, Phenylketonuria, and Fructose-1 phosphate aldolase deficiency involve defects on metabolic enzymes.

Autosomal dominant diseases are caused when an alteration in only one of the two copies of the gene is sufficient to produce the disease phenotype. Examples are Huntington disease, familial hypercholesterolaemia, Marfan syndrome, Retinoblastoma, and myotonic dystrophy.

X-linked recessive diseases manifest when a genetic alteration in one gene within the X chromosome causes the phenotype to be observed in all males, or in females that are homozygous for the alteration. Examples of X-linked recessive conditions include: Haemophilia A, Duchenne muscular dystrophy, and Glucose-6-phosphate dehydrogenase deficiency.

X-linked dominant diseases require only one defective copy of the causing gene located on the X chromosome for the phenotype to manifest. Both males and females are thus affected, although in males the phenotype may be more severe because the affected gene is presented in a single dose in their only copy of the X chromosome. Some X-linked dominant disorders are lethal in males. Examples of diseases with X-linked dominant inheritance include hypophosphatemic rickets, oral-facial-digital syndrome type I, and Rett syndrome.

Y Chromosome–Linked diseases are expected to be very rare, because the chromosome Y does not contain many genes. Given that males only carry one copy of the Y chromosome, it does not make sense to talk about dominant or recessive, because only one copy of the gene is present. The existence of Y chromosome-linked traits, is currently questioned (Jobling and Tyler-Smith 2000).

It is important to bear in mind that the classification of diseases into recessive or dominant refers to the mode of inheritance of the trait, rather than qualifying an allele, or gene. Nevertheless, the underlying alleles, and genes themselves are widely referred to as dominant or recessive if they cause the phenotype in heterozygous or homozygous state, respectively.

With the advance of our knowledge about the genetic alterations underpinning diseases, has come the recognition that only in a limited number of disorders the phenotype can be satisfactorily explained by mutations at a single locus (Badano and Katsanis 2002).

1.1.2. Oligogenic disorders

Several disorders initially classified as monogenic, are actually either caused or modulated by the action of a few genes. These disorders are described as ‘oligogenic’, a term that groups a broad spectrum of phenotypes that are neither monogenic nor complex. These disorders are primarily genetic in etiology, and occur due to the synergistic action of mutant alleles at a small number of loci (Badano and Katsanis 2002). Some examples of oligogenic disorders are the digenic interaction of ROM1 and RDS that causes Retinitis Pigmentosa (Kajiwara, Berson, and Dryja 1994) and Hirschsprung disease caused by mutations genes RET and GDNF (Angrist et al. 1996). One of the most used examples is Bardet–Biedl Syndrome (BBS), a genetically heterogeneous disease that is thought to be caused by alterations in several of 19 genes (Novas et al. 2015) .

1.1.3. Mitochondrial disorders

Mitochondrial disorders are a group of clinically diverse phenotypes that arise as a consequence of a dysfunction in the mitochondrial respiratory chain. They are caused by alterations in the mitochondrial DNA (mtDNA) or in the nuclear DNA encoding mitochondrial proteins, and according to this, they may have different patterns of inheritance (autosomal-dominant or recessive, X-linked, or maternal) (Ylikallio and Suomalainen 2012). Some examples of this type of diseases are the Mitochondrial encephalopathy, lactic acidosis, stroke-like episodes (MELAS) syndrome, caused by mutations in the mitochondrial transfer RNA for leucine (UUR) gene (Goto, Nonaka, and Horai 1990), and Leber hereditary optic neuropathy (LHON), caused, in over 90% of cases, by one of three mtDNA point mutations in genes encoding subunits of the complex I of the mitochondrial respiratory chain (Man 2002)).

1.1.4. Chromosomal abnormalities

Diseases caused by chromosomal abnormalities involve changes in chromosome number (aneuploidies), or large physical changes in chromosomal structure. The first chromosomal disorder described was Down syndrome, in 1959, caused by a trisomy of chromosome 21 (Lejeune, Gautier, and Turpin 1959). Other examples of diseases involving alterations in the number of chromosomes are Klinefelter syndrome (individuals with two X chromosomes and one Y chromosome) and Turner Syndrome (individuals with only one sex chromosome, an X).

Chromosomal abnormalities involving structural genetic alterations happen by chromosomal breakage or unequal crossing-over, which may ultimately result in deletion, duplication, translocation, or inversion of chromosomal segments. Examples of this type of diseases are Prader-Willi syndrome, caused by deletion of several genes in chromosome 15 (Horsthemke and Wagstaff 2008); Charcot-Marie-Tooth disease type 1A, caused by the duplication of a region in the short arm of chromosome 17 (Raeymaekers et al. 1991); Chronic myelogenous leukemia arises as a result of a chromosomal translocation, between segments of chromosomes 9 and 22 (Nowell 2007); and severe hemophilia A is triggered in half of the patients by an inversion involving intron 22 of the factor VIII gene (Antonarakis et al. 1995).

1.1.5. Complex diseases

Shortly after Mendel and Garrod delineated the first rules of the inheritance of certain visible traits, statisticians such as Galton and Pearson were starting to realize that some human features exhibited inheritance patterns that did not follow Mendel’s first law. In 1918, Ronald Fisher solved this dilemma, when he published a paper (Fisher 1918) that reconciled the discontinuous nature of Mendelian genetics and the more continuous variations seen for other traits such as height, establishing a new paradigm for quantitative genetics. These traits, in which the observed variation is not explained by the behavior of a single gene, but instead result from the interaction among many different alleles at several loci, and which are greatly influenced by the environment, are called multifactorial, complex or polygenic traits. Most human diseases show this behavior (Robinson, Wray, and Visscher 2014). They affect large number of people with varying degrees of severity. Examples of this type of diseases are coronary artery disease, hypertension, asthma, diabetes, Alzheimer's disease, cancer, and many psychiatric disorders.

While the cause of monogenic diseases are single variants with large functional effects and high penetrance[3], with the environment playing only a minor role in the resulting phenotype, complex diseases manifest due to the interplay of many variants with small additive effects, and the influence of the environment on the final outcome is larger.

Currently, there are two alternative hypotheses to explain the genetic architecture underpinning complex diseases. The **common disease, common variant** (CD/CV) hypothesis states that complex diseases arise as consequence of the existence of disease-predisposing alleles that occur at high frequencies (variants with a frequency higher than 1% in the population), and display relatively low penetrance (Reich and Lander 2001). The second hypothesis, the **common disease, rare variant** (CD/RV), attributes complex diseases to the occurrence of rare variants (variants with a frequency below 1% in the population) with larger penetrance (Manolio et al. 2009; Schork et al. 2009).

It is generally assumed that a conceptual continuum of phenotypes exists between classical Mendelian and complex traits (Figure 2). The position of any given disorder along this continuum depends on whether a major locus contributes markedly to the phenotype, the number of loci involved and the influence of the environment on the phenotype (Badano and Katsanis 2002).

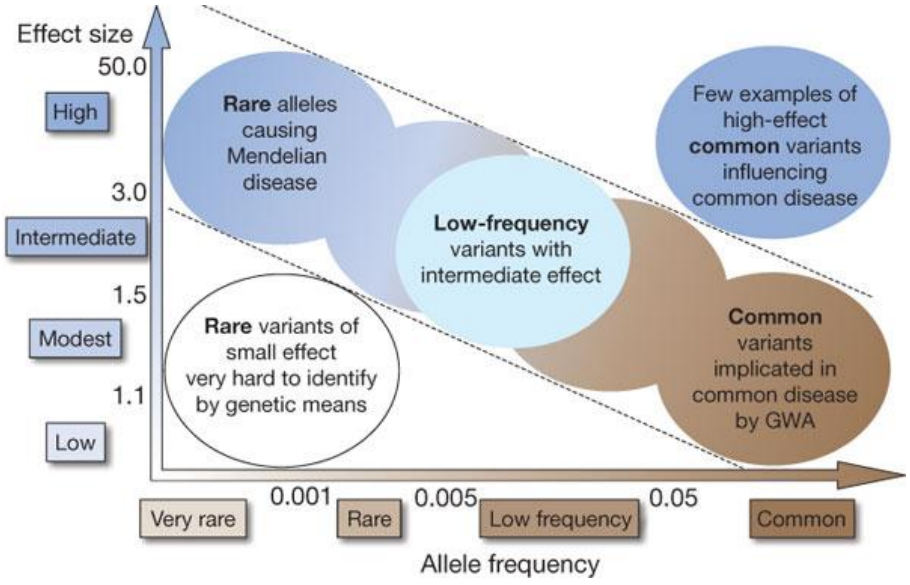

Figure 2: The continuum between classical Mendelian diseases and complex diseases. A few common variants with small effect size underlie complex diseases while rare variants implicated in Mendelian disease are highly penetrant and have larger effects. Allele frequency and effect size are generally inversely related, with common variants with large effects being rare and subject to strong purifying selection, and rare variants with small effects being difficult to detect. Taken from (Manolio et al. 2009)

[3] Penetrance: the proportion of individuals that carry a disease-causing variant and exhibit disease phenotype

1.1.6. Cancer

Neoplastic diseases are a collection of complex pathological entities characterized by the uncontrolled growth of cells and tissues. A separate disease classification, the ICD-O (Kleihues and Sobin 2000) has been developed for cancer, which follows the hierarchical rationale to classify them according to the organ, or site and specific tissue type, or histology, involved. While Mendelian, complex, and mitochondrial diseases are caused by germline mutations, cancer arises as a consequence of the accumulation of mutations occurring in somatic cells (somatic mutations). The landscape is further complicated by the fact that some germline mutations confer susceptibility to certain types of cancer, the most common examples of which are BRCA1 and BRCA2, and tumor suppressors RB1 and TP53.

Cancer has been often described as a disease of a limited number of genes (commonly referred to as drivers) or pathways (called the cancer hallmarks) (Hanahan and Weinberg 2000, 2011; Stratton, Campbell, and Futreal 2009; Vogelstein et al. 2013). After several decades of identifying the genes underlying tumorigenesis on the basis of individual biochemical or genetic assays, the advent of Next Generation Sequencing made possible the high throughput interrogation of thousands of tumor samples taken from cohorts of patients suffering from different malignancies. Several large initiatives, such as the International Cancer Genome Consortium (Hudson et al. 2010) and The Cancer Genome Atlas (Weinstein et al. 2013) were assembled to carry out this process across dozens of tumor types. As a result of the impulse received by cancer genomics in the past decade, a set of a few hundreds of driver genes (Futreal et al. 2004; Kandoth et al. 2013; Lawrence et al. 2014; Rubio-Perez et al. 2015; Tamborero et al. 2013; Vogelstein et al. 2013) have been identified as causative of several dozens of the most common malignancies. This process has kicked off in recent years the development of new therapeutic agents targeting the key driver genes in different malignancies, such as the inhibitors Erlotinib, Gefitinib, and Lapatinib for EGFR, Vemurafenib for BRAF, and Bosutinib, Dasatinib, Imatinib for ABL. Due to the better understanding of the molecular mechanisms underlying cellular malignancies, and the relatively small number of important actionable alterations in a cancer genome, this is the area where personalized medicine has made the greatest advancement in recent years (Rubio-Perez et al. 2015).

1.1.7. Epigenetics and disease

Diseases can also arise as a consequence of alterations in the epigenetic cellular landscape. Epigenetics is defined as inherited alterations of gene expression or function that are not explained by variations in the DNA sequence. These alterations are due to 'chromatin marks' on top of the DNA sequence that change the structure of chromatin or interfere with the transcriptional/translational machinery, thus changing the expression of neighboring genes. Chromatin marks entail chemical modifications to both nucleotides –such as cytosine methylation–and nucleosomes –such as methylations and acetylations of histone tails. Specially, the involvement of epigenetic alterations has been recently recognized as key in the emergence of several types of cancer, neurologic conditions such as schizophrenia and bipolar disorders, and cardiovascular diseases (Rodenhiser and Mann 2006). Many other diseases have also been linked to alterations of epigenetic modifiers (reviewed in (Brookes and Shi 2014)). The study of epigenetic mechanisms and their de-regulation will help to understand key issues related to certain

diseases which have not been satisfactorily explained to date: their dependence with age, their quantitative nature, and the impact of the environment on them (Bjornsson, Fallin, and Feinberg 2004) .

1.1.8. Genetic susceptibility to infectious diseases

Hereditary factors also play a role in the inter-individual variation of the susceptibility to infectious diseases. Studies in tuberculosis, leprosy, *Helicobacter pylori* infection, and chronic hepatitis B infection have shown greater concordance of the incidence of these diseases in monozygotic compared with dizygotic twin pairs, which provides a measure of heritability (Hill 2012). Variants in the interferon gamma receptor gene (IFNGR1), for example, have been linked to the susceptibility to mycobacterial infections (Newport et al. 1996), and variants in parkin RBR E3 ubiquitin protein ligase (PARK2) and PARK2 co-regulated (PACRG) are recognized as genetic risk factors for leprosy (Mira et al. 2004). There are also cases of protective alleles, as the sickle cell allele of hemoglobin which reduces the risk of malaria (Allison 1954), and the chemokine (C-C motif) receptor 5 (CCR5) deletion allele and HIV infection (Dean et al. 1996). Several studies document the relationships between the Toll-like receptors (TLRs) and the susceptibility to pathogens such as *Meningococcus*, *Legionella*, *Borrelia*, and *Mycobacteria* (Misch and Hawn 2008). Finally, the Human Leucocyte Antigens (HLAs) possess a number of variants that have been shown to both protect or confer susceptibility to several infectious diseases such as Chronic Hepatitis B, Hepatitis C, H1V/AIDS, Leprosy, Tuberculosis, and Leishmaniasis (reviewed in (Blackwell, Jamieson, and Burgner 2009)).

1.1.9. Identifying the genetic determinants of human diseases: an historical view

Detecting the DNA alterations responsible for specific traits in humans remains particularly challenging. In 1949, sickle cell anemia became the first genetic disease to have a known molecular basis (Pauling et al. 1949). Later, it was found that the biochemical differences between healthy and sickle hemoglobin were caused by a single amino acid change (Ingram 1956, 1957). It would take two decades for the identification of the genes causing Mendelian diseases to become routine. Before the 1980's, the identification of genes related with diseases was done by direct analysis of the candidate gene. In the 1980's, with the advent of genome-wide linkage analysis using anonymous DNA polymorphisms, the basis for the construction of a genetic linkage map of the human genome were laid (Botstein et al. 1980). The earliest genetic linkage maps were based on restriction-fragment length polymorphism (RFLP) markers, but very soon, abundant highly polymorphic microsatellite loci were employed (Litt and Luty 1989; Weber and May 1989). During the 80's and 90's, around 1,200 genes were associated to human disorders and traits, using positional cloning (Botstein and Risch 2003). The results of these studies with genes associated to Mendelian diseases produced several important lessons. First, many disease genes were unexpected, according to inferences from prior biological knowledge. Second, genetic alterations causing single gene diseases occur mainly in protein-coding regions. Third, disease alleles are, in most instances, rare in the population. Finally, there are several phenomena that complicate the interpretation of this type of analysis: locus

heterogeneity, incomplete penetrance of variants, and variable expressivity (Altshuler, Daly, and Lander 2008).

Applying the same approaches to complex diseases has not proven to be as successful as in the case of Mendelian diseases. Although genes that explain some rare forms, or Mendelian subtypes of common diseases such as diabetes, breast cancer, colon cancer, hypertension, or Alzheimer's disease have been found, causal variants found in those genes are not able to explain the existence of the disease in the majority of the patients (Altshuler et al. 2008; Risch 2000). The complex disease riddle has then started to be approached using genetic association studies. In this new approach, rather than mapping the transmission of a trait across pedigrees in families, the frequencies of genetic variants (single nucleotide polymorphisms, SNPs) are compared between affected and unaffected individuals. When there is a higher frequency in affected individuals compared to controls, this is interpreted as the allele or genotype being associated with increased risk of disease, or being a marker for other nearby SNP that is the causal variant (Altshuler et al. 2008). In 1996, (Risch and Merikangas 1996) demonstrated that linkage studies are well-powered to detect variants with large effects and high penetrance, but underpowered for detection of variants of small effect. They hypothesized that small genetic effects produced by common variants could be detected with greater power by a population-based alternative mapping approach. This is how the CD/CV hypothesis was born (Collins 1997; Lander 1996; Risch and Merikangas 1996). Under the CD/CV hypothesis, if many different common SNPs have small effect size[4] on each disease, some will be found by testing enough SNPs in enough people. It took ten years to develop large enough SNPs catalogs, techniques to analyze the variants in large cohorts, and the statistical framework to pinpoint the right variants (Altshuler et al. 2008). These advances laid the grounds for Genome Wide Association Studies (GWAS). GWAS analysis nowadays performs tests for the association of SNPs (up to 10 million) with diseases or traits in hundreds to tens of thousands of individuals (Stranger, Stahl, and Raj 2011).

To date, hundreds of GWAS have been performed, identifying hundreds of thousands of disease-associated variants and there are several repositories devoted to gather this type of results (See chapter 1.3.2). Most of the variants identified by these studies lie within non-coding regions, and concentrate in regulatory DNA, where they might alter transcription factor bind sites, chromatin states, etc. (Maurano et al. 2012). Although GWAS approaches have produced thousands of loci that are significantly and robustly associated with one or more complex trait (Visscher et al. 2012; Zuk et al. 2012), the associated SNPs usually explain a small proportion of the genetic variation in the population, which has been dubbed as the problem of "missing heritability"[5] (Manolio et al. 2009). The CD/RV hypothesis could explain the failure of the GWAS approach: because of their low frequencies, they are very poorly assessed with available GWAS arrays.

[4] Effect size: the increase in risk (or proportion of population variation) that is conferred by a given causal variant.

[5] Heritability of a trait: the proportion of total phenotypic variation that attributable to additive genetic factors

The advances in next generation sequencing (NGS) technologies are making large-scale clinical genomics a reality. Several large consortia across the world are sequencing exomes or whole genomes of hundreds and thousands of patients, searching for disease-causing genetic mutations. NGS are useful in the study of complex diseases, because given their unbiased nature and deep coverage, common and rare variants may be identified, in coding and non-coding genome regions. In cancer, NGS pipelines of normal and tumor tissues of a patient allow the identification of driver mutations, which could in turn, be used for choosing the therapy.

1.2 Clinical classification of diseases

In the elucidation of the relationships between the molecular origins of diseases and the resulting phenotype, it is important to employ a unified disease nomenclature. The need for a common medical terminology, nevertheless, is not new. Nosology is the branch of medicine that deals with the classification of diseases. The oldest medical taxonomies can be traced back to the ancient Greeks. Hippocrates classified diseases according to their signs and symptoms and to their anatomical location into four humors: black bile, yellow bile, phlegm, and blood (Balint, Buchanan, and Dequeker 2006; Hyman). In the XVII century, London health authorities started to classify diseases in order to describe, and keep records of the causes of death. In the following century, Carl von Linné, who developed the taxonomic system that is still used to classify living organisms, published *Genera Morborum*, which is one of the first recorded attempts to perform a scientific classification of disease (Linné 1763). Diseases were classified according to their symptomatology in 11 classes, 37 orders and 325 species. With the accumulation of medical knowledge, this system based on symptoms became less useful. This is why, in 1839, William Farr developed a new classification system, based on the site of the body where the diseases manifested, which was improved by Jacques Bertillon in 1899, becoming the predecessor of the International Classifications of Diseases (Moriyama et al. 2010).

A large number of vocabularies, terminologies, and ontologies are currently in use for the representation of our knowledge about diseases (Box 1). They are usually based on signs and symptoms, and on the organ or organ system that the disease affects. Many of them include not only diseases, but also signs, syndromes, symptoms, traits, disease manifestations, and other phenotypes that constitute deviation from the healthy status. Some of them also include other health-related concepts, besides diseases, and signs and symptoms, such as anatomy, drugs, procedures, etc. The following chapter reviews the most commonly employed disease classification systems.

The **International Statistical Classification of Diseases and Related Health Problems** (ICD, http://apps.who.int/classifications/icd10/browse/2015/en) developed and maintained by the World Health Organization (WHO), is one of the most widely used tools for categorizing diseases, disorders, injuries and other health-related problems (Organization 2004). Its first version was developed by Jacques Bertillon in 1893, and it was adopted by the WHO in 1948 (Moriyama et al. 2010). The ICD is currently used by WHO member states for reporting and comparing morbidity and mortality. Additionally, it enables storage and retrieval of diagnostic information for epidemiological purposes. The ICD current version is the 10th revision (ICD-10) and it contains codes for diseases, signs and symptoms, abnormal findings, complaints, social

circumstances, and external causes of injury or diseases, divided in XXII chapters (version 2015).

Box 1

Concept: unit of thought

Term: linguistic label use to designate a particular concept.

Codes: letters, numerals or a combination thereof, can be used to designate concepts in a computerized system.

Terminology: list of terms referring to concepts in a particular domain.

Classification: arrangement of based on their essential characteristics into groups of concepts, called classes.

Thesaurus: terminology in which terms are ordered, e. g., alphabetically or systematically and in which concepts can possibly be described by more than one (synonymous) term.

Vocabulary: system of terms plus explanations of the meanings. When a concept in a terminology or thesaurus is accompanied by a definition, it is called a vocabulary or glossary

Controlled vocabulary: closed list of named concepts, which can be used for classification. The constituents of a controlled vocabulary are terms. The purpose of controlling vocabulary is to avoid authors defining meaningless terms, terms which are too broad, or terms which are too narrow, and to prevent different authors from misspelling and choosing slightly different forms of the same term.

Coding system: A terminology, thesaurus, vocabulary, nomenclature or classification is called a coding system when the system uses codes for designating concepts.

Ontology: A formal representation of the concepts and relations in a given domain.

Taxonomy or hierarchy: the simplest kind of ontology in which concepts are arranged according to only one relation: "is a kind of" (is_a).

The **Medical Subject Headings** (MeSH, http://www.nlm.nih.gov/mesh/MBrowser.html) is a vocabulary created by the U.S. National Library of Medicine in the 1960's (Lipscomb 2000; Rogers 1963) to catalogue, index and retrieve biomedical literature. It consists of a set of terms organized in a hierarchical structure with 16 top-level categories. These categories include Anatomy [A], Organisms [B], Diseases [C], Chemicals and Drugs [D], Psychiatry and Psychology [F], and other concepts. The MeSH tree structure is polyhierarchic, which means that a heading can appear in more than one category: for instance, "Breast Neoplams" is a child of "Neoplasms" and "Skin and Connective Tissue Diseases".

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT, http://www.ihtsdo.org/snomed-ct/) is a comprehensive clinical vocabulary, developed by the College of American Pathologists (CAP) in 2002 (Stearns et al. 2001; Wang, Sable, and Spackman 2002), and since 2007 maintained, and distributed by the International Health Terminology Standards Development Organization (IHTSDO). SNOMED CT includes concept codes, definitions, synonymous, and relationships between concepts. It was designed to capture clinical information in electronic health records. SNOMED CT concepts are classified in 19 independent hierarchies. One of them is "Clinical Finding", that is the parent of "Disease". The concept of "Disease"contains 85 child concepts (version July, 2105). SNOMED CT concepts may have multiple parents, for example,

"Neoplasm of breast" is a child of "Breast lump (finding)", "Disorder of breast (disorder)" and "Neoplasm of thorax (disorder)".

The **National Cancer Institute** (NCI) thesaurus (https://ncit.nci.nih.gov/ncitbrowser/) is an ontology containing the reference terminology for cancer, and cancer-related information. The NCI thesaurus was created by the NCI as part of the Enterprise Vocabulary Services (EVS) Project, that aimed at integrating and unifying cancer-related concepts in areas such as cancer types, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and model organisms (Sioutos et al. 2007). It has 20 logically distinct "kinds", similar to disjoint classes, that include "Disease, Disorder or Finding", "Drug, Food, Chemical or Biomedical Material", "Anatomic Structure", "System, or Substance", etc. The "Disease, Disorder or Finding" class contains categories such as "Hyperplasia", "Rare Disorder", "Genetic Disorder", "Neoplasm", and "Childhood Disorder". The NCI thesaurus contains information, definitions and synonyms for 10,000 cancers and related diseases (release June, 2015).

The **Disease Ontology** (DO, http://disease-ontology.org/) is developed in collaboration among researchers at Northwestern University, the Center for Genetic Medicine and the University of Maryland School of Medicine, Institute for Genome Sciences. The DO is a standardized ontology for human diseases (Kibbe et al. 2015; Schriml et al. 2012). The DO integrates clinical and biomedical vocabularies and ontologies, by cross-mapping MeSH, OMIM, SNOMED-CT, ICD, and NCI thesaurus. It includes common and rare disease concepts, and it pays special attention to standards, providing stable identifiers, and organizing diseases according to etiology (Schriml and Mitraka 2015; Schriml et al. 2012). The disease concept has 8 children, classified according to etiology: "disease by infectious agent", "disease of anatomical entity", "disease of cellular proliferation", "disease of mental health", "disease of metabolism", "genetic disease", "physical disorder", and "syndrome" (update from 2015-09-03). The disease "breast cancer" is inside "diseases of cellular proliferation".

The **Human Phenotype Ontology** (HPO, http://www.human-phenotype-ontology.org/) describes human abnormal phenotypes (Köhler et al. 2014). It was originally developed by the Institute for Medical Genetics, at Le Charitéin Berlin (Robinson et al. 2008), but it is now being developed in collaboration with members of the OBO Foundry (Open Biological and Biomedical Ontologies). Currently, it contains over 11,000 terms and 115,000 annotations about hereditary diseases, organized as a directed acyclic graph, with nodes connected by "is_a" relationships. The HPO is an attempt to describe clinical phenotypes in an unambiguous way, using a controlled vocabulary that could be readily analyzed using computational algorithms, to find similarity between diseases, to ease interoperability between laboratories, and to provide a translational bridge between the genomic and phenotypic levels. This resource focuses on mapping clinical alterations described in the Online Mendelian Inheritance in Man database, although it has been recently extended to contain also annotations of common diseases, using text mining approaches (Groza et al. 2015). Phenotypic abnormality is the top level category of the HPO, while subontologies such as Mode of inheritance, Mortality/Aging and Clinical modifier, describe other features of the phenotype, such as severity, age of onset, etc. The disease "Neoplasm of the breast" is a child of the concepts "Abnormality of the breast" and "Neoplasms".

The **Unified Medical Language System** (UMLS, www.nlm.nih.gov/research/umls/), was developed by the U.S. National Library of Medicine in 1986 (Lindberg, Humphreys, and McCray 1993; Lindberg and Humphreys 1989), to facilitate the exchange of information of different types of machine readable biomedical sources, such as electronic health records, and bibliographic databases. The UMLS consists of a Metathesaurus, a semantic network and a specialist lexicon with tools. The Metathesaurus is a collection of more than a million medical concepts integrating more than 100 sources that include several well established medical ontologies such as SNOMED CT, MeSH, ICD, NCI Thesaurus, etc. The Metathesaurus is composed of mappings of synonymous concepts among these different vocabularies. The semantic network includes a set of categories to classify all concepts in the Metathesaurus and a set of relationships (semantic relations) among these concepts. Several semantic types related to disease phenotypes are organized inside the "Event" branch, such as "Disease or Syndrome" (further subdivided in "Mental or Behavioral Dysfunction" and "Neoplastic Process"), but some disease-related semantic types are inside the "Entity" branch, for example "Congenital abnormality", "Acquired Abnormality".

In general, current systems of disease classification, for historical reasons, rely heavily on disease signs and symptoms, which are not the best descriptors of a disease, because they are rarely unambiguous, and because they manifest usually at late stages of the disease. Gradually, nevertheless, these systems have incorporated disease mechanisms (pathophysiology) or causes of disease (etiology), and so they have 'chapters' defined on an etiological basis (infectious diseases, external causes), others on a pathophysiological basis (neoplasms, endocrine disorders) and others on an anatomical basis (cardiovascular diseases, respiratory diseases) (Mackenbach 2004).

These systems have been created by independent organizations, and employed for different purposes, which rather than being an advantage, may constitute an impediment to the effective communication among different systems. ICD codes, for example, are widely used for encoding disease diagnosis on electronic health records, for sharing statistics of morbidity and mortality, and for billing and reimbursement systems. Other classifications, such as MeSH and the UMLS Metathesaurus, provide comprehensive collections of names, in order to assist information retrieval, indexing, mappings across different resources, sharing and integrating different kinds of data, and automatic extraction of knowledge by means of natural language processing systems. NCI thesaurus is research-oriented and focused in cancer. While several of the most employed classification systems are included in the UMLS Metathesaurus, only a handful of resources provide mappings to other classification systems (for example, the DO provides mappings to UMLS, MeSH, ICD and OMIM). Furthermore, the number of concepts in common between different vocabularies varies (Figure 3). Currently, there are only 252 diseases included in all the vocabularies, and a rather large number of concepts are only covered by only one resource (Figure 3, panel A). In panel B, we show the overlaps for pairs of vocabularies. There is no clear trend for any of the vocabularies to contain another one. Rather, each vocabulary has its own definition of a particular disease. The second reason underlying this lack of overlap is the different level of granularity employed by each disease vocabulary.

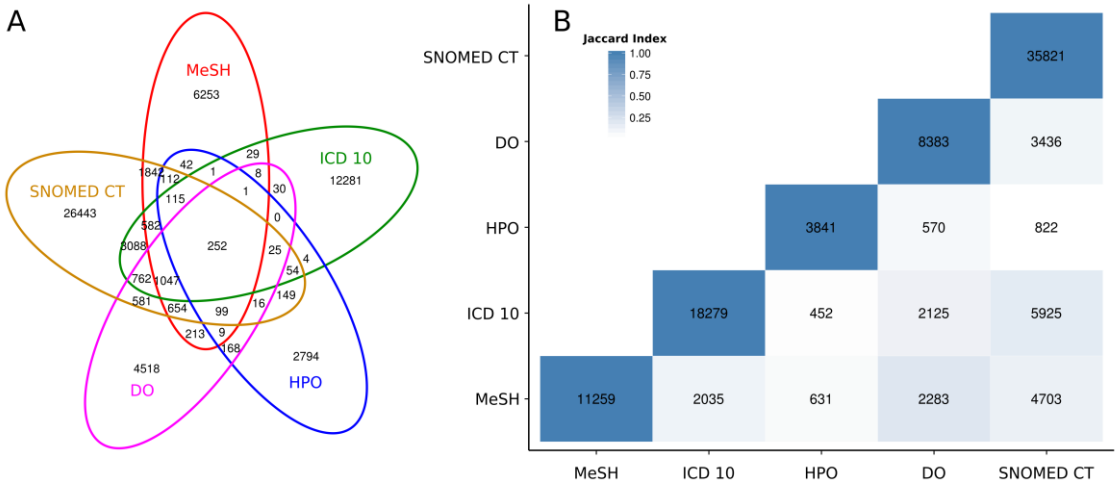

Figure 3: Overlaps among several of the more commonly employed disease vocabularies. All concepts from the UMLS Metathesaurus corresponding to the semantic type "Disease or Syndrome" (around 69,000 concepts, UMLS release 2015 AA), were mapped to other disease vocabularies, via UMLS (for MeSH, ICD, and SNOMED CT) or via the original resource (for DO and HPO). Panel A: Venn diagram of the concepts in each vocabulary. Panel B: Overlaps in pairs of all vocabularies. Note that not every concept in the vocabulary will be included in the UMLS Metathesaurus.

