In Silico Edgetic Profiling and Network Analysis of Human Genetic Variants, with an Application to Disease Module Detection

by

Hongzhu Cui

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Bioinformatics and Computational Biology

by

May 2020

APPROVED:

Professor Dmitry Korkin Worcester Polytechnic Institute Advisor	Professor Amity L. Manning Worcester Polytechnic Institute Committee Member
Professor Zheyang Wu Worcester Polytechnic Institute Committee Member	Professor Manoj Bhasin Emory University External Committee Member
Professor Dmitry Korkin Worcester Polytechnic Institute Head of Department	

Home Rooms – S4E3

"I love the first day, man. Everybody all friendly an' shit"
– Namond Brice

The Detail – S1E2

"You cannot lose if you do not play."
– Marla Daniels

The Pager – S1E5

"…a little slow, a little late."
– Avon Barksdale

Took – S5E7

"They don't teach it in law school."
– Pearlman

The Wire – S1E6

"…and all the pieces matter."
– Freamon

Final Grades – S4E13

"If animal trapped call 410-844-6286"
– Baltimore, traditional

Slapstick – S3E9

"…while you're waiting for moments that never come."
– Freamon

Collateral Damage – S2E2

"They can chew you up, but they gotta spit you out."
– McNulty

Hard Cases – S2E4

"If I hear music, I'm gonna dance."
– Greggs

Time After Time – S3E1

"Don't matter how many times you get burnt, you just keep doin' the same."
– Bodie

Port in a Storm – S2E12

"Business. Always business."
– The Greek

Moral Midgetry – S3E8

"Crawl, walk, and then run."
– Clay Davis

Backwash – S2E7

"Don't worry kid. You're still on the clock."
– Horseface

Unto Others – S4E7

"Aw yeah. That golden rule."
– Bunk

Boys of Summer – S4E1

"Lambs to the slaughter here."
– Marcia Donnelly

Cleaning Up – S1E12

"This is me, yo, right here."
– Wallace

The Buys – S1E3

"The king stay the king."
– D'Angelo

The Dickensian Aspect – S5E6

"If you have a problem with this, I understand completely."
– Freamon

- Some opening credit quotes from my favorite TV show *The Wire*

https://www.youtube.com/watch?v=QqNNpGlUKHw

Abstract

In the past several decades, Next Generation Sequencing (NGS) methods have produced large amounts of genomic data at the exponentially increasing rate. It has also enabled tremendous advancements in the quest to understand the molecular mechanisms underlying human complex traits. Along with the development of the NGS technology, many genetic variation and genotype–phenotype databases and functional annotation tools have been developed to assist scientists to better understand the intricacy of the data. Together, the above findings bring us one step closer towards mechanistic understanding of the complex phenotypes. However, it has rarely been possible to translate such a massive amount of information on mutations and their associations with phenotypes into biological or therapeutic insights, and the mechanisms underlying genotype-phenotype relationships remain partially explained. Meanwhile, increasing evidence shows that biological networks are essential, albeit not sufficient, for the better understanding of these mechanisms. Among them, protein-protein interaction (PPI) network studies have attracted perhaps most attention. Our overarching goal of this dissertation is to (i) perform a systematic study to investigate the role of pathogenic human genetic variant in the interactome; (ii) examine how common population-specific SNVs affect PPI network and how they contribute to population phenotypic variance and disease susceptibility; and (iii) develop a novel framework to incorporate the functional effect of mutations for disease module detection.

In this dissertation, we first present a systematic multi-level characterization of human mutations associated with genetic disorders by determining their individual and combined interaction-rewiring effects on the human interactome. Our *in-silico* analysis highlights the intrinsic differences and important similarities between the pathogenic single nucleotide variants (SNVs) and frameshift mutations. Functional

profiling of SNVs indicates widespread disruption of the protein-protein interactions and synergistic effects of SNVs. The coverage of our approach is several times greater than the recently published experimental study and has the minimal overlap with it, while the distributions of determined edgotypes between the two sets of profiled mutations are remarkably similar. Case studies reveal the central role of interaction-disrupting mutations in type 2 diabetes mellitus and suggest the importance of studying mutations that abnormally strengthen the protein interactions in cancer.

Second, aided with our SNP-IN tool, we performed a systematic edgetic profiling of population specific non-synonymous SNVs and interrogate their role in the human interactome. Our results demonstrated that a considerable amount of normal nsSNVs can cause disruptive impact to the interactome. We also showed that genes enriched with disruptive mutations associated with diverse functions and have implications in various diseases. Further analysis indicates that distinct gene edgetic profiles among major populations can help explain the population phenotypic variance. Finally, network analysis reveals phenotype-associated modules are enriched with disruptive mutations and the difference of the accumulated damage in such modules may suggest population-specific disease susceptibility.

Lastly, we propose and develop a computational framework, Discovering most IMpacted SUbnetworks in interactoMe (DIMSUM), which enables the integration of genome-wide association studies (GWAS) and functional effects of mutations into the protein–protein interaction (PPI) network to improve disease module detection. Specifically, our approach incorporates and propagates the functional impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on PPIs to implicate the genes that are most likely influenced by the disruptive mutations, and to identify the module with the greatest functional impact. Comparison against state-of-the-art seed-based module detection methods shows that our approach could yield modules that are biologically more relevant and have stronger association with the studied disease.

With the advancement of next-generation sequencing technology that drives precision medicine, there is an increasing demand in understanding the changes in molecular mechanisms caused by the specific genetic variation. The current and future *in-silico* edgotyping tools present a cheap and fast solution to deal with the rapidly growing datasets of discovered mutations. Our work shows the feasibility of a large-scale *in-silico* edgetic study and revealing insights into the orchestrated play of mutations inside a complex PPI network. We also expect for our module detection method to become a part of the common toolbox for the disease module analysis, facilitating the discovery of new disease markers.

Acknowledgements

I would like to extend my sincere gratitude to my dissertation advisor Professor Dmitry Korkin. He is a respectful and resourceful scholar. He has also exemplified to me what entails to do academic research: hard work, open-mindedness, scientific attitude, and perseverance. These have become invaluable assets in my life.

I owe my gratitude to my dissertation committee members: Professor Zheyang Wu, Professor Amity Manning and Professor Manoj Bhasin, for their help, patience and encouragement, which support me keeping improving this dissertation.

A special thanks to my fellow Korkin lab members over the years, including but not limited to: Nan Zhao, Andi Dhroso, Nathan Johnson, Katie Hughes, Oleksandr Narykov, Pavel Terentiev, Suhas Srinivasan, Ana Leshchyk, Nan Hu, Ziyang Gao, Senbao Lu. You all have been an invaluable resource and good friends.

Finally, I would like to thank my parents and my big brother. I thank them for their unconditional love and support, and all the precious values they taught me that have become part of my life.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Since the first evidence of the genetic complexity of cancer [1], numerous research efforts have been dedicated to deciphering the mechanistic nature of polygenic diseases. Due to the rapid advancement of the next generation sequencing (NGS) technologies, including whole-genome [2] and whole-exome sequencing [3], and most recently single-cell transcriptomics [4], we have been able to sequence and analyze thousands of genomes at a much lower cost. As result, the high-throughput experiments produce large amounts of genomic data at the exponentially increasing rate. These data have also transformed the design of genome-wide association studies and enabled us to do comprehensive analyses of the genotype–phenotype relationships [5]. Most importantly, the studies have provided us with an extensive list of susceptible alleles and genes associated with complex diseases, as well as with catalogs of disease-relevant mutations [6]. The list includes the majority of common and many rare complex diseases. Furthermore, it is expected that a comprehensive catalog of nearly all human genomic variations will be available soon [7].

Next-generation sequencing (NGS) technologies have also enabled advances in human population genetics and comparative genomics and have made it possible to gain increasing insight into the nature of genetic diversity[8-11]. The 1000 Genomes Project[12] have produced first large population-scale sequencing data and established by far the most detailed catalogue of human genetic variation. Since January 2008, scientists at the Sanger Institute, BGI Shenzhen and the National Human Genome Research Institute planned to sequence a minimum of 1,000 human genomes in the era of next-generation sequencing (NGS). After that, more and more large-scale sequencing

consortium sequencing effort have been launched to generate more local data sets. Notably, the UK10K Project aims to sequencing 4000 genomes from the UK, along with 6000 exomes from individuals with selected extreme phenotypes from National Health Service systems. Japan also initiated a similar genomic cohort study (1KJPN) to identify genetic variants affecting health, disease and responses to drugs and environmental factors[13]. It produced the whole-genome sequences of 1,070 healthy Japanese individuals and a Japanese population genetic variation reference panel. These regional cohort sequencing projects have shown that many genetic variations are population-specific. The results suggest that the genetic variance across/within populations are important to uncover the mechanistic details of complex diseases and need to be considered to interpret their relevance to certain phenotypes.

Rapid progress in high-throughput -omics technologies moves us one step closer to the datacalypse in life sciences. In spite of the already generated volumes of data, our knowledge of the molecular mechanisms underlying complex genetic diseases remains limited. Increasing evidence shows that biological networks are essential, albeit not sufficient, for the better understanding of these mechanisms[6, 14]. Among them, protein-protein interaction (PPI) network studies have attracted perhaps most attention[15]. These complex networks have been utilized to explore the genotype-to-phenotype relationships. However, the mechanisms underlying genotype-phenotype relationships remain partially explained[16, 17].

Since a disease phenotype could be linked to a synergistic effect of multiple genetic variations targeting a common component of the reference interactome [18], module-based approaches are promising for studying complex diseases. Integrating PPI data with genetic variation data related to disease helps in determining modules and pathways perturbed in a disease of interest [19]. One basic way to uncover the perturbed modules or pathways is mapping the disease genes to an interaction network and searching for the modules enriched with genetic variations [19]. Several complex disease studies have demonstrated the utility of this idea [20].

The dissertation addresses several challenges in current research communities. The first

important question, how disease-associated mutations impair protein activities in the context of biological networks remains mostly unclear, let alone how to quantify the rewiring effects of a group of genetic variations on the interactome and interpret their clinical relevance. Secondly, studies have suggested that many variants are population-specific. However, it is still up in the air whether these normal population-specific mutations can cause distinct rewiring effects in the interactome and are related to the phenotypic variance across populations. Lastly, many studies have investigated and confirmed the important roles of interaction rewiring and network rewiring caused by mutation in complex diseases[21-23]. However, there has not been a computational strategy to incorporate the functional impact of mutations on protein-protein interaction for identifying the disease module in the interactome. We will address these challenges in the Research Objectives section below.

1.2 Research Objectives

Our overarching goal of this dissertation is to (i) perform a systematic study to investigate the role of pathogenic human genetic variant in the interactome, especially how they impact the protein-protein interaction activities and cause the network rewiring; (ii) examine how common population-specific SNVs affect PPI network and how they contribute to population phenotypic variance and disease susceptibility; and (iii) develop a novel computational framework to incorporate the functional effect of mutations on protein-protein interactions for disease module detection. Each of these research objectives is discussed in detail below.