1.3 Available resources characterizing genotype-phenotype relationships within the context of diseases

Several publicly available resources collect our current knowledge on the genetic determinants of disease. This chapter presents a revision of the more popular of them.

1.3.1 Catalogs of human gene-disease associations

The definition of a gene-disease association in the context of this thesis encompasses both a statistical association between the gene and the disease, (for instance a co-occurrence of the two entities in a scientific publication), or a causal relationship between a particular mutation and the disease phenotype.

The **Online Mendelian Inheritance in Man®** (OMIM, http://www.omim.org/) is probably the most widely used resource by clinicians and researchers of human diseases. Originally published in a printed edition, in 1966, developed by Dr. Victor A. McKusick as *Mendelian Inheritance in Man* (MIM) (McKusick 1998), it has an online version since 1987, initially maintained by the Johns Hopkins University, and later, from 1995 on, by the NCBI (Hamosh 2004). OMIM identifiers for phenotypes and genetic loci are 6 digit codes. This resource does not use a controlled vocabulary to annotate the clinical features of diseases. Entries in OMIM are organized following a structured text-free format, because according to its maintainers, this is the best way to deal with the complexity of the relationships between gene and disease phenotypes, and because so far, free text is the best way to provide the flexibility they need to be able to describe the biological and pathological processes behind this relationship (Amberger et al. 2015). The main source of information of OMIM is the wealth of biomedical

literature, reviewed by experts. It is an authoritative resource in naming, classifying, and annotating genetic phenotypes, and describing the relationships between them, and human genes. Currently, OMIM's online catalog contains data for all known Mendelian disorders (around 8,000 diseases and phenotypes, including susceptibility traits and quantitative trait loci), and for more than 15,000 genes (Update from July 7th, 2015).

The **Comparative Toxicogenomics Database** (CTD, http://ctdbase.org/) is a resource devoted to connect chemicals, genes, and diseases (Mattingly et al. 2003). It was developed by the Mount Desert Island Biological Laboratory, to promote research in environmental health, toxicology, and molecular biology, with a focus on annotating data of nucleotide and protein sequences of toxicologically relevant genes in aquatic and mammalian species (Mattingly et al. 2003). Nevertheless, it has expanded, and it currently provides information on the relationships between chemicals and genes (1,166,89 curated interactions, version July, 2015), genes and diseases (33,814 curated interactions), and chemicals and diseases (197,288 curated interactions). The data is extracted by biocurators from the scientific literature and it is annotated using controlled vocabularies and ontologies. CTD data also contains inferences, linking diseases to chemicals via genes, or genes to diseases, via chemicals. It provides information about diseases with shared toxicogenomic profiles, as well as annotations and inferences of Gene Ontology terms (Ashburner et al. 2000), and biological pathways (from Reactome (Croft et al. 2014) and KEGG databases (Kanehisa et al. 2014)). The CTD disease vocabulary is MEDIC (MErged DIsease voCabulary), resulting from merging OMIM and the MeSH 'Diseases' branch. CTD curators review and integrate OMIM identifiers into the MeSH disease hierarchy, following a series of rules (Davis et al. 2012). CTD is the resource of choice for scientists researching the relations between environmental chemicals and human health.

The **Universal Protein Resource** (UniProt, http://www.uniprot.org) is a comprehensive, high-quality and freely accessible hub of knowledge on proteins, containing more than a half million sequence entries (UniProtKB/Swiss-Prot protein knowledgebase release, July 2015) of proteins of from more than 100 different species (The UniProt Consortium 2014). UniProt was launched on 2004, as a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR) (Leinonen et al. 2004). UniProt intends to describe all proteins in a given species, with their names, functions, post-translational modifications, catalytic activities, co-factors, pattern of expression, domains, etc. All data is captured from the literature by curators, who also annotate diseases associated to deficiencies of proteins, and variants thereof, either common polymorphisms or disease-causing mutations. UniProt uses its own "in-house" controlled vocabulary for diseases, built by manually cross-referencing MeSH and OMIM vocabularies. UniProt contains more than 26,000 gene-disease associations, linking 2,300 genes and 3,000 diseases (Human polymorphisms and disease mutations, humsavar file, Release: 2015_07)

The **Orphanet** portal (http://www.orpha.net) keeps an inventory of rare disorders since 1997 (Aymé et al. 1998). The resource is maintained by a consortium of some 40 countries, coordinated by the French INSERM team. The Orphanet definition of "rare disorder" includes diseases, malformations and clinical syndromes, as well as morphological or biological anomalies. In Orphanet, each disorder is represented by a code, an ORPHA number, given a preferred name, synonyms, and it is indexed with

identifiers from OMIM, ICD-10, UMLS, and MeSH. Interestingly, the resource also provides qualifiers for these mappings, in a way that explicitly indicates whether the mappings are exact, or reflect other types of relationship, such as from broader to narrower terms or from narrower to broader terms. The data in Orphanet includes genes associated to rare diseases, which are classified as causative, modifiers (both from germline or somatic mutations), major susceptibility factors or playing a role in the phenotype (for chromosomal anomalies). Orphanet contains around 6,000 gene-disease associations, for more than 3,000 disorders, and 3,000 genes (September, 2015).

Malacards (http://www.malacards.org) (Rappaport et al. 2013) and **Genecards** (http://www.genecards.org/) (Rebhan et al. 1998), two disease-centric and gene-centric platforms, integrate information on genes and diseases from more than 60 data sources, maintained by the Weizmann Institute of Science. Currently, they contain information on more than 150,000 genes (Genecards) and 18,000 diseases (Malacards) (accessed from the web, July 2015). The platforms are a compendium of all the information available in the wealth of scientific literature about genes and diseases, organized as "disease cards" or "gene cards", each having several sections. The disease section within Genecards integrates information from several databases, including OMIM, UniProtKB, Novoseek, Genatlas, GeneReviews, GeneTests, GAD, HuGENavigator. It presents the user also with text mining disease associations obtained from the biomedical literature. The information about diseases in Malacards includes data on aliases, clinical features, drugs and therapeutics, animal models, genes (extracted from Genecards), genetic variations, etc. Malacards offers information about a gene set "affiliated" to a disease, but this affiliation does not imply causation. Additionally, the resource offers a group of "elite" genes associated to a disease, which being reported by curated sources, constitute the most likely set of genes underlying the phenotype. Malacards catalog contains over 18,000 diseases, but not all of them contain gene annotations. Genecards contains information for over 100,000 human genes and loci, and disease associations for around 9,000 of them (September, 2015). Unlike previously described catalogs, the data contained in these two resources is not available for download.

1.3.2 Catalogs of human variant-phenotype associations

The increase of data emerging from genetic association studies, in particular GWAS, has kicked off the development of novel repositories that specifically annotate the relationships between variants and disease phenotypes.

The **Genetic Association Database** (GAD; http://geneticassociationdb.nih.gov) was created in 2004, with the aim of becoming a public archive of genetic association study data of complex diseases and disorders (Becker et al. 2004). Its main goal was the standardization of the annotations for GWAS studies. Instead of employing a controlled vocabulary for the annotation of diseases, GAD employs directly disease names. Furthermore, it annotates traits (high altitude adaptation, Body Weight), as well as pharmacogenomic-related information. Biochemical or physiological indicators, such as cholesterol or glucose levels, or Echocardiography results are annotated under the label of diseases. All these issues make it very difficult for automatic pipelines to capture and use its information. Recently, GAD has been retired from the public realm and all data is "frozen" as of September 2014. It was hosted by the National Institute on Aging, at

the NIH. GAD contains around 66,000 gene-disease associations between 10,000 genes and 2,700 diseases (via DisGeNET, v3.0)

The **Human Genome Epidemiology Network** (HuGENet) Navigator (www.hugenavigator.net) is a database containing information about the impact of genetic variants on human health at the population level (Yu et al. 2008). Its developers curate population-based epidemiologic studies extracted from PubMed using text-mining approaches. The resource contains information about the relationship between variants-disease, gene-environment interactions, the variant prevalence in US, etc. Genopedia and Phenopedia constitute later additions (Yu et al. 2010), which provide a gene-centric and disease-centric summary of the curated genetic association studies mined from Pubmed. It contains around 400,000 associations between 2,700 diseases, traits, and phenotypes, and 13,000 genes and loci (September, 2015).

Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/) is a public archive of reports addressing the relationships between human variation and phenotypes (Landrum et al. 2014). It was released in 2013, hosted by the NCBI. Phenotypes are annotated using the MedGen vocabulary. MedGen is the NCBI portal for human diseases, disorders and phenotypes. It is based on UMLS concepts, but it also integrates terms from sources not included in this repository. Variants are normalized using dbSNP identifiers. ClinVar annotates the clinical relevance of the variants within the context of disease, using its own vocabulary: benign, protective, risk factor, association, protective, pathogenic, etc. It currently contains information concerning more than 100,000 variants, in approximately 25,000 genes (accessed from the web in August, 2015). Of this, only around 3,000 genes are associated to about 6,000 phenotypes with clinical significance pathogenic or likely pathogenic.

The National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) **GWAS Catalog** (https://www.ebi.ac.uk/gwas/) contains a manually curated collection of the results of all published genome-wide association (GWAS) studies, assaying at least 100,000 SNPs, with p value $<1\times10^{-5}$ (Hindorff et al. 2009). It was developed by the NHGRI in 2008, and it is currently a collaborative project between the NHGRI and the EBI, hosted by the EMBL-EBI (Welter et al. 2014). Traits are modeled in the GWAS catalog using the Experimental Factor Ontology (EFO) (Malone et al. 2010), which contains terms not only for diseases, but also for other phenotypic descriptions, and laboratory measurements. The GWAS Catalog contains information for more than 15,000 SNPs in over 8,000 genes, and around 1,300 traits (August, 2015).

The **GWAS Central** database (http://www.gwascentral.org/) integrates genetic association data and metadata from many different sources (Beck et al. 2014). It was previously known as the Human Genome Variation database of Genotype to Phenotype information (HGVbaseG2P) (Thorisson et al. 2009), and it is maintained by the University of Leicester. Phenotypes are annotated using the Human Phenotype Ontology (HPO) and Medical Subject Headings (MeSH). Statistics are not easily obtained, because there are limits to the size of the downloadable data. As of September 2013, the database contained data for almost 3,000,000 dbSNP rs numbers.

The **Human Gene Mutation Database** (HGMD®) goal is to gather all disease-associated genetic alterations, including not only single-base pair substitutions, but also micro-deletions, micro-insertions, repeat variations, and chromosomal structural

abnormalities (Stenson et al. 2014). HGMD data are obtained through curation of scientific literature. HGMD classifies genotype-phenotype associations into: i) disease-causing mutations (DM, mutation involved in conferring the associated clinical phenotype); ii) disease-associated polymorphisms (DP, evidence for a significant association with a clinical phenotype along with additional evidence that the polymorphism is itself likely to be of functional relevance); iii) functional polymorphisms (FP, the polymorphism has a direct functional effect but no disease association has been reported yet); iv) disease-associated polymorphisms with supporting functional evidence (DFP, the polymorphism has not been reported to be significantly associated with disease but displays evidence of being of direct functional relevance); and v) frameshift or truncating variants (FTV, variants reported in the literature that are predicted to truncate or alter the length of the gene product, with no disease association reported as yet). The data is available in two versions: one public, freely available to registered users from academic institutions/non-profit organizations (http://www.hgmd.org) and a commercial version. The public version of HGMD contains over 120,000 mutations in more than 4,000 genes.

There are other repositories of GWAS results such as database of Genotypes and Phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/gap/) (Tryka et al. 2014) and the European Genome-phenome Archive (EGA) (http://www.ebi.ac.uk/ega/), which provide controlled access to individual-level GWAS data and open access to some categories of summary-level data.

1.3.3 Databases of gene-phenotype associations in animal models

The use of animal models has provided great and valuable insights in our understanding of disease mechanisms. The availability of the sequence of the genomes of several of these organisms, combined with our increased capability to manipulate the models genetic background, has produced more accurate models of human diseases. Increasing wealth of available data on this subject has made it possible to start integrating information about animal genetics and phenotypes into catalogs dedicated to the annotation of the genetic basis of phenotypes.

The **Mouse Genome Database** (MGD, http://www.informatics.jax.org/) hosted at The Jackson Laboratory is an authoritative international resource devoted to the laboratory mouse (*Mus musculus*), created more than 25 years ago (Eppig et al. 2015). It covers genetics, genomics, and other types of biological data for the lab mouse. MGD curates and integrates data on mutations (spontaneous, induced and genetically engineered) and their phenotypic impact from the biomedical literature, researcher submissions and large scale projects, and standardizes it, using the Mammalian Phenotype Ontology (Smith and Eppig 2012). Phenotypes that model human diseases are associated with OMIM terms. The relationships between the gene and the phenotype are classified as "susceptibility", "induced", and "resistance". MGD covers over 1,000 diseases, in around 1,000 genes (with human orthologs) (July, 2015).

The **Rat Genome Database** (RGD, http://rgd.mcw.edu/) created in 2000 by the NIH annotates genomic, genetic, biochemistry, nutrition, phenotype and disease data on the laboratory rat, *Rattus norvegicus* (Shimoyama et al. 2015). RGD includes several disease portals, with information about genes, QTLs and strains, pathways, and

biological processes. RGD disease information is curated from the literature. Additionally, RGD integrates information from human databases such as OMIM. The relationships between gene and phenotypes have an evidence code from the Gene Ontology, such as IEE, ISS, IAGP, etc. Additionally, RGD curators add semi-structured text notes to further characterize the relationships, such as mRNA, protein:increased expression:liver (rat); protein:increased expression:serum; or DNA:amplification. Diseases are annotated using the Rat Disease Ontology, derived from MEDIC (Davis et al. 2012). RGD contains annotations for more than 1,700 genes (with human orthologs) and around 800 disease phenotypes (July, 2015).

The **Monarch Initiative** (http://monarchinitiative.org) is a collaboration between Oregon Health & Science University, University of California, San Diego, Lawrence Berkeley National Laboratory, the University of Pittsburgh, Sanger Institute, and Charité - Universitätsmedizin Berlin. It is a resource devoted to extract genetic and phenotypic information from animal models, that makes an special emphasis in the formal representation of the data, using ontologies such as the HPO, the DO, and the Mammalian Phenotype Ontology (Smith and Eppig 2009; Smith, Goldsmith, and Eppig 2005). Monarch integrates curated data sources, primarily focused on genotype- and disease-phenotype associations, for example CTD, OMIM, Orphanet, and ClinVar. Their first, and so far only release (July, 2014) includes data from fly, worm, Zebrafish, rat and mouse models. The resource has not yet been published, and therefore its statistics and more details are not available.

In general, there are several resources available aiming at covering different aspects of the relationships between genes, or their variants, and phenotypes. Many of them constitute indeed a pocket of valuable but partial information. In effect, some resources cover only one specific type of diseases (such as OMIM for Mendelian diseases), or are organized around one type of approach to detect the association (such as GWAS studies catalogs), or gather information on one particular aspect of the relationship (such as CTD with the impact of the environment in human health). They may do more emphasis on genes, or diseases, of variants. They annotate primary data, mainly by curating biomedical literature, or integrate data from several resources. Each of them follows its own annotation criteria, and applies different degree of standardization of the information. It is worth noting that data standardization is a pre-requisite for the integration of the data with other resources and its further exploitation. For example, for the cases of diseases (OMIM, UMLS, MeSH, DO, HPO) most resources use different vocabularies. For genes, this is not a problem, given that official gene symbols from HUGO, and NCBI Gene identifiers are the two terminologies used in all resources. Another problem is the standardization of the type of relation between the gene and the disease. Each resource uses its own criteria for defining and annotating the gene-disease association. More importantly, this information is not always available from these resources. In most cases, no further characterizing is available, whereas, in others, there's only free text generated by biocurators, or semi-structured notes. There is no current standard for characterizing the relationship between a genetic alteration and the phenotype. Furthermore, currently there is no available vocabulary to reflect the different types of association between genetic alterations and phenotypes.

Gathering the information scattered across several resources is also burdensome because it is necessary to understand the type of data each resource offers, how it has been produced and how it is structured, etc. This is why resources that integrate,

homogeneously annotate and make available to researchers all this scattered information in several formats, suitable for different kinds of users are necessary.

1.4 Network biology approaches to understand human disease

To understand the molecular mechanisms explaining diseases, the identification of the causal genes is only the first step. Although genetic alterations affect individual genes, biological systems are complex, and their behavior emerges from the orchestrated activity of many components. Biological systems possess properties that cannot be explained solely from a full description of the activity of all their individual components. The scenario is further complicated when several biological phenomena such as pleiotropy[6], epistasis[7], the effect of modifier genes[8] and the variable extent of the influence of environmental factors are considered. In general, human diseases are characterized by far greater genetic heterogeneity than previously suspected.

Let's take "simple" monogenic Mendelian diseases as an example. It is known that the resulting phenotypes are produced by a complex interplay between the causative gene and other modifier genes (Cooper et al. 2013; Dipple & McCabe 2000). One of the earliest successes of linkage mapping analysis is finding the gene causing cystic fibrosis, CTFR (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)) (Kerem et al. 1989; Riordan et al. 1989). CTFR is a chloride channel expressed in the apical membrane of epithelial cells. Mutations in CFTR result in abnormalities in transepithelial electrolyte transport, which in turn alters the balance of fluid in the epithelium, ultimately producing respiratory airway disease, pancreatic failure, meconium ileus, male infertility and salty sweat (Wang et al. 2014). So far, more than two thousand mutations in human populations are reported for this gene, and are listed in the CFTR mutation database (http://www.genet.sickkids.on.ca/) (Tsui and Dorfman 2013). Approximately 70% of cystic fibrosis patients carry both alleles of the CTFR gene with the F508del mutation, while 90% carry one (Wang et al. 2014). The CTFR genotype has been shown to correlate well with the preservation of the pancreatic function, nevertheless, minimal correlation has been found with the severity of lung disease, the major cause of death in CF patients (Cutting 2010). The variability in the phenotypes of CF patients bearing the same CTFR genetic background has kicked off the search for "modifier genes". Despite the fact that several studies have been conducted, and dozens of possible modifiers have been evaluated, the success has been moderate, and the reproducibility of the results, low. Although Cystic Fibrosis is a disease of high incidence, and has been the subject of extensive research, it is not clear yet why some patients will develop lung disease, or pancreatic insufficiency, or diabetes, or intestinal obstruction, or other traits.

Extrapolating from the example of cystic fibrosis to the wealth of our current knowledge on the genetic basis of disease, we know that more than 3,000 genes are

[6] Pleiotropy: the ability of some genes to produce multiple phenotypes

[7] Epistasis: originally defined to describe a masking effect whereby a variant at one locus prevents the variant at another locus from manifesting its effect.

[8] Modifier genes: genes that have small quantitative effects on the level of expression of another gene

associated to more than 5,000 Mendelian diseases (OMIM, August 2015), and almost 3 million of disease-associated variants are stored in GWAS Central (GWAS central, August 2015). Whole genome and exome sequencing projects continuously add new variants to the catalogs of both, normal human variation, and possible pathogenic variants. To illustrate the daunting task before us to build upon these genetic basis all the way up to the manifestation of diseases at the phenotypic level, just bear in mind that the analysis of the catalogs of variants in human populations has revealed that the genomes of human "healthy" individuals contains up to 100 loss-of-function variants, and about twenty genes completely inactivated (MacArthur et al. 2012). Understanding how the effect of the perturbations caused by the variants interacts with its cellular context is another challenge, because the effects of most of these variants are likely to be small, and they become apparent only in the light of the concerted action with the genetic background of the individual.

Network analysis is particularly suited to model the complexity underlying human diseases. Different types of network representations have been employed in the last two decades to understand the organization of the cell and disease phenotypes. According to the graph theory, the nodes in these networks represent some kind of biological entity, while the edges are the relationships between pairs of nodes. Common examples of networks in biological research are protein interaction networks (nodes are proteins and edges are physical interactions), metabolic networks (nodes are metabolites and proteins and edges are metabolic reactions), gene regulatory networks (nodes are transcription factors and genes, and edges are regulatory interactions), and disease networks (nodes are diseases, and edges represent different types of relationships, such as shared genes, or risk ratios).

In particular, protein interaction networks have been a useful tool to explore the complexity of the molecular processes associated with diseases. The central hypothesis underlying the relationship between protein interaction networks and disease phenotypes is that disturbances in the local or global structure of protein interaction networks underlie the pathological changes that drive disease.

1.4.1 The human interactome

Protein interaction networks constitute an essential framework to visualize and study cellular functions. Very important biological processes are mediated by the interaction between proteins, such as metabolic pathways, the transduction of extracellular signals, the immune response, protein synthesis, etc. Our technological ability to produce protein interactions maps has significantly increased in the last ten years. The first human interactome maps were obtained in 2005, using the yeast two-hybrid system (Rual et al. 2005; Stelzl et al. 2005). Currently, there are over a hundred repositories that collect protein interaction data in a variety of organisms, obtained using different experimental techniques (Bader, Cary, and Sander 2006), for example IntAct (Orchard et al. 2014), the Human Protein Reference Database (HPRD) (Keshava Prasad et al. 2009), the Human Integrated Protein-Protein Interaction rEference database (HIPPIE) (Schaefer et al. 2012), IrefIndex (Razick, Magklaras, and Donaldson 2008), Biana (Garcia-Garcia et al. 2010), the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al. 2015), etc. For a review, see (Klingstrom and Plewczynski 2010; Mosca, Pons, et al. 2013). Noticeably, the information collected in these resources has low overlaps (Lopes et al. 2011; Rolland et al. 2014; Wodak et al.

2013) which has been attributed mainly to the complementary nature of different protein interactions detection methods (Jensen and Bork 2008).

The human interactome is estimated to contain between 130,000 (Venkatesan et al. 2009) and 650,000 (Stumpf et al. 2008) interactions. The human interactome behaves as a scale-free network characterized by a power law degree distribution in which the vast majority of nodes has a low degree, while some nodes, usually referred as hubs, have a very large degree of connectivity (Barabasi and Albert 1999; Barabási and Oltvai 2004; Barshir et al. 2014; Huttlin et al. 2015; Janjić and Pržulj 2014), although these observations have been questioned by other authors (Lima-Mendez and van Helden 2009). The interactome displays what is known as small world effect, characteristic of complex networks, where any two nodes can be connected with a path of a few links only (Barabási and Oltvai 2004).

Protein interaction networks have been employed successfully to predict protein function (Deng et al. 2003; Freschi 2007; Letovsky and Kasif 2003; Sharan, Ulitsky, and Shamir 2007), to find and prioritize disease candidate genes (Chen, Aronow, and Jegga 2009; Guney and Oliva 2012; Karni, Soreq, and Sharan 2009; Köhler et al. 2008; Moreau and Tranchevent 2012; Navlakha and Kingsford 2010; Oti et al. 2006; Smedley et al. 2014; Vanunu et al. 2010; Wu et al. 2008), to study disease comorbidities (Menche et al. 2015; Paik et al. 2014; J. Park et al. 2009; Park et al. 2011; Zhou et al. 2014), to pinpoint therapeutic targets (Gottlieb et al. 2011; Paik et al. 2014; Suthram et al. 2010; Zhao and Li 2010), to classify diseases (Chuang et al. 2007), and find similarity between them (Hamaneh and Yu 2014, 2015), to unravel the molecular basis underlying disease and its treatment (Sharma et al. 2015; Zhao and Li 2012), to understand the effect of genetic variants in relation to phenotypes (Mosca et al. 2015; Sunyaev et al. 2013; Zhong et al. 2009), to find disease biomarkers (Chen et al. 2015), among other applications.

1.4.2 Modular organization of the interactome

The distribution of edges in the human interactome is not homogeneous. On the contrary, there are regions more densely connected, that are associated to each other by loose links, in a type of structure known as modular (Girvan and Newman 2002; Rives and Galitski 2003). In a pioneer work in 1999, Hartwell proposed that cellular functions are carried out by "modules" composed of different types of molecules, carrying out discrete biological functions (Hartwell et al. 1999). Ideally, these modules (also referred as network communities, or clusters) should overlap with cellular functions. Modules may be interpreted as the functional building blocks of the cell (Mitra et al. 2013). The modular organization of biological networks has attracted attention because modularity may confer advantages in terms of robustness (because modules would limit the number of components of the system affected by a given perturbation), and in terms of adaptability to new conditions (modular networks are easily reconfigured to adapt to new conditions) (Alon 2003; Kitano 2004).

The interest in network communities has also been fueled by the fact that several studies have shown that proteins associated to the same diseases, or to similar diseases, are in the same vicinity of the interactome, show a high propensity to interact between them (Gandhi et al. 2006; Goh et al. 2007; Lage et al. 2007; Oti and Brunner 2007; Oti et al. 2006), and also tend to localize in the same subcellular compartments (S. Park et al. 2009). This has led to the hypothesis of the existence of disease modules, which

means that genes involved in the same disease tend to cluster together in protein interaction networks (Barabási, Gulbahce, and Loscalzo 2011; Ghiassian, Menche, and Barabási 2015; Menche et al. 2015). These disease modules are not expected to be identical to functional or topological modules, but rather to overlap to some extent with them. Diseases may affect or perturb several modules at once, causing developmental and/or physiological abnormalities (Barabási et al. 2011; Goh et al. 2007). Since there is no common mathematical definition of these modules or clusters, the determination of network modules has been a notoriously difficult problem (Fortunato 2010). Many popular algorithms used for community detection employ the maximization of modularity (Girvan and Newman 2002) as the optimization criteria. Nevertheless, it has been previously shown that algorithms based on this principle tend to merge small, well-defined clusters together (Fortunato and Barthélemy 2007). Recently, a review of twelve different popular clustering algorithms showed that *Infomap* –an information-based algorithm that optimizes the minimum description length of a random walk taking place in the graph – (Rosvall and Bergstrom 2008) was the best performing method, achieving the best partition on different synthetic networks (Lancichinetti and Fortunato 2009). However, its performance in the context of biological networks has been less explored. Furthermore, no studies to date have focused on analyzing the impact of the choice of the clustering algorithm in the subsequent biological analysis of the network. It remains to be clarified how different the results would be by applying different clustering algorithms.

In spite of these important issues regarding the methods to detect network modules, the knowledge of the modular structure of the network can be exploited to gain insight on biological processes. For example, network modules have been employed to predict proteins belonging to functional complexes (Cui et al. 2008; Spirin and Mirny 2003), protein function (reviewed in (Sharan et al. 2007)), disease genes (Milenkovic et al. 2010; Sharma et al. 2013), disease subnetworks (García-Alonso et al. 2012; Stevens et al. 2014), and to describe the organization and conservation of the metabolic networks in different organisms (Guimerà and Amaral 2005; Guimerà, Sales-Pardo, and Amaral 2007), among others.

1.4.3 Network properties of disease genes

Understanding the molecular features that characterize disease genes has been the subject of intense research in the past decade, with the underlying goal of using disease gene features to identify novel candidates. Seminal studies determined that disease genes tend to be longer, older, more conserved and to have more paralogs than non-disease genes (López-Bigas and Ouzounis 2004). There is a body of literature addressing the properties of disease genes in the context of protein interaction networks, that were motivated in part by the finding that yeast essential genes tend to have a higher degree in the yeast interactome (Jeong et al. 2001). The most common topological properties assessed for disease genes in protein interaction networks are degree, betweenness, and clustering coefficient (See Appendix 6.4).

The earliest studies showed that cancer genes tend to have a higher degree than non-cancer genes (Jonsson and Bates 2006; Wachi, Yoneda, and Wu 2005). Similarly, Goh and coworkers found that disease genes tend to be hubs in the interactome, but after separating essential genes from the disease genes set, they showed that non-

essential disease genes actually occupy peripheral positions in the network (Goh et al. 2007).

Following these first reports, several groups probed the topological features of different sets of disease genes in the context of protein interaction networks. (Feldman, Rzhetsky, and Vitkup 2008), for example, using genes from (Jimenez-Sanchez, Childs, and Valle 2001) show that the degree of polygenic disease genes is significantly higher than that of monogenic disease genes, while the degree of disease genes in general showed no differences respect to the average gene in a yeast-to-hybrid protein interaction network. Later, (Barrenas et al. 2009) using (Hindorff et al. 2009) as a dataset for complex disease genes and (Jimenez-Sanchez et al. 2001) to derive a dataset for monogenic diseases, and showed that complex disease genes have a higher degrees than non-disease genes, while monogenic disease genes have a higher degree than this two groups.

Cai *et al.* assessed the degree, betweenness, and clustering coefficient for Mendelian (obtained from OMIM), complex (obtained from GAD) and GWAS (obtained from (Hindorff et al. 2009)) disease genes (Cai, Borenstein, and Petrov 2010). They found that the degree of Mendelian disease genes is not different from that of non-disease genes, and for complex disease genes, the degree is only marginally significantly higher than that of non-disease genes. Nevertheless, both groups show higher betweenness and clustering coefficient than non-disease genes.