1.2.1 Systematically characterize mutations associated with genetic diseases in the interactome and explore their functional role

Experimentally profiling missense mutations using the interaction assays remains costly and laborious. Armed with our recently developed SNP-IN tool [24], we will bypass the bottleneck and systematically characterize the genetic mutations at a much lower cost. We intend to present a systematic multi-level characterization of human mutations

associated with genetic disorders by determining their individual and combined interaction-rewiring, “edgetic”, effects on the human interactome. Also, it is intriguing to study the effect the removal of a set of disease-associated protein and PPIs on the performance of the whole networked system and quantify the rewiring effects of a certain group of variations. Further, we seek to perturb and rewire the PPI network and study how they can be linked to the phenotypes. Case studies will be performed to reveal the central role of interaction-disrupting mutations in complex diseases, such as type 2 diabetes mellitus. With the advancement of next-generation sequencing technology that drives precision medicine, there is an increasing demand in understanding the changes in molecular mechanisms caused by the patient-specific genetic variation. Our in-silico edgotyping tools present a cheap and fast solution to deal with the rapidly growing datasets from various ongoing sequencing projects.

1.2.2 Perform an edgotype based analysis of population specific SNV in human interactome

Classical ‘one-gene/one-disease’ models have the flaw that it cannot fully interpret the complicated genotype-to-phenotype associations in human disease, as genes and their products function not in isolation but as components of intricate networks. Accordingly, ‘edgetics’ is proposed to uncover how disease-causing mutations affect systems or interactome properties. However, it is unknown whether the edgetic property of common genetic variations could be helpful to understand the diverse phenotype across different population. With the functional annotation from SNP-IN, we are able to provide the population-specific edgetic landscape of the human interactome. We will create a comprehensive catalog of population-specific edgetic effects at the whole-interactome level. We will also compare the functional impact of population-specific variations with pathogenic mutations studied in the first research objective. These functional characterizations will allow us to examine whether genes enriched with disruptive mutations obtained from the healthy populations are associated with certain biological processes or molecular functions. Finally, we will investigate whether the difference of the accumulated damage in the interacome can be linked with different disease susceptibility and population phenotypic variance. We expect that our approach will provide a more

accurate and in-depth characterization of the functional consequences of ethnic-specific alleles, leading to a better understanding of the clinical and phenotypic outcome.

1.2.3 Develop a computational framework incorporating the functional impact of the mutations for disease module detection.

The identification of disease-specific functional modules in the human interactome can provide a more focused insight into the mechanistic nature of the disease. However, carving a disease associated module from the whole interactome is a challenging task. Here, we will develop a novel computational framework that allows for flexible integration of genome-wide association studies (GWAS) and functional effects of mutations into the protein–protein interaction (PPI) network. We propose to then propagate the network rewiring effect of SNVs to detect the disease specific module. It is noteworthy to mention that the detected module is not only disease-associated but also mutation-specific. Specifically, our approach incorporates and propagates the functional impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on PPIs to implicate the genes that are most likely influenced by the disruptive mutations, and to identify the module with the greatest functional impact. We will compare our method against state-of-the-art seed-based module detection methods to show that our approach could yield modules that are biologically more relevant and have stronger association with the studied disease. We expect for our method to become a part of the common toolbox for the disease module analysis, facilitating the discovery of new disease markers.

1.3 Dissertation Organization

The rest of this proposal is organized as follows. Chapter 2 first provides the background knowledge and some related research works needed for this dissertation. In following three chapters (Chapter 3-5), we discuss in detail the three main research topics in this dissertation, namely “multilayer view of pathogenic mutations in human interactome” (research topic 1), “edgotype based analysis of population-specific mutations” (research topic 2) and “DIMSUM: Discovering most IMpacted SUbnetworks in interactoMe” (research topic 3), respectively. The discussion of each of the three research topics

includes three main sections: methods section explains how the study is carried out; results section objectively presents the main findings; and each chapter ends with a discussion of some key results and related work. Chapter 6 concludes this dissertation and discusses promising future work.

Chapter 2 Background and Related Work

In this chapter, we give a brief review of some basic concepts and background knowledge related to this dissertation. We also discuss some recently published research works and their relevance to this dissertation.

2.1 Human Genetic Variants

This section will summarize the current knowledge of human genetic variants. Genetic variants are the difference in DNA sequences from the reference DNA sequence. Single nucleotide variants (SNVs) are DNA sequence variations that occur when a single nucleotide differs from the reference DNA sequence. This is the most common source of human genetic variations. It is also the main genetic variants studied in this dissertation. Besides, insertion-deletion mutations (indels) are another group of common genetic variations. Relative to the reference genome, insertions are when additional nucleotides inserted in a DNA sequence; deletions are when there are missing nucleotides. Structural variation (SV) are is generally defined as a region of DNA approximately 1 kb and larger in size where large sections of a chromosome or even whole chromosomes are inserted, deleted, duplicated or rearranged in some manner. In this section, we mainly focus on the single nucleotide variants (Fig 2-1). We will survey how current experimental techniques, especially Next Generation Sequencing (NGS) experiments, are applied to uncover these genetic variations. We will also talk about many variation and genotype-phenotype databases developed to store the huge amount of genetic variation data and the computational tools that are used to characterize these genetic changes.

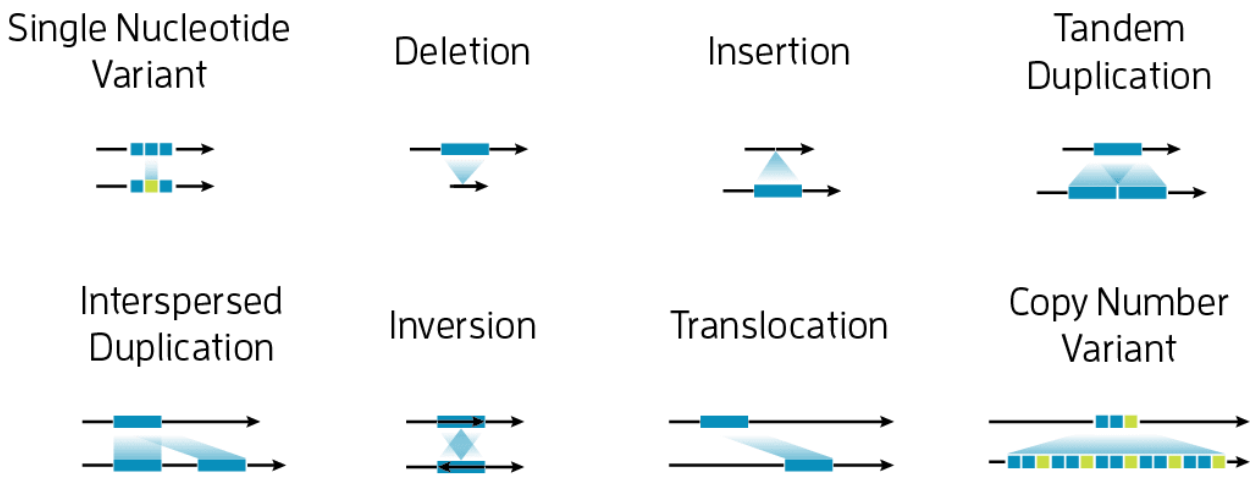


Figure 2-1 Different types of human genetic variants: SNVs, indels, duplications and CNVs. The image is adapted from Wikipedia.

2.1.1 Single Nucleotide Variants (SNVs)

Single Nucleotide Variants are the most common type of genetic variation among people. Each SNV represents a substitution of a single nucleotide that occurs at a specific position in the genome. SNV having a Minor Allele Frequency (MAF) larger than 0.5% in the population is also called Single Nucleotide Polymorphism. However, in terms of characterizing genetic variation's impact on the protein-protein interaction and the interactome, there is no distinction between the two. Thus, SNV and SNP sometimes are interchangeable through this dissertation. Being one of the most prevalent types of genetic variation in humans, SNVs can occur in both coding and non-coding regions of the genome. SNVs within a coding sequence do not necessarily cause the residue change of the protein sequence. These SNVs are called synonymous SNVs. On the other hand, we call SNVs causing residue changes non-synonymous SNVs. The nonsynonymous SNVs can be further divided into two categories: missense SNVs and nonsense SNVs. Missense SNVs substitute an amino acid residue; whereas nonsense SNVs results in a premature stop codon and in an incomplete, and usually nonfunctional protein product. Synonymous SNVs and SNVs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, messenger RNA splicing *etc*. Thus, they could also exert some biological functions. But these variations are not in the study scope of this dissertation.

SNVs occur throughout an individual's genome. It is estimated that there are roughly 4 to 5 million SNPs in a person's genome. An average gene is estimated to have several nsSNVs. With the technological advancements of Next Generation Sequencing, millions of SNPs have been determined. Most SNVs, such as synonymous SNVs, are expected to have no significant impact on individual's health or development[25]. However, some of these genetic mutations are known to play an important role in human health and diseases[26]. Many studies have shown that SNVs are associated with an individual's normal phenotypes[27, 28], susceptibility and risk of developing particular diseases[29] and response to certain drugs[27, 30]. Nevertheless, our knowledge of Single Nucleotide Variants is limited and incomplete[31, 32].

2.1.2 Genetic Variation Detection Techniques

In the past decade, the technological advancements have propelled genome sequencing with the rate that has surpassed the Moore's Law[33], generating petabytes of information and making genomics one of the first scientific areas that entered the era of Big Data. Due to the reduced cost of DNA sequencing and NGS's superior coverage and resolution, the NGS technology is taking over the traditional array-based detection methods [34]. The technology provides geneticists and bioinformatics researchers with new sequence-based reference datasets and necessitates revisiting the tools of the genome-wide association studies (GWAS) era [35]. Armored with the rapidly growing NGS data, scientists are now reaching beyond the GWAS methods, which primarily focus on genetic markers that are intended to represent causal variation indirectly, with the goal of identifying causal variants directly. This possibility is often considered the key advantage of the new sequencing approaches over genotyping methods [36], especially given the widely accepted hypothesis that many complex genetic diseases could be influenced by rare variants in many different genes [37].

The diversity of Next Generation Sequencing (NGS) methods [38], ranging from whole-genome [39] to whole-exome [40] to RNA-sequencing [41] and reaching a single-cell precision [42] has allowed investigating the genetic material between the healthy and disease tissues of an individual and across populations [43]. A typical large-scale NGS-

based study reports several million genetic variants [5]. However, not all mutations, even with statistically significant correlations with the disease, would contribute to the disease phenotype [44]. For example, in cancer genomics, many mutations are defined as "passenger" mutations. Unlike the "driver" mutations, which induce the clonal expansion, the passenger mutations do not provide any functional advantage to the development of cancer cells [45, 46]. Thus, distinguishing between the functional and non-functional mutations is usually the first step in genetics studies.

2.1.3 Databases on Genetic Variation

Along with the development of NGS technologies and their applications studying human diseases, many variation and genotype-phenotype databases have been developed to help us make sense of the huge amount NGS data. In a typical setting, a clinical geneticist first filters preliminary variants calling results based on genotype quality and variant frequency that are harbored in the centralized databases, such as 1000 Genomes [47] or Database of short Genetic variations (dbSNP) [48]. In addition, several databases include genotype-phenotype information, which can also be used for further filtering and annotation. HGMD [49] is a unique resource providing comprehensive data on human inherited disease mutations. It is fee-based (also, there is a free academic version of a limited coverage) and focuses on published variants present in genes known or suspected to be associated with a human disease. ClinVar [50] is a recently launched freely accessible public database for reports of the relationship between the genetic variants and phenotypes. Unlike HGMD, ClinVar also serves as a central archive for predictions for causality. All mutations collected by ClinVar are grouped into five clinical significance categories, including Benign, Likely benign, Uncertain significance, Likely pathogenic, and Pathogenic. OMIM [51] is a comprehensive, authoritative compendium of human genes and genetic phenotypes OMIM is most useful in linking a candidate gene to a disease. In addition to general repositories, numerous LSDBs exist; HGVS (Human Genome Variation Society) maintains a list of LSDBs [52]. They are often curated by gene experts, but lack centralized editing.