Another study compared the degree for five categories of disease genes: Mendelian and complex disease (MC) genes, Mendelian but not complex disease (MNC) genes, complex but not Mendelian disease (CNM) genes, essential genes and OTHER genes. The disease genes were obtained from OMIM and GAD (Jin et al. 2012). They found that MC genes had a higher degree compared to other disease gene groups and non-diseases genes and that the degree of CNM genes was significantly higher than that of MNC. Their results indicate that proteins involved in both complex and Mendelian disorders have more interacting partners than proteins participating only in Mendelian disorders.

It has been recently shown that the mode of inheritance of the genes may be related also to the position of the proteins in the interactome (Hao, Li, et al. 2014; Hao, Wang, et al. 2014). Specifically, autosomal dominant disease genes have significantly higher degree than that of non-disease genes while no significant difference was found for the degree between autosomal recessive disease genes and non-disease genes.

In general, the studies addressing the network properties of disease and non-disease genes have reported contradictory results. For instance, it is not clear if disease genes occupy central, peripheral o are homogeneously distributed in the interactome. Mendelian disease genes have been found more connected (Jin et al. 2012) or equally connected (Cai et al. 2010) than non-disease genes. It is not know to what extent the contradictions arise from methodological issues, such as the source of disease genes, or differences in the protein interaction networks, or if they are actually reflecting differences in the properties of disease genes associated to different disease classifications.

Several studies have shown the presence of potentially deleterious variants in the genome or exome of apparently healthy individuals (Durbin et al. 2010; MacArthur et al. 2012; Xue et al. 2012). The reasons stated to explain this observation include

recessive alleles in heterozygosis, low penetrance of the variant, gene redundancy, or sequencing artifacts (MacArthur et al. 2012). It has recently been suggested that the interactome could play a role in mitigating the effect of these deleterious variants (Garcia-Alonso et al. 2014). Using exome and whole genome data, from healthy individuals as well as from chronic lymphocytic leukaemia patients, they showed that deleterious mutations in the healthy population tend to accumulate in the periphery of the interactome, while cancer somatic variants concentrate in the internal regions. This study illustrates the usefulness of the interactome in modeling disease-associated processes, and highlights the importance of a network perspective in the analysis of NGS data.

1.4.4 Network biology in the study of disease comorbidities

The notion of "comorbidity" was introduced in 1970 by Feinstein as "any distinct clinical entity that has co-existed or that may occur during the clinical course of a patient who has the index disease under study" (Feinstein 1970). The definition has undergone a lot of changes, and it usually refers to the co-occurrence of different medical conditions in the same patient, sometimes requiring that the two diseases appear simultaneously in a patient more than expected by chance alone, and that they are related through their pathogenic mechanisms. Several models have been proposed to explain the etiological association between two diseases: direct causation (one disease predisposes to the other), associated risk factors (diseases have the same risk factor, or correlated risk factors), etc. For more details see (Valderas et al. 2009).

The use of network biology approaches to study disease comorbidity has the potential to reveal hidden genetic connections between diseases, and to highlight molecular mechanisms underlying this complex phenomenon. Network approaches have been widely used to gain understanding on the disease comorbidity phenomena (Goh et al. 2007; Menche et al. 2015; J. Park et al. 2009; Roque et al. 2011; Rzhetsky et al. 2007). The networks built for these studies usually represent nodes as diseases or phenotypes, with edges representing a variety of processes, for example, disease co-occurrence (Hidalgo et al. 2009; Rzhetsky et al. 2007), or shared genes, or pathways. Bipartite networks, where nodes are genes and diseases, and edges, their associations, have also been used (Bauer-Mehren et al. 2011; Goh et al. 2007).

In a pioneer study, Goh et al. 2007 constructed the first disease network, the human diseasome, a map that connected diseases with shared genetic components, based on gene-disease associations extracted from OMIM. The underlying assumption in this study was that comorbidities occur when a pair of diseases shares the same genetic alterations (*shared gene formalism*). A second approach, the *shared pathway formalism*, states that diseases are connected via biological modules such as protein–protein interactions or molecular pathways (Lage et al. 2007; J. Park et al. 2009). Park 2009 found significant correlations between comorbidity patterns and cellular interactions, measured not only in terms of shared genes, but also, in terms of protein-protein interactions, and the degree of correlation of gene expression (J. Park et al. 2009). Additionally, comorbidity measures are also correlated to the similarity of pairs of diseases in the subcellular localization (Park et al. 2011). The number of shared genes and protein interactions is also strongly correlated to symptom-based disease similarity (Zhou et al. 2014). These two formalisms employ only molecular data to connect diseases, which allows generating new hypothesis about the underlying pathogenesis. A

third approach to the study of disease comorbidities is the "*disease comorbidity formalism*" that links diseases based on statistically significant co-occurrences in clinical data. For example (Rzhetsky et al. 2007) constructed a phenotypic disease network incorporating 161 diseases from 1.5 million of Medicare patients, identifying positive and negative correlations between diseases using a probabilistic model. This study showed that this approach allows finding well-known connections between diseases, and also novel ones. Interestingly, their results also raise the possibility of a genetic model where complex phenotypes are probably rooted in genetic variation that is significantly shared by other disease phenotypes, hypothesis that has recently further explored by a similar study, that ranged 110 million patients (Blair et al. 2013). Similarly, (Hidalgo et al. 2009) constructed a phenotypic disease network from the disease history of over 30 million of Medicare patients. The pairs of diseases were linked using two different comorbidity measures. Their results illustrated the dynamics of disease progression, directionality and morbidity. Recently, Menche and cols. demonstrated that despite the incompleteness of the interactome, and of our knowledge of the genetic causes of diseases, the magnitude of the overlap of the disease network modules is indicative of the similarity of their pathobiology and their comorbidity (Menche et al. 2015).

The use of network biology approaches to study disease comorbidity has the potential to reveal hidden genetic connections between diseases, and to highlight molecular mechanisms underlying this complex phenomenon.

2. OBJECTIVES

The era of massive whole genome and exome sequencing has bring the hope that in the near future it will be possible to make accurate predictions about the potential development of diseases along a lifetime based on genomic information. This capability will be essential to interpret individual genomes with the aim of carrying out personalized diagnoses and eventually, personalized treatments. We are nevertheless still far from that capacity. Moreover, the development of new therapies lags behind the discovery of disease-related genes. While we now know that the majority of human protein-coding genes may be involved in diseases, we are currently able to target ~200-400 of them for treatments. Another, deeper reason is that modeling the complexity of the human phenotypes requires the integration of several layers of biological data, which is usually scattered across different specialized resources, annotated with a myriad of different standards. New resources are urgently needed that gather the information on the genes underlying human diseases under standard vocabularies and unified annotations. Furthermore, new bioinformatics tools and algorithms are necessary to mine these resources in search for the molecular mechanisms underpinning different types of diseases and their comorbidities. With this background, we set the following objectives for the thesis:

General Objective: To develop computational tools and methodologies to gain insight on the molecular underpinnings of human diseases, using network biology approaches.

Specifically, our goals were:

- ✓ To develop a comprehensive, integrative platform, that integrates information on the genetic basis of diseases, allowing easy access to a broad range of users
- ✓ To analyze the molecular and network features of disease genes
    - o To study the impact of clustering algorithms on the biological analysis of protein interaction networks
    - o To assess local, mesoscale and global network properties of disease genes
    - o To study the tolerance to likely deleterious variants identified by NGS of genes in different disease classifications from a network biology perspective
- ✓ To develop and apply a network medicine strategy to explain the molecular mechanisms underlying disease comorbidities

3. RESULTS

3.1 DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes

The information on the genetic causes of human diseases is scattered across several resources, each one using different schemas and standards. This is why the creation of resources that collect and homogeneously annotate our current knowledge on the genetic basis of diseases is a key starting point to many translational bioinformatics applications. With this aim, we developed DisGeNET, a discovery platform that aims to collect all available information on gene-disease associations, covering the whole landscape of human diseases. DisGeNET possesses several unique features that make it a very useful platform for biomedical researchers. First, it contains a very large collection of gene-disease associations arising from both expert-curated knowledge, and information extracted from the scientific literature using refined text-mining techniques, with special attention paid to the explicit provenance of the association. Second, mappings to different biomedical vocabularies annotating diseases are provided for gene-disease associations, thus facilitating the work of clinical and biomedical researchers. Third, a score developed to rate the confidence of each association is available to users. Finally, several ways to access the data are available, to serve better the purposes of different types of users.

Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* (Oxford). 2015 15;2015:bav028. doi: 10.1093/database/bav028.

3.2 Mining the modular structure of protein interaction networks.

Biological networks are organized following a modular structure that has been hypothesized to reproduce cellular functions, and as such is key to study the functionality of biological systems (Barabási et al. 2011; Hartwell et al. 1999). Breaking down biological networks into modules allows us both reducing their complexity, and describing them using a more coarse-grained scale. Modules are usually defined as zones where the nodes are more densely connected between them, and sparsely with their neighborhood, but they lack common mathematical definition. Several clustering algorithms are currently used to partition protein interaction networks. Some of the most popular use the principle of optimization of the modularity (Newman 2006), whereas other well performing algorithms rely on more information-theoretical considerations (Rosvall and Bergstrom 2008). In this work, we compared the network partitions produced by two clustering algorithms that define modules in different ways. Then, we explored how the differences in the partitions influence the results of subsequent biological analysis of the network.

Berenstein AJ*, Piñero J*, Furlong LI, Chernomoretz A. Mining the Modular Structure of Protein Interaction Networks. *PLoS ONE* 2015, 10(4): e0122477. doi: 10.1371/journal.pone.0122477

3.3 Uncovering disease mechanisms through network biology in the era of next generation sequencing

The network properties of disease genes have been in debate for the last decade due to the lack of consistency of the result of this type of study. Most of the previous works have been performed on disease genes from OMIM, and using only one resource of protein interaction network data. In this study we aimed to provide a definitive answer [9] to the question of the network properties of diseases genes. We used DisGeNET, one of the most complete repository of disease genes, and the state of the art resources of protein interaction data. We studied the network properties of disease genes at different scales of the protein interaction network (local, mesoscale and global). We also analyzed the molecular and network properties of disease genes classified from different perspectives (clinical, genetic and molecular classifications). Finally, in the light of recent evidence suggesting that the interactome plays a role in buffering likely deleterious mutations from whole genome and exome data, we evaluated the relationship between the network properties of different classes of disease genes and their tolerance to likely deleterious mutations.

Piñero J*, Berenstein AJ*, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of next generation sequencing. *Submitted*

[9] At least with the most current data and methodology available today. We envision that this kind of studies should be revisited once the knowledge of the interactome of disease genes is complete.

Abstract

Characterizing the behavior of disease genes in the context of biological networks has the potential to shed light on disease mechanisms, and to identify new candidate disease genes and therapeutic targets. Previous studies addressing the network properties of disease genes have produced contradictory results. Here we have explored the causes of these discrepancies and assessed the relationship between the network roles of disease genes and their tolerance to deleterious germline variants in human populations leveraging on: the abundance of interactome resources, a comprehensive catalog of disease genes and exome variation data. We found that the most salient network features of disease genes are driven by cancer genes and that genes related to different types of diseases play network roles whose centrality is inversely correlated to their tolerance to likely deleterious germline mutations. This proved to be a network multiscale signature, including global, mesoscopic and local network centrality features. Cancer driver genes, the most sensitive to deleterious variants, occupy the most central positions, followed by dominant disease genes and then by recessive disease genes, which are tolerant to variants and isolated within their network modules.

Background

With the application of next generation sequencing technologies to the identification of both germline and somatic variants across cohorts of patients as well as healthy individuals, the catalog of potential pathogenic variants is expanding rapidly (Collins, 2015). Recent findings have shown a large number of potentially damaging germline variants, but for most of them there is no information on the functional effect and its relation to disease. Currently, our ability to interpret the effect of the variants discovered by genome sequence projects and how this leads to phenotypic variation and diseases is very limited (MacArthur *et al*, 2014; Chong *et al*, 2015). The fact that disease phenotypes are not caused by the individual action of genes but by their interaction in the context of biological networks further complicates the identification of clinically relevant variants (del Sol *et al*, 2010). The discovery of somatic variants causing tumorigenesis is also a challenge. This field has advanced in recent years with the development of bioinformatics methods that detect signals of positive selection in genes across tumor samples, thus identifying the most likely driver candidates (Lawrence *et al*, 2014; Tamborero *et al*, 2013; Kandoth *et al*, 2013). Nevertheless, uncovering driver mutations that occur at frequencies below the level of detection of these methods would require either sequencing even larger cohorts of tumors or new approaches incorporating prior knowledge of cancer genes features.

Biological networks are useful tools to model the complexity of the genotype-phenotype relation underpinning disease. In the last decade, protein-protein interaction networks (PINs) have extensively been exploited with the goal of unraveling the molecular mechanisms of a variety of human diseases (Taylor *et al*, 2009; Liu *et al*, 2007; Santiago & Potashkin, 2013; Jonsson & Bates, 2006; Sarajlić *et al*, 2013; Zhong *et al*, 2009), and identifying novel disease genes candidates (Guney & Oliva, 2012; Nivit *et al*, 2014; Köhler *et al*, 2008; Oti *et al*, 2006; Lage *et al*, 2007; Franke *et al*, 2006; Navlakha & Kingsford, 2010; Lee *et al*, 2011; Chen *et al*, 2009; Vanunu *et al*, 2010; Wu *et al*, 2014b; Li *et al*, 2014; Wu *et al*, 2008; Luo & Liang, 2015). There is, thus, a body of scientific literature concerning the network properties of disease-related genes. They have been reported to possess distinctive topological properties that cannot

be attributed solely to the fact that they have been extensively studied (Lim *et al*, 2006; Xu & Li, 2006; Ideker & Sharan, 2008; Feldman *et al*, 2008; Cai *et al*, 2010; Barrenas *et al*, 2009; Ghersi & Singh, 2013). While cancer genes tend to occupy central positions in a PIN (Taylor *et al*, 2009; Wachi *et al*, 2005; Jonsson & Bates, 2006), the scenario is less clear for other classes of disease genes. For example, (Goh *et al*, 2007) found that if essential genes are excluded from the analysis of Mendelian disease genes, these do not show a tendency to occupy hub positions in a PIN (Goh *et al*, 2007). Nevertheless, (Xu & Li, 2006) reported that Mendelian disease genes are more central in a literature curated PIN (Xu & Li, 2006). Later, Mendelian and complex disease genes were found to possess higher degrees and lower clustering coefficients than non-disease genes (Cai *et al*, 2010). Currently it is not clear to what extent these seemingly contradictory results are caused by i) methodological issues, such as the sources of disease genes and/or PIN, ii) data incompleteness in any of these sources (disease genes, PIN), iii) discrepant topological definitions, or iv) actual variations in the network properties of genes belonging to different disease classes (Furlong, 2012).

We have carried out a study to systematically determine whether and to what extent groups of genes resulting from different disease classifications –for example, clinical taxonomies vs. molecular genetics– possess distinct network properties. First, we have used a comprehensive catalog of disease genes, DisGeNET (Piñero *et al*, 2015), and culled different sets of genes corresponding to diverse disease classifications. Second, we carried out all our analyses across six different PINs to investigate the impact of the selection of the network in the results. Third, we have analyzed local, mesoscale and global network properties of disease genes within the PIN. The analysis at the mesoscale level of the network provides insight into the modular organization of the PIN, potentially shedding light onto the mechanisms and regulation of cellular processes. Finally, we have also explored the relationship between the network indices that probe connectivity features at different scales of disease genes and their tolerance to likely deleterious germline variants. We were motivated by the lack of systematic studies addressing the question of whether the network location of different classes of disease genes correlates with their tolerance to possibly deleterious germline mutations. In probably the closest precedent, a recent study found that deleterious variants in the exomes of 1,330 healthy individuals are located in the periphery of the PINs, while cancer somatic mutations appear in internal regions and monogenic disease variants are at intermediate network positions (Garcia-Alonso *et al*, 2014).

We discuss our results in the context of a model that combines genetics (type of variation and mode of inheritance), genomic and interactomics (protein interactions through networks) to understand the mechanisms underlying human diseases.

Results

In order to assess the network properties of different classes of disease genes, we obtained six high quality PINs (See Methods for details) and analyzed them using an approach that combines topology and the modular structure of the network. Throughout the paper, we illustrate the results obtained with the HIPPIE protein interaction network (Schaefer *et al*, 2012), but all analyses were replicated in other five PINs (see overlaps between the PINs in Supplementary Fig S1).

The HIPPIE protein interaction network contains 9,580 proteins and 47,102 interactions. We found that 4,122 out of 7,412 disease genes, DGs (extracted from DisGeNET, see details in Methods) are included in HIPPIE (the number of genes mapping to other networks are in Supplementary Table 1). Notice that all PINs contain between 40-60% of DG, highlighting the incompleteness of our knowledge of the human interactome. Roughly half of the genes in the DG set are related to neoplasms (cancer related genes, CGs), while the other half are associated to other diseases (non-cancer disease genes or NCDGs). Out of the 4,122 DGs, 1,607 are related to Mendelian diseases (MGs), while 1,839 are related to complex diseases (CxDGs).

Cancer genes are responsible for the network centrality of disease genes

We found that DG have a higher degree and betweenness than non-disease genes, but this trend is driven by cancer genes, as indicated by the decrease in the observed differences when CG are removed from the DG set (that even disappear in the case of the degree in 3 of the 6 PINs, Supplementary Fig S2 and S3). MGs and CxDGs have higher degree and betweenness than non-disease genes. The results for clustering coefficient are inconclusive (Supplementary Fig S4). The behavior of this parameter seems to be linked to the nature of the method of detection of protein-protein interactions. Methods that capture indirect interactions, for example affinity purification with a bait protein, tend to produce more clustered networks, depending on the way that the interactions are annotated. This could lead to zones of artificially inflated estimates of clustering coefficients (Rual *et al*, 2005).

Overall, these results provide systematic support to the observation made by various groups that cancer genes are more central to PINs, and present different local environments than NCDG (Taylor *et al*, 2009; Jonsson & Bates, 2006; Wachi *et al*, 2005; Ghersi & Singh, 2013). Non-Cancer Disease Genes, on the other hand, do not show distinctive topological features, once the contribution of cancer genes is disregarded.

The cartographic analysis of a PIN highlights mesoscale connectivity patterns

We next explored the connectivity of different groups of disease genes at the mesoscale level of the network –i.e., pertaining to its organization in clusters or modules. Despite the widespread use of network topological metrics like the betweenness or the degree of a node, it is worth noting that these features are not intended to explicitly mine mesoscale connectivity patterns. Betweenness related centrality indices may unveil interesting connectivity features at the global level, but they might not reflect them at a more local scale. On the other hand, as the degree of a node does not reflect the modular organization of the network, the use of degree-based centrality metrics might confers similar importance to genes that link different modules or, on the other hand, are confined within a module. In addition, genes presenting low degree might be disregarded from a degree-centric point of view, even if they do play relevant connectivity roles in the biological network. Examples from HIPPIE include ADRBK2 (56^{th} and 43^{rd} percentile of degree and betweenness, respectively), a protein kinase involved in several signaling pathways (Croft et al. 2014), the phospholipase PLCB1 (48^{th} and 37^{th}), that plays an important role in the intracellular transduction of many extracellular signals and in metabolism, the protease FURIN (48^{th} and 47^{th}), and MAPK12 (61^{st} percentile for both), involved in pathways such as Signal Transduction

(NGF, VEGF), Cell-Cell communication, Developmental Biology and Innate Immune System and Organelle biogenesis and maintenance.

The mesoscale organization of the network is linked to its organization in clusters or modules. To unveil the modular organization of the PINs, we employed the *Infomap* procedure (Rosvall & Bergstrom, 2008), one of the best performing network community recognition methodologies, which has produced sensible partitions of different types of complex networks (Lancichinetti & Fortunato, 2009; Berenstein *et al*, 2015; Liu *et al*, 2014). After partitioning the PINs with *Infomap*, we characterized the mesoscale connectivity features for each network node in terms of two parameters: the *intra-cluster connectivity*, *z*, and the *participation* coefficient, *P* (Guimerà & Amaral, 2005b). The *z* parameter standardizes the degree of a node in relation with the degree of nodes that belong to the same community, and the *P* parameter quantifies the fraction of links that a given node projects to other communities (see methods). We further categorized each network node according to the universal cartographic role classification scheme established by Guimerà and Amaral (Guimerà & Amaral, 2005b, 2005a). Figure 1 shows the distribution of *HIPPIE* nodes over the *z-P* plane. Dashed lines in the figure delimit regions corresponding to the seven cartographic roles (Guimerà & Amaral, 2005b).

The classification of HIPPIE nodes into cartographic roles revealed a majority of poorly connected nodes playing *peripheral* (2,596) or *ultra-peripheral* (3,346) roles. More densely connected nodes were either *connectors* (2,193), with links more or less evenly distributed between the genes in their cluster and genes of other clusters, *kinless* nodes (1176), displaying fewer than 35% of intra-cluster links, or *kinless hubs* (188) with more than half of their connections established with members of different clusters. The remaining nodes were either *connector hubs* (71), or *provincial hubs* (10). Reassuringly, we found that despite the differences in the PINs (Supplementary Fig S1), the proportion of proteins in each cartographic role was very similar across all the six high-quality studied PINs (Supplementary Table S2).

The cartographic analysis summarized relevant mesoscale interconnectivity features that might serve to highlight biologically sensitive patterns. For instance, the cartographic classification of the 6,608 proteins in HIPPIE annotated to Panther protein classes (Thomas, 2003) recapitulated major features of the architecture of cellular signaling pathways (Supplementary Fig S5). Signaling molecules, membrane receptors and transporters were significantly enriched for *ultra-peripheral* and *peripheral* HIPPIE nodes. Furthermore, nodes in these cartographic roles exhibited a clear enrichment for Gene Ontology terms related to the activity of membrane receptors –many are ligands, receptors or receptor modulators– and transporters (Supplementary Table S3). On the other hand, proteins with high participation roles (*kinless* and *kinless hubs*) were most significantly enriched for chaperone functioning and regulatory classes, such as kinase, ligase, transferase and nucleic acid binding activities from Panther (Supplementary Fig S5), and chromatin maintenance, regulation of transcription and regulation of ubiquitin mediated proteolysis from the Gene Ontology molecular function (Supplementary Table S3).

Different types of disease genes show distinctive connectivity patterns at the mesoscale level

We analyzed the mesoscale features of disease genes by looking at the overrepresentation of each set of disease genes across the seven network cartographic roles (Fisher exact Test, Table 1) for the different gene sets analyzed. Disease genes as a group (see DG column, Table 1) exhibited significant enrichment for *connector*, *kinless* and *kinless hub* roles. However, this signal disappears when CGs are removed from the set of DGs (see NCDGs column, Table 1). CGs are enriched for nodes with high participation roles, such as kinless (p-value 10^{-38}), kinless hub (10^{-10}) and connector (10^{-9}). This result is consistent across the six PINs (p-values 10^{-27}-10^{-49}, 10^{-6}-10^{-15}, and 10^{-2}-10^{-9}, respectively; Supplementary Fig S6). NCDGs, on the other hand, are homogeneously distributed amongst roles, which underlines that CGs, again, are solely responsible for the observed enrichment of DGs for high participation roles.

The fact that cancer genes have high degree values in PINs could partially explain the observed enrichment for the kinless-hub role, given the existent role-to-degree relationship (Supplementary Fig S7). However, the enrichment for non-hub roles (*connector* and *kinless roles*) revealed a qualitatively different participation-based bias, more directly linked to inter-modular connectivity patterns. Moreover, the fact that the enrichment of CGs for *kinless* nodes was several fold greater than their enrichment for nodes in *connector* roles suggests that cancer genes tend to connect many separate modules of the network, rather than genes in their close vicinity which belong to the same cluster. To further validate this hypothesis we decided to de-convolve the *degree* signal from the enrichment results observed for CG genes performing a degree-aware bootstrap test for the cartographic role enrichment calculation (see Methods). Interestingly, we found that *connector* and *kinless*, but not the *kinless hub* category enrichments remain significant under the bootstrap analysis ($p_{conn} < 10^{-3}$, $p_{kin} < 10^{-3}$, $p_{kinless\text{-}hub}=0.92$). These results suggest that CG genes display a non-trivial enrichment for non-hub, high participation cartographic roles, which cannot be explained by the effect of the degree distribution, but is cemented on mesoscale connectivity patterns.

When assessing the difference between genes related to complex and Mendelian diseases, we found that CxDGs are overrepresented amongst *kinless* and *kinless hub* genes, whereas MGs are enriched for *kinless* and *connector* genes. However, both trends disappeared after removing the CGs from each gene set (Table 1). We observed a similar behavior when genes are grouped according to MeSH disease classes. With the exception of Parasitic and Eye diseases, the corresponding gene sets are similarly enriched for *kinless* and *kinless hub* nodes (Supplementary Fig S8B). Nevertheless, this trend disappears when CG are removed from each of the MeSH disease genes.

We reasoned that other disease categorizations, more homogeneous in terms of genetic or molecular mechanisms might result in gene sets with clearer network trends. In order to further investigate this hypothesis, we subdivided MGs according to their inheritance mode into autosomal dominant (AD) and autosomal recessive (AR) disease genes. In addition, we filtered the set of cancer related genes (CGs) to keep only genes related to tumorigenesis upon somatic alterations (drivers). The mapping of the gene sets into the different PINs is summarized in Supplementary Table S1. There is a certain degree of overlap between these sets because some genes may contribute to different diseases. For instance, some germline variants in several well-known loss of function cancer driver genes such as DNM2, SMAD4, NF1, PTCH1, PTEN, SMARCB1, TSC1, cause dominant negative Mendelian diseases (Zhu *et al*, 2014).

While AD genes are enriched for *kinless* and *kinless hub* roles (Table 2), AR genes are not significantly overrepresented within any role in HIPPIE. Nevertheless, they do exhibit enrichment for *ultra-peripheral* nodes, in two of the PINs (BioGRID, p-value 10^{-2} and IntAct p-value 10^{-3}, Supplementary Fig S9). Driver genes show significant enrichment for *kinless* and *kinless hub* roles (Table 2). Removing the driver genes from AR and AD sets resulted in a decrease of the enrichment for nodes of high participation roles for AD genes and, inversely, an increase of the enrichment of the AR genes for nodes of low participation roles. Noticeably, the enrichment of drivers, AD, and AD non-driver genes for *kinless* nodes remains significant under the degree-aware bootstrap analysis (p_{driver}=0.001, p_{AD}=0.001, $p_{AD\text{-}non\ driver}$=0.026). This last result stresses that these gene sets display non-trivial connectivity patterns at the network mesoscale level.

The tolerance of different types of disease genes to likely deleterious germline variants reflects the heterogeneity of their network roles

We hypothesized that disease genes with higher-than-average participation in the network must be under strong purifying selection and therefore, be less tolerant to likely deleterious variants across human populations. To the best of our knowledge, there is no previous study that systematically addresses the relationship between the tolerance to germline mutations with the network properties of disease genes. Therefore, we asked next whether genes involved in diseases of different classes, which display distinctive network roles, exhibit different sensitivity to likely deleterious germline variants. To answer this question, we retrieved the germline variants detected across 60706 exomes (ExAC) and kept those falling into one of two groups. In the first group, we included protein sequence affecting variants –missense, stop gained, stop lost, frameshift, splice donor and splice acceptor variants– with CADD score > 15 (Kircher *et al*, 2014) and considered them as *likely deleterious variants*. The second group comprised synonymous variants, which were considered *non-deleterious*. We then computed, for each gene, a High-impact-to-Synonymous variants Ratio (HS Ratio) as the quotient between the number of *likely deleterious variants* and the number of *non-deleterious variants* (see Methods). We use this HS Ratio as a proxy of the sensitivity of genes to likely deleterious germline variants.

Cancer drivers and AD genes exhibit lower HS Ratio than the average genes (Table 3), denoting a higher-than-average sensitivity to likely deleterious variants in human populations. The trend of AD genes towards lower-than-average HS Ratio becomes less significant in the AD_{ND} set. AR genes, on the other hand, show significant less sensitivity to such deleterious variants than average genes in the PINs.

Figure 2 illustrates the double separation that takes place between groups of disease genes and non-disease genes in terms of network features and the tolerance to likely deleterious germline variants. We have plotted the z-score of each parameter resulting from 10,000 randomizations for each set of genes (see Methods). It can be observed from the figure that the more significantly lower the HS Ratio for a given disease gene set, the more significantly higher the corresponding centrality indices. Differences in network centrality metrics were particularly large for drivers and AD genes. A moderate bias toward high values of mesoscale centrality features could still be recognized for the set of non-driver AD genes (AD_{ND} set). On the other hand, AR genes show almost no differences in their network features with respect to the average

gene in the network across all PINs but exhibit significantly-higher-than-average HS Ratio. They do exhibit smaller degree (in 4 out of 6 PINs, Supplementary Fig S10) and smaller participation coefficient than the average node in the BioGRID and IntAct PINs (Supplementary Fig S10). Again, this trend became stronger when known driver genes were removed from the AR set, so that it became significant for HIPPIE as well (Supplementary Fig S10). In summary this analysis uncovers a relationship between a gene's centrality in the PIN and its sensitivity to likely deleterious germline variants. Noticeably this proved to be a network multiscale signature, as the same trend is observed when global (i.e. betweenness), mesoscopic (within-module degree or participation coefficient) and local (degree) network centrality features are considered.

Finally, we focused on the subset of genes that are both cancer drivers and associated to Mendelian diseases. These genes behave collectively –in terms of centrality, molecular activities and sensitivity to probably deleterious germline variants– like driver genes. We hypothesized that the deleterious germline variants in these genes that cause Mendelian disorders affect positions in the protein sequence that are different from those affected by somatic mutations that turn the gene into a cancer driver – because changes at these positions would likely result in lethal phenotypes. We tested this hypothesis on a group of 81 driver genes (35 loss-of-function and 46 gain-of-function drivers) on which deleterious germline variants causing a Mendelian disorder have been mapped to at least three separate positions. Specifically, we asked whether somatic mutations with tumorigenic potential (non-synonymous mutations on oncogenes and non-synonymous and truncating mutations in tumor suppressors) and disease related germline variants tend to occur at different positions. We found that this is the case for the majority of these 81 genes –65 of them possess Fisher's odds-ratios below 0.1 (Supplementary Table S4). In Figure 3 we show examples of loss-of-function (NF2 and KDM5C) and gain-of-function (GATA2, PAX8, and PTPN11) driver genes with little or no overlap between germline and somatic mutations. Exceptions to this trend are genes that suffer germline variants that confer susceptibility to cancer, such as the von Hippel-Lindau syndrome caused by some variants in VHL; Cowden disease 5 caused by mutations in PIK3CA; Li-Fraumeni syndrome 1 and TP53; and proteins related to the RAS family, or belonging to RAS pathways, whose germline mutations produce developmental diseases that frequently increase the risk of cancer (Fernández-Medarde & Santos, 2011).