Table 2-1 Popular genetic variation databases

Database	Purpose	Ref
1000 genome	Database of human genetic variation from different ethnic groups	[47]
CLINVAR	Database of human variation-phenotype relationships, with supporting evidence	[50]
dbSNP	Database of short variations in nucleotide sequences from a wide range of organisms	[48]
HGMD	Database of published gene lesions responsible for human inherited disease	[49]
LOVD	Open-source database on patient-centered DNA variations	[53]
OMIM	Comprehensive database of human genes and genetic phenotypes	[51]

2.1.4 Functional Annotation of Genetic Variations

Additionally, computational approaches for functional annotation are increasingly important, since many variants are not previously described in the literature [6, 54]. There are a plethora of functional annotations tools for genetic variations. Several recent reviews [55-58] give a comprehensive survey of state-of-art variant annotation tools. Most of the tools focus on the annotation of SNVs, as they are easier to capture and analyze, while some tools also cover indels. Some tools such as ANNOVAR [59] and VAAST [60] could be applied for whole genome level annotation. These tools employ different approaches, ranging from assessment of the evolutionary conservation to functional genomics. Here, we briefly review current methods of annotating nsSNVs. Concerning non protein coding variation, Ward et al [57] did a comprehensive survey about all available annotation methods for non-coding variations.

I. Evolutionary conservation based annotation

The effect of a nsSNV can be evaluated by studying properties of the residue substitution. Researchers have come up with different amino acid substitution metrics, such as PAM [61] and BLOSUM [62], to estimate the expected evolutionary distance between each possible amino acid residue pair. The main idea behind the evolutionary distance based

approach is that conservative substitutions, which are more consistent with the evolutionary trends, are less likely to be disruptive. However, Ng et al noted that the importance of the evolutionary distance between a pair of amino acids depends on the position where an amino acid substitution occurs [49]. Based on this, a variety of scores have been designed to quantify the idea including SIFT score [63], AGVGD score [64], PolyPhen score [65], etc. Currently, many annotation tools rely on these scores to predict the potential deleterious impact.

II. Sequence based annotation

Many bioinformatics tools and web servers provide information about the sequence–function relationship. This information can, in turn, be useful in assessing whether a genetic variation is functional. UniProt database [66] is one of the best known resources. Based on published study results, UniProt maintains a feature table for each curated protein sequence with annotated positions and regions of interest. These features can be used to identify whether a genetic variant occurs at a location that may be sensitive to residue changes.

III. Structure based annotation

If an nsSNV can be mapped on the experimentally determined protein structure or a corresponding homology model, then one can compute a number of properties using the structure information which could improve the accuracy of predicting the functional impact of this mutation [67]. In addition, many proteins are structurally solved with their interacting protein partners. Thus, if we could map an nsSNV to a protein-protein interaction complex, we could also assess whether the change occurs at or near a binding site or at a protein-protein interaction interface in the complex, and evaluate the effect of the nsSNV on the protein-protein interaction. Recently, our group has developed a new computational method [24], called the SNP-IN tool. SNP-IN tool predicts the effects of nsSNVs on PPIs, provided the interaction's structure or structural model. It leverages supervised and semi-supervised feature-based classifiers, including a new Random Forest self-learning protocol. The accurate and balanced performance of SNP-IN tool makes it useful for functional annotation of disease-associated SNPs.

Table 2-2 Popular SNV annotation tools and web-servers

Software	Purpose	Input	Ref
Align-GVGD	Estimates SNP risk	protein sequence, substitutions	[68]
ANNOVAR	Integrated tools providing gene annotation and various score	VCF4, GFF3-SOLiD, ANNOVAR format	[59]
AnnTools	Integrated annotation toolset for SNVs and indels	VCF, SamTools pileup, tabular files	[69]
FOLD-X	Performs protein stability analysis	PDB file, substitution	[70]
GERP++	Produces evolutionary conservation scores	csv, VCF, pileup, variant identifier	[71]
PANTHER	Provides likelihood of residue substitution (subPSEC score)	protein sequence, substitution	[72]
PolyPhen-2	Predicts damaging effect of a missense mutation	Uniprot ID, protein sequence, dbSNP ID	[65]
SIFT	Predicts if a residue substitution affects protein function	dbSNP ID, NCBI GI number, protein sequence, alignment, Pileup, VCF4, *etc*	[63]
SNAP	Predicts effect of nsSNP on function	protein sequence, substitutions	[73]
SNIP-IN tool	Predicts the effect of nsSNPs on PPI using structural information	pdb file of interaction, substitution	[24]
VARIANT	Provides comprehensive set of tools to analyze genetic variants	VCF, ANNOVAR format, BEDTools format	[74]
VEP	Integrated tools providing gene annotation and various score	csv, VCF, pileup, variant identifier	[75]

2.2 Complex Genetic Diseases

In the beginning of the 21st century, we are witnessing a truly pandemic growth of common diseases that are molecularly and genetically complex. It is estimated that 12.7 million cancer cases including 7.6 million deaths occurred only in 2008, with more than half of the cases and 64% of deaths coming from the economically developing countries [76]. Being the 7th leading cause of death in the U.S. in 2010, diabetes affected 25.8 million children and adults in the U.S. (8.3% of the population) [77]. The number of neurodevelopmental, psychotic, and neurodegenerative disorder cases are also on the

rise: for instance, the number of U.S. children aged 3 to 17 diagnosed with developmental disabilities, such as autism or ADHD, has reached a staggering 10 million affecting 15% of children of this age [78]. To cope with the complex diseases, doctors and scientists have been relentlessly trying to improve diagnostics and therapeutic intervention through the use of experimental and computational approaches. However, for many of these diseases the tasks of early diagnostics and successful treatment are challenging and, in some cases, still unfeasible, impeded by our lack of knowledge of the disease at the molecular level.

2.2.1 Complex Diseases and Human Mutations

Many common and rare genetic variants have been associated with complex diseases. According to the National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI) [79] catalog of published genome-wide association studies, as of November 2014, there are 2,060 publications describing 14,876 SNVs. Almost all common and many rare complex diseases have been addressed, including various types of cancer, cardiovascular diseases, neurological disorders, and immune system diseases. More importantly, this knowledge base has provided insights to the key molecular mechanisms underlying complex diseases [80]. For example, Multiple Sclerosis (MS) disease is an autoimmune demyelinating disease, whose mechanism is still not fully understood. After integrating different source of association study data, the interleukin 7 receptor (IL7R) gene stands out as a strong candidate gene with promising insights into the underling pathogenesis mechanism [81]. The nsSNV rs6897932(T244I) on IL7R has a strong association with MS, and its interplay with the alternative splicing of IL7R suggested a reliable hypothesis [81] for MS，which deserves further investigation.

The complex genotype-phenotype relationships among diseases are much more complex than was previously expected. Specifically, these diseases involve multiple genes and may have multiple sets of mutations associated with the same disease. Another interesting finding is that one genetic locus could be associated with multiple clinically distinct diseases (gene pleiotropy). For example, different interleukin receptor genes that are associated with Crohn's disease, multiple sclerosis, systemic lupus erythematosus and rheumatoid arthritis, suggesting that autoimmune diseases may share a common

mechanism[82]. Thus, the traditional “one gene/one enzyme/one function” concept assuming a simple, direct, and linear connection between the genotype of an organism and its phenotype often no longer holds [83].

2.2.2 Genetic Intricacy Underling Complex Diseases

Complex diseases have been found to exhibit molecular complexity at different levels, making them very challenging to study both experimentally and computationally. Understanding of the molecular mechanisms driving complex genetic diseases is in turn hindered by the multiple layers of complexity due to dozens, often hundreds, of pathogenic mutations affecting many genes, targeting multiple regulatory mechanisms and perturbing multiple pathways and systems. Complex diseases commonly manifest changes at the genetic, post-transcriptional, and epigenetic levels [35-37, 84-87]. Single nucleotide variations (SNVs) and indel mutations occurring in coding as well as non-coding regions of genomes are perhaps the most widely studied class of genetic changes owing to the recent progress in next generation sequencing [35-37]. Other genetic defects include larger structural variations such as copy number variations (CNVs) [84]. The transcriptional complexity of complex diseases is further complicated by post-transcriptional diversity—one of the most recent discoveries is the intrinsic role of post-transcriptional variations, such as alternative splicing variations (ASVs), in a number of diseases [85, 86]. Finally, another recent finding supported by the rapidly increasing volume of evidence is the link between the epigenetic variations and complex diseases [87]. Lastly, it has become evident that in many cases, not a single gene but a group of genes, often associated with a specific pathway or biomolecular network, are targeted by the mutations [88, 89]. Thus, the network and pathways information could be useful in identifying sets of genes (rather than individual genes) implicated in the disease. For instance, by applying pathway-based analysis to the whole genome association studies Askland et al found that multiple ion channel structural and regulatory genes are likely to contribute to the susceptibility of bipolar disorder [90]. Even more importantly, they propose that the heterogeneity of these gene sets across multiple studies could be the key feature of the genetic mechanism behind the susceptibility to this complex genetic disease.

2.3 Network Biology as an Emerging Approach

It is widely accepted that network analysis is the key to understanding disease biology[15]. Genes and proteins in cells do not work individually: they exert their functions through communication and coordination with others. Therefore, we need systems approaches to uncover the structure and the dynamics of the complex interaction networks that are essential to the structure and function of a living cell (See Fig 2-2). Network systems biology as an emerging approach has gain more and more attention in recent years. Network, or graphical model, not only provide us a theoretical model for representing a biological system, is also a conceptual framework to investigate and understand the organizing principles that govern cellular networks and the implications of these principles for understanding disease[17]. More importantly, biological networks are a natural platform to integrate different sources of data and incorporate prior knowledge[91]. Among various biological interaction networks, protein–protein interaction (PPI) gain most attention and are the most widely studied networks in biology. In this section, we mainly focus on protein-protein interaction network. We briefly review current knowledge about protein-protein interaction network and some new proposed ideas in this field.

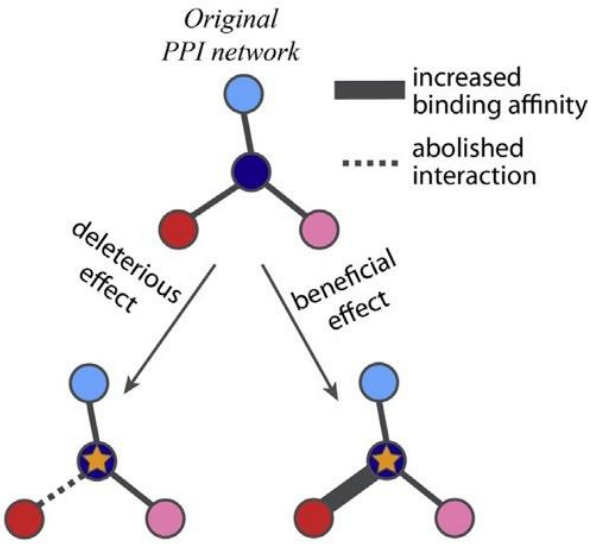


Figure 2-2 Genetic variation and its effects on PPI network. Effects of nsSNVs can be observed at the systems level, for instance by studying a PPI network centered around the mutant proteins.

2.3.1 Human Protein-Protein Interaction Network

Among different kinds of biomolecular networks, protein-protein interaction (PPI) networks (Fig 2-3) have attracted most intention and have been studied intensively [92]. It is estimated that, given present experimental methods and ignoring multiple splice variants of proteins, the human interactome to contain ≈650,000 protein interactions[93]. The mapping of interactome networks was essential for further network analysis. There are three main distinct data sources when constructing interactome networks: curation of already existing interactions available in the literature, systematic high-throughput experiments and computational prediction of potential interactions[94].