Discussion

Our results show that the network centrality of different classes of disease genes, including complex, Mendelian and clinical-oriented classifications, is mostly attributable to cancer genes. Cancer genes are central not only in terms of number of neighbors, but also in terms of the clusters they connect. Remarkably, we found that high Participation roles played by cancer genes are not explained by their higher number of interactions, but by their unique inter-modular connectivity patterns. We also found that these connectivity patterns differ for disease genes with contrasting inheritance modes: while autosomal dominant genes play high participation roles, autosomal recessive genes are more confined to their own modules. Interestingly, the network roles of these two different types of disease genes relate to their tolerance to likely deleterious germline variants: the more central the disease genes are, the more sensitive to damaging germline variants.

Our findings may explain some of the seemingly contradictory results reported so far (Xu & Li, 2006; Goh *et al*, 2007; Feldman *et al*, 2008; Jin *et al*, 2012; Cai *et al*, 2010), which found that "disease", "complex disease", and "Mendelian disease" genes occupy central network positions, but are probably measuring the centrality of cancer genes. On the other hand, our study found that only autosomal recessive Mendelian genes are overrepresented in the periphery of the network, contrasting previous reports (Goh *et al*, 2007). Additionally, we found that autosomal dominant Mendelian genes possess network properties that are in part driven by a small subset of driver genes overlapping with them. This also might explain part of the differences between the results of previous studies, which have largely ignored that some genes linked to dominant and recessive Mendelian diseases are also cancer drivers. In summary, our results put in perspective previous observations regarding the properties of disease genes as a whole and even question the rationale behind the analysis of such heterogeneous sets of disease genes. Our findings are constrained to our current knowledge on the interactome of disease genes. However, as they are consistent across six different PINs, we confidently propose that our results are not caused by any bias of a particular protein-protein interaction dataset.

We found compelling evidence both from topological and cartographic analysis of PINs that among all disease related genes, cancer drivers occupy the most central roles, significantly expanding a previous report focused on 21 Chronic Lymphocytic Leukemia drivers (Garcia-Alonso *et al*, 2014). In particular, the cartographic analysis showed that cancer drivers are very significantly overrepresented among the proteins that connect several modules of the PIN (*kinless* and *kinless hub* nodes). Genes that play these roles are frequently involved in very core cellular processes, such as signal transduction through several pathways, the regulation of transcription, chromatin maintenance, and ubiquitin mediated proteolysis. The enrichment of driver genes for these two central roles also explains why they are significantly more sensitive to likely deleterious germline variants. Probably an important fraction of deleterious germline variants affecting these genes are filtered out by purifying selection. On the other hand, somatic mutations that affect them have a high likelihood of causing tumorigenesis probably because they impact on key cellular functions (Hanahan & Weinberg, 2000, 2011; Vogelstein *et al*, 2013). A fraction of driver genes, nevertheless bear deleterious variants that are not lethal, but cause Mendelian diseases. Interestingly, the majority of the driver genes that play a role in Mendelian disease have a dominant inheritance mode. This might explain why autosomal dominant genes resemble driver genes. Some of these genes increase predisposition to cancer, for example, in rasopathies, and BRAF, KRAS, HRAS, NF1, NRAS PTPN11, RAF1 (Fernández-Medarde & Santos, 2011). These genes behave like drivers both in the roles they play in the PIN and in their sensitivity to likely deleterious variants, rather than as Mendelian disease genes. In summary, our findings suggest that if a gene plays high participation roles in the network, deleterious germline variants affecting it will have high probability to be filtered out by purifying selection, while somatic mutations impacting its function, will cause tumorigenesis with high probability. This finding could be incorporated into automatic systems aimed at identifying novel candidate driver genes.

Autosomal recessive genes show a behavior entirely opposite to that of cancer drivers. They tend to be less central (as observed by (Hao *et al*, 2014b, 2014a)), enclosed within a network module and exhibiting low or null participation and low

degree, and they are significantly more tolerant to likely deleterious variants than other genes in the network. The effect of variants in one such gene would be confined to its own module, as in the case of metabolic enzymes. Autosomal dominant genes occupy an intermediate position between autosomal recessive and drivers. Proportionally, they represent a smaller share of kinless and kinless hubs than drivers; and coherently, they are less sensitive to likely deleterious germline variants, although still more than the average gene in the interactome.

Deleterious variants in genes associated to dominant diseases might be dominant negative or produce haploinsufficiency (Veitia, 2002; Wilkie, 1994) resulting in the disease phenotype. In the first case, the protein product of a mutated allele ultimately produces an aberrant protein complex, whereas in the second the protein level produced by the normal allele is not sufficient to fulfill the entire functionality of the protein. The former mechanism may fit better the behavior of Autosomal dominant genes encoding structural proteins, which are enriched for non-truncating variants, while the latter may explain better the case of transcription factors, enriched for truncating variants (Zhong *et al*, 2009). These two mechanisms could explain why some loss-of-function drivers, such as TP53, PTEN, RB1, and APC may also behave as Mendelian dominant genes. In the first case, while both alleles of the gene may be rendered inactive by alterations in cancer, deleterious germline variants in only one allele may produce a defective copy of the protein which in turn produces a faulty multimer composed of both active and inactive subunits of the protein. As for haploinsufficiency, the decrease in the level of active protein caused by a deleterious germline variant on one allele of the gene, determines the disruption of at least some of the functions –maybe by the failure to fulfill all its interactions, or to maintain signaling through certain pathways at homeostatic levels– carried out by the protein. In the haploinsufficiency scenario, the more complexes a protein is involved in, the more likely it is that a decrease of its level results in disease.

We can explain our results within the framework proposed by (Zhong *et al*, 2009). In this model, because Autosomal recessive genes are involved in very few interactions, confined to their own pathways or modules, only damaging variants in both alleles, which can be regarded as a removal of the node from the network are able to cause Autosomal recessive diseases. On the other hand, node removal of Autosomal dominant genes –and some driver cancer genes– is probably lethal to the cell. Nevertheless, less damaging variants which do not abolish the function of the protein but affect one or some of its many interactions –i.e, edge perturbations— might trigger an Autosomal dominant disease. Most drivers, on the other hand, are intolerant to these edge perturbation events, probably because they are incompatible with the development of a viable organism. On the other hand, somatic cells may acquire growth advantages and eventually become malignant from mutations that cause either edge perturbations (new interactions, or their hyper activation, in the case of oncogenes) or node removal (tumor suppressors) of cancer driver genes.

Finally, we believe that computational methods aimed at the prioritization of candidate disease genes –from exome sequencing data (Smedley *et al*, 2014; Itan *et al*, 2014) or exploiting the guilty-by-association principle (Wu *et al*, 2014a; Guney & Oliva, 2012)– or at the identification of driver genes from somatic mutations across cohorts of tumors could benefit from this knowledge to improve their performance. For instance, the aforementioned differences in the sensitivity of driver, Autosomal

dominant and Autosomal recessive genes to likely deleterious germline variants may refine approaches like the one proposed by (Petrovski *et al*, 2013), based on the residual variation intolerance score (RVIS) to quantify gene intolerance to functional mutations, genome-wide, or (Shyr *et al*, 2014), that ranked genes based on their frequency of rare non-synonymous/splice-site variants in general populations. Our results can also be applied to a scoring system of gene-disease associations inferred through text mining. For example, such scoring system would take into account the classification of genes into cartographic roles to weight its decisions. Higher scores would be awarded to candidates that resemble more the features expected for genes connected to the type of disease in the association.

Materials and Methods

Assembling the Protein Interaction Networks

Protein interaction data were retrieved from HIPPIE (Human Integrated Protein–Protein Interaction rEference, (Schaefer *et al*, 2012)), BIANA (Garcia-Garcia *et al*, 2010), BioGRID (Stark *et al*, 2006), IntAct (Orchard *et al*, 2014), IrefIndex (Razick *et al*, 2008) and data from Human Interactome Project (HBI) (Rual *et al*, 2005; Rolland *et al*, 2014). All files were obtained in December, 2014.

The source, number of genes and set of interactions in each PIN, as well as the overlaps between them are shown in Supplementary Fig S1. The giant component of the six PINs contained 7,000-12,000 proteins, and 25,000-70,000 interactions after filtering to retain only high-confidence interactions (Janjić & Pržulj, 2012). The overlap between pairs of PINs (Jaccard's index) ranged between 0.5 and -0.7 for genes and 0.1-0.35 for interactions. Similarly small overlaps have been reported before (Wodak *et al*, 2013; Lopes *et al*, 2011; Rolland *et al*, 2014; Jensen & Bork, 2008), and are mainly attributed to the complementary nature of different protein interactions detection methods (Jensen & Bork, 2008). Throughout all the paper we illustrate the results with HIPPIE PIN, and we show the results for the rest of the networks as Supplementary material.

HIPPIE: We used only the interactions with a score greater than, or equal to 0.72, corresponding to the 25% of the highest scoring interactions, as suggested by the authors (Schaefer *et al*, 2012). Interactions involving genes UBC, SUMO1, SUMO2, SUMO3, SUMO4, RPS27A, UBA52 were excluded, following the criteria in (Rolland *et al*, 2014). The file was downloaded from http://cbdm.mdc-berlin.de/tools/hippie/index.php, v1.7.

BIANA: We did not include interactions obtained by methods producing co-complex. We removed interactions involving genes UBC, SUMO1, SUMO2, SUMO3, SUMO4, RPS27A, UBA52.

The data was obtained from http://sbi.imim.es/web/index.php/research/servers/biana

IntAct: We only kept interactions annotated as human. Using the same quality criteria as in Hippie HC, we kept the 25% of the top scoring interactions. The data was obtained from ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact.zip

Human Binary Interactome (HBI): We merged the files, HI-I-05, HI-II-14, Lit-BM-13, downloaded from http://interactome.dfci.harvard.edu/H_sapiens/index.php

iRefIndex: We only kept interactions annotated as human, and detected by more than one method. We removed interactions involving genes UBC, SUMO1, SUMO2, SUMO3, SUMO4, RPS27A, and UBA52. The data was obtained from

http://irefindex.org/download/irefindex/data/archive/release_13.0/psi_mitab/MITAB2.6/9606.mitab.08122013.txt.zip

BioGRID: We only kept interactions annotated as human, and reported by at least one experimental system, or at least two different publications. We removed pairs containing genes UBC, SUMO1, SUMO2, SUMO3, SUMO4, RPS27A, and UBA52. The data was obtained from http://thebiogrid.org/downloads/archives/Release%20Archive/BIOGRID-3.2.120/BIOGRID-ALL-3.2.120.tab2.zip.

Mapping disease genes to the networks

Disease Genes: We used DisGeNET (http://disgenet.org, version 2.1, 5/5/2014), as source of disease genes (Piñero *et al*, 2015). We filter the disease phenotypes contained in DisGeNET using their UMLS® Metathesaurus® semantic type and their MeSH Class. We kept only semantic types T019, T047, T048, and T191 corresponding to Congenital Abnormality, Disease or Syndrome, Mental or Behavioral Dysfunction, and Neoplastic Process, respectively, and disease classes from C01 to C20, C25, F01 and F03 (see Supplementary Fig S8 for details of the MeSH disease class). We restricted the disease genes to those reported by CURATED sources.

Cancer-related genes: Genes annotated to diseases classified as Neoplasms in DisGeNET (curated sources). The involvement of these genes in neoplastic diseases varies: they might be driver genes, or may have been reported as altered in some tumor type and behave as passengers.

Cancer Driver Genes: We downloaded Cancer Genes Census genes on August 25, 2014 (Futreal *et al*, 2004). A second driver list of 464 genes was obtained from (Rubio-Perez *et al*, 2015). After merging both lists, there were 781 driver genes.

Mendelian disease genes and inheritance modes: Mendelian disease genes were retrieved from OMIM (Amberger *et al*, 2009) on August, 2013. We excluded genes with associations to disease marked as susceptibility. Inheritance modes of the genes were obtained from two datasets 1) (Singh *et al*, 2014), who manually curated inheritance information from OMIM and from (Blekhman *et al*, 2008) 2) from (Hao *et al*, 2014b) who also manually curated inheritance information from OMIM. We removed the genes with contradictory annotations. In total, we obtained 1,153 AR genes and 954 AD genes.

Complex Disease genes: We manually compiled a list of complex diseases, using DisGeNET diseases (CURATED). From the Disease Genes List, we excluded Bacterial Infections and Mycoses (C01), Virus Diseases (C02), Parasitic Diseases (C03), Neoplasms (C04) MeSH disease classes, general disease terms (such as Brain diseases, Kidney diseases, Autoimmune diseases), phenotypes, signs and symptoms, and Mendelian diseases. At the end, we obtain 2863 genes associated to 644 complex diseases.

Mutation rate and Functional Impact assessment

Germline variants were detected across 60,706 exomes. The data was downloaded from Exome Aggregation Consortium (ExAC), Cambridge, MA (http://exac.broadinstitute.org)

[accessed November, 2014]. We used Gencode annotations (ENSEMBL) to find canonical transcripts and a deleteriousness score was computed using Combined Annotation Dependent Depletion

(CADD) scores http://cadd.gs.washington.edu/ (Kircher *et al*, 2014). Extremely rare variants (less than 10^{-5}) were excluded from the analysis. Only mutations in coding regions, of type synonymous, non-synonymous, splicing site, stop gain and stop lose and frameshift were analyzed. A CADD (scaled) score of more than 15 was used to classify a variant as high impacting. We then calculated the high impacting to synonymous ratio for all genes (17,438 genes). We did not include in the analysis genes with less than four synonym or protein sequence affecting variants (missense, stop gained, stop lost, frameshift, splice donor and splice acceptor variant)

Positional Analysis of Mutations

We obtained the deleterious variants associated to disease from the Human polymorphisms and disease mutations file (http://www.uniprot.org/docs/humsavar, release 2015_01 of 07-Jan-2015) from UniProt (The UniProt Consortium, 2014) and from ClinVar file ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz downloaded on January, 2015. We kept variants annotated as "disease" in UniProt and as "Pathogenic" in ClinVar (Landrum *et al*, 2014). We obtained cancer mutations from (Rubio-Perez *et al*, 2015). We kept genes where at least three mutations of both types were found. We obtained the length of the coding sequence from UniProt.

Network Clustering

In order to assign roles to the genes, we partitioned the PIN in clusters using *Infomap* algorithm (Rosvall & Bergstrom, 2008). *Infomap* algorithm was performed with the code from: http://www.tp.umu.se/~rosvall/code.html.

Cartography

According to (Guimerà & Amaral, 2005b) the genes were assigned to one of the following roles: ultra-peripheral nodes, peripheral, non-hub connector, non-hub kinless, provincial hubs, connector hubs, kinless hubs. These seven different roles are heuristically defined, using their localization in the different regions of the z–P parameter space, where z (within-module degree) and P (participation coefficient) are calculated according to:

$$z_i = \frac{k_i - \bar{k}_{si}}{\sigma \bar{k}_{si}}$$

where k_i is the number of links of node i to other nodes in its module, k_{s_i} is the mean degree of all nodes in cluster s_i, and σk_{s_i} is the standard deviation of the degree in the cluster s_i

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{is}}{k_i} \right)^2$$

where k_{is} is the number of links of node i to nodes in the module s, and k_i is the total degree of node i.

Nodes with $z > 2.5$ are classified as module hubs and nodes with $z < 2.5$ as non-hubs. Both hub and non-hub nodes are then further characterized by using their participation coefficient.

Non-hub nodes can be divided into four different roles: (R1) ultra-peripheral nodes; that is, nodes with all their links within their module ($P \leq 0.05$); (R2) peripheral nodes; that is, nodes with most links within their module ($0.05 < P \leq 0.62$); (R3) non-hub connector nodes; that is, nodes with many links to other modules ($0.62 < P \leq 0.80$); and (R4) non-hub kinless nodes; that is, nodes with links homogeneously distributed among all modules ($P > 0.80$). Similarly, hub nodes are assigned to: (R5) provincial hubs; that is, hub nodes with the vast majority of links within their module ($P \leq 0.30$); (R6) connector hubs; that is, hubs with many links to most of the other modules ($0.30 < P \leq 0.75$); and (R7) kinless hubs; that is, hubs with links homogeneously distributed among all modules ($P > 0.75$).

Panther Database

We downloaded the file containing family/subfamily name, and the molecular function, biological process, and pathway classifications corresponding to Release 9.0 (Mi & Thomas, 2009)

ftp://ftp.pantherdb.org//hmm_classifications/current_release/PANTHER9.0_HMM_classifications

and mapped the UniProt identifiers to Entrez gene identifiers.

Network and Statistical Analysis

The network analysis was carried out using R (version 3.1.0) and the iGraph Library (version igraph_0.7.1) (Csardi & Nepusz, 2006). We used package Gostats_2.30.0 (Falcon & Gentleman, 2007) to perform the Molecular Function GO enrichment analysis. Other statistical test, such as Fisher and Mann Whitney U test were also performed in R. All multiple testings were corrected using Benjamini & Hochberg method (Benjamini & Hochberg, 1995)

Degree control for role enrichment estimation

A bootstrapping procedure was devised to control the node's degree distribution confounding factor for the role enrichment analysis. For each enrichment test we considered an ensemble of 1,000 control random gene-sets having the same degree distribution than the genes under study (we disregarded from the analysis the top 5% of nodes presenting the highest degree values, i.e. $k>50$). A p-value level was assigned according to the number of random realizations displaying the same or larger effects (over/under representation significance) than the ones observed in the original data. Each random realization was built blindly selecting genes from pools of given degree levels in order to follow the degree distribution displayed by the original gene set.

Statistical significance of gene sets features

For each gene set (AD, AD_{ND}, AR, AR_{ND} and driver), we generated 10,000 randomly selected samples of genes from the network of the same size of the gene set. Then, we computed the mean value of each sampled feature (degree, betweenness, clustering coefficient, participation coefficient, within-module degree, and HS Ratio) for the 10,000 randomizations. From this distribution of means a z-score was calculated for every gene set and feature pair.

Figures

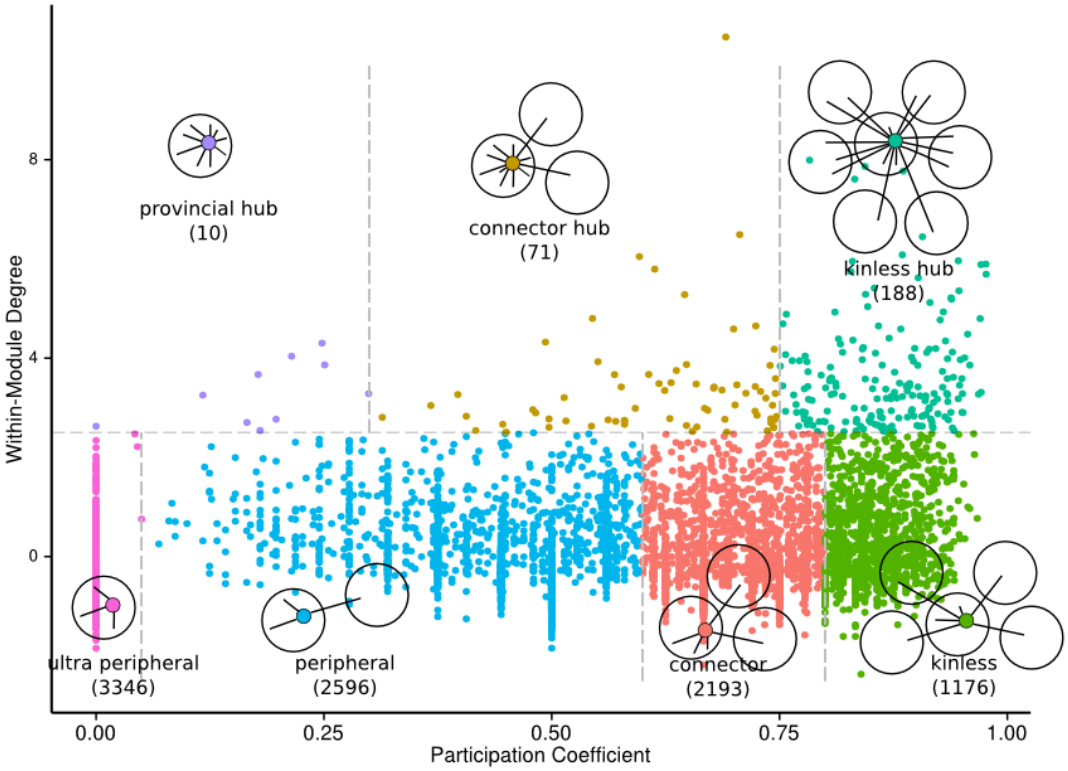

Figure 1: Cartographic partition of the nodes in the human protein interaction network (HIPPIE) in the z-P plane. The cartographic roles are represented with different colors. In parenthesis, we show the number of proteins in each role. Dashed lines in the figure delineate regions corresponding to the seven cartographic roles. We show a schematic representation of the type of connection for each role.

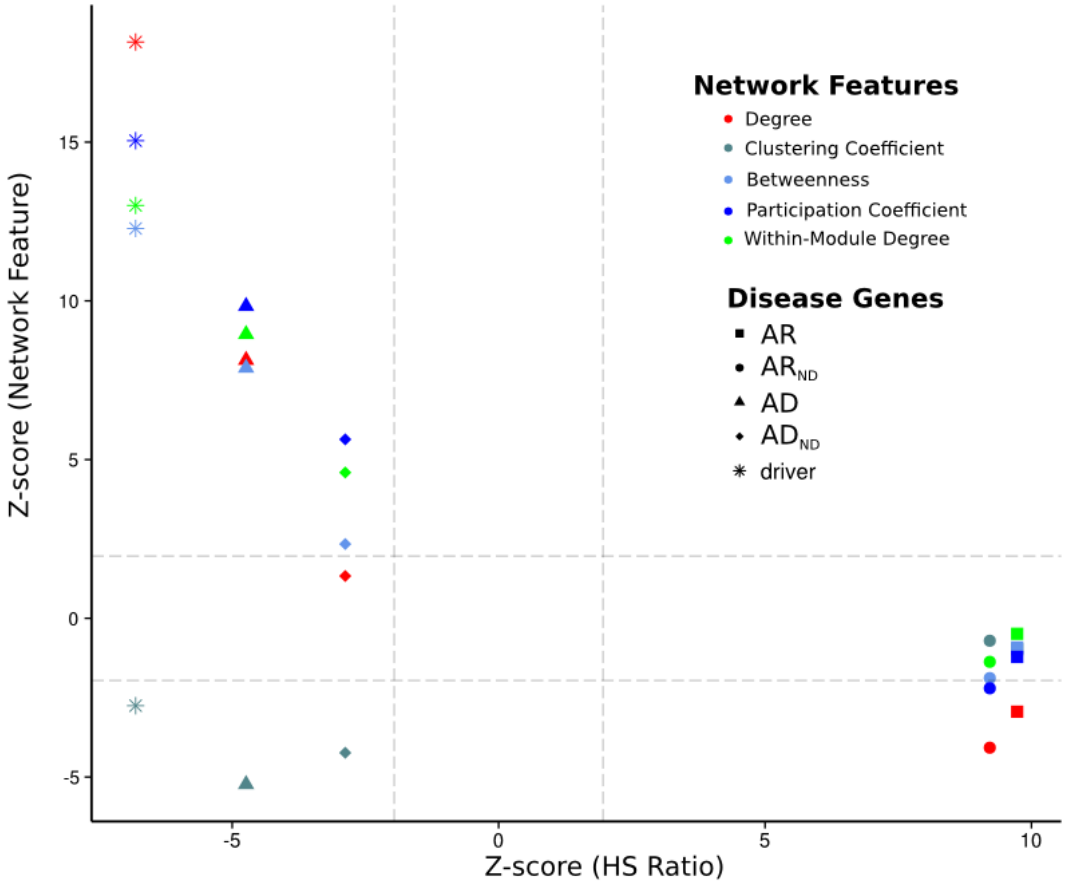

Figure 2: Relationship between the network features (degree, betweenness, clustering coefficient, participation coefficient, and within-module degree) and the HS Ratio for the disease gene sets in HIPPIE. We plot the z-scores resulting from 10000 randomizations. AD: Autosomal Dominant, AD_{ND}: AD genes without driver genes, AR: Autosomal Recessive, AR_{ND}: AR genes without driver genes, driver: cancer driver genes.

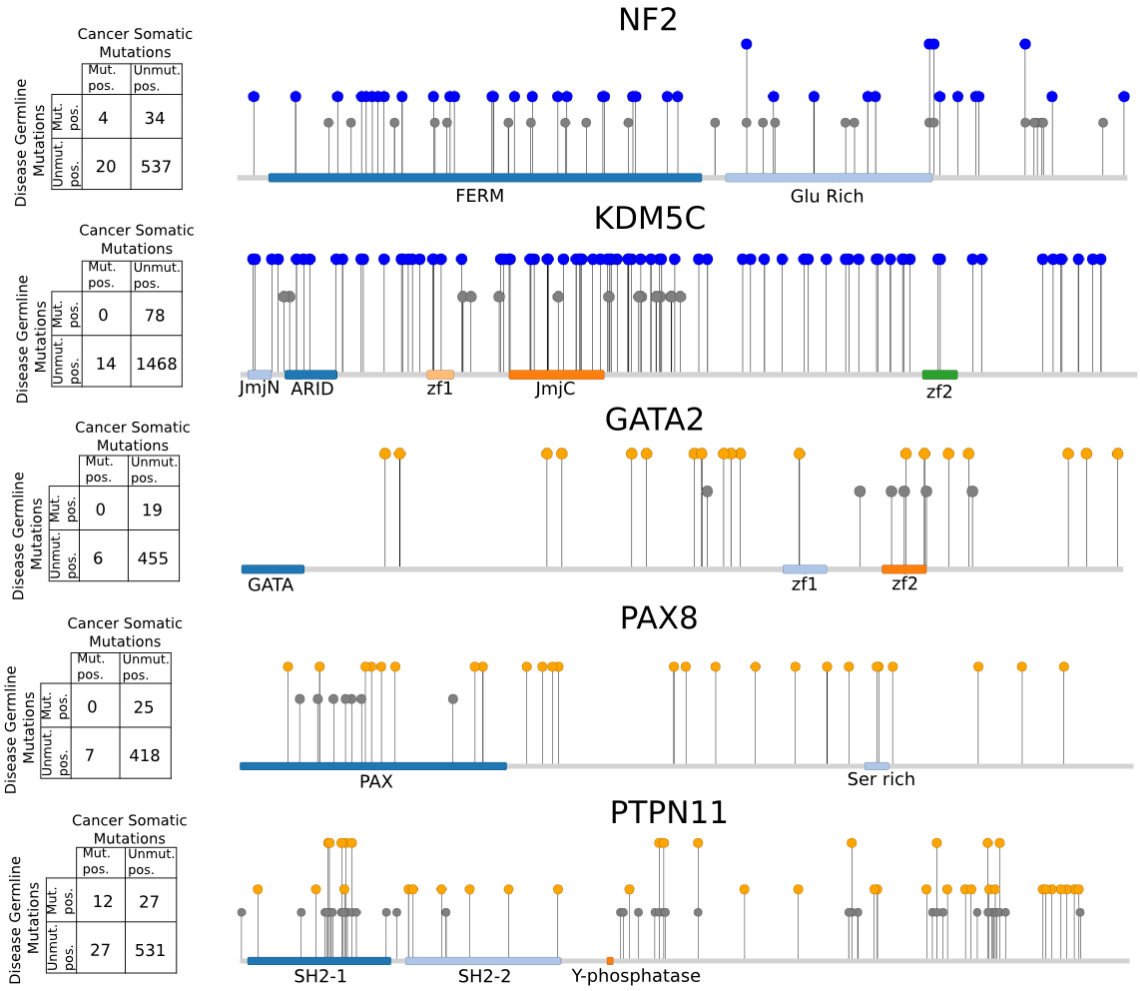

Figure 3: Examples of distribution of disease-associated germline and cancer somatic mutations in genes NF2, KDM5C, GATA2, PAX8, and PTPN11. In panel A, the contingency table of the Fisher test is shown. Mut pos: mutated position, Unmut pos: position where there are not annotated mutations. In panel B, a schematic representation of the position of the germline and somatic mutations is shown.

Tables

Table 1: Overrepresentation of the cartographic roles in each disease gene set in HIPPIE.

Cartographic role	DG	CG	NCDG	CxDG	$CxDG_{NC}$	MG	MG_{NC}
kinless hub	1,84E-07	3,75E-10	1	1,82E-09	0,175	0,122	1
connector hub	0,100	0,137	0,570	0,002	0,122	1	1
provincial hub	0,076	0,134	0,627	1	1	0,065	0,562
kinless	2,41E-26	2,55E-38*	1	1,76E-10	1	1,84E-07	1
connector	3,41E-06	1,82E-09*	1	0,142	1	0,016	0,627
peripheral	1	1	0,562	1	0,172	1	1
ultra-peripheral	1	1	0,562	1	1	1	0,122

* set of genes that remains significant under the bootstrap analysis

We show the p-values of the Fisher test, corrected by multiple testing according to Benjamini & Hochberg method. DG: all disease genes, NCDG: non-cancer disease genes, CG: cancer genes, CxDG: complex disease genes, $CxDG_{NG}$: CxDG without cancer genes, MG: Mendelian disease genes, MG_{NG}: MG without cancer genes.

Table 2 Overrepresentation of the cartographic roles in each disease gene set in HIPPIE.

Cartographic role	AD	AD_{ND}	AR	AR_{ND}	driver
kinless hub	3,51E-05	0,012	1	1	3,51E-05
connector hub	1	1	1	1	0,003
provincial hub	0,103	1	0,343	0,329	0,097
kinless	1,32E-10*	0,002*	0,868	1	4,15E-28*
connector	0,015	0,103	0,932	0,932	0,005
peripheral	1	1	1	1	1
ultra-peripheral	1	1	0,281	0,063	1

* set of genes that remains significant under the bootstrap analysis

We show the p-values of the Fisher test, corrected by multiple testing according to Benjamini & Hochberg method.AD: Autosomal Dominant, AD_{ND}: AD genes without driver genes, AR: Autosomal Recessive, AR_{ND}: AR genes without driver genes, driver: cancer driver genes.

Table 3: Average high-impact to synonymous ratio (HS Ratio) of the disease gene sets.

Disease gene set	N	HS Ratio	z-score	p-value	Corrected p-value
driver	691	0,759	-6.881	5,93E-12	9,89E-12
AD	750	0,809	-4.720	2,36E-06	2,95E-06
AD_{ND}	589	0,838	-2.462	1,38E-02	1,38E-02
AR	684	1.108	9.690	3,34E-22	1,67E-21
AR_{ND}	641	1.112	8.754	2,06E-18	5,15E-18

We show the number of genes in each set (N) that maps to HIPPIE, and the z-score resulting of 10,000 randomizations. The p-values are computed from the z-score. AD: Autosomal Dominant, AD_{ND}: AD genes without driver genes, AR: Autosomal Recessive, AR_{ND}: AR genes without driver genes, driver: cancer driver genes.

Supplementary Figures

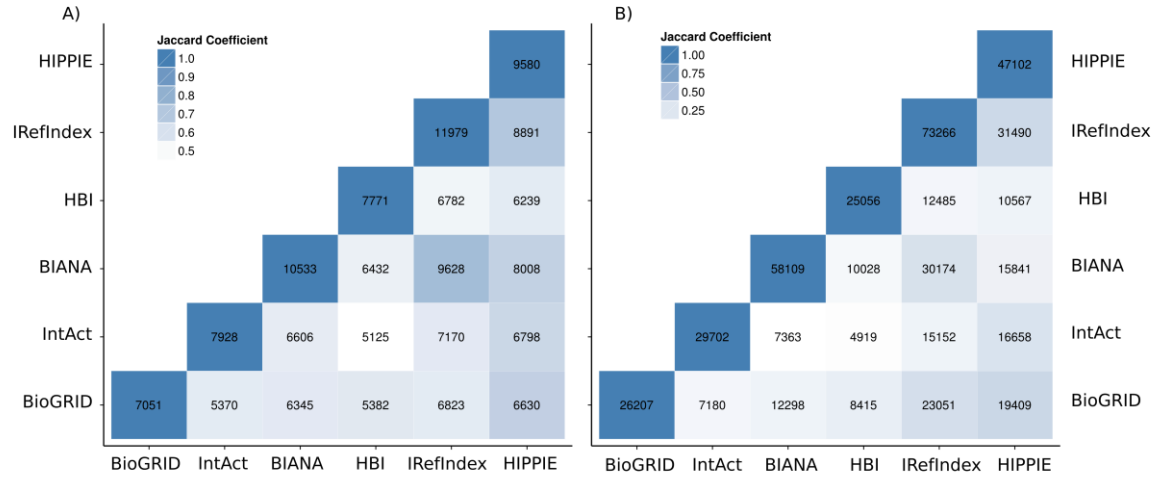

Figure S1: Overlaps between the PINs in terms of proteins (Panel A) and interactions (Panel B), measured with the Jaccard coefficient.

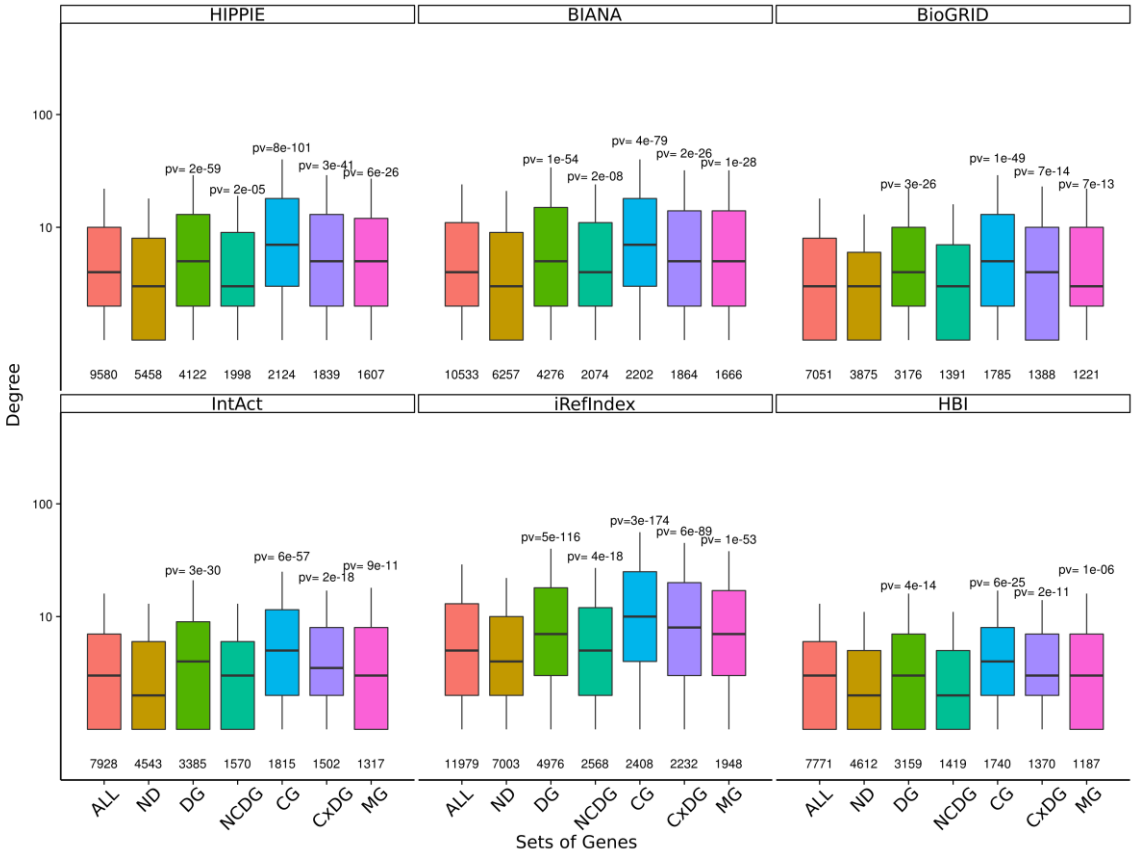

Figure S2: Degree distributions of the disease gene sets in the six PINs. We show the p-values of Man-Whitney test resulting from the comparisons of each group of disease genes with non-disease genes, corrected by multiple testing according to Benjamini & Hochberg method. The sets of disease genes are: ALL (all genes), ND (non-disease genes), DG (all disease genes), NCDG (non-cancer disease genes), CG (cancer genes), CxDG (complex disease genes), MG (Mendelian disease genes).

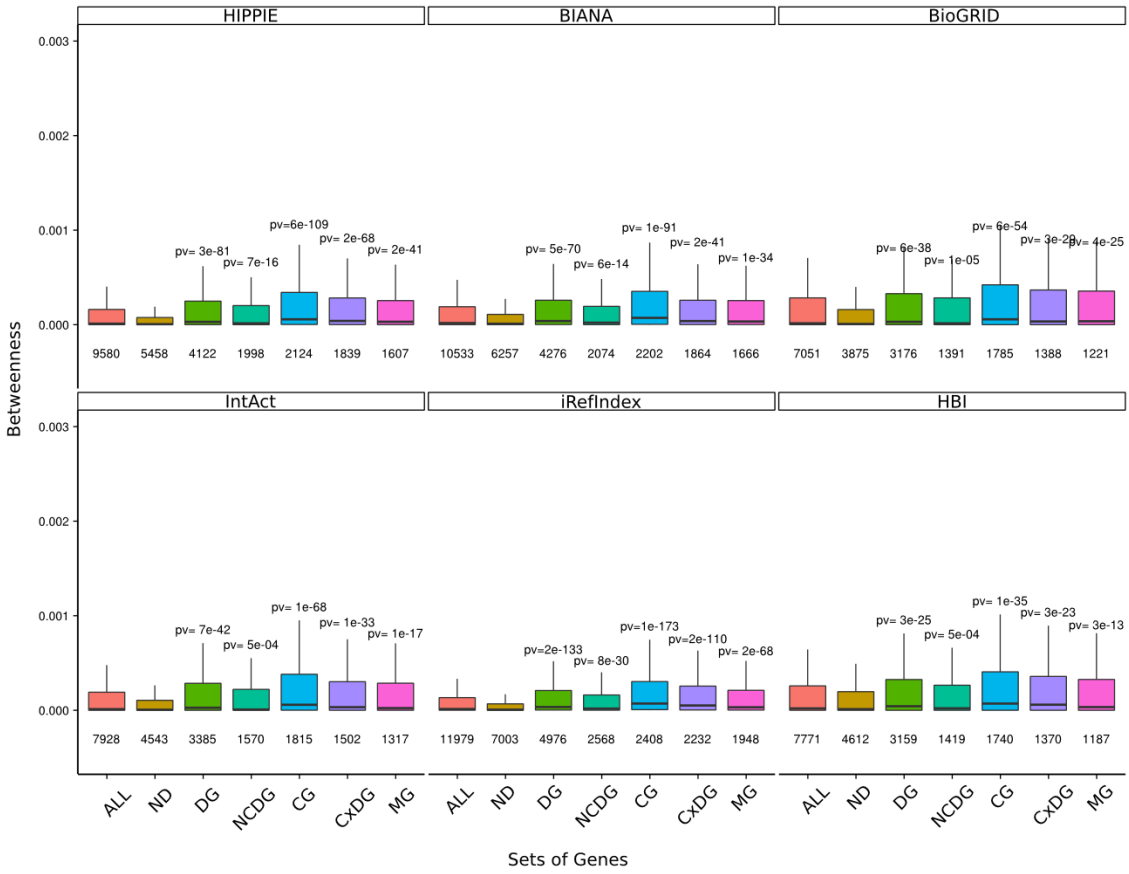

Figure S3: Distribution of betweenness of the disease gene sets in the six PINs. We show the p-values of Man-Whitney test resulting from the comparisons of each group of disease genes with non-disease genes, corrected by multiple testing according to Benjamini & Hochberg method. The sets of disease genes are: ALL (all genes), ND (non-disease genes), DG (all disease genes), NCDG (non-cancer disease genes), CG (cancer genes), CxDG (complex disease genes), MG (Mendelian disease genes).

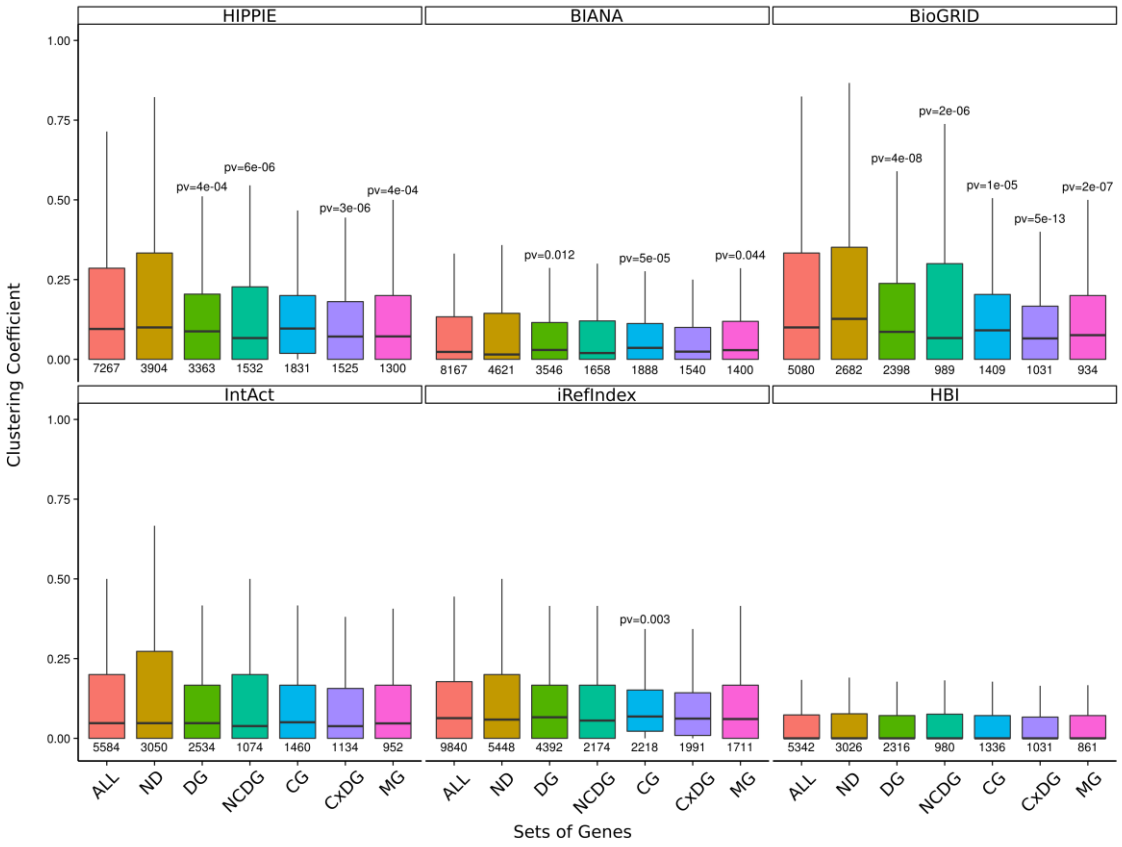

Figure S4: Distribution of the clustering coefficient (CC) of different sets of disease genes in the six PINs. We show the p-values of Man-Whitney test resulting from the comparisons of each group of disease genes with non-disease genes, corrected by multiple testing according to Benjamini & Hochberg method. The sets of disease genes are: ALL (all genes), ND (non-disease genes), DG (all disease genes), NCDG (non-cancer disease genes), CG (cancer genes), CxDG (complex disease genes), MG (Mendelian disease genes).

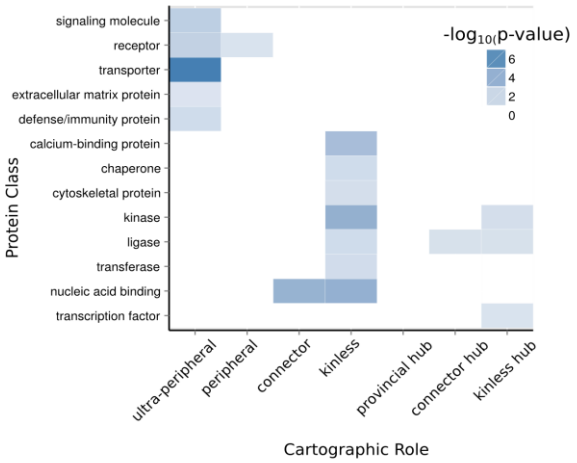

Figure S5: Overrepresentation of the cartographic roles in each Panther protein class. We show the p-value of the exact Fisher test, corrected for multiple testing by Benjamini & Hochberg

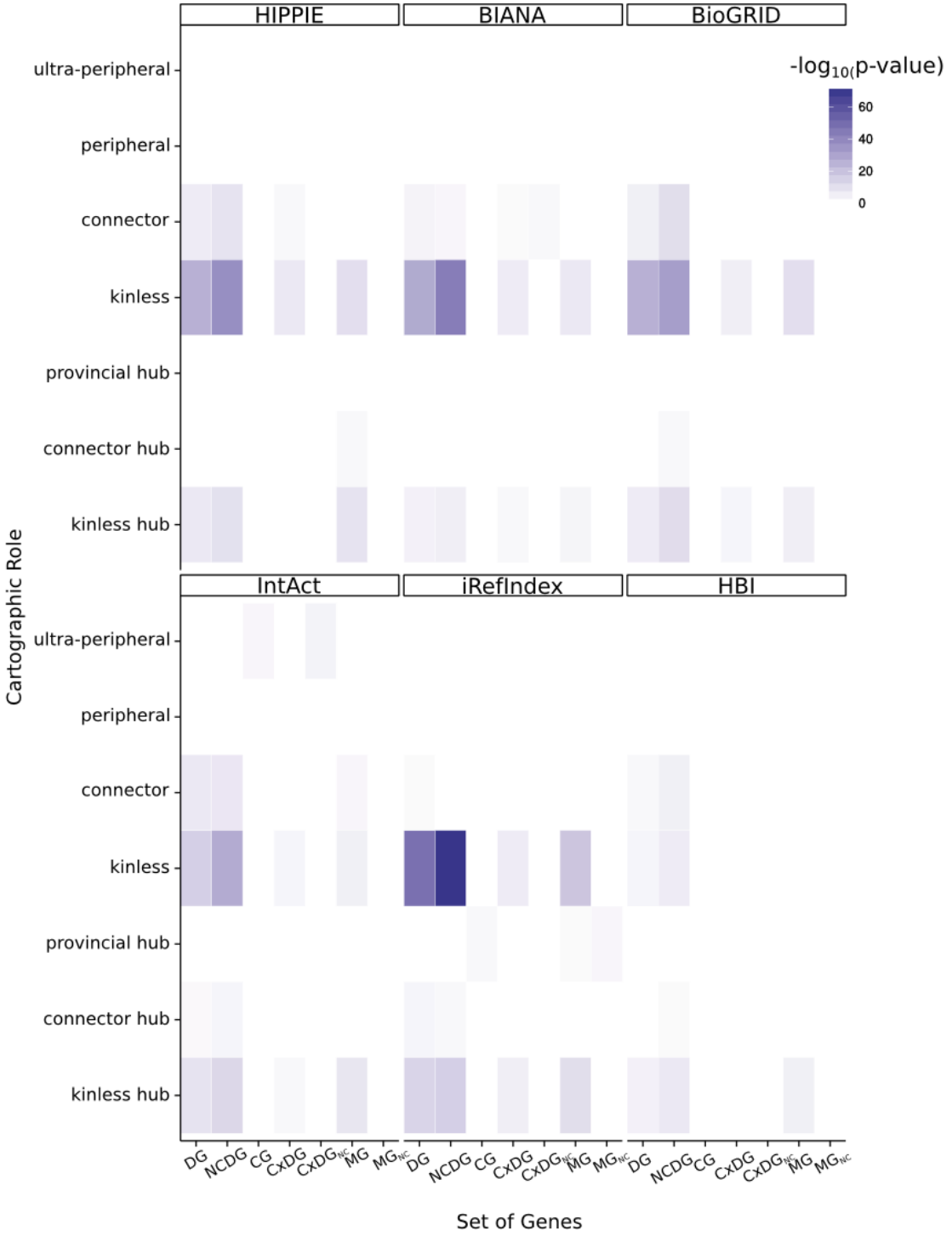

Figure S6: Overrepresentation of the cartographic roles in each disease gene set in the six PINs. DG: all disease genes, NCDG: non-cancer disease genes, CG: cancer genes, CxDG: complex disease genes, $CxDG_{NC}$: CxDG without cancer genes, MG: Mendelian disease genes, MG_{NC}: MG without cancer genes. The color is proportional to logarithm of p-value of the exact Fisher test, corrected for multiple testing by Benjamini & Hochberg.

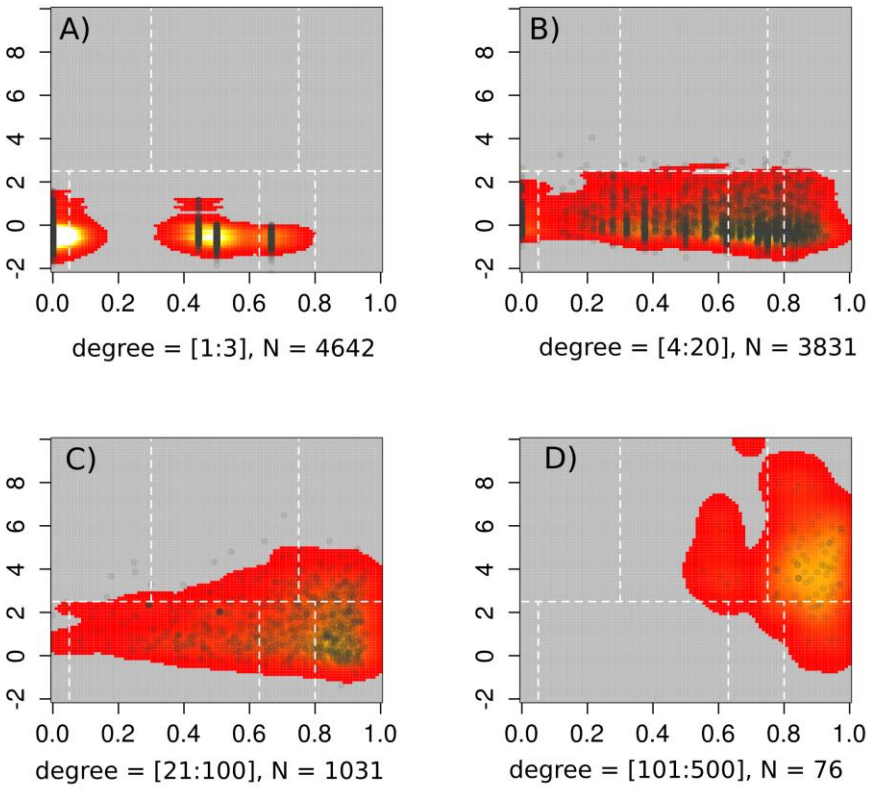

Figure S7: The distribution of PIN nodes over the z-P plane for nodes with degree values within the [1,3], [4,20], [21,100] and [101,500] intervals are shown in panels (A), (B), (C) and (D) respectively. A color-coded kernel density estimation was calculated for the 4642, 3831, 1031 and 76 gene nodes included in each panel. Dashed lines in the figures delineate regions corresponding to the seven cartographic roles.

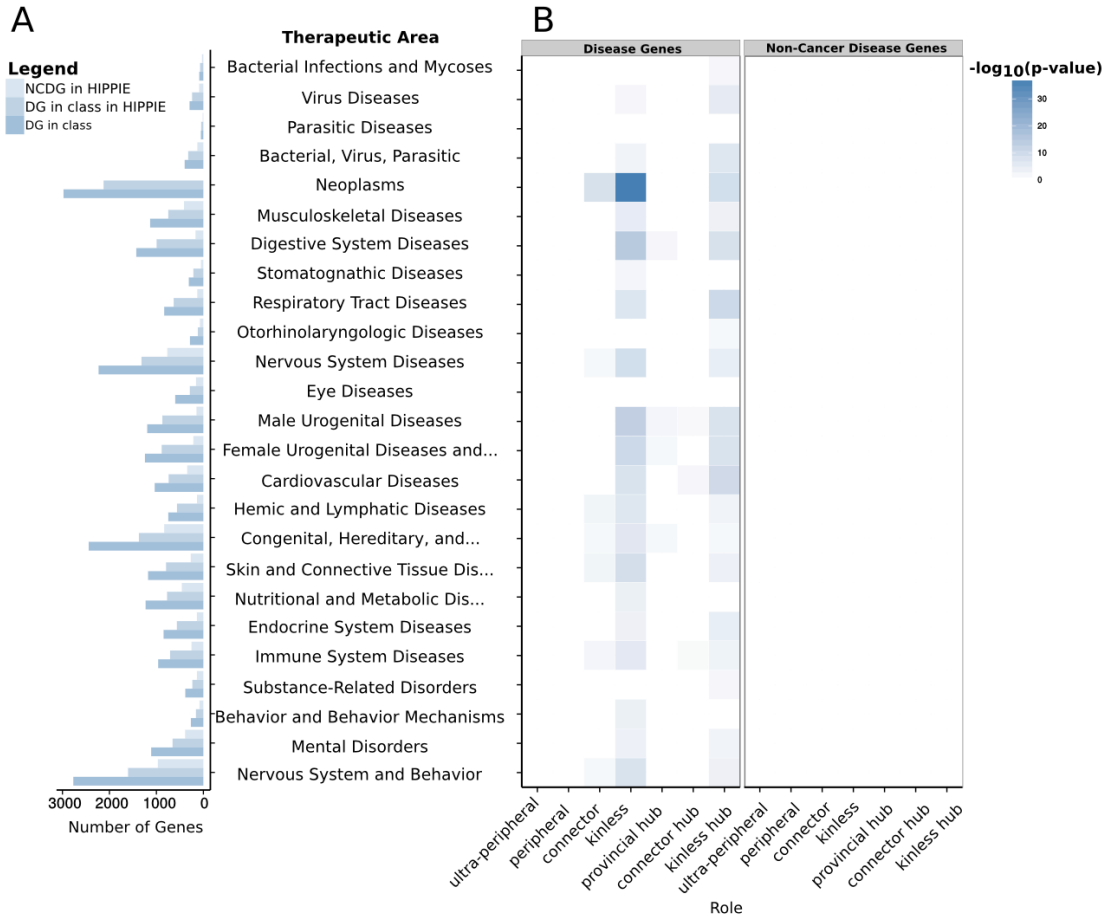

Figure S8: Overrepresentation of the different sets of genes belonging to the different MeSH disease classifications in each cartographic role in HIPPIE. Panel A) Distribution of genes in each MeSH category. DG in class: Total number of genes in each MeSH category, DG in class in HIPPIE: DG in class that maps to HIPPIE, NCDG: non-cancer disease genes in class that maps to HIPPIE. Panel B) Overrepresentation of the cartographic roles in each disease gene set in HIPPIE. The color is proportional to logarithm of the p-value of the exact Fisher test, corrected for multiple testing by Benjamini & Hochberg.

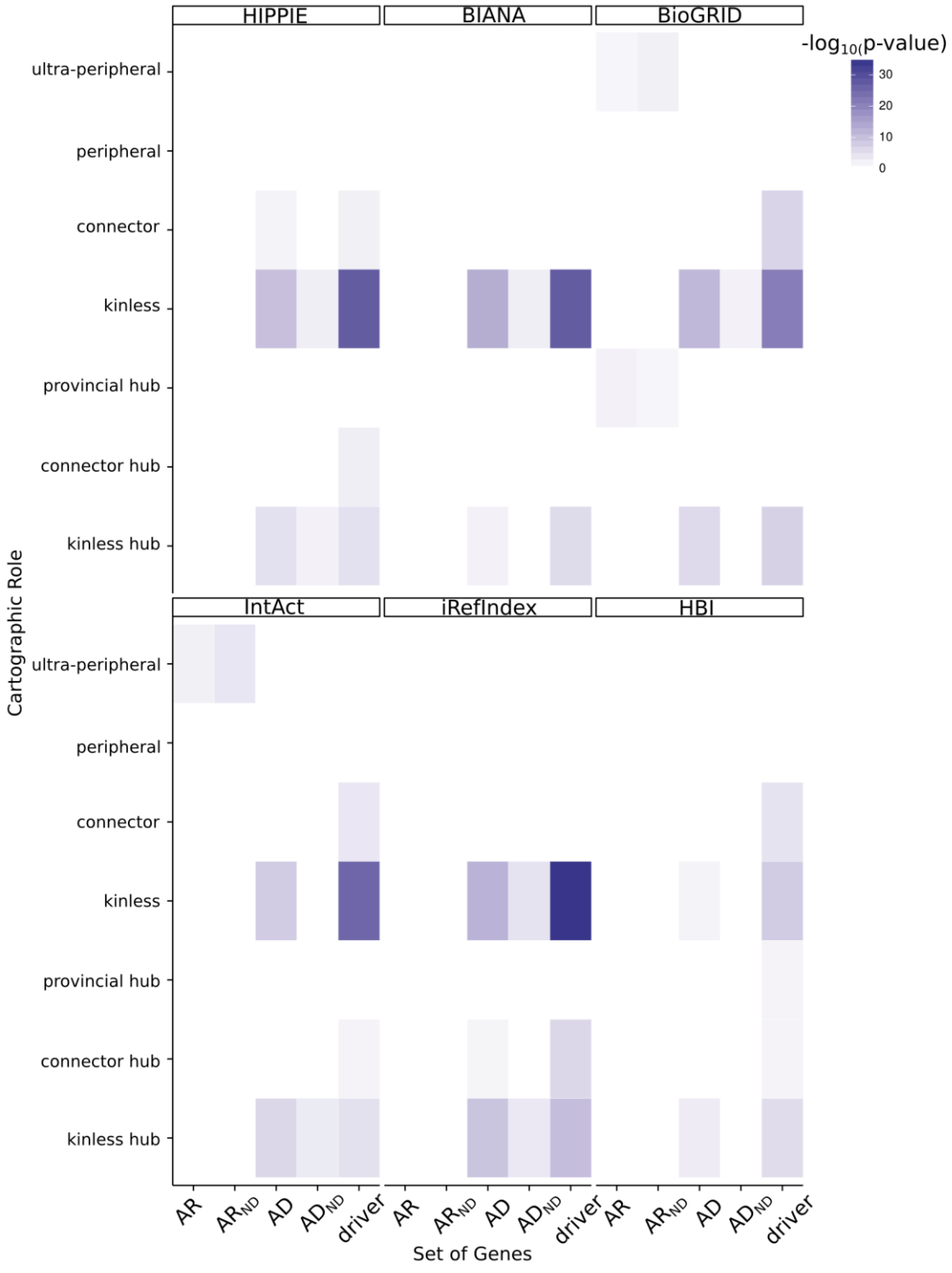

Figure S9: Overrepresentation of the cartographic roles in each disease gene set in the six PINs. AD: Autosomal Dominant, AD_{ND}: AD genes without driver genes, AR: Autosomal Recessive, AR_{ND}: AR genes without driver genes, driver: cancer driver genes. The color is proportional to logarithm of p-value of the exact Fisher test, corrected for multiple testing by Benjamini & Hochberg.