Systematically collecting interaction data from literature present the advantage of using already available information. Many databases[95-101] have been developed to construct the entire repository of interactions from numerous small-scale studies. For example, Human Protein Reference Database (HPRD)[101] is one of the earliest databased built for this purpose. All interactions in HPRD are manually extracted from the literature by expert biologists who read, interpret and analyze the published data. However, Literature-curated maps are limited by the inherently variable quality of the published data, the lack of systematization, and the absence of reporting of negative data[94]. At the same time, substantial improvements have been made in high-throughput data-collection techniques[102]. High-throughput experimental mapping strategies applied at the scale of whole genomes or proteomes have the advantage of producing unbiased, systematic and well-controlled data[94]. With the release of several large-scale human interactomes [103-105], these complex PPI networks have been utilized to explore the genotype-to-phenotype relationships on the basis that many proteins function by interacting with other proteins, and thus the network-rewiring genetic effects may lead to the disease phenotype [16, 94]. Besides, potential interactions can be predicted based on sequence similarities[106], phylogenetic profiling[107], statistical network inference[108] and text/literature mining[109]. Lastly, we note that all three approaches discussed above complement each other, but differ greatly in the possible interpretations of the resulting maps[15, 94].

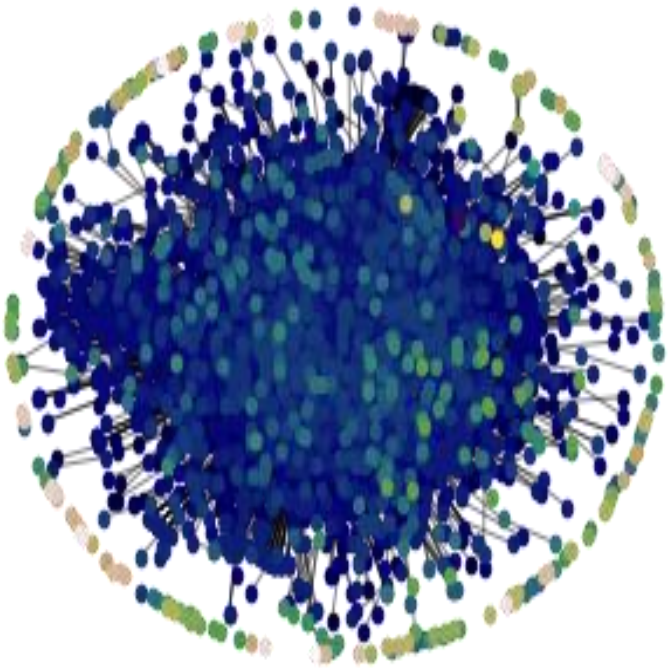


Figure 2-3 Visualization of the HINT protein-protein interaction network studied in this dissertation

2.3.2 Network-based approaches to study complex diseases

Network based approaches are particularly valuable to study complex genetic diseases like diabetes mellitus and cancers. Complex diseases are rarely consequences of a single genetic culprit, it usually involves various molecular aberrations and environmental factors[110, 111]. More importantly, these genetic factors interacting in the biological network further amplify the complexity[112]. We next review the current network-based approaches to study complex diseases, focusing on integrating the genetic variation and PPI data. The reviews [18, 92, 112, 113] give a more comprehensive account of the network base-based methods to study molecular mechanisms underlying the disease.

Since a disease phenotype could be linked to a synergistic effect of multiple genetic variations targeting a common component of the reference interactome [18], module-based approaches are promising for studying complex genetic diseases. Integrating PPI data with genetic variation data related to disease helps in determining modules and pathways perturbed in a disease of interest [19]. One basic way to uncover the perturbed

modules or pathways is mapping the disease genes to an interaction network and searching for the modules enriched with genetic variations [19]. Several complex disease studies have demonstrated the utility of this idea. Barrenäs et al work [20] shows that modules of highly interconnected complex disease genes were enriched for disease-associated SNPs, and could be used to find novel genes for functional studies. Reimand et al [114] applied novel algorithms to identify genes with significant phosphorylation-associated SNVs, phospho-mutated pathways, kinase networks, and clinically correlated signaling modules. By performing survival analysis, they identified signaling modules associated with increased patient survival in ovarian cancer.

2.3.3 Network Topological Analysis

Network topology is the arrangement of the nodes and edges within a biological network. Network properties, and particularly topological properties, can help us untangle the network 'hairy ball' and uncover the meaningful information encoded inside the network. Network topology can be characterized at different level; topological properties can apply to individual nodes and edges, or to the network as a whole. There are a range of very useful topolectal parameters or graph meters: node degree, clustering coefficient, betweenness centrality and average path length *etc* [115]. These graph measures provide quantifiable tools of network theory to understand the cell's internal organization and evolution [17] .

Some of the topological structures of biological network across different species, especially protein-protein interaction network, have been well studied. Interestingly, despite the remarkable diversity of networks in nature, their architecture is governed by several simple principles[17, 113, 116]. First, it is known that biological networks are scale-free, which means that some hub proteins have a huge proportion of the interactions while most proteins only interact with a small fraction of proteins[117]. In other words, their degree distribution approximates a power law. Another common feature of many biological networks is the so-called "small world effect" [116, 118]. Essentially, it means that two nodes can be connected with a path of a few links only. This "small world effect" is also known as "six degrees of separation" in social networks[116]. Recently, Santolini

et al. [119] show that an accurate knowledge of the network topology captures on average 65% of the influence patterns of the full biochemical model. They further suggest that mapping out the topology of biological networks opens avenues for accurate perturbation spread modeling and has direct implications for medicine and drug development [119].

2.3.4 3D Interacome

Usually, a PPI network is defined as an undirected graph, and network analyses treat proteins simply as the labeled nodes, ignoring structural details of the individual proteins. However, the missing structure information may play a key role in understanding the biological mechanism underlying disease process. Recently, Wang et al integrated the atomic-level protein structure information with high-quality large-scale PPI data and mapped genetic variations to the PPI interfaces [120]. They found that the interaction interfaces are enriched with in-frame mutations associated with the corresponding disorders. Based on this framework, they proposed a molecular mechanism hypothesis for complex disease mutations enriched on a specific interaction interface. In a similar study, the authors assigned SNVs related to Hemolytic Uremic Syndrome(HUS) to the corresponding proteins and their interactions, and classified these SNVs as buried, surface and interface mutations [121]. The study revealed that most of the mutations related to HUS on CFH and C3 genes are co-localized on the same interaction interface shared between the two proteins. This structure-informed result rationalized the hypothesis of a common mechanism for mutations causing HUS. Moreover, it is expected that PPI networks complemented with the structural information will be an essential component for the next generation of drug development strategies [122].

2.3.5 Edgotype

Recently, a concept of “edgotype” has been proposed [123], which is concerned with the functional outcomes of genetic variants on a PPI. Unlike the traditional view that a genetic variation causes a complete loss of gene product, the edgetic perturbation model treats such variation as interaction-specific ‘edgetic’ perturbation: the variation may cause the removal or addition of specific interactions while other edges remain unperturbed. The

traditional “one-gene/one-enzyme/one-function” model assumes a simple and linear connection between the genotype and its phenotype. A mutation in a gene leads to loss of the gene product, which leads to disease phenotype. Correspondingly, a mutation is usually modeled as the removal of a node and all of its edges in the network. However, recent studies[22, 124, 125] showed that different mutations leading to different molecular defects to proteins, and it may cause distinct perturbations of biological networks. A mutation targets at only specific interactions and the interaction-specific “perturbation” results in the removal or addition of specific edges while other edges remain undisrupted. It has been reported that a considerable portion of known disease-causing missense mutations are edgetic[22]. Recent studies showed edgetics can help interpreting the underlying genetic complexity of human disease and shed new insights into the mechanistic connections from genotype to phenotype[126-128]. For instance, most of the mutations associated with Type 1 von Hippel-Lindau (VHL) syndrome are frame-shift mutations. However, mutations associated with Type 2 VHL syndrome are typically missense mutations[129]. Research has shown that the missense mutations on the protein surfaces are disrupting specific edges (PPIs), which are responsible for the development of the syndrome[129]. Mutations in TPM3 can also exemplify the “edgotype” concept well. TPM3 gene encodes slow muscle alpha-tropomyosin. Three TPM3 edgetic mutations L100M, R168G and R245G are known to be associated with fiber-type disproportion myopathy [65, 130]. These edgetic mutations perturb 5 of the 10 interaction partners of TPM3 gene. (See Fig 2-4). In contrast, mutation M9R causes a different disease, nemaline myopathy. M9R might affect actin binding, thus leading to the formation of abnormal nemaline rods[131].

Characterization of PPI perturbations associated with disease mutations has been done by high throughput experiments. However, experimentally profiling several thousand missense mutations using the interaction assays remains costly and laborious. To support this new model, the SNP-IN tool can be used as an in silico edgetic profiling tool.

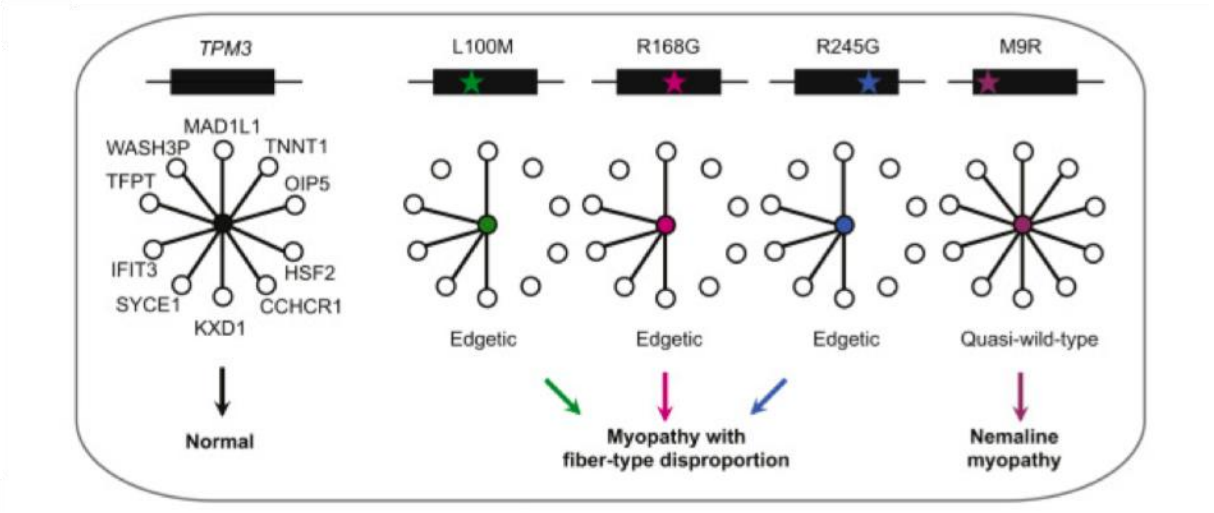


Figure 2-4 Illustration of the edgotype concept by mutations on gene TPM3. The image is adapted from Sahni N, et al. Cell (2015)

2.3.6 Network Propagation

Along with the increasing availability of the high-throughput human protein interactomics data [103, 132], new computational approaches have been developed. In particular, network propagation has recently emerged as a prominent approach in network biology [133]. In network propagation, genes/proteins of interest correspond to the nodes in the biological network, the edges represent pair-wise protein–protein interactions, and the information is propagated through the edges to nearby nodes in an iterative fashion. Thus, it could amplify the weaker disease association signals from the genes interacting with the "seed" genes that carry the stronger source signal [133]. Network propagation have been applied for various purposes, including predicting gene function, identifying disease related subnetworks, and drug target prediction [14, 133]. PRINCE [134] is a pioneering work of applying network propagation for prioritizing disease genes and inferring protein complex associations. It is shown to outperform previous methods in both the gene prioritization task and the protein complex task. Recently, Li et al. [135] developed a biological network propagation–based algorithm for large-scale drug synergy prediction. This model integrates the indirect drug targeting effects by interrogating the gene–gene network structure. Their method achieved the best performance in AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge [136]and is considered as the best solutions for cancer drug synergy prediction.