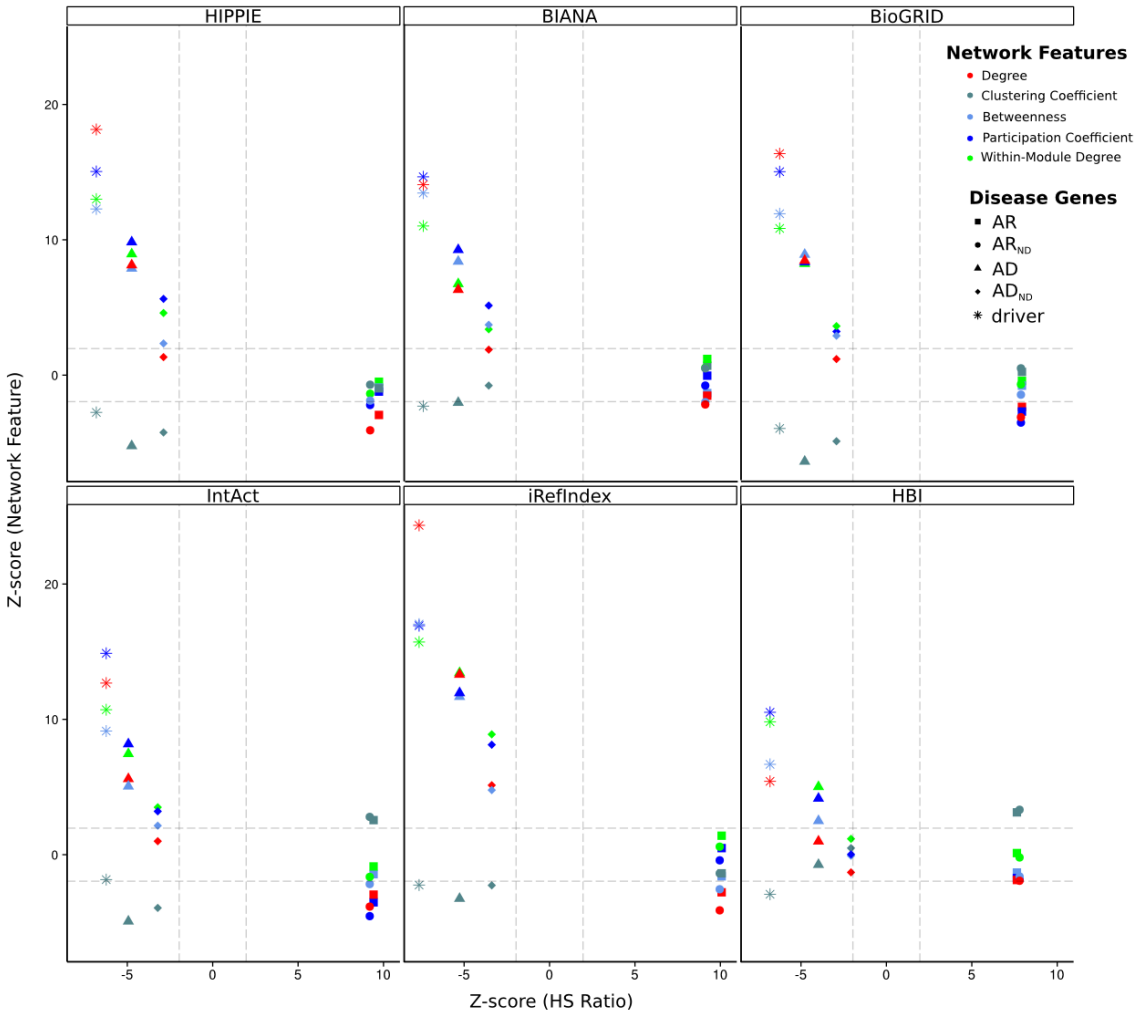

Figure S10: Relationship between the network features (degree, betweenness, clustering coefficient, participation coefficient, and within-module degree) and the HS Ratio for the disease gene sets across all PINs. We plot the z-scores resulting from 10,000 randomizations. AD: Autosomal Dominant, AD_{ND}: AD genes without driver genes, AR: Autosomal Recessive, AR_{ND}: AR genes without driver genes, driver: cancer driver genes.

Supplementary Tables

Table S1: Total number of genes in the different disease classification, and coverage in each PIN.

Gene Set	Abbr.	size	HIPPIE	BioGRID	IntAct	BIANA	HBI	iRefIndex
Disease Genes	DG	7412	55,6	42,8	45,7	57,7	42,6	67,1
Cancer related Genes	CG	2977	71,3	60	61	74	58,4	80,9
Non-cancer disease genes	NCDG	4435	45,1	31,4	35,4	46,8	32	57,9
Mendelian Genes	MG	3114	51,6	39,2	42,3	53,5	38,1	62,6
Complex Disease Genes	CxDG	2863	64,2	48,5	52,5	65,1	47,9	78
Autosomal Recessive Disease Genes	AR	1153	59,3	41,5	47,4	65,6	39	78,5
Autosomal Dominant Disease Genes	AD	954	78,6	63,2	63,5	77,1	64,4	89,2
Driver Genes	drivers	781	88,5	81,6	78,7	86,8	76,3	92,7

Table S2: Percentage of genes in each cartographic role in all PINs.

role	HIPPIE	BioGRID	IntAct	BIANA	HBI	iRefIndex
ultra peripheral	34,9	42,2	42,3	29,9	43,0	25,4
peripheral	27,1	27,0	26,1	25,6	27,4	25,5
connector	22,9	18,9	20,6	25,1	20,7	26,5
kinless	12,3	8,7	7,9	17,0	6,2	19,9
provincial hub	0,1	0,2	0,1	0,1	0,1	0,1
connector hub	0,7	1,1	1,2	0,5	1,2	0,7
kinless hub	2,0	1,8	1,8	1,8	1,4	1,9
totals	100,0	100,0	100,0	100,0	100,0	100,0

Acknowledgements

The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at

http://exac.broadinstitute.org/about.

Funding

This work was supported by UBACyT (20020130100582BA), Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (CP10/00524 and PI13/00082), the Innovative Medicines Initiative Joint Undertaking (115002 (eTOX), 115191 [Open PHACTS]), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme [FP7/2007-2013] and EFPIA companies' in kind contribution. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

Author Contributions

Conceived and designed the experiments: AC LIF. Performed the experiments: AB JP AGP AC LIF. Analyzed the data: AB JP AGP AC LIF. Wrote the paper: AC LIF JP AGP.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Amberger J, Bocchini CA, Scott AF & Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37:** D793–6

Barrenas F, Chavali S, Holme P, Mobini R & Benson M (2009) Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS One* **4:** 6

Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57:** 289–300

Berenstein AJ, Piñero J, Furlong LI & Chernomoretz A (2015) Mining the modular structure of protein interaction networks. *PLoS One* **10:** e0122477

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM & Przeworski M (2008) Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18:** 883–9

Cai JJ, Borenstein E & Petrov DA (2010) Broker genes in human disease. *Genome Biol. Evol.* **2:** 815–25

Chen J, Aronow BJ & Jegga AG (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* **10:** 73

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, et al (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.*

Collins A (2015) The genomic and functional characteristics of disease genes. *Brief. Bioinform.* **16:** 16–23

Csardi G & Nepusz T (2006) The igraph software package for complex network research. *InterJournal* **Complex Sy:** 1695

Falcon S & Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* **23:** 257–8

Feldman I, Rzhetsky A & Vitkup D (2008) Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U. S. A.* **105:** 4323–4328

Fernández-Medarde A & Santos E (2011) Ras in cancer and developmental diseases. *Genes Cancer* **2:** 344–58

Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M & Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78:** 1011–25

Furlong LI (2012) Human diseases through the lens of network biology. *Trends Genet.* **null:**

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N & Stratton MR (2004) A census of human cancer genes. *Nat. Rev. Cancer* **4:** 177–83

Garcia-Alonso L, Jiménez-Almazán J, Carbonell-Caballero J, Vela-Boza A, Santoyo-López J, Antiñolo G & Dopazo J (2014) The role of the interactome in the maintenance of deleterious variability in human populations. *Mol. Syst. Biol.* **10:** 752

Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J & Oliva B (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics* **11:** 56

Ghersi D & Singh M (2013) Disentangling function from topology to infer the network properties of disease genes. *BMC Syst. Biol.* **7:** 5

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M & Barabási A-L (2007) The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* **104:** 8685–90

Guimerà R & Amaral LAN (2005a) Functional cartography of complex metabolic networks. *Nature* **433:** 895–900

Guimerà R & Amaral LAN (2005b) Cartography of complex networks: modules and universal roles. *J. Stat. Mech. Online* **2005:** nihpa35573

Guney E & Oliva B (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS One* **7:** e43557

Hanahan D & Weinberg RA (2000) The Hallmarks of Cancer. *Cell* **100:** 57–70

Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144:** 646–74

Hao D, Li C, Zhang S, Lu J, Jiang Y, Wang S & Zhou M (2014a) Network-based analysis of genotype-phenotype correlations between different inheritance modes. *Bioinformatics* **30:** 3223–31

Hao D, Wang G, Yin Z, Li C, Cui Y & Zhou M (2014b) Systematic large-scale study of the inheritance mode of Mendelian disorders provides new insight into human diseasome. *Eur. J. Hum. Genet.*

Ideker T & Sharan R (2008) Protein networks in disease. *Genome Res.* **18:** 644–52

Itan Y, Mazel M, Mazel B, Abhyankar A, Nitschke P, Quintana-Murci L, Boisson-Dupuis S, Boisson B, Abel L, Zhang S-Y & Casanova J-L (2014) HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* **15:** 256

Janjić V & Pržulj N (2012) Biological function through network topology: a survey of the human diseasome. *Brief. Funct. Genomics* **11:** 522–32

Jensen LJ & Bork P (2008) Biochemistry. Not comparable, but complementary. *Science* **322:** 56–7

Jin W, Qin P, Lou H, Jin L & Xu S (2012) A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum. Mol. Genet.* **21:** 1611–24

Jonsson PF & Bates PA (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22:** 2291–7

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ & Ding L (2013) Mutational landscape and significance across 12 major cancer types. *Nature* **502:** 333–9

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM & Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46:** 310–5

Köhler S, Bauer S, Horn D & Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82:** 949–58

Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y & Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25:** 309–16

Lancichinetti A & Fortunato S (2009) Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80:** 056117

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM & Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42:** D980–5

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES & Getz G (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505:** 495–501

Lee I, Blom UM, Wang PI, Shim JE & Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21:** 1109–21

Li Z-C, Lai Y-H, Chen L-L, Xie Y, Dai Z & Zou X-Y (2014) Identifying and prioritizing disease-related genes based on the network topological features. *Biochim. Biophys. Acta* **1844:** 2214–2221

Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J-F, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási A-L, Vidal M & Zoghbi HY (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125:** 801–14

Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS & Kasif S (2007) Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* **3:** e96

Liu W, Pellegrini M & Wang X (2014) Detecting communities based on network topology. *Sci. Rep.* **4:** 5739

Lopes TJS, Schaefer M, Shoemaker J, Matsuoka Y, Fontaine J-F, Neumann G, Andrade-Navarro MA, Kawaoka Y & Kitano H (2011) Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* **27:** 2414–21

Luo J & Liang S (2015) Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data. *J. Biomed. Inform.* **53:** 229–36

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, et al (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* **508:** 469–76

Mi H & Thomas P (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol. Clift. Nj* **563:** 123–140

Navlakha S & Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26:** 1057–63

Nivit G, Shailendra S & C. AT (2014) Computational Disease Gene Prioritization: An Appraisal. *J. Comput. Biol.*

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, et al (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42:** D358–63

Oti M, Snel B, Huynen MA & Brunner HG (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43:** 691–8

Petrovski S, Wang Q, Heinzen EL, Allen AS & Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9:** e1003709

Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F & Furlong LI (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015:** bav028–bav028

Razick S, Magklaras G & Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9:** 405

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis A-R, et al (2014) A Proteome-Scale Map of the Human Interactome Network. *Cell* **159:** 1212–1226

Rosvall M & Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.* **105:** 1118–23

Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang L V, Wong SL, Franklin G, Li S, et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437:** 1173–8

Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A & Lopez-Bigas N (2015) In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27:** 382–396

Santiago JA & Potashkin JA (2013) Integrative network analysis unveils convergent molecular pathways in Parkinson's disease and diabetes. *PLoS One* **8:** e83940

Sarajlić A, Janjić V, Stojković N, Radak D & Pržulj N (2013) Network topology reveals key cardiovascular disease genes. *PLoS One* **8:** e71537

Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE & Andrade-Navarro MA (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One* **7:** e31826

Shyr C, Tarailo-Graovac M, Gottlieb M, Lee J, van Karnebeek C & Wasserman WW (2014) FLAGS, frequently mutated genes in public exomes. *BMC Med. Genomics* **7:** 64

Singh PP, Affeldt S, Malaguti G & Isambert H (2014) Human dominant disease genes are enriched in paralogs originating from whole genome duplication. *PLoS Comput. Biol.* **10:** e1003754

Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, Veldboer J, Zemojtel T & Robinson PN (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* **30:** 3215–22

Del Sol A, Balling R, Hood L & Galas D (2010) Diseases as network perturbations. *Curr. Opin. Biotechnol.* **21:** 566–71

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A & Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34:** D535–9

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L & Lopez-Bigas N (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3:** 2650

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q & Wrana JL (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27:** 199–204

The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42:** D191–8

Thomas PD (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31:** 334–341

Vanunu O, Magger O, Ruppin E, Shlomi T & Sharan R (2010) Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput. Biol.* **6:** 9

Veitia RA (2002) Exploring the etiology of haploinsufficiency. *Bioessays* **24:** 175–84

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA & Kinzler KW (2013) Cancer genome landscapes. *Science* **339:** 1546–58

Wachi S, Yoneda K & Wu R (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21:** 4205–8

Wilkie AO (1994) The molecular basis of genetic dominance. *J. Med. Genet.* **31:** 89–98

Wodak SJ, Vlasblom J, Turinsky AL & Pu S (2013) Protein-protein interaction networks: the puzzling riches. *Curr. Opin. Struct. Biol.* **23:** 941–53

Wu J, Li Y & Jiang R (2014a) Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **10:** e1004237

Wu X, Jiang R, Zhang MQ & Li S (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.* **4:** 189

Wu Y, Jing R, Jiang L, Jiang Y, Kuang Q, Ye L, Yang L, Li Y & Li M (2014b) Combination use of protein-protein interaction network topological features improves the predictive scores of deleterious non-synonymous single-nucleotide polymorphisms. *Amino Acids*

Xu J & Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* **22:** 2800–5

Zhong Q, Simonis N, Li Q-R, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, Yildirim M a, Yan H, Dricot A, Szeto D, Lin C, Hao T, Fan C, Milstein S, Dupuy D, et al (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5:** 321

Zhu X, Need AC, Petrovski S & Goldstein DB (2014) One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. *Nat. Neurosci.* **17:** 773–81

3.4 Network medicine analysis of COPD multimorbidities

Chronic obstructive pulmonary disease (COPD) has been the third leading causes of death in the world during the last decade, according to the World Health Organization (http://www.who.int/mediacentre/factsheets/fs310/en/). COPD is characterized by persistent airflow limitation that is usually progressive and associated with an enhanced chronic inflammatory response in the airways and the lung to noxious particles or gases (Vestbo et al. 2013). Several comorbidities have been described for COPD, and the clinical course and prognosis of COPD patients is greatly influenced by them. The COPD diseasome includes different disease groups, such as cardiovascular diseases (coronary artery disease), neoplasms (lung cancer), metabolic diseases (type II diabetes mellitus), psychiatric diseases (depression), among others (Divo et al. 2012). Some of the causes of these comorbidities are shared risk factors, such as smoking and age. Alternatively, comorbidities may be produced by COPD treatments (Chatila et al. 2008). Nevertheless, it has been increasingly recognized that even after adjusting for these confounding factors, epidemiological evidences link COPD to some of its associated diseases (Chatila et al. 2008; Vestbo et al. 2013). Understanding the molecular mechanisms underlying the COPD diseasome will allow designing non-pulmonary interventions that might improve the course and outcome of the disease. In this paper we used DisGeNET to probe the "*shared genetic component*", and "*shared molecular pathways*" hypotheses to gain insight into the pathobiology of COPD and 16 of its associated diseases.

Grosdidier S, Ferrer A, Faner R, Piñero J, Roca J, Cosío B, Agustí A, Gea J, Sanz F, Furlong LI. Network medicine analysis of COPD multimorbidities. *Respiratory Research.* 2014;15:111. doi: 10.1186/s12931-014-0111-4.

4. DISCUSSION

4.1 Data integration in the research of the genetic causes of human diseases

Modern biomedical research could not be thought of without two key elements: a) the immediate and easy access to different types of clinical, molecular and genetic information; b) the ability to interrogate these sources of information by means of bioinformatics tools and approaches to extract knowledge from them (Burgun and Bodenreider 2008). The diversity and quantity of information available in databases on the genetic basis of diseases has increased in recent years and it continues to grow. Nevertheless, it is dispersed across several resources, and annotated following different criteria, vocabularies and standards. As biomedical research evolves from the traditional 'one gene at a time' approach to modern high throughput techniques, platforms that integrate the corpus of knowledge on the genetic basis of disease, facilitate the access to different types of users, and support in-depth, reproducible and rapid analysis of new data become essential. With this in mind, we developed the DisGeNET discovery platform, with the aim of integrating gene-disease associations covering all the landscape of human diseases, and identified by different types of experimental approaches that detect associations between human genes and disease, as well as those extrapolated from animal models. DisGeNET, whose first version had been developed as a Cytoscape plugin (Bauer-Mehren et al. 2010), is now available as a web platform, a Semantic Web resource, and it is also accessible through customizable R, Perl and Python scripts.

DisGeNET integrates data from several repositories, and also contains its own dataset of gene-disease associations obtained by text-mining the scientific literature using the BeFree system (Bravo et al. 2014). Text mining approaches are an invaluable tool for assisting the work of biocuration teams, or for directly populating databases. The BeFree system is based on a supervised learning approach that achieved a performance of 74 % Precision and of 90 % Recall in the identification of gene-disease associations from Medline abstracts (Bravo et al. 2014). Text mining methodologies are able to unlock the information hidden in the near 22 million of publications currently available in Medline, and also, produce standardized and normalized data, with clear provenance. The current BeFree dataset in DisGeNET (v3.0, May 2015) contains more than 300,000 associations. The 60% of the papers supporting these associations have been published in the last ten years. In general, around 80% of all associations in DisGeNET are obtained using different text mining approaches. This highlights the incompleteness of the repositories that only include data curated by experts, which results in a slower speed to incorporate the most recent scientific findings. Overall, including this type of data is of key importance for resources that aim to keep up with the most recent findings in the area of disease genetics.

The DisGeNET platform was launched in July 2012, and it has had over 22,000 sessions, from all over the world in the period comprised between August 2014 and August 2015 (according to Google Analytics, Figure 4). The Cytoscape plugin has been downloaded more than 1,000 times during the last year, and the data files has been downloaded more than 10,000 times (See Appendix 6.5.1). More than 30 papers use DisGeNET, either for the analysis of specific diseases, or as part of automatic pipelines

with different aims (Appendix 6.5.2). Additionally, DisGeNET is increasingly being recognized as a valuable resource in reviews on disease genetics (Appendix 6.5.3).

The current release of DisGeNET (v3.0) contains over 400,000 gene-disease associations, which makes it one of the largest available repositories of its kind. We find a rather small overlap between the different data sources, which highlights still today the need for data integration. The associations are ranked using the DisGeNET score that takes into account the sources and the number of papers reporting the association. Thanks to this score, users can easily identify consolidated knowledge on disease genes, but also novel gene-disease associations recently reported in the literature and often not still collated by authoritative resources.

Gene-disease associations are classified according to the *DisGeNET association type* ontology, recently integrated into the Semantic science Integrated Ontology (Dumontier et al. 2014). This ontology provides a framework to classify gene-disease associations, which can be extended to include other types of phenomena such as epigenetics, or associations due to the new players that are gradually incorporated to the disease panorama, such as miRNA, lncRNAs, etc.

DisGeNET includes over 17,000 genes associated to disease. In the current release, 85% of the disease-associated loci that map to transcripts are protein-coding genes. Since the number of protein coding genes has been estimated recently in around 19,000 (Ezkurdia et al. 2014), the current version of DisGeNET contains information for 77% of the human proteome. The rest of disease-associated genes in DisGeNET are non-coding RNAs (4%), or other types of RNAs (rRNA, snoRNA, snRNA, tRNA, representing less than 1%) while around 7% of the associations map to loci with unknown function. The current version of DisGeNET has doubled the number of disease genes with respect to the first release. The aforementioned numbers suggest that any gene could potentially bear mutations that produce pathological consequences, and hence, be considered a disease gene. In agreement with this idea, it has recently been suggested that the majority of human protein coding genes may bear alterations that could produce Mendelian phenotypes (Chong et al. 2015).

DisGeNET has annotations for over 14,000 diseases. This number includes phenotypes, laboratory findings, as well as signs and symptoms. DisGeNET currently allows searching by diseases using different disease vocabularies. Nevertheless, there are several issues regarding the coverage and annotation criteria employed by these different vocabularies. For example, only 40% of the CUIs in DisGeNET have coverage in MeSH identifiers. This number is around 10-15 % for the HPO, which is expected because the HPO focuses mainly in phenotypes, not diseases. An additional problem is that for some cases, the disease identifiers map in a one-to-many, or many-to-one relationship. For instance, in the Disease Ontology, the disease "pharyngitis" (DOID:2275) maps to 4 UMLS, 2 MeSH, 2 ICD-10, 17 SNOMED CT, and 3 NCI concepts. On the other hand, 14 concepts in SNOMED CT are mapped to "Alcoholic Intoxication, Chronic" (UMLS CUI C0001973), but only 1 MeSH, 1 ICD-10 and 1 NCI concept map this concept. Furthermore, very similar disease entities –as apparent from their definitions– are encoded by the UMLS using different concepts. For example, the term "breast cancer" is found as a synonym both for "Breast Carcinoma" (UMLS CUI C0678222) and for "Malignant neoplasm of breast" (UMLS CUI C0006142). This ambiguity poses challenges both to text mining workflows, manual curation and data

integration. Moreover, different vocabularies describe diseases with different level of granularity, which represents another challenge when mapping from one vocabulary to another. This is particularly important when we want to translate research findings into the clinical area, as this often requires mapping a disease term from a publication (for instance encoded in MeSH) to a clinical record (that will be encoded in ICD-9CM, for example). All these problems are intrinsic to biomedical vocabularies, and need to be confronted by researchers interested in the study of the molecular mechanisms of disease.

Throughout the work presented in this thesis, we have illustrated the usefulness of DisGeNET to support bioinformatics studies aimed at unraveling the molecular mechanisms of disease. In particular, DisGeNET data has been employed to explore the network properties of disease genes, and to study the molecular mechanisms underlying disease comorbidities. Furthermore, DisGeNET platform is starting to be recognized as a reference resource in the biomedical research community (see Appendix 6.5). It features several characteristics that provide flexibility to adapt to new demands and paradigms, and absorb new types of information in the era of big data in biomedicine.

4.2 Network biology approaches in the study of the molecular mechanisms underlying human diseases

The explosion of data produced by a series of technological breakthroughs in biomedicine, health care, and computing systems can no longer be analyzed using traditional approaches. New methods are required to interrogate and explore this information in order to decipher the complex molecular interplay underlying human physiology and pathology. There is consensus in the scientific literature that network approaches are the ideal framework to tame this biomedical data deluge. The last two decades have widely exploited protein interaction networks for a variety of purposes, which has allowed adding pieces of information to the very incomplete puzzle of human disease. Even when these works have contributed to our understanding of the basic principles of cellular organization, the studies addressing the network properties of disease genes have produced contradictory results, which may be attributed to several causes. In our work, we aimed at providing a definitive answer[9] to the question of distinctive network properties of disease genes, by consider some key methodological aspects.

First, we assessed the impact of the choice of the protein interaction data in the network analysis. It has been recognized that this type of data is subject to different biases: those imposed by our current technological limitations and those associated to the reporting bias for certain groups of proteins (Jensen and Bork 2008; Rolland et al. 2014). To address these issues, initiatives such as the Human Binary Interactome (Rolland et al. 2014; Rual et al. 2005; Venkatesan et al. 2009) that systematically screens the interactome space with minimal inspection bias are important for providing a more complete picture of the human interactome. Furthermore, given the lack of overlap in protein interaction data across different resources, we performed our analysis on six different protein interaction networks, one of them being the Human Binary Interactome (Rolland et al. 2014; Rual et al. 2005).

A second issue may be the source of disease genes. Many previous studies addressing the properties of disease genes have been performed using data extracted

mainly from OMIM, a catalog of Mendelian diseases. Although OMIM is a reference resource for Mendelian and rare diseases, it lacks information on many common disorders. This could bias the result for only this type of diseases. In the last few years, the catalogs of genes associated to disease have increased significantly, so revisiting the results of the early studies in the light of the new data (gene-disease associations, and protein-protein interactions), while taking into account the different disease classifications, is important. We used DisGeNET data to assess the network properties of disease genes, both as a whole, and grouped according to different classifications. By employing DisGeNET we aimed at covering the whole landscape of human diseases.

Thirdly, to clarify whether discordant results have arisen from different definitions of network properties, in addition to re-evaluating the classical network properties, we characterized different groups of disease genes using the mesoscale structure of the network. To this end, we first conducted a study to choose an appropriate clustering algorithm for the problem of partitioning the network into discrete and biologically relevant modules. We compared the performance of two clustering algorithms that are based on different optimization criteria, and chose the one that produced network partitions with higher levels of biological homogeneity.

We also paid special attention to the statistical analysis. Most of the previous reports have employed Mann-Whitney tests to compare the distribution of the topological parameters. Since it has been shown that p-values are sensitive to large sample sizes (Lin, Lucas, and Shmueli 2013), we have also used randomization to compute z-scores and bootstrap analysis to control for the node's degree distribution confounding factor in the enrichment analysis for the cartographic roles.

Finally, we sought to explain the differences in the network properties of different groups of disease genes from a genomics perspective, measuring their tolerance to likely deleterious germline variants across human populations.

Our goal was to study the network features of all disease-associated genes, using some of the most common classification of diseases. We found that disease genes exhibit heterogeneous network properties, attributable to subsets of genes associated to different diseases. Additionally, our results indicate that the subgroups of disease genes that show a pattern that differs the most from the bulk of genes in the network are those related to Mendelian diseases and cancer drivers. The behavior of complex diseases genes is primarily driven by cancer genes, and their properties are similar to the non-disease genes once this overlap is removed. This lack of a clear trend of genes associated to complex diseases may be due to several reasons. First, to the heterogeneity of complex diseases as a group and the current disease classification that is mostly based on signs and symptoms, and therefore could group together diseases with different molecular basis. Second, to the nature of GWAS, which detect associations involving SNPs, which are not necessarily causative but a marker of the real causal genomic variant. Alternatively, it might also occur that the complex disease proteins play a variety of roles in the network and therefore it is not possible to identify a clear pattern. Our findings could be employed in different ways. In the case of Mendelian diseases, bioinformatics pipelines aiming at the identification of disease genes that are candidate to be associated to the large number of Mendelian diseases with still unknown molecular basis. According to the mode of inheritance of the phenotype, researchers may prioritize genes based on their network features and tolerance to mutations. For

cancer, pipelines for the analysis of somatic mutations may add as a feature the centrality and low tolerance displayed by putative genes to rank them first as cancer drivers.

In this thesis, we have also employed DisGeNET data and protein interaction networks to probe the molecular mechanisms underlying disease comorbidities. Disease comorbidity is a major health issue, but it has received less attention than other issues involving human diseases. In our approach, the COPD diseasome was constructed by linking diseases sharing not only the proteins directly associated to the disease, but also, disease proteins that are connected in the human interactome. The use of protein interaction data allowed expanding the molecular genetic connections underlying the two diseases. We defined the Molecular Comorbidity Index to quantify the strength of this genetic association, and to correct biases for diseases that have been more extensively studied. Additionally, to study the comorbidities from a higher level of organization, we performed a pathway enrichment analysis, and we compared common molecular pathways associated to different comorbidity pairs using the Jaccard coefficient. All in all, we have presented a novel network medicine method to investigate the molecular basis of disease comorbidities, which can be applied to many other case studies.

4.3 Future perspectives

The DisGeNET platform is designed to fulfill the needs of a wide variety of biomedical researchers. It provides standardized data that allows integration with bioinformatics pipelines and a set of tools to facilitate the exploitation of its data in different research scenarios. Nevertheless, it shares an unmet need with many currently available computational resources: to be easily accessible to medical practitioners. Thus, it remains a challenge to extend the scope of DisGeNET to this type of users too. In the near future, medical doctors will need to exploit this kind of information to interpret the genomics information of patients, and tools like DisGeNET should aid them in this goal.

In the case of network biology, our current models need refining. With the accumulation of context-specific molecular expression profiles, cell and tissue specific maps of the interactome are becoming more available, and studies addressing diseases, in this specific context are beginning to appear (Barshir et al. 2014; Cornish et al. 2015; Greene et al. 2015; Guan et al. 2012; Lage et al. 2008). Nevertheless, they are not yet systematical, partially because they are proof of concept studies or they address particular diseases. Furthermore, the context specific networks should also include isoform information, regulatory modifications, and cellular compartments, in order to be an accurate portrait of the cell. Other challenge of current network biology consists in modelling the precise impact of mutations on the structure of the network (Khurana et al. 2013; Mosca et al. 2015; Vázquez, Valencia, and Pons 2015; Wang et al. 2012). Results in this direction remain limited, because the number of protein complexes with known tri-dimensional structure is still small (Mosca, Céol, and Aloy 2013).

5. CONCLUSIONS

1. We have presented DisGeNET, a discovery platform that integrates information on the genetic causes of disease, covering the whole landscape of human diseases, available through a web interface, a Cytoscape plugin, a Semantic Web resource, and customizable scripts in several programming languages.
2. We have employed DisGeNET to explore the relationship between the network features of disease genes and their tolerance to deleterious variants, and to investigate the molecular basis underlying disease comorbidities.
3. We have shown that the choice of the clustering algorithm of a protein interaction network has an impact in the subsequent biological analysis, an aspect that has not been taken into account before and deserves special attention in network biology studies.
4. We have shown that the trends resulting from the analysis of the properties of disease genes as a whole are actually caused by the behavior of specific subsets of genes inside the larger set of "disease genes".
5. We have assessed the network properties of disease genes associated to different disease classifications, and found that while clinical taxonomy-based classification of diseases does not correlate with the network properties of disease genes, both the molecular biology and the mode of inheritance of the disease capture better major differences in their network properties.
6. We have found that network centrality and tolerance to deleterious mutations show opposite trends for different classes of disease genes.
7. We have developed a novel network medicine approach based on the *shared components formalism* to explore the mechanisms underlying disease comorbidities, and showed its value in a case study of COPD comorbidities.