2.4 Population Genetics

Population genetics is a genetic subfield of studying the distribution of genetic variations within and between populations. It seeks to understand how and why the frequencies of genotypes and phenotypes change over time across different populations. Modern population genetics more emphasize the genetic phenomena[137], such as mutation and epistasis, and this separate it from phenotypic approaches studying evolution, such as evolutionary game theory[138]. The difference of underlying genetic architecture is the main source of the phenotypic variance across different populations. Population genetics is also relevant nowadays when studying the genetic basis of complex disease susceptibility of different populations[139]. Genetic disease susceptibility is the likelihood of developing a particular disease based on a person's underlying genetic structure. It is established that genetic variations can have effects on the likelihood of developing a particular disease.

2.4.1 Phenotypic Variance across Populations

Phenotypic variance is the observed variance in the phenotype of interest. The phenotype of interest can be binary, discrete or continuous. Take the height as an example, then the phenotypic variance is simply the observed variation of the height distribution in the population. Phenotypic variance usually combines the variation due to genetic reasons and the variation related to environmental factors [140]. Genetic variance has three major sources: the additive genetic variance, dominance variance, and epistatic variance. Additive genetic effect means that two or more genes, or alleles of a single gene work synergistically, and their combined effects equal the sum of their individual effects. Non-additive genetic effects involve dominance (of alleles at a single locus) or epistasis (of alleles at different loci). The ratio of genetic variance to phenotypic variance gives the proportion of observed variation that can be attributed to genetic reasons. This is called heritability [141, 142]. In other words, heritability explains how much of the phenotypic variance is due to variance in genetic factors. The Human Genome Project and follow-up large consortium sequencing efforts made researchers believe that the large genetic contributions to many traits and diseases would soon be mapped. However, single genetic

variations cannot account for much of the heritability of diseases, behaviors, and other phenotypes. This is often called the “missing heritability” problem [143, 144]. Many present studies regarding phenotypic variance focus on examining the SNP kinds of variants simultaneously. For example, Yang et al. [145] showed that 45% of the phenotypic variance of human height can be explained by considering SNPs simultaneously in a linear model analysis. Still, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability.

On the other hand, edgetics, or edgotype, provides alternative molecular explanations for mutation’s impact and why they underlie many complex genotype-to-phenotype relationships [146]. Edgetic perturbation models view mutations as and interaction-specific or edge-specific (‘edgetic’) alterations in the human interactome, rather than a complete loss of gene product (‘node removal’). An edgetic alteration can cause the removal of one or a few interactions but leaving the rest intact and functioning. It might have subtler impact on the network, and does not necessarily result in disease phenotype[147]. Mutations on the same gene might cause specific loss or gain of distinct molecular interaction(s), and lead to different phenotypic outcome. More importantly, edgetic perturbation model can easily explain confounding genetic phenomena, such as genetic heterogeneity [21, 94]. In sum, edgetics is a new approach to interpret genotype-to-phenotype relationships in the context of the biological network. It also shows us a way of studying population genetics. Meanwhile, large consortiums, like 1000 genome project[148] and ENCODE[149], have generated numerous amount of genetic variation data from different populations around the world. Altogether, it provides us a great opportunity to shift from traditional population genetics to population edgetics and apply the new methodology to study the genetic differences within and between populations.

2.4.2 Population Genetic Disease Susceptibility

Disease susceptibility is a condition that the individual is likely to get infected by a disease. Most diseases involve many risk factors, both environmental factors and genetic factors. Genetic disease susceptibility refers to a genetic predisposition to a health problem. An individual with high genetic disease susceptibility may not be born with a disease, but

she/he is more likely to acquire it. And the disease progress is often triggered or compounded by particular environmental influence or lifestyle factors. Many studies [110, 150-152] have established the critical role of genetic factors in determining health and disease. Perhaps, the most famous example about genetic disease susceptibility might be the mutations in the BRCA1 or BRCA2 genes. A person with mutated BRCA1 or BRAC2 genes have serious risks of developing breast cancer and ovarian cancer[153]. At the population level, many diseases differ in frequency between different populations. Different population with distinct underlying genetic make-up shows different disease predisposition to certain diseases [154-158]. Moreover, if some genetic risk factors are linked to the disease susceptibility, it does not necessarily mean they are abnormal or rare. The presence of one or more genetic mutations contributing to disease susceptibility might be relatively prevalent in normal population [159].

There are many ways that genetic variations can cause phenotypic differences. For example, non-synonymous SNPs can change the amino acid sequence of a protein and, thereby, cause the alteration of structure and function of the gene products. Genetic variation may also result in changes in the expression levels of gene products. Edgotype provides a new perspective to study how genetic variations is linked to disease phenotype [123]. It has acquired a lot of attention recently. We expect that distinct edgetic profiles, rather than one or several mutations, harbored by populations or individuals, can better explain why there is different disease frequency patterns and susceptibility across different populations. On the other hand, complex diseases often involve multiple genetic risk factors. Studying the interplay of these genetic risk often leads to a better understanding of complex disease. Furthermore, biological network provides a unique framework to interrogate a group of genes affecting heritable phenotypes. For example, Nayak et al. [160] constructed co-expression networks using correlations in expression levels of more than 8.5 million gene pairs. The construction was based on the expression profiles of African, European, and Asian ancestries in the HapMap project. They found that the subnetwork structures are not random but relevant to biological pathways and disease susceptibility. In sum, network-based analysis of the genetic architectures may not only shed light on biological mechanisms underlying complex diseases, but also yield better ways of measuring the genetic predisposition to a certain disease.

2.5 Network Module Detection

It is widely accepted that biological networks are not random graph but follow some principles that are common to most networks, such as scale-free topology, hierarchical organization and modular structure[161]. Modular structure is one of the essential characteristics of biological networks (See Fig 2-5). It has been also suggested that there exist specific disease modules for complex diseases[15, 16]. The identification of these modules is a crucial step in network analysis towards elucidating the biological mechanisms of a disease. There is a plethora of module detection methods in network science field[115]. However, it is unknown how these methods perform on biological network, as they are usually tested on artificially generated network. At the same time, a variety of computational methods have been developed to identify disease-specific modules in biological network by integrating other data source[162].

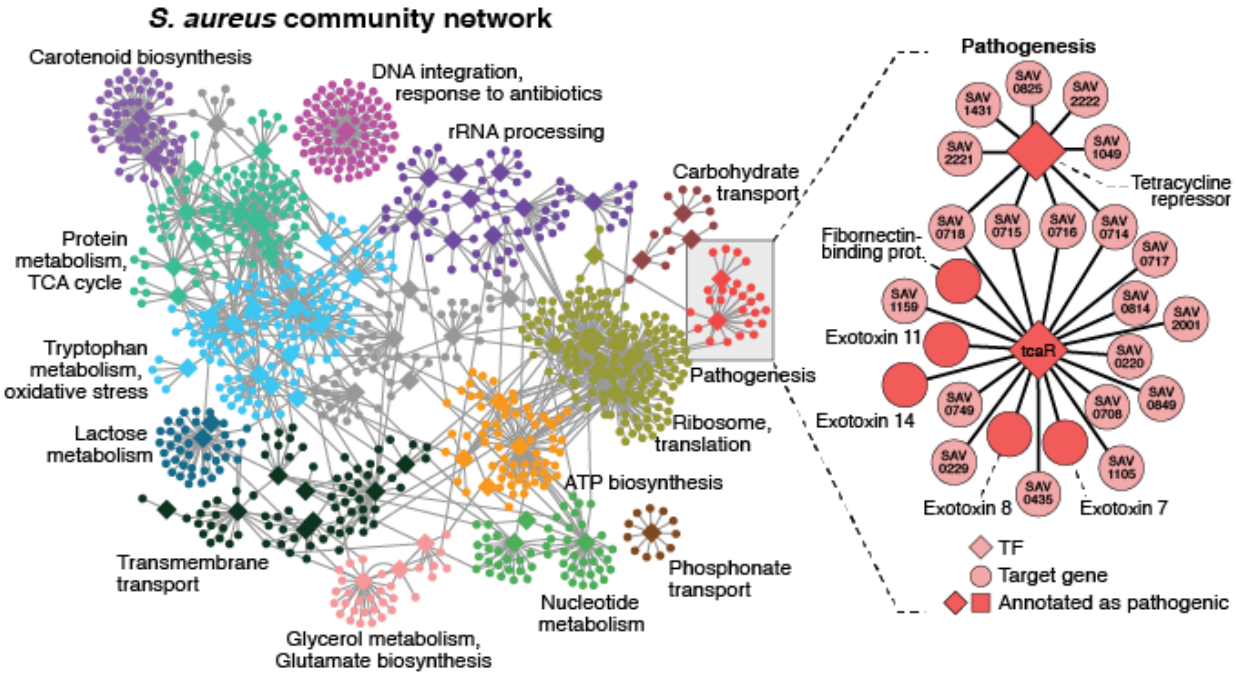


Figure 2-5 Examples of network modules in transcription regulatory network. The image is adapted from Disease Module Identification DREAM Challenge.

2.5.1 Network Modules

Biological network works in a modular way. A topological module is a set of genes (nodes) with dense interactions between each other; these groups of nodes are also referred to as communities, or clusters, in network science [163]. A topological module can be functional, since the constituting proteins are often shown to pertain to the same biological function or be involved in a similar biological process. A functional module refers to a group of physically or functionally connected biological molecular entities that work together to achieve a biological function. It is well known that functional modules have a high degree of modularity, which means subsets of nodes are more densely connected than expected randomly. A disease module is a sub-network of proteins enriched with the disease-relevant proteins and responsible for the disease phenotype. Complex diseases usually involve multiple genes and their products. And they interact with each other to fulfill a specific function within cellular networks[164]. Because of the functional interdependencies between the molecular entities in a human cell, complex disease should reflect the perturbation of the intricate network, rather than simply a consequence of an abnormality in a single gene.

2.5.2 Module Detection Methods

Module detection aims to find the functional units in the biological network. Module identification is a central problem in network biology [165]. Various module identification approaches have been proposed, presenting a wide range of theoretical perspectives and implementations. These methods primarily come in two different flavors. The first group of methods identify the modules in a biological network by relying exclusively on the network's topology. This is a challenging task due to the lack of information about specific genes/proteins contributing to biological functionality or disease phenotype. The recent open community DREAM challenge [165] provided a good review and benchmark of existing methods falling in this category. Methods from the second group start with the "seed genes", and gradually extract additional genes in the network to grow the module. For example, DIseAse MOdule Detection (DIAMOnD) [166] is a disease module detection algorithm that utilizes known seed genes to identify disease modules according to the

number of connections to the seed proteins. The algorithm outputs a connected disease module with a list of candidate disease-associated proteins ranked by their connectivity significance.

Discovering biologically relevant modules (disease modules or functional modules) is a challenging task [167, 168]. To tackle the mechanistic intricacy underlying complex diseases, it is necessary to couple disparate sources of data, each informing about a different aspect of the biological function. Computational approaches integrating molecular networks with different types of -omics data have demonstrated considerable power in bioinformatics studies [169]. For example, Ideker et.al. [170] incorporate mRNA expression data to identify differentially expressed sub-networks in PPI network. Surprisingly, integrating interactomics data with GWAS data has not yet gained wide attention in the bioinformatics community, and functional annotation information regarding genetic variants and mutated genes are not included during such integrations.

Chapter 3 Multilayer View of Pathogenic Mutations in Human Interactome

In the past several decades, tremendous efforts and vast resources invested in the quest to understand the molecular mechanisms underlying human genetic disorders. Next Generation Sequencing (NGS) methods, ranging from the whole-genome sequencing [39] to single-cell transcriptomics [171], have played an instrumental role in this quest, allowing the researchers to investigate genetic determinants in the healthy and disease tissues or cells of the individuals and across populations [6]. Along with the development of the NGS-driven applications studying human diseases, many genetic variation and genotype–phenotype databases and functional annotation tools have been developed to assist scientists to better understand the intricacy of the data [6]. Together, the above findings bring us one step closer towards mechanistic understanding of the complex genetic disease. However, it has rarely been possible to translate such a massive amount of information on mutations and their associations with disease into biological or therapeutic insights, and the mechanisms underlying genotype-phenotype relationships remain partially explained [17].

Non-synonymous mutations linked to the complex diseases often have a global impact on a biological system, affecting large biomolecular networks and pathways. However, the magnitude of the mutation-driven effects on the macromolecular network is yet to be fully explored. In this chapter, we provide a systematic in-silico edgetic analysis of the genes associated with various diseases and carrying mutations that potentially impact protein-protein interactions. We determine important differences and similarities between the

pathogenic single nucleotide variants (SNVs) and frameshift mutations with respect to how they affect the human interactome. We then identify three major groups of SNVs with respect to the PPIs they may affect: neutral to a PPI, significantly strengthening a PPI (beneficial), or eliminating a PPI (detrimental). To quantify the overall network rewiring caused by a group of mutations associated with a disease, the concept of cumulative network damage is introduced, and the importance of employing an edge-based rather than a traditional node-based measure is demonstrated. We also compare our analysis with the recently published experimental edgetic profiling study showing greater coverage of our approach and the minimal overlap with the existing experimentally obtained dataset. Finally, we apply our approach to the case-studies of interaction-affecting SNVs in type 2 diabetes mellitus and cancer. By combining our edgetic analysis with the clinical data on cancer patients, we demonstrate the critical roles of beneficial mutations in the disease progress and determine the link between the disruptive mutations in the cancer driver genes and the decreased patient relapse time and survival time.

3.1 Methods and Materials

The system-wide edgetic characterization of the pathogenic mutations in the human interactome in this work can be broken down to several stages (Fig. 3-1). First, we studied the topological properties of the pathogenic non-synonymous single nucleotide variants (nsSNVs) in the network. The main goal of this stage was to determine whether one could differentiate between the different types of human genetic variations, either in terms of clinical importance, i.e., pathogenic nsSNVs versus non-pathogenic nsSNVs, or in terms of structural properties, i.e., pathogenic nsSNVs versus pathogenic frameshift mutations. Next, we wanted to check which of the two most common genetic variations was more likely to be the cause of gene pleiotropy and whether perturbations of specific PPIs caused by these variations play a role in gene pleiotropy. Second, we examined these mutations in a structurally resolved PPI network, INstruct [172]. Specifically, we leveraged structural information on PPI complexes and utilized our SNP-IN tool to annotate the edgetic effects of nsSNVs at the systems level and evaluate the widespread perturbations of pathogenic SNVs in the human interactome, thus adding another layer of functional information in this work. Further, we adopted a concept of network robustness from the field of physics in order to quantify the overall, or cumulative, damage induced by the disease-associated mutations and to correctly characterize the network rewiring behavior. Finally, we collected cancer patients' clinical data, with the goal to link the network damage caused by mutations with the clinical outcome.

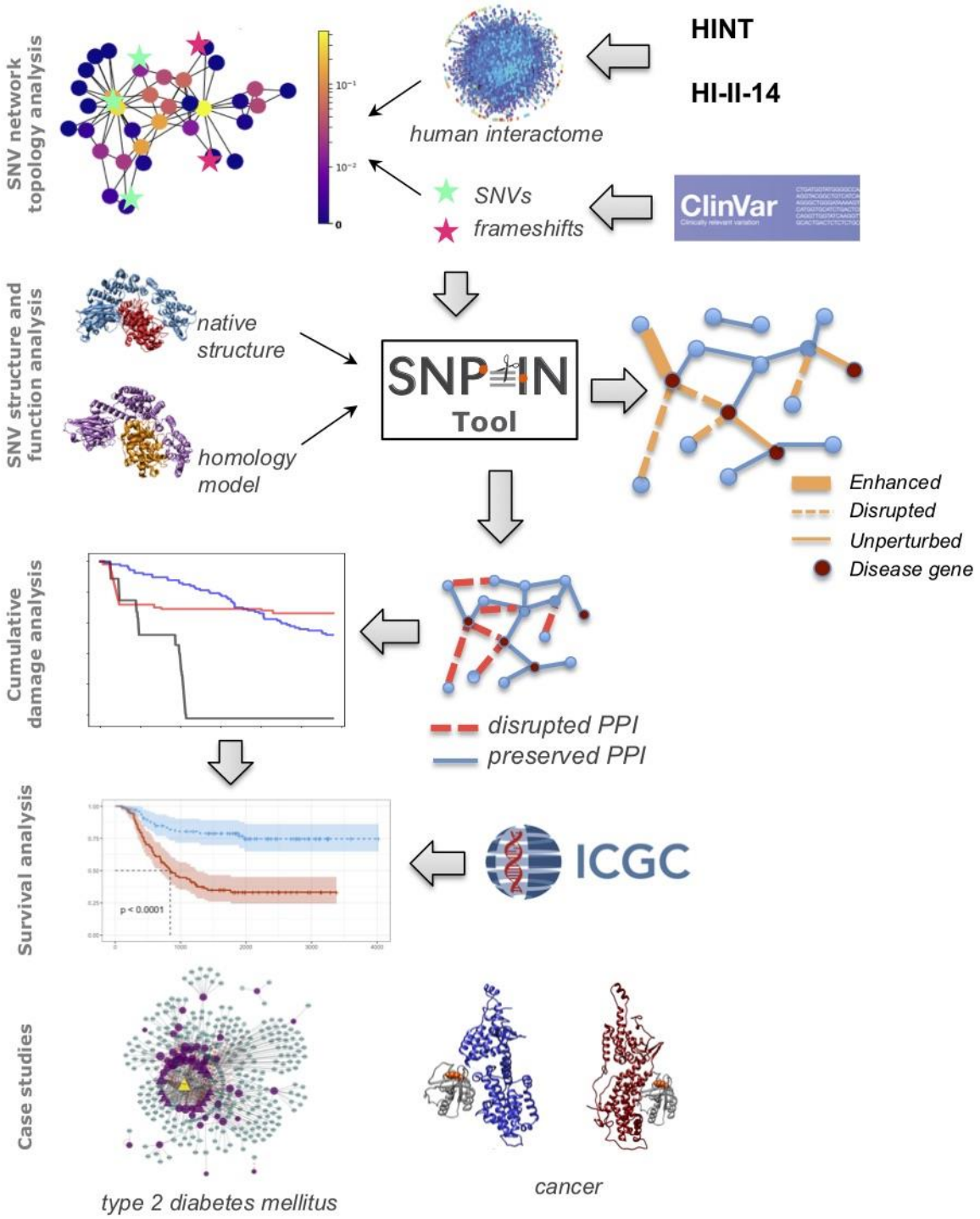


Figure 3-1 Overview of our computational workflow in Chapter 3. First, the topological properties of the non-pathogenic SNVs as well as pathogenic SNVs and frameshift mutations are compared for two interactomes, HINT and HI-II-14. Second, the edgetic profiling of disease SNVs is obtained by applying our SNP-IN tool. Third, the cumulative network damage is studied by applying the principles of the network robustness theory. Fourth, the survival analysis is performed to understand the role of mutations that affect PPI. Last, two large-scale case studies are considered.

3.1.1 A dataset of disease genes and genetic mutations

We collect a list of disease genes and their pathogenic non-synonymous SNVs from ClinVar database [173]. ClinVar database [173] was used as a source for genetic variants associated with the diseases. ClinVar is a public database where each genetic variant is annotated with some clinical significance for the reported conditions. All mutations collected by ClinVar are grouped into five clinical significance categories, including Benign, Likely benign, Uncertain significance, Likely pathogenic, and Pathogenic, following the guidance by The American College of Medical Genetics and Genomics (ACMG) [174]. It contains both germline and somatic variants of different types, sizes, or locations. Specifically, the clinvar_00-latest.vcf file is used, followed by data preprocessing, to get a paired gene and mutation list. The nsSNVs that belong to categories Likely pathogenic or Pathogenic are defined as pathogenic for this work, while nsSNVs that belong to the remaining three categories are defined as non-pathogenic. In addition, we extracted from SNV another group of genetic variants, the pathogenic frameshift mutations. In this work, we curated 11,487 disease nsSNVs distributed across 2,240 genes, 2,719 non-pathogenic nsSNVs in 807 genes, and 6,498 pathogenic frameshift mutations in 1,537 genes. We note that the same gene can carry mutations of different types: for instance, there are 1,039 genes carrying both pathogenic SNVs and frameshifts and 388 genes carrying both pathogenic and non-pathogenic SNVs.

3.1.2 Extraction of PPI data and construction of PPI network

For the disease network analysis, two human PPI networks are used: HINT network [175] and the experimental human interactome project network (HI-II-14) [103] . HINT (http://hint.yulab.org) is organized as a centralized database of high-quality human PPIs collected from several databases and annotated using both, an automated protocol and manual curation. The human interactome project is another recently released PPI source. It includes a set of binary PPIs that were constructed through by systematically interrogating all pairwise combinations of predicted gene products using yeast-two-hybrid experiments. As a result, HINT and HI-II-14 represent two distinct networks with a small overlap. Instead of merging two networks, we analyze HINT and HI-II-14 networks independently, treating them as the complementary rather than competing

views of the human interactome. We expect that different groups of false positive and false negative PPIs exist for each network. The networks present complementary information: the interaction overlap between HINT network (45,226 PPIs), and HI-II-14 network (32,465 PPIs) is only 13,223 PPIs.

3.1.3 Topological Analysis of pathogenic SNVs in human interactome

In our analysis of the pathogenic SNVs, we first investigate their topological importance, that is, whether these mutations are located on the proteins that occupy critical positions in the human interactome. To do so, for each interactome we calculate and examine the centrality measures associated with the proteins that carry those SNVs. Specifically, we investigate three major centrality measures in the graph theory: node degree, betweenness and closeness (Figure 3-2). Previous works suggest that the PPI network topology could encode information about how molecular interactions contribute to the disease phenotypes [91]. These centrality measures are useful to explore the shared properties of genetic architectures underlying genetic diseases.