6. APPENDIX

6.1 PsyGeNET: a knowledge platform on psychiatric disorders and their genes

PsyGeNET (Psychiatric disorders and Genes association NETwork) is a knowledge platform for the exploratory analysis of psychiatric diseases and their associated genes. PsyGeNET is composed of a database and a web interface supporting data search, visualization, filtering and sharing. PsyGeNET integrates information from DisGeNET and data extracted from the literature by text mining, which has been curated by domain experts. It currently contains 2,642 associations between 1,271 genes and 37 psychiatric disease concepts. In its first release, PsyGeNET is focused on three psychiatric disorders: major depression, alcohol and cocaine use disorders. PsyGeNET represents a comprehensive, open access resource for the analysis of the molecular mechanisms underpinning psychiatric disorders and their comorbidities.

Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, Sanz F, Furlong LI. PsyGeNET: a knowledge platform on psychiatric disorders and their genes *Bioinformatics* 2015 15;31(18):3075-7.
doi:10.1093/bioinformatics/btv301

6.2 Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research.

Background: Current biomedical research needs to leverage and exploit the large amount of information reported in scientific publications. Automated text mining approaches, in particular those aimed at finding relationships between entities, are key for identification of actionable knowledge from free text repositories. We present the BeFree system aimed at identifying relationships between biomedical entities with a special focus on genes and their associated diseases.

Results: By exploiting morpho-syntactic information of the text, BeFree is able to identify gene-disease, drug-disease and drug-target associations with state-of-the-art performance. The application of BeFree to real-case scenarios shows its effectiveness in extracting information relevant for translational research. We show the value of the gene-disease associations extracted by BeFree through a number of analyses and integration with other data sources. BeFree succeeds in identifying genes associated to a major cause of morbidity worldwide, depression, which are not present in other public resources. Moreover, large-scale extraction and analysis of gene-disease associations, and integration with current biomedical knowledge, provided interesting insights on the kind of information that can be found in the literature, and raised challenges regarding data prioritization and curation. We found that only a small proportion of the gene-disease associations discovered by using BeFree is collected in expert-curated databases. Thus, there is a pressing need to find alternative strategies to manual curation, in order to review, prioritize and curate text-mining data and incorporate it into domain-specific databases. We present our strategy for data prioritization and discuss its implications for supporting biomedical research and applications.

Conclusions: BeFree is a novel text mining system that performs competitively for the identification of gene-disease, drug-disease and drug-target associations. Our analyses show that mining only a small fraction of MEDLINE results in a large dataset of gene-disease associations, and only a small proportion of this dataset is actually recorded in curated resources (2%), raising several issues on data prioritization and curation. We propose that joint analysis of text mined data with data curated by experts appears as a suitable approach to both assess data quality and highlight novel and interesting information.

Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics* 2015; 16:55 doi:10.1186/s12859-015-0472-9

6.3 Automatic Filtering and Substantiation of Drug Safety Signals.

Drug safety issues pose serious health threats to the population and constitute a major cause of mortality worldwide. Due to the prominent implications to both public health and the pharmaceutical industry, it is of great importance to unravel the molecular mechanisms by which an adverse drug reaction can be potentially elicited. These mechanisms can be investigated by placing the pharmaco-epidemiologically detected adverse drug reaction in an information-rich context and by exploiting all currently available biomedical knowledge to substantiate it. We present a computational framework for the biological annotation of potential adverse drug reactions. First, the proposed framework investigates previous evidences on the drug-event association in the context of biomedical literature (signal filtering). Then, it seeks to provide a biological explanation (signal substantiation) by exploring mechanistic connections that might explain why a drug produces a specific adverse reaction. The mechanistic connections include the activity of the drug, related compounds and drug metabolites on protein targets, the association of protein targets to clinical events, and the annotation of proteins (both protein targets and proteins associated with clinical events) to biological pathways. Hence, the workflows for signal filtering and substantiation integrate modules for literature and database mining, in silico drug-target profiling, and analyses based on gene-disease networks and biological pathways. Application examples of these workflows carried out on selected cases of drug safety signals are discussed. The methodology and workflows presented offer a novel approach to explore the molecular mechanisms underlying adverse drug reactions.

Bauer-Mehren A, van Mullingen EM, Avillach P, Carrascosa MdC, Garcia-Serna R, Piñero J, Singh B, Lopes P, Oliveira JL, Diallo G, Helgee EA, Boyer S, Mestres J, Sanz F, Kors JA, Furlong LI. Automatic Filtering and Substantiation of Drug Safety Signals. (2012) *PLoS Comput Biol* 8(4): e1002457. doi:10.1371/journal.pcbi.1002457

6.4 Fundamentals of Network Biology

Degree: number of connections of a node.

Distance: the shortest path length between two vertices.

Betweenness: number of shortest paths traversing a given node.

Clustering Coefficient: number of links between the neighbors connected to a node divided by the number of links that are possible between them.

Topological module represents a locally dense neighborhood in a network, such that nodes have a higher tendency to link to nodes within the same local neighborhood than to nodes outside of it.

Participation Coefficient: quantifies the fraction of links that a given node projects to other communities. Computed according to:

$$z_i = \frac{k_i - \bar{k}_{si}}{\sigma \bar{k}_{si}}$$

where k_i is the number of links of node i to other nodes in its module, k_{s_i} is the mean degree of all nodes in cluster s_i, and σk_{s_i} is the standard deviation of the degree in the cluster s_i

Within-module degree: standardizes the degree of a node in relation with the degree of nodes that belong to the same community. Computed according to:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{is}}{k_i} \right)^2$$

where k_{is} is the number of links of node i to nodes in the module s, and k_i is the total degree of node i.

6.5 DisGeNET impact on the scientific community

6.5.1. Statistics of use of DisGeNET

We present the statistics of use of DisGeNET for the period between August 2014 and August 2015.

a) Web interface: accessed 25,513 times (4:35 minutes/session), by 13,573 users
b) Data: the database has been downloaded 8,498 times. Other files (tabulated files with DisGeNET data) have been downloaded 4,911 times.
c) Cytoscape plugin: 1,035 downloads
d) RDF version: 50 downloads

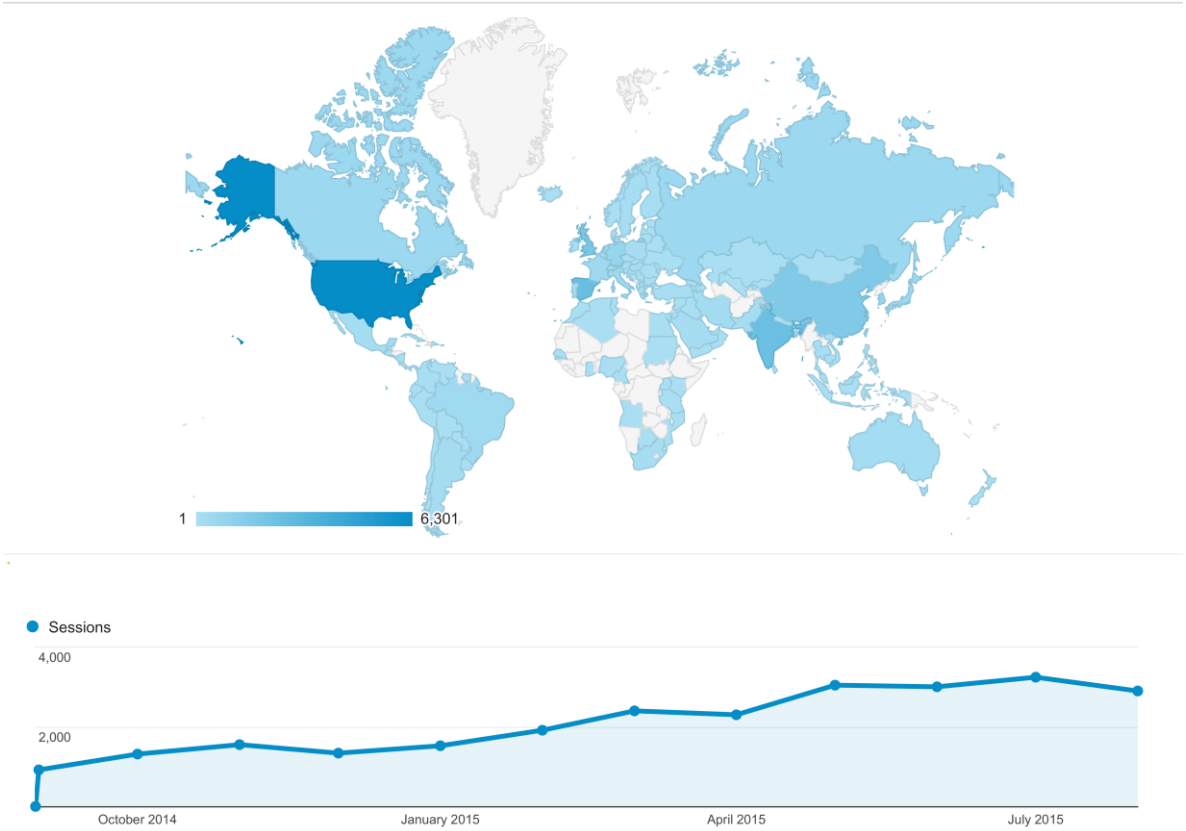

Figure 4: Geographic and temporal distribution of DisGeNET web interface users

6.5.2. Papers that use DisGeNET

1. **Cell type-selective disease-association of genes under high regulatory load**. Galhardo M, Berninger P, Nguyen T, Sauter T, Sinkkonen L. *Nucl. Acids Res.* (2015) doi:10.1093/nar/gkv863
2. **Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types.** Cornish AJ, Filippis I, David Asternberg MJE. *Genome Med* (2015) doi:10.1186/s13073-015-0212-9
3. **Dissecting Xuesaitong's mechanisms on preventing stroke based on the microarray and connectivity map** Wang L, Yu Y, Yang J, Zhao X, Li Z. *Mol Biosyst.* (2015) doi:10.1039/c5mb00379b
4. **MicroRNA and Transcription Factor Mediated Regulatory Network Analysis Reveals Critical Regulators and Regulatory Modules in Myocardial Infarction.** Zhang G, Shi H, Wang L, Zhou M, Wang Z, Liu X, Cheng L, Li W, & Li X. *PLoS One* (2015) doi:10.1371/journal.pone.0135339
5. **Insights from Chromosome-Centric Mapping of Disease-Associated Genes: Chromosome 12 Perspective.** Jayaram S, Gupta, MK, Shivakumar BM, Ghatge M, Sharma A, Vangala RK, & Sirdeshmukh R. *J Proteome Res* (2015) doi:10.1021/acs.jproteome.5b00488
6. **Inferring disease associations of the long non-coding RNAs through non-negative matrix factorization.** Biswas AK, Kang M, Kim DC, Ding CH, Zhang B, Wu X, & Gao, JX. *Netw Model Anal Health Inform Bioinforma* (2015) doi:10.1007/s13721-015-0081-6
7. **Genetic mutations associated with status epilepticus.** Bhatnagar M, & Shorvon S. *Epilepsy Behav.* (2015) doi:10.1016/j.yebeh.2015.04.013
8. **Novel scripts for improved annotation and selection of variants from whole exome sequencing in cancer research.** Hansen, MC, Nederby L, Roug A, Villesen P, Kjeldsen E, Nyvold CG, & Hokland P. *MethodsX* (2015) doi:10.1016/j.mex.2015.03.003
9. **Molecular Architecture of Spinal Cord Injury Protein Interaction Network.** Alawieh A, Sabra M, Sabra Z, Tomlinson S, & Zaraket FA *PLoS One* (2015) doi:10.1371/journal.pone.0135024
10. **A pipeline for the systematic identification of non-redundant full-ORF cDNAs for polymorphic and evolutionary divergent genomes: Application to the ascidian Ciona intestinalis.** Gilchrist MJ, Sobral D, Khoueiry P, Daian F, Laporte B, Patrushev I, Matsumoto J, Dewar K, Hastings KEM, Satou Y, Lemairea P & Rothbächer U. *Dev Biol.* (2015) doi:10.1016/j.ydbio.2015.05.014
11. **Analysis of Deregulated microRNAs and Their Target Genes in Gastric Cancer** Juzėnas S, Saltenienė V, Kupcinskas J, Link A, Kiudelis G, Jonaitis G, Jarmalaite S, Kupcinskas L, Malfertheiner P, Skieceviciene J. *PLoS One* (2015) doi:10.1371/journal.pone.0132327
12. **Nature and nurture: a case of transcending haematological pre-malignancies in a pair of monozygotic twins adding possible clues on the pathogenesis of B-cell proliferations.** Hansen MC, Nyvold CG, Roug AS, Kjeldsen E, Villesen P, Nederby L, Hokland P. *Br J Haematol.* (2015) doi:10.1111/bjh.13305
13. **Pathway reporter genes define molecular phenotypes of human cells** Zhang JD, Küng E, Boess F, Certa U and Ebeling M. *BMC Genomics* (2015) doi:10.1186/s12864-015-1532-2
14. **Global Mapping of Herpesvirus-Host Protein Complexes Reveals a Transcription Strategy for Late Genes.** Davis ZH, Verschueren E, Jang GM, Kleffman K, Johnson JR, Park J, Von Dollen J, Maher MC, Johnson T, Newton W, Jäger S, Shales M, Horner J, Hernandez RD, Krogan NJ, Glaunsinger BA *Mol Cel.l* (2015) doi:10.1016/j.molcel.2014.11.026
15. **Integromics network meta-analysis on cardiac aging offers robust multi-layer modular signatures and reveals micronome synergism.** Dimitrakopoulou K, Vrahatis AG, and Bezerianos A. *BMC Genomics* (2015) doi:10.1186/s12864-015-1256-3

16. **Discovery of new candidate genes related to brain development using protein interaction information.** Chen L, Chu C, Kong X, Huang T, Cai YD. *PLoS One*. (2015) doi:10.1371/journal.pone.0118003
17. **ncRNA-Disease association prediction through tripartite network based inference** Alaimo S, Giugno R, & Pulvirenti A *Front. Bioeng. Biotechnol.* (2014) doi:10.3389/fbioe.2014.00071
18. **A network approach to clinical intervention in neurodegenerative diseases.** Santiago, JA, Potashkin JA. *Trends Mol Med.* (2014) doi:10.1016/j.molmed.2014.10.002
19. **Control of VEGF-A transcriptional programs by pausing and genomic compartmentalization.** Kaikkonen MU, Niskanen H, Romanoski CE, Kansanen E, Kivelä AM, Laitalainen J, Heinz S, Benner C, Glass CK, Ylä-Herttuala S. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku1036
20. **Network medicine analysis of COPD multimorbidities.** Grosdidier S, Ferrer A, Faner R, Piñero J, Roca J, Cosío B, Agustí A, Gea J, Sanz F, Furlong LI. *Respir Res.* (2014) doi:10.1186/s12931-014-0111-4
21. **An R-based tool for miRNA data analysis and correlation with clinical ontologies.** Cristiano F, Veltri P. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2014) doi:10.1145/2649387.2660847
22. **Using 2-node hypergraph clustering coefficients to analyze disease-gene networks.** Renick Gallagher S, Dombrower M, Goldberg DS. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2014) doi:10.1145/2649387.2660817
23. **Organ system heterogeneity DB: a database for the visualization of phenotypes at the organ system level.** Mannil D, Vogt I, Prinz J, Campillos M. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku948
24. **Molecularly and clinically related drugs and diseases are enriched in phenotypically similar drug-disease pairs.** Vogt I, Prinz J, Campillos M. *Genome Med.* (2014) doi:10.1186/s13073-014-0052-z
25. **System-based approaches to decode the molecular links in Parkinson's disease and diabetes.** Santiago, JA, Potashkin JA. *Neurobiol Dis.* (2014) doi:10.1016/j.nbd.2014.03.019
26. **Prioritizing Disease-Linked Variants, Genes, and Pathways with an Interactive whole Genome Analysis Pipeline.** Lee IH, Lee K, Hsing M, Choe Y, Park JH, Kim SH, Bohn JM, Neu MB, Hwang KB, Green RC, Kohane IS, Kong SW. *Hum Mutat.* (2014) doi:10.1002/humu.22520
27. **A Computational Framework to Infer Human Disease-Associated Long Noncoding RNAs.** Liu MX, Chen X, Chen G, Cui QH, Yan GY. *PloS One* 9.1 (2014) doi:10.1371/journal.pone.0084408
28. **Choline protects against cardiac hypertrophy induced by increased after-load.** Zhao Y, Wang C, Wu J, Wang Y, Zhu W, Zhang Y, Du Z. *Int J Biol Sci.* (2013) doi:10.7150/ijbs.5976
29. **Detection of differentially methylated gene promoters in failing and nonfailing human left ventricle myocardium using computation analysis.** Koczor CA, Lee EK, Torres RA, Boyd A, Vega JD, Uppal K, Yuan F, Fields EJ, Samarel AM, Lewis W. *Physiol Genomics.* (2013) doi:10.1152/physiolgenomics.00013.2013
30. **Global DNA methylation and transcriptional analyses of human ESC-derived cardiomyocytes.** Gu Y, Liu GH, Plongthongkum N, Benner C, Yi F, Qu J, Suzuki K, Yang J, Zhang W, Li M, Montserrat N, Crespo I, Del Sol A, Esteban CR, Zhang K, Belmonte JC. *Protein Cell.* (2013) doi:10.1007/s13238-013-0016-x
31. **Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network.** Galhardo M1, Sinkkonen L, Berninger P, Lin J, Sauter T, Heinäniemi M. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gkt989

32. **Charting the NF-κB Pathway Interactome Map.** Tieri P, Termanini A, Bellavista E, Salvioli S, Capri M, Franceschi C. *PLoS One* (2012) doi:10.1371/journal.pone.0032678

6.5.3. Papers that cite DisGeNET

1. **Allele, phenotype and disease data at Mouse Genome Informatics: improving access and analysis.** Bello, SM, Smith, CL, and Eppig, JT. *Mammalian Genome* (2015) doi:10.1007/s00335-015-9582-y
2. **Interoperability of text corpus annotations with the semantic web.** Verspoor, K, Kim, JD, and Dumontier, M. *BMC Proceedings* (2015) doi:10.1186/1753-6561-9-S5-A2
3. **How to build personalised multi-omics comorbidity profiles.** Moni MA, and Lio P. *Front. Cell Dev. Biol.* (2015) doi:10.3389/fcell.2015.00028
4. **Integrating proteomics profiling data sets: a network perspective.** Bhat A, Dakna M, Mischak H. *Methods Mol Biol.* (2015) doi:10.1007/978-1-4939-1872-0_14
5. **Network Analysis in the Investigation of Chronic Respiratory Diseases. From Basics to Application.** Diez D, Agustí A, and Wheelock CE. American Journal of Respiratory and Critical Care Medicine, (2014) doi:10.1164/rccm.201403-0421PP

33. **Associating disease-related genetic variants in intergenic regions to the genes they impact.** Macintyre G, Jimeno Yepes A, Ong CS, Verspoor K. PeerJ. (2014) doi:10.7717/peerj.639
6. **Clinical proteomic biomarkers: relevant issues on study design and technical considerations in biomarker development.** Frantzi M, Bhat A, Latosinska A. *Clin Transl Med.* (2014) doi:10.1186/2001-1326-3-7
7. **Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review.** Csermely P, Korcsmáros T, Kiss HJ, London G, Nussinov R. *Pharmacol Ther.* (2013) doi:10.1016/j.pharmthera.2013.01.016
8. **ChemProt-2.0: visual navigation in a disease chemical biology database.** Kjærulff SK, Wich L, Kringelum J, Jacobsen UP, Kouskoumvekaki I, Audouze K, Lund O, Brunak S, Oprea TI, Taboureau O. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gks1166
9. **State of the art *in silico* tools for the study of signaling pathways in cancer.** Villaamil, VM, Gallego, GA, Cainzos, IS, Valladares-Ayerbes, M, and Aparicio, LMA. *Int J Mol Sci*. (2012) doi:10.3390/ijms13066561
10. **iCTNet: a Cytoscape plugin to produce and analyze integrative complex traits networks.** Wang L, Khankhanian P, Baranzini SE, Mousavi P. *BMC Bioinformatics* (2011) doi:10.1186/1471-2105-12-380.

6.5.4. Companies that include DisGeNET data in their services

https://www.solvebio.com/

http://www.edgeleap.com/company/

7. BIBLIOGRAPHY

Allison, A. C. 1954. “Protection Afforded by Sickle-Cell Trait against Subtertian Malareal Infection.” *British medical journal* 1(4857):290–94.

Alon, U. 2003. “Biological Networks: The Tinkerer as an Engineer.” *Science (New York, N.Y.)* 301(5641):1866–67.

Altshuler, David, Mark J. Daly, and Eric S. Lander. 2008. “Genetic Mapping in Human Disease.” *Science* 322(5903):881–88.

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders.” *Nucleic acids research* 43(Database issue):D789–98.

Angrist, M., S. Bolk, M. Halushka, P. A. Lapchak, and A. Chakravarti. 1996. “Germline Mutations in Glial Cell Line-Derived Neurotrophic Factor (GDNF) and RET in a Hirschsprung Disease Patient.” *Nature genetics* 14(3):341–44.

Antonarakis, S. E. et al. 1995. “Factor VIII Gene Inversions in Severe Hemophilia A: Results of an International Consortium Study.” *Blood* 86(6):2206–12.

Ashburner, M. et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature genetics* 25(1):25–29.

Aymé, S., B. Urbero, D. Oziel, E. Lecouturier, and A. C. Biscarat. 1998. “[Information on Rare Diseases: The Orphanet Project].” *La Revue de médecine interne / fondée ... par la Société nationale francaise de médecine interne* 19 Suppl 3:376S – 377S.

Badano, Jose L. and Nicholas Katsanis. 2002. “Beyond Mendel: An Evolving View of Human Genetic Disease Transmission.” *Nature reviews. Genetics* 3(10):779–89.

Bader, Gary D., Michael P. Cary, and Chris Sander. 2006. “Pathguide: A Pathway Resource List.” *Nucleic acids research* 34(Database issue):D504–6.

Balint, Geza P., W. Watson Buchanan, and Jan Dequeker. 2006. “A Brief History of Medical Taxonomy and Diagnosis.” *Clinical rheumatology* 25(2):132–35.

Barabasi, Albert-Laszlo and Reka Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286(5439):11.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. “Network Medicine: A Network-Based Approach to Human Disease.” *Nature Reviews Genetics* 12(1):56–68.

Barabási, Albert-László and Zoltán N. Oltvai. 2004. “Network Biology: Understanding the Cell’s Functional Organization.” *Nature reviews. Genetics* 5(2):101–13.

Barrenas, Fredrik, Sreenivas Chavali, Petter Holme, Reza Mobini, and Mikael Benson. 2009. “Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies” edited by Thomas Mailund. *PLoS ONE* 4(11):6.

Barshir, Ruth, Omer Shwartz, Ilan Y. Smoly, and Esti Yeger-Lotem. 2014. “Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases.” *PLoS computational biology* 10(6):e1003632.

Bauer-Mehren, Anna et al. 2011. “Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases” edited by Raya Khanin. *PLoS ONE* 6(6):13.

Bauer-Mehren, Anna, Michael Rautschka, Ferran Sanz, and Laura I. Furlong. 2010. “DisGeNET: A Cytoscape Plugin to Visualize, Integrate, Search and Analyze Gene-Disease Networks.” *Bioinformatics (Oxford, England)* 26(22):2924–26.

Beck, Tim, Robert K. Hastings, Sirisha Gollapudi, Robert C. Free, and Anthony J. Brookes. 2014. “GWAS Central: A Comprehensive Resource for the Comparison and Interrogation of Genome-Wide Association Studies.” *European journal of human genetics : EJHG* 22(7):949–52.

Becker, Kevin G., Kathleen C. Barnes, Tiffani J. Bright, and S. Alex Wang. 2004. “The Genetic Association Database.” *Nature genetics* 36(5):431–32.

Bjornsson, Hans T., M. Daniele Fallin, and Andrew P. Feinberg. 2004. “An Integrated Epigenetic and Genetic Approach to Common Human Disease.” *Trends in genetics : TIG* 20(8):350–58.

Blackwell, Jenefer M., Sarra E. Jamieson, and David Burgner. 2009. “HLA and Infectious Diseases.” *Clinical microbiology reviews* 22(2):370–85, Table of Contents.

Blair, David R. et al. 2013. “A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk.” *Cell* 155(1):70–80.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis. 1980. “Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms.” *American journal of human genetics* 32(3):314–31.

Botstein, David and Neil Risch. 2003. “Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease.” *Nature genetics* 33 Suppl:228–37.

Bravo, A., J. Pinero, N. Queralt, M. Rautschka, and L. I. Furlong. 2014. *Extraction of Relations between Genes and Diseases from Text and Large-Scale Data Analysis: Implications for Translational Research*. Cold Spring Harbor Labs Journals.

Brookes, Emily and Yang Shi. 2014. “Diverse Epigenetic Mechanisms of Human Disease.” *Annual review of genetics* 48:237–68.

Burgun, A. and O. Bodenreider. 2008. “Accessing and Integrating Data and Knowledge for Biomedical Research.” *Yearbook of medical informatics* 91–101.

Cai, James J., Elhanan Borenstein, and Dmitri A. Petrov. 2010. “Broker Genes in Human Disease.” *Genome biology and evolution* 2:815–25.

Chatila, Wissam M., Byron M. Thomashow, Omar A. Minai, Gerard J. Criner, and Barry J. Make. 2008. “Comorbidities in Chronic Obstructive Pulmonary Disease.” *Proceedings of the American Thoracic Society* 5(4):549–55.

Chatr-Aryamontri, Andrew et al. 2015. “The BioGRID Interaction Database: 2015 Update.” *Nucleic acids research* 43(Database issue):D470–78.

Chen, Hao et al. 2015. “Pathway Mapping and Development of Disease-Specific Biomarkers: Protein-Based Network Biomarkers.” *Journal of cellular and molecular medicine* 19(2):297–314.

Chen, Jing, Bruce J. Aronow, and Anil G. Jegga. 2009. “Disease Candidate Gene Identification and Prioritization Using Protein Interaction Networks.” *BMC bioinformatics* 10(1):73.

Chong, Jessica X. et al. 2015. “The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.” *The American Journal of Human Genetics*.

Chuang, Han-Yu, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. 2007. “Network-Based Classification of Breast Cancer Metastasis.” *Molecular systems biology* 3(1):140.

Collins, F. S. 1997. “Variations on a Theme: Cataloging Human DNA Sequence Variation.” *Science* 278(5343):1580–81.

Cornish, Alex J., Ioannis Filippis, Alessia David, and Michael J. E. Sternberg. 2015. “Exploring the Cellular Basis of Human Disease through a Large-Scale Mapping of Deleterious Genes to Cell Types.” *Genome Medicine* 7(1):95.

Croft, David et al. 2014. “The Reactome Pathway Knowledgebase.” *Nucleic acids research* 42(Database issue):D472–77.

Cui, Guangyu, Yu Chen, De-Shuang Huang, and Kyungsook Han. 2008. “An Algorithm for Finding Functional Modules and Protein Complexes in Protein-Protein Interaction Networks.” *Journal of biomedicine & biotechnology* 2008:860270.

Cutting, Garry R. 2010. “Modifier Genes in Mendelian Disorders: The Example of Cystic Fibrosis.” *Annals of the New York Academy of Sciences* 1214:57–69.

Davis, Allan Peter, Thomas C. Wiegers, Michael C. Rosenstein, and Carolyn J. Mattingly. 2012. “MEDIC: A Practical Disease Vocabulary Used at the Comparative Toxicogenomics Database.” *Database : the journal of biological databases and curation* 2012:bar065.

Dean, M. et al. 1996. “Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE .” *Science (New York, N.Y.)* 273(5283):1856–62.

Deng, Minghua, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. 2003. “Prediction of Protein Function Using Protein-Protein Interaction Data.” *Journal of computational biology : a journal of computational molecular cell biology* 10(6):947–60.

Divo, Miguel et al. 2012. “Comorbidities and Risk of Mortality in Patients with Chronic Obstructive Pulmonary Disease.” *American journal of respiratory and critical care medicine* 186(2):155–61.

Dumontier, Michel et al. 2014. “The Semanticscience Integrated Ontology (SIO) for Biomedical Research and Knowledge Discovery.” *Journal of biomedical semantics* 5(1):14.

Durbin, Richard M. et al. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467(7319):1061–73.

Eppig, Janan T., Judith A. Blake, Carol J. Bult, James A. Kadin, and Joel E. Richardson. 2015. “The Mouse Genome Database (MGD): Facilitating Mouse as a Model for Human Biology and Disease.” *Nucleic acids research* 43(Database issue):D726–36.

Ezkurdia, Iakes et al. 2014. “Multiple Evidence Strands Suggest That There May Be as Few as 19 000 Human Protein-Coding Genes.” *Human molecular genetics*.

Feinstein, AR. 1970. “Pre-Therapeutic Classification of Co-Morbidity in Chronic Disease.” *J Chronic Dis.* 23(7):455–68.

Feldman, Igor, Andrey Rzhetsky, and Dennis Vitkup. 2008. “Network Properties of Genes Harboring Inherited Disease Mutations.” *Proceedings of the National Academy of Sciences of the United States of America* 105(11):4323–28.

Fisher, Ronald Aylmer. 1918. “The Correlation Between Relatives on the Supposition of Mendelian Inheritance.” *Transactions of the Royal Society of Edinburgh* 52:399–433.

Fortunato, Santo. 2010. “Community Detection in Graphs.” *Physics Reports* 486(3-5):75–174.

Fortunato, Santo and Marc Barthélemy. 2007. “Resolution Limit in Community Detection.” *Proceedings of the National Academy of Sciences of the United States of America* 104(1):36–41.

Freschi, Valerio. 2007. “Protein Function Prediction from Interaction Networks Using a Random Walk Ranking Algorithm.” *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering BIBE 2007* 42–48.

Futreal, P. Andrew et al. 2004. “A Census of Human Cancer Genes.” *Nature reviews. Cancer* 4(3):177–83.

Gandhi, T. K. B. et al. 2006. “Analysis of the Human Protein Interactome and Comparison with Yeast, Worm and Fly Interaction Datasets.” *Nature Genetics* 38(3):285–93.