The simplest measure of centrality in a network is the node degree. For a protein in an interactome, the node degree specifies the number of direct interaction partners this protein has. Betweenness is another global centrality measure, which determines the number of shortest paths that connect any pair of nodes in the network and also pass through a given node. Formally, the betweenness centrality measure, $C_B(v)$, of a vertex v is defined as:

$$C_B(v) = \sum_{u,w \in V} \frac{\pi_{uw}(v)}{\pi_{uw}}$$

where π_{uw} is the number of shortest paths between vertices u and w, and $\pi_{uw}(v)$ is the number of such shortest paths that come through vertex v. Thus, nodes that occur on many shortest paths connecting pairs of nodes have the higher betweenness.

Closeness centrality provides a rather different view of centrality compared to the above two measures, because it is based on the mean distance between a given node and all other nodes in the network. It is defined as the reciprocal average distance to every other node:

$$C_c(v) = \sum_{u \in V} \frac{1}{d(u,v)}$$

where *d(u,v)* is a graph-based distance between the selected node v and any other node in the network, u. A node with high closeness centrality is, on average, close to the other nodes when using the graph distance. The calculation of the centrality is done using the python Networkx package [176]. To check whether the pathogenic SNVs have higher centrality compared against frameshift mutations, we formulate this question as a statistical test, and apply Wilcoxon test for this task since no prior information about the underlying distribution is known.

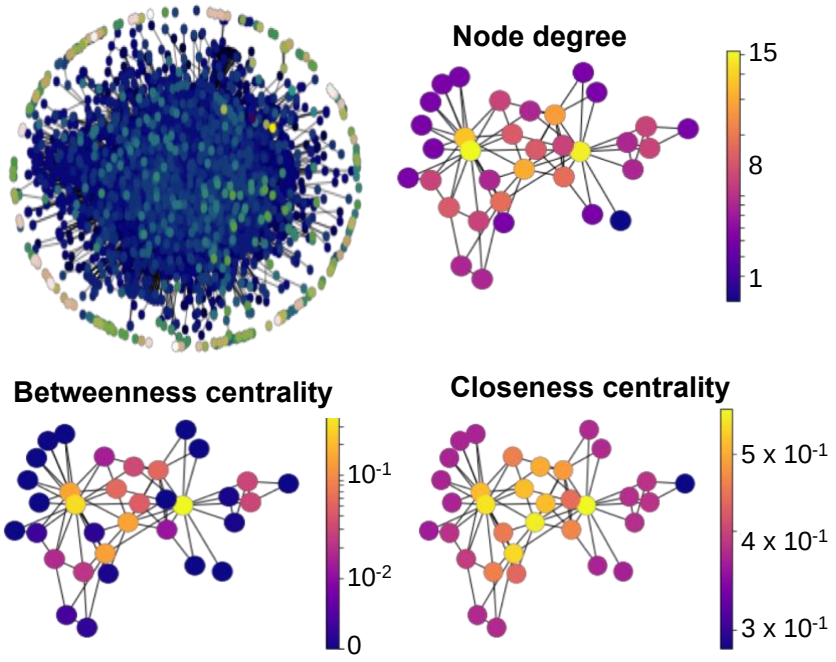


Figure 3-2 Three basic topological characteristics of mutations in the network are calculated: node degree, betweenness centrality, and closeness centrality. Shown are examples of the three measures calculated for the same small network, and the overall scale-free topology of HINT interactome.

3.1.4 Linking pathogenic SNVs to gene pleiotropy

Pathogenic SNVs and pathogenic frameshift mutations have been suggested to affect the phenotype in different ways [21]. A missense mutation is likely to affect one or several specific interactions. On contrary, a frameshift mutation often causes the loss of all the interactions in which the mutated protein is involved. The mutation-induced perturbations of the network properties give rise to the altered phenotypes, which are often linked to a disease. The distinct interaction profiles caused by genetic variants could provide a more accurate link between genotype and phenotype [123]. We then formulate and test two hypotheses. The first hypothesis states that the phenotypes caused by pathogenic SNVs should be more diverse than the phenotypes caused by frameshift mutations. Our second hypothesis is that the average phenotype similarity score between a pair of a pathogenic SNV and a pathogenic frameshift mutation will be higher than the corresponding similarities between the pairs of pathogenic SNVs.

Each of the two hypotheses is statistically tested with the disease phenotype similarity identified for each pair of mutations [177]. The disease phenotype dissimilarity spans 5,080 diseases in OMIM. The similarity between OMIM records is calculated by comparing the feature vectors, in which an entry represents a MeSH concept. For this work, we annotate only those pairs of disease genes where each gene has at least one pathogenic nsSNV and at least one pathogenic frameshift mutation. For each disease gene, we define the average disease similarity between all nsSNVs on this gene ($S1$), between all frameshift mutations associated with the gene ($S2$), and between each nsSNV and each frameshift mutation from this gene ($S3$) as following:

$$S_1 = \frac{\sum_{i,j} s_{i,j}}{n_1}, S_2 = \frac{\sum_{k,l} s_{k,l}}{n_2}, S_3 = \frac{\sum_{m,n} s_{m,n}}{n_3}$$

where $s_{i,j}$ is the similarity score between the phenotype corresponding to SNV i and SNV j and n_1 is the number of total SNV pairs for the gene; $S2$ is defined in the same way for frameshift mutation k and frameshift mutation l, as well as $S3$ is defined for an SNV m and a frameshift mutation n. To verify each hypothesis, we compare the average similarities defined above, and the statistical significance is calculated using a Mann–Whitney test.

3.1.5 Examination of pathogenic SNVs in a structurally resolved PPI network

When studying molecular networks centered around a complex disease, the proteins implicated in the disease are often treated as mere network nodes. It has been suggested that adding structural details about the mutations and the corresponding proteins could help in understanding the mutations' roles in the complex disease. We then examine the distribution of SNVs in a structurally resolved PPI network, INstruct [172]. INstruct is a database of high quality, structurally resolved protein-protein interactions for human. The database includes high-quality binary PPI data and the structural information about the PPI complex at the atomic or near-atomic resolution level derived from the experimental data using a tested interaction interface inference method [172]. For each PPI from INstruct, the PPI interface and the binding sites of each interacting protein have been structurally characterized.

Combining the domain information collected from Uniprot and the PPI interface information collected from INstruct, we divide a disease-related protein into the following three regions: "interface domain", "non-interface domain" and "non-domain". If the distributions of pathogenic mutations are not influenced by the domain architecture of the protein, then one should expect for the numbers of SNVs across the three regions to correlate with the lengths of these three regions. Then, the odds ratio (OR) for pathogenic SNVs on each of these three regions is calculated. The odds ratio is a statistical measure of association between an exposure and an outcome defined as:

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)},$$

where $p1$ is the number of observed mutations in each region across all proteins divided by the total number of mutations and $p2$ is the total sequence length of each region across all proteins divided by the length of all proteins combined.

3.1.6 Functional annotation of SNV's effect on PPI using SNP-IN tool

Determining whether an nsSNV disrupts or preserves a PPI is a challenging task. We have previously formulated this task as a classification problem, and developed a computational method, SNP-IN tool (non-synonymous SNP INteraction effect predictor tool) [24]. SNP-IN tool predicts the effects of nsSNVs on PPIs, given the interaction's experimental structure or accurate comparative model. There are three classes of edgetic effects predicted by the SNP-IN tool: beneficial, neutral, and detrimental (Supplementary Table S5). The effects are assigned based on the difference between the binding free energies of the mutant and wild-type complexes ($\Delta\Delta G$). The beneficial, neutral, or detrimental types of mutations are then determined by applying two previously established thresholds to $\Delta\Delta G$ [178, 179]:

$$Beneficial: \Delta\Delta G < -0.5\ kcal/mol$$

$$Neutral: -0.5\ kcal/mol \leq \Delta\Delta G < 0.5\ kcal/mol$$

$$Detrimental: \Delta\Delta G \geq 0.5\ kcal/mol.$$

To apply SNP-IN tool, we first structurally characterize, when possible, each PPI from one of the two interaction networks in which a disease protein is involved. If a PPI already has a native structure in PDB, we extract the interaction structure directly using the recently launched INstruct database [172]. Specifically, we identify an interacting chain pair for each PPI in the corresponding PDB file, to make sure that the two chains physically interact. During the process, 3did database [180] is utilized, which maintains the information about the two interacting domains with physical interfaces. If a PPI does not have a structure in PDB, two options are explored. First, if a structural template for such interaction (i.e. a homologous protein complex) exists, a comparative model of this interaction can be obtained. Alternatively, if one cannot structurally resolve the full-length PPI, one can try to model only the domain-domain interaction on which the mutation can be mapped to. Homology modeling is done through Interactome3D [181], a web service for structural modeling of PPI network. When modeling a PPI involving either the full-length proteins or partial, domain-domain, structures, the template with the

highest sequence identity with respect to the target sequences is selected. Finally, for each PPI that is structurally resolved, the mutated residue is mapped to the protein structure by indicating the position of the mutated residue in the PDB file of the modeled PPI.

3.1.7 Network cumulative damage analysis

We intend to quantify the cumulative damage effect of a group of pathogenic SNVs on a PPI network associated with a certain disease. To do so, we adapt the methodology from the network robustness theory. The first step of our cumulative damage analysis is to define the "attacking strategy" for the genetic variants. In the traditional network robustness analysis [182], a simple way to perturb a network is to randomly remove its nodes, which we refer to as the "random failure". Another way is to remove nodes in order of their degrees, from the highest to the lowest, which we call the "malicious attack". However, these two strategies can be viewed as two extremes. Neither of the two strategies can realistically model the damage caused by the disease genes and pathogenic mutations disrupting the PPIs, and hence these strategies cannot characterize the biological network rewiring behavior. In real networks, the failure or attack could also occur on the edges. In our case, disrupted interactions caused by pathogenic SNV are more likely to fall into this category. The pathogenic mutations would only disrupt limited number of the interactions. It suggests us that an "edge-based" attack strategy might be more suitable to characterize the network rewiring behavior. On the contrary, since frameshift mutations usually result in the polypeptide abnormally short or abnormally long, and the final product will most likely not be functional. They more conform to the complete removal of the nodes. We also included frameshift mutations in this analysis as a comparison. We tried different "attack strategies" (both node-based and edge-based) to find out how pathogenic SNVs perturb the network. For node-based attach strategy, we pick the node degree to guide the node removal process, as previous work reports that degree centrality is proven to be superior to other centrality measures at exposing the vulnerability under malicious attacks and random failures [183]. For edge-based attack strategy, a link-robustness concept based on the highest edge-betweenness attack was recently proposed [184]. This concept captures the network behavior for any fraction of link removal. In this work, we focus on the disease-associated proteins and PPIs disrupted by the pathogenic

SNVs according to the SNP-IN tool annotation. And we remove the corresponding edges in the PPI network based on their betweenness centrality.

After we select the "attack strategy", a quantitative metric to measure the cumulative damage caused by the pathogenic SNVs is defined. One of the key aspects of studying the robustness of a physical networked system against the failure of their component parts is to understand how the size of the largest component changes as nodes and/or edges are removed from the network. If the size of the largest component shrinks sufficiently after the failure, when compared to the original size of the network, then it is reasonable to assume that the networked system is unlikely to function [185]. For an initial network of size N with a largest component $S0$, removing a fraction of the nodes or edges according to some specified procedure described above would result in a new network, in which the largest component would be $S1$. The key quantity that we will study here is the size of $S1$ relative to the initial size of the network: $|S1|/ N$, where $|S1|$ denotes the number of vertices in $S1$. Given a set of pathogenic mutations, the cumulative damage to the network caused by the mutations is then defined as: $(|S0|-|S1|)/N$. Once a suitable centrality measure and the attack strategy have been fixed, we can compute $|S1|/N$ as a function of the fraction for removed nodes/edges in decreasing order of that centrality measure to characterize the network rewiring behaviors.