Garcia-Alonso, Luz et al. 2014. “The Role of the Interactome in the Maintenance of Deleterious Variability in Human Populations.” *Molecular systems biology* 10(9):752.

García-Alonso, Luz et al. 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.

Garcia-Garcia, Javier, Emre Guney, Ramon Aragues, Joan Planas-Iglesias, and Baldo Oliva. 2010. “Biana: A Software Framework for Compiling Biological Interactions and Analyzing Networks.” *BMC bioinformatics* 11(1):56.

Garrod, ArchibaldE. 1902. “The Incidence of Alkaptonuria: A Study in Chemical Individuality.” *The Lancet* 160(4137):1616–20.

Ghiassian, Susan Dina, Jörg Menche, and Albert-László Barabási. 2015. “A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of

Connectivity Patterns of Disease Proteins in the Human Interactome" edited by Andrey Rzhetsky. *PLOS Computational Biology* 11(4):e1004120.

Girvan, M. and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences of the United States of America* 99(12):7821–26.

Goh, Kwang-Il et al. 2007. "The Human Disease Network." *Proceedings of the National Academy of Sciences of the United States of America* 104(21):8685–90.

Goto, Y., I. Nonaka, and S. Horai. 1990. "A Mutation in the tRNA(Leu)(UUR) Gene Associated with the MELAS Subgroup of Mitochondrial Encephalomyopathies." *Nature* 348(6302):651–53.

Gottlieb, Assaf, Gideon Y. Stein, Eytan Ruppin, and Roded Sharan. 2011. "PREDICT: A Method for Inferring Novel Drug Indications with Application to Personalized Medicine." *Molecular systems biology* 7(496):496.

Greene, Casey S. et al. 2015. "Understanding Multicellular Function and Disease with Human Tissue-Specific Networks." *Nature Genetics* 47(6):569–76.

Groza, Tudor et al. 2015. "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease." *American journal of human genetics* 97(1):111–24.

Guan, Yuanfang et al. 2012. "Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes." edited by Christos A. Ouzounis. *PLoS computational biology* 8(9):e1002694.

Guimerà, R., M. Sales-Pardo, and L. A. N. Amaral. 2007. "A Network-Based Method for Target Selection in Metabolic Networks." *Bioinformatics (Oxford, England)* 23(13):1616–22.

Guimerà, Roger and Luís A. Nunes Amaral. 2005. "Functional Cartography of Complex Metabolic Networks." *Nature* 433(7028):895–900.

Guney, Emre and Baldo Oliva. 2012. "Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization." *PloS one* 7(9):e43557.

Hamaneh, Mehdi B. and Yi-Kuo Yu. 2015. "DeCoaD: Determining Correlations among Diseases Using Protein Interaction Networks." *BMC research notes* 8:226.

Hamaneh, Mehdi Bagheri and Yi-Kuo Yu. 2014. "Relating Diseases by Integrating Gene Associations and Information Flow through Protein Interaction Network." *PloS one* 9(10):e110936.

Hamosh, A. 2004. "Online Mendelian Inheritance in Man {(OMIM),} a Knowledgebase of Human Genes and Genetic Disorders." *Nucleic Acids Research* 33(Database issue):D514–17.

Hanahan, Douglas and Robert A. Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100(1):57–70.

Hanahan, Douglas and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144(5):646–74.

Hao, Dapeng, Chuanxing Li, et al. 2014. “Network-Based Analysis of Genotype-Phenotype Correlations between Different Inheritance Modes.” *Bioinformatics (Oxford, England)* 30(22):3223–31.

Hao, Dapeng, Guangyu Wang, et al. 2014. “Systematic Large-Scale Study of the Inheritance Mode of Mendelian Disorders Provides New Insight into Human Diseasome.” *European journal of human genetics : EJHG*.

Hartwell, Leland H., John J. Hopfield, Stanislas Leibler, and Andrew W. Murray. 1999. “From Molecular to Modular Cell Biology.” *Nature* 402(6761 Suppl):C47–52.

Hidalgo, César A., Nicholas Blumm, Albert-László Barabási, and Nicholas A. Christakis. 2009. “A Dynamic Network Approach for the Study of Human Phenotypes.” *PLoS computational biology* 5(4):e1000353.

Hill, Adrian V. S. 2012. “Evolution, Revolution and Heresy in the Genetics of Infectious Disease Susceptibility.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367(1590):840–49.

Hindorff, Lucia A. et al. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106(23):9362–67.

Horsthemke, Bernhard and Joseph Wagstaff. 2008. “Mechanisms of Imprinting of the Prader-Willi/Angelman Region.” *American journal of medical genetics. Part A* 146A(16):2041–52.

Hudson, Thomas J. et al. 2010. “International Network of Cancer Genome Projects.” *Nature* 464(7291):993–98.

Huttlin, Edward L. et al. 2015. “The BioPlex Network: A Systematic Exploration of the Human Interactome.” *Cell* 162(2):425–40.

Hyman, Mark A. “Functional Diagnostics: Redefining Disease.” *Alternative therapies in health and medicine* 14(4):10–14.

Ingram, V. M. 1956. “A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin.” *Nature* 178(4537):792–94.

Ingram, V. M. 1957. “Gene Mutations in Human Hæmoglobin: The Chemical Difference Between Normal and Sickle Cell Hæmoglobin.” *Nature* 180(4581):326–28.

Janjić, Vuk and Nataša Pržulj. 2014. “The Topology of the Growing Human Interactome Data.” *Journal of integrative bioinformatics* 11(2):238.

Jensen, Lars Juhl and Peer Bork. 2008. “Biochemistry. Not Comparable, but Complementary.” *Science (New York, N.Y.)* 322(5898):56–57.

Jeong, Hawoong, Sean P. Mason, Albert-Laszlo Barabasi, and Zoltan N. Oltvai. 2001. “Lethality and Centrality in Protein Networks.” *Nature* 411(6833):41–42.

Jimenez-Sanchez, G., B. Childs, and D. Valle. 2001. “Human Disease Genes.” *Nature* 409(6822):853–55.

Jin, Wenfei, Pengfei Qin, Haiyi Lou, Li Jin, and Shuhua Xu. 2012. "A Systematic Characterization of Genes Underlying Both Complex and Mendelian Diseases." *Human molecular genetics* 21(7):1611–24.

Jobling, Mark A. and Chris Tyler-Smith. 2000. "New Uses for New Haplotypes." *Trends in Genetics* 16(8):356–62.

Jonsson, Pall F. and Paul A. Bates. 2006. "Global Topological Features of Cancer Proteins in the Human Interactome." *Bioinformatics (Oxford, England)* 22(18):2291–97.

Kajiwara, K., E. L. Berson, and T. P. Dryja. 1994. "Digenic Retinitis Pigmentosa due to Mutations at the Unlinked peripherin/RDS and ROM1 Loci." *Science (New York, N.Y.)* 264(5165):1604–8.

Kandoth, Cyriac et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502(7471):333–39.

Kanehisa, Minoru et al. 2014. "Data, Information, Knowledge and Principle: Back to Metabolism in KEGG." *Nucleic acids research* 42(Database issue):D199–205.

Karni, Shaul, Hermona Soreq, and Roded Sharan. 2009. "A Network-Based Method for Predicting Disease-Causing Genes." *Journal of computational biology a journal of computational molecular cell biology* 16(2):181–89.

Kerem, B. et al. 1989. "Identification of the Cystic Fibrosis Gene: Genetic Analysis." *Science* 245(4922):1073–80.

Keshava Prasad, T. S. et al. 2009. "Human Protein Reference Database--2009 Update." *Nucleic acids research* 37(Database issue):D767–72.

Khurana, Ekta, Yao Fu, Jieming Chen, and Mark Gerstein. 2013. "Interpretation of Genomic Variants Using a Unified Biological Network Approach." edited by Andrey Rzhetsky. *PLoS computational biology* 9(3):e1002886.

Kibbe, Warren A. et al. 2015. "Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge through Disease Data." *Nucleic acids research* 43(Database issue):D1071–78.

Kitano, Hiroaki. 2004. "Biological Robustness." *Nature reviews. Genetics* 5(11):826–37.

Kleihues, P. and L. H. Sobin. 2000. "World Health Organization Classification of Tumors." *Cancer* 88(12):2887.

Klingstrom, T. and D. Plewczynski. 2010. "Protein-Protein Interaction and Pathway Databases, a Graphical Review." *Briefings in Bioinformatics* 12(6):702–13.

Köhler, Sebastian et al. 2014. "The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data." *Nucleic acids research* 42(Database issue):D966–74.

Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. "Walking the Interactome for Prioritization of Candidate Disease Genes." *American journal of human genetics* 82(4):949–58.

Lage, Kasper et al. 2007. “A Human Phenome-Interactome Network of Protein Complexes Implicated in Genetic Disorders.” *Nature biotechnology* 25(3):309–16.

Lage, Kasper et al. 2008. “A Large-Scale Analysis of Tissue-Specific Pathology and Gene Expression of Human Disease Genes and Complexes.” *Proceedings of the National Academy of Sciences of the United States of America* 105(52):20870–75.

Lancichinetti, Andrea and Santo Fortunato. 2009. “Community Detection Algorithms: A Comparative Analysis.” *Physical Review E* 80(5):056117.

Lander, E. S. 1996. “The New Genomics: Global Views of Biology.” *Science (New York, N.Y.)* 274(5287):536–39.

Lander, E. S. et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409(6822):860–921.

Landrum, Melissa J. et al. 2014. “ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype.” *Nucleic acids research* 42(Database issue):D980–85.

Lawrence, Michael S. et al. 2014. “Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types.” *Nature* 505(7484):495–501.

Leinonen, Rasko et al. 2004. “UniProt Archive.” *Bioinformatics (Oxford, England)* 20(17):3236–37.

Lejeune, J., M. Gautier, and R. Turpin. 1959. “{Etude Des Chromosomes Somatiques de Neuf Enfants Mongoliens}.” *Comptes rendus hebdomadaires des séances de l’Académie des sciences*.

Letovsky, S. and S. Kasif. 2003. “Predicting Protein Function from Protein/protein Interaction Data: A Probabilistic Approach.” *Bioinformatics* 19(Suppl 1):i197–204.

Lima-Mendez, Gipsi and Jacques van Helden. 2009. “The Powerful Law of the Power Law and Other Myths in Network Biology.” *Molecular bioSystems* 5(12):1482–93.

Lin, Mingfeng, Henry C. Lucas, and Galit Shmueli. 2013. “Research Commentary — Too Big to Fail: Large Samples and the P -Value Problem.” *Information Systems Research* 24(4):906–17.

Lindberg, D. A. B., B. L. Humphreys, and A. T. McCray. 1993. “The Unified Medical Language System.” *Methods of Information in Medicine* 32(4):281–91.

Lindberg, Donald A. B. and Betsy L. Humphreys. 1989. “Computer Systems That Understand Medical Meaning.” *Scherrer JR, C6t6 RA, Mandil SH, eds.: Computerized Natural Medical Language Processing for Knowledge Representation. Amsterdam: Elsevier* 5–17.

Linné, Carl von. 1763. *Genera Morborum*. edited by Steinert. Uppsala.

Lipscomb, C. E. 2000. “Medical Subject Headings (MeSH).” *Bulletin of the Medical Library Association* 88(3):265–66.

Litt, M. and J. A. Luty. 1989. “A Hypervariable Microsatellite Revealed by in Vitro Amplification of a Dinucleotide Repeat within the Cardiac Muscle Actin Gene.” *American journal of human genetics* 44(3):397–401.

Lopes, Tiago J. S. et al. 2011. "Tissue-Specific Subnetworks and Characteristics of Publicly Available Human Protein Interaction Databases." *Bioinformatics (Oxford, England)* 27(17):2414–21.

López-Bigas, Núria and Christos A. Ouzounis. 2004. "Genome-Wide Identification of Genes Likely to Be Involved in Human Genetic Disease." *Nucleic acids research* 32(10):3108–14.

MacArthur, Daniel G. et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science (New York, N.Y.)* 335(6070):823–28.

Mackenbach, Johan P. 2004. "Carl von Linné, Thomas McKeown, and the Inadequacy of Disease Classifications." *European journal of public health* 14(3):225.

Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado. 2011. "What Can Exome Sequencing Do for You?" *Journal of Medical Genetics* 48(9):580–89.

Malone, James et al. 2010. "Modeling Sample Variables with an Experimental Factor Ontology." *Bioinformatics (Oxford, England)* 26(8):1112–18.

Man, P. Y. W. 2002. "Leber Hereditary Optic Neuropathy." *Journal of Medical Genetics* 39(3):162–69.

Manolio, Teri A. et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461(7265):747–53.

Mattingly, Carolyn J., Glenn T. Colby, John N. Forrest, and James L. Boyer. 2003. "The Comparative Toxicogenomics Database (CTD)." *Environmental health perspectives* 111(6):793–95.

Maurano, Matthew T. et al. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science (New York, N.Y.)* 337(6099):1190–95.

McKusick, Victor A. 1998. *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*. JHU Press.

Menche, Jörg et al. 2015. "Disease Networks. Uncovering Disease-Disease Relationships through the Incomplete Interactome." *Science (New York, N.Y.)* 347(6224):1257601.

Milenkovic, Tijana, Vesna Memisevic, Anand K. Ganesan, and Natasa Przulj. 2010. "Systems-Level Cancer Gene Identification from Protein Interaction Network Topology Applied to Melanogenesis-Related Functional Genomics Data." *Journal of the Royal Society, Interface / the Royal Society* 7(44):423–37.

Mira, Marcelo T. et al. 2004. "Susceptibility to Leprosy Is Associated with PARK2 and PACRG." *Nature* 427(6975):636–40.

Misch, E. Ann and Thomas R. Hawn. 2008. "Toll-like Receptor Polymorphisms and Susceptibility to Human Disease." *Clinical science (London, England : 1979)* 114(5):347–60.

Mitra, Koyel, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. 2013. "Integrative Approaches for Finding Modular Structure in Biological Networks." *Nature reviews. Genetics* 14(10):719–32.

Moreau, Yves and Léon-Charles Tranchevent. 2012. "Computational Tools for Prioritizing Candidate Genes: Boosting Disease Gene Discovery." *Nature reviews. Genetics* 13(8):523–36.

Moriyama, Iwao Milton Moriyama et al. 2010. *History of the Statistical Classification of Diseases and Causes of Death*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Mosca, Roberto et al. 2015. "dSysMap: Exploring the Edgetic Role of Disease Mutations." *Nature methods* 12(3):167–68.

Mosca, Roberto, Arnaud Céol, and Patrick Aloy. 2013. "Interactome3D: Adding Structural Details to Protein Networks." *Nature methods* 10(1):47–53.

Mosca, Roberto, Tirso Pons, Arnaud Céol, Alfonso Valencia, and Patrick Aloy. 2013. "Towards a Detailed Atlas of Protein-Protein Interactions." *Current opinion in structural biology* 23(6):929–40.

Nachtomy, Ohad, Ayelet Shavit, and Zohar Yakhini. 2007. "Gene Expression and the Concept of the Phenotype." *Studies in history and philosophy of biological and biomedical sciences* 38(1):238–54.

Navlakha, Saket and Carl Kingsford. 2010. "The Power of Protein Interaction Networks for Associating Genes with Diseases." *Bioinformatics (Oxford, England)* 26(8):1057–63.

Newman, M. E. J. 2006. "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences of the United States of America* 103(23):8577–82.

Newport, M. J. et al. 1996. "A Mutation in the Interferon-Gamma-Receptor Gene and Susceptibility to Mycobacterial Infection." *The New England journal of medicine* 335(26):1941–49.

Novas, Rossina, Magdalena Cardenas-Rodriguez, Florencia Irigoín, and Jose L. Badano. 2015. "Bardet-Biedl Syndrome: Is It Only Cilia Dysfunction?" *FEBS letters*.

Nowell, Peter C. 2007. "Discovery of the Philadelphia Chromosome: A Personal Perspective." *The Journal of clinical investigation* 117(8):2033–35.

Nussbaum, Robert, Roderick R. McInnes, and Huntington F. Willard. 2007. *Thompson & Thompson Genetics in Medicine*. Elsevier Health Sciences.

Orchard, Sandra et al. 2014. "The MIntAct Project--IntAct as a Common Curation Platform for 11 Molecular Interaction Databases." *Nucleic acids research* 42(Database issue):D358–63.

Organization, World Health. 2004. *International Statistical Classification of Diseases and Related Health Problems*. World Health Organization.

Oti, M. and H. G. Brunner. 2007. "The Modular Nature of Genetic Diseases." *Clinical Genetics* 71(1):1–11.

Oti, M., B. Snel, M. A. Huynen, and H. G. Brunner. 2006. "Predicting Disease Genes Using Protein-Protein Interactions." *Journal of medical genetics* 43(8):691–98.

Paik, Hyojung, Hyoung-Sam Heo, Hyo-jeong Ban, and Seong Beom Cho. 2014. "Unraveling Human Protein Interaction Networks Underlying Co-Occurrences of Diseases and Pathological Conditions." *Journal of translational medicine* 12(1):99.

Park, Juyong, Deok-Sun Lee, Nicholas A. Christakis, and Albert-Laszlo Barabasi. 2009. "The Impact of Cellular Networks on Disease Comorbidity." *Molecular Systems Biology* 5(262):262.

Park, Solip et al. 2011. "Protein Localization as a Principal Feature of the Etiology and Comorbidity of Genetic Diseases." *Molecular systems biology* 7:494.

Park, Solip, Jae-Seong Yang, Sung Key Jang, and Sanguk Kim. 2009. "Construction of Functional Interaction Networks through Consensus Localization Predictions of the Human Proteome." *Journal of proteome research* 8(7):3367–76.

Pauling, L., H. A. Itano, S. J. Singer, and I. C. Wells. 1949. "Sickle Cell Anemia, a Molecular Disease." *Science* 110(2865):543–48.

Raeymaekers, P. et al. 1991. "Duplication in Chromosome 17p11.2 in Charcot-Marie-Tooth Neuropathy Type 1a (CMT 1a). The HMSN Collaborative Research Group." *Neuromuscular disorders : NMD* 1(2):93–97.

Rappaport, Noa et al. 2013. "MalaCards: An Integrated Compendium for Diseases and Their Annotation." *Database : the journal of biological databases and curation* 2013(0):bat018.

Razick, Sabry, George Magklaras, and Ian M. Donaldson. 2008. "iRefIndex: A Consolidated Protein Interaction Database with Provenance." *BMC bioinformatics* 9:405.

Rebhan, M., V. Chalifa-Caspi, J. Prilusky, and D. Lancet. 1998. "GeneCards: A Novel Functional Genomics Compendium with Automated Data Mining and Query Reformulation Support." *Bioinformatics* 14(8):656–64.

Reich, David E. and Eric S. Lander. 2001. "On the Allelic Spectrum of Human Disease." *Trends in Genetics* 17(9):502–10.

Riordan, J. R. et al. 1989. "Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA." *Science* 245(4922):1066–73.

Risch, N. J. 2000. "Searching for Genetic Determinants in the New Millennium." *Nature* 405(6788):847–56.

Risch, N. and K. Merikangas. 1996. "The Future of Genetic Studies of Complex Human Diseases." *Science (New York, N.Y.)* 273(5281):1516–17.

Rives, Alexander W. and Timothy Galitski. 2003. "Modular Organization of Cellular Networks." *Proceedings of the National Academy of Sciences of the United States of America* 100(3):1128–33.

Robinson, Matthew R., Naomi R. Wray, and Peter M. Visscher. 2014. "Explaining Additional Genetic Variation in Complex Traits." *Trends in genetics : TIG* 30(4):124–32.

Robinson, Peter N. et al. 2008. "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease." *American journal of human genetics* 83(5):610–15.

Rodenhiser, David and Mellissa Mann. 2006. “Epigenetics and Human Disease: Translating Basic Biology into Clinical Applications.” *CMAJ : Canadian Medical Association journal = journal de l’Association medicale canadienne* 174(3):341–48.

Rogers, F. B. 1963. “Medical Subject Headings.” *Bulletin of the Medical Library Association* 51:114–16.

Rolland, Thomas et al. 2014. “A Proteome-Scale Map of the Human Interactome Network.” *Cell* 159(5):1212–26.

Roque, Francisco S. et al. 2011. “Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts” edited by Marylyn D. Ritchie. *PLoS Computational Biology* 7(8):e1002141.

Rosvall, Martin and Carl T. Bergstrom. 2008. “Maps of Random Walks on Complex Networks Reveal Community Structure.” *Proceedings of the National Academy of Sciences of the United States of America* 105(4):1118–23.

Rual, Jean-François et al. 2005. “Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network.” *Nature* 437(7062):1173–78.

Rubio-Perez, Carlota et al. 2015. “In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities.” *Cancer Cell* 27(3):382–96.

Rzhetsky, Andrey, David Wajngurt, Naeun Park, and Tian Zheng. 2007. “Probing Genetic Overlap among Complex Human Phenotypes.” *Proceedings of the National Academy of Sciences of the United States of America* 104(28):11694–99.

Schaefer, Martin H. et al. 2012. “HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores.” edited by Charlotte M. Deane. *PloS one* 7(2):e31826.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. “Common vs. Rare Allele Hypotheses for Complex Diseases.” *Current opinion in genetics & development* 19(3):212–19.

Schriml, Lynn M. and Elvira Mitraka. 2015. “The Disease Ontology: Fostering Interoperability between Biological and Clinical Human Disease-Related Data.” *Mammalian genome : official journal of the International Mammalian Genome Society*.

Schriml, Lynn Marie et al. 2012. “Disease Ontology: A Backbone for Disease Semantic Integration.” *Nucleic acids research* 40(Database issue):D940–46.

Sharan, Roded, Igor Ulitsky, and Ron Shamir. 2007. “Network-Based Prediction of Protein Function.” *Molecular systems biology* 3(1):88.

Sharma, Amitabh et al. 2013. “Network-Based Analysis of Genome Wide Association Data Provides Novel Candidate Genes for Lipid and Lipoprotein Traits.” *Molecular & cellular proteomics : MCP* 12(11):3398–3408.

Sharma, Amitabh et al. 2015. “A Disease Module in the Interactome Explains Disease Heterogeneity, Drug Response and Captures Novel Pathways and Genes in Asthma.” *Human molecular genetics* 24(11):3005–20.

Shimoyama, Mary et al. 2015. “The Rat Genome Database 2015: Genomic, Phenotypic and Environmental Variations and Disease.” *Nucleic acids research* 43(Database issue):D743–50.

Sioutos, Nicholas et al. 2007. “NCI Thesaurus: A Semantic Model Integrating Cancer-Related Clinical and Molecular Information.” *Journal of biomedical informatics* 40(1):30–43.

Smedley, Damian et al. 2014. “Walking the Interactome for Candidate Prioritization in Exome Sequencing Studies of Mendelian Diseases.” *Bioinformatics (Oxford, England)* 30(22):3215–22.

Smith, Cynthia L. and Janan T. Eppig. 2009. “The Mammalian Phenotype Ontology: Enabling Robust Annotation and Comparative Analysis.” *Wiley interdisciplinary reviews. Systems biology and medicine* 1(3):390–99.

Smith, Cynthia L. and Janan T. Eppig. 2012. “The Mammalian Phenotype Ontology as a Unifying Standard for Experimental and High-Throughput Phenotyping Data.” *Mammalian Genome* 23(9-10):653–68.

Smith, Cynthia L., Carroll-Ann W. Goldsmith, and Janan T. Eppig. 2005. “The Mammalian Phenotype Ontology as a Tool for Annotating, Analyzing and Comparing Phenotypic Information.” *Genome biology* 6(1):R7.

Spirin, Victor and Leonid a Mirny. 2003. “Protein Complexes and Functional Modules in Molecular Networks.” *Proceedings of the National Academy of Sciences of the United States of America* 100(21):12123–28.

Stearns, M. Q., C. Price, K. A. Spackman, and A. Y. Wang. 2001. “SNOMED Clinical Terms: Overview of the Development Process and Project Status.” *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* 662–66.

Stelzl, Ulrich et al. 2005. “A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome.” *Cell* 122(6):957–68.

Stenson, Peter D. et al. 2014. “The Human Gene Mutation Database: Building a Comprehensive Mutation Repository for Clinical and Molecular Genetics, Diagnostic Testing and Personalized Genomic Medicine.” *Human genetics* 133(1):1–9.

Stevens, Adam, Stefan Meyer, Daniel Hanson, Peter Clayton, and Rachelle P. Donn. 2014. “Network Analysis Identifies Protein Clusters of Functional Importance in Juvenile Idiopathic Arthritis.” *Arthritis research & therapy* 16(3):R109.

Stranger, Barbara E., Eli A. Stahl, and Towfique Raj. 2011. “Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics.” *Genetics* 187(2):367–83.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. “The Cancer Genome.” *Nature* 458(7239):719–24.

Stumpf, Michael P. H. et al. 2008. “Estimating the Size of the Human Interactome.” *Proceedings of the National Academy of Sciences of the United States of America* 105(19):6959–64.

Sunyaev, Shamil et al. 2013. “Edgotype: A Fundamental Link between Genotype and Phenotype.” *Current Opinion in Genetics & Development* 23(6):649–57.

Suthram, Silpa et al. 2010. “Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets” edited by Yanay Ofran. *PLoS Computational Biology* 6(2):10.

Tamborero, David et al. 2013. “Comprehensive Identification of Mutational Cancer Driver Genes across 12 Tumor Types.” *Scientific reports* 3:2650.

The UniProt Consortium. 2014. “UniProt: A Hub for Protein Information.” *Nucleic Acids Research* 43(D1):D204–12.

Thorisson, Gudmundur A. et al. 2009. “HGVbaseG2P: A Central Genetic Association Database.” *Nucleic acids research* 37(Database issue):D797–802.

Tryka, Kimberly A. et al. 2014. “NCBI’s Database of Genotypes and Phenotypes: dbGaP.” *Nucleic acids research* 42(Database issue):D975–79.

Tsui, Lap-Chee and Ruslan Dorfman. 2013. “The Cystic Fibrosis Gene: A Molecular Genetic Perspective.” *Cold Spring Harbor perspectives in medicine* 3(2):a009472.

Valderas, Jose M., Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. 2009. “Defining Comorbidity: Implications for Understanding Health and Health Services.” *Annals of family medicine* 7(4):357–63.

Vanunu, Oron, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. 2010. “Associating Genes and Protein Complexes with Disease via Network Propagation” edited by Wyeth W Wasserman. *PLoS Computational Biology* 6(1):9.

Vázquez, Miguel, Alfonso Valencia, and Tirso Pons. 2015. “Structure-PPi: A Module for the Annotation of Cancer-Related Single-Nucleotide Variants at Protein-Protein Interfaces.” *Bioinformatics (Oxford, England)* 31(14):2397–99.

Venkatesan, Kavitha et al. 2009. “An Empirical Framework for Binary Interactome Mapping.” *Nature methods* 6(1):83–90.

Vestbo, Jørgen et al. 2013. “Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease.” *American Journal of Respiratory and Critical Care Medicine*.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. “Five Years of GWAS Discovery.” *American journal of human genetics* 90(1):7–24.

Vogelstein, Bert et al. 2013. “Cancer Genome Landscapes.” *Science (New York, N.Y.)* 339(6127):1546–58.

Wachi, Shinichiro, Ken Yoneda, and Reen Wu. 2005. “Interactome-Transcriptome Analysis Reveals the High Centrality of Genes Differentially Expressed in Lung Cancer Tissues.” *Bioinformatics (Oxford, England)* 21(23):4205–8.

Wang, Amy Y., Jeremiah H. Sable, and Kent A. Spackman. 2002. “The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content.” *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* 845–49.

Wang, Xiujuan et al. 2012. "Three-Dimensional Reconstruction of Protein Networks Provides Insight into Human Genetic Disease." *Nature biotechnology* 30(2):159–64.

Wang, Yiting, Joe A. Wrennall, Zhiwei Cai, Hongyu Li, and David N. Sheppard. 2014. "Understanding How Cystic Fibrosis Mutations Disrupt CFTR Function: From Single Molecules to Animal Models." *The international journal of biochemistry & cell biology* 52:47–57.

Weber, J. L. and P. E. May. 1989. "Abundant Class of Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction." *American journal of human genetics* 44(3):388–96.

Weinstein, John N. et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature genetics* 45(10):1113–20.

Welter, Danielle et al. 2014. "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations." *Nucleic acids research* 42(Database issue):D1001–6.

Wodak, Shoshana J., James Vlasblom, Andrei L. Turinsky, and Shuye Pu. 2013. "Protein-Protein Interaction Networks: The Puzzling Riches." *Current opinion in structural biology* 23(6):941–53.

Wu, Xuebing, Rui Jiang, Michael Q. Zhang, and Shao Li. 2008. "Network-Based Global Inference of Human Disease Genes." *Molecular Systems Biology* 4(189):189.

Xue, Yali et al. 2012. "Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing." *American journal of human genetics* 91(6):1022–32.

Ylikallio, Emil and Anu Suomalainen. 2012. "Mechanisms of Mitochondrial Diseases." *Annals of medicine* 44(1):41–59.

Yu, W., M. Clyne, M. J. Khoury, and M. Gwinn. 2010. "Phenopedia and Genopedia: Disease-Centered and Gene-Centered Views of the Evolving Knowledge of Human Genetic Associations." *Bioinformatics (Oxford, England)* 26(1):145–46.

Yu, Wei, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J. Khoury. 2008. "A Navigator for Human Genome Epidemiology." *Nature genetics* 40(2):124–25.

Zhao, Shiwen and Shao Li. 2010. "Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification." *PloS one* 5(7):e11764.

Zhao, Shiwen and Shao Li. 2012. "A Co-Module Approach for Elucidating Drug-Disease Associations and Revealing Their Molecular Basis." *Bioinformatics (Oxford, England)* 28(7):955–61.

Zhong, Quan et al. 2009. "Edgetic Perturbation Models of Human Inherited Disorders." *Molecular systems biology* 5(321):321.

Zhou, XueZhong, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. "Human Symptoms-Disease Network." *Nature communications* 5:4212.

Zuk, Or, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. 2012. "The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability."

Proceedings of the National Academy of Sciences of the United States of America 109(4):1193–98.

La impressió d'aquesta tesi ha estat possible gràcies a l'ajut per a la finalització de tesis doctorals de la Fundació IMIM