3.1.8 Correlation between disruptive mutations and decreased survival in cancer patients

We next study the relationship between the mutations predicted as disruptive and the survival time and relapse time in the cancer patients. To do so we first assemble a list of well-known cancer genes. Specifically, a high-confidence collection of 869 cancer genes is defined as a union of genes in the Cancer Gene Census [186] and recent literature [187]. The Cancer Gene Census catalogs the genes for which mutations have been found implicated in cancer. The recently published dataset [187] was derived using computational methods and includes a list of 291 high-confidence cancer driver genes from 3,205 tumors and 12 different cancer types. Thus, the two datasets are complementary, and we consider their union to be our final dataset. To explore the

functional and clinical significance of disruptive mutations on these cancer drivers, we curate the genomic and clinical data for the cancer patients from ICGC. Somatic mutations from several cancer genomics projects are downloaded from the International Cancer Genome Consortium (ICGC) data portal [188]. A subset of mutations mapped to the human genome build 37 are annotated with ANNOVAR [59]. We discard all non-coding and silent mutations, short insertions and deletions, and retain only non-synonymous, missense SNVs.

Finally, we study somatic mutations occurring on the above set of the cancer drivers with high mutation frequency. Specifically, we consider the cancer driver genes with the mutation rate higher than the background mutation rates plus the standard deviation. For those pathogenic mutations, we annotate their effects on the corresponding protein-protein interactions to investigate whether the rewiring of the PPI network would play a role in cancer. Then, based on the mutation annotation results, we divide the patients into two groups: the cancer patients with mutations potentially affecting the interactions and cancer patients without such mutations. Because SNP-IN tool can only be applied to a limited number of these mutations, the obtained groups using SNP-IN tool based annotation are limited in size, preventing application of statistical tests. Therefore, we resort to a more general annotation of somatic mutations by using the information stored in INstruct. More specifically, the interaction interface data for each cancer driver, when available, is extracted from INstruct database, and we check whether the mutation is located in the interface. Lastly, given the annotation results, the comparison of the survival distributions of two groups is performed using the log-rank test [189]. The log-rank test is among the most popular methods for comparing the survival of groups. To do so, the method computes the observed and expected numbers of events in one group at each observed event time, and then obtains the overall summary across all event times as a hazard ratio.

3.2 Results

3.2.1 Pathogenic SNVs share similar centrality properties as frameshift mutations, but are more likely to cause gene pleiotropy

We examined the topological properties of the genes carrying pathogenic SNVs in the two human interactomes by comparing these properties with (i) genes carrying pathogenic frameshift mutations, and (ii) genes carrying non-pathogenic SNVs. SNVs and frameshifts are the two most common genetic variations associated with human disorders [190], so it is natural to study them first in the context of network topology. We evaluated three basic topological properties: node degrees, betweenness, and closeness for the HINT interactome and HI-14 interactome separately. The average node degree, betweenness, and closeness for the proteins carrying pathogenic SNVs were 8.6, 1.1 x 10-4, and 0.23 in HINT, and 8.3, 1.3 x 10-4 and 0.26 in HI-14 interactome. When comparing with the pathogenic frameshift mutations, we found the values for all three properties to be similar (Fig. 3-3, Supplementary Fig. S3-1).

Intriguingly, the comparison of the network properties between the genes carrying pathogenic and non-pathogenic SNVs (Fig. 3-3, Suppl. Fig. S3-2) revealed that the former had significantly higher node degree in HINT interactome (P-value is 0.046, Wilcoxon test) and significantly higher betweenness in both interactomes (P-values are 0.01 and 0.043 for HINT and HI-14 interactomes, respectively, Wilcoxon test). We therefore concluded that the genes with pathogenic SNVs are more central in the network, compared to the genes carrying non-pathogenic SNVs and the changes in the former group of genes are likely to have greater impact on the interactome than similar changes in the latter group. However, the obtained results also implied that these plain topological properties could not differentiate the pathogenic SNVs from pathogenic frameshift mutations. Thus, we next investigated if one could differentiate these two groups of pathogenic variations in a structurally resolved interactome.

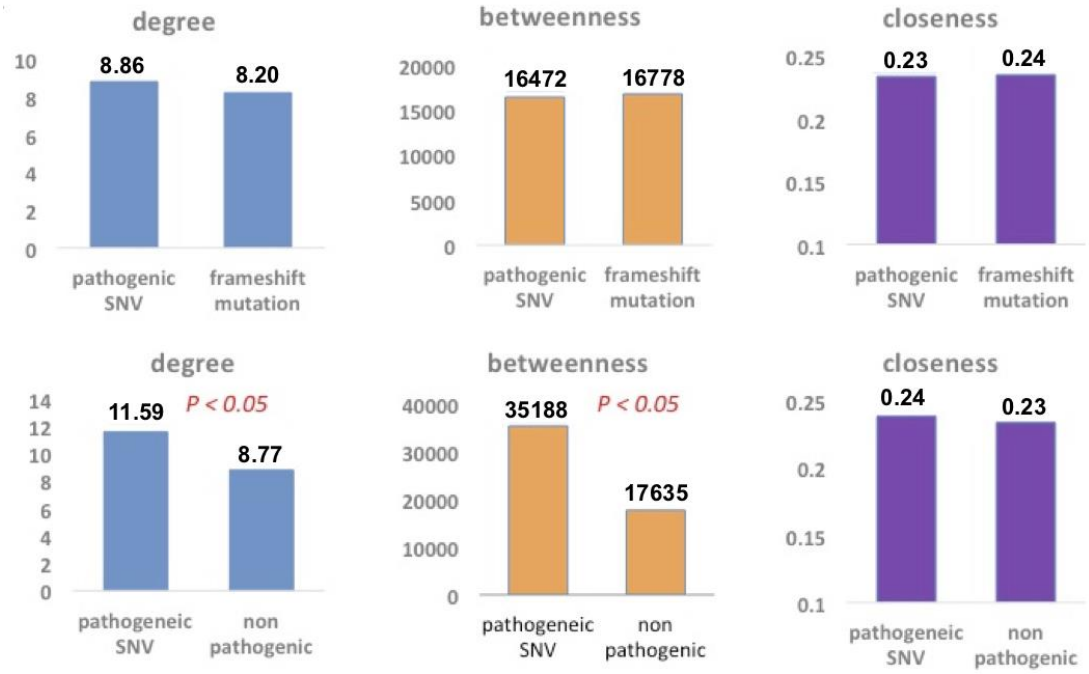


Figure 3-3 Comparison of centralities between two groups of mutations calculated for HINT network: pathogenic SNVs versus pathogenic frameshift mutations, and pathogenic SNVs versus non-pathogenic SNVs.

Pathogenic mutations are believed to cause the disease in multiple ways [21]. While a frameshift mutation often results in an incomplete protein fragment that is likely to be unfolded or misfolded and thus degraded by the proteosome, a pathogenic nsSNV is likely to produce a full-length protein with a local defect. However, the question of whether such different structural effects on a protein can result in similar effects on a PPI mediated by this protein, the function carried out by this interaction, and a result, the phenotypic change caused by the functional changes, remains largely unanswered. To answer this question, we leveraged the concept of disease phenotype similarity score [177]. The score determines if the two disease phenotypes are similar; the higher the score the more similar two phenotypes are. We calculated the disease similarity caused by above two kinds of pathogenic mutations in each gene and found that the average disease similarity between a pair of pathogenic frameshift mutations is significantly higher than that the similarity between an nsSNV and a frameshift mutation (Fig. 3-4). Furthermore, the

average disease similarity between a pair of pathogenic nsSNVs in a gene was significantly lower compared to that one between a pair of frameshift mutations.

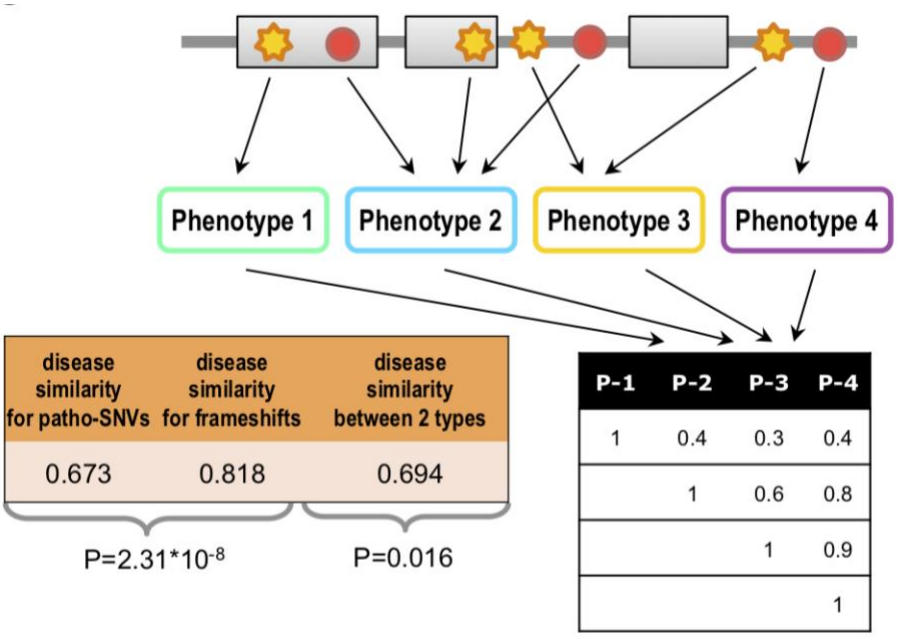


Figure 3-4 Basic principles of the analysis of the relationship between gene pleiotropy and mutation source. Red circles represent pathogenic SNVs, while yellow stars represent the frameshift mutations. Both mutation types can be associated with different disease phenotypes. For each pair of phenotypes, their similarity is calculated based on the number of common genes associated with both phenotypes.

3.2.2 Pathogenic SNVs are enriched on the interaction interfaces

As we observed above, the network topology itself could help providing only a high-level view on the effects of the pathogenic nsSNVs. In the recent years, several works proposed to complement the topological network with the structural information [21, 181]. Following the same strategy, we compared the two types of pathogenic mutations extracted in this work by mapping them into a structurally resolved network, INstruct [172]. Specifically, we wanted to find whether the pathogenic mutations have a tendency to accumulate on the protein binding site, and thus contributing to the interaction interface, as opposed to the rest of the protein. The interface enrichment of the pathogenic SNVs would indicate that they are likely to cause the disease through rewiring the corresponding protein-protein interactions (Fig. 3-5). The set of disease genes (672

genes) containing at least one pathogenic nsSNV and at least one protein binding site from INstruct was selected to calculate the interface mutation enrichment. We found that the pathogenic nsSNVs were indeed significantly enriched in the protein interaction interfaces: among all the 4,108 disease-associated nsSNVs in the structurally resolved network, 2,781 of them were observed on a PPI interface (Table 3-1). Furthermore, the pathogenic nsSNVs were found to be under-represented on the other domains not involved in the PPIs (Fig. 3-5, Table 3-1). In contrast, we found that the non-pathogenic SNPs from the same set of the disease genes were not enriched on the PPI interfaces (Table 3-1). These observations provide strong support for the proposed mechanism causing the disease phenotype through PPI rewiring by the pathogenic SNVs.

Table 3-1 Distribution of SNVs across the protein sequence: 1. pathogenic SNV group; 2. non-pathogenic SNV group. N corresponds to the number of SNVs. OR corresponds to odds ratio.

	Interface Domain		Other Domain		Non-domain	
	N	OR	N	OR	N	OR
1.	2,781	2.41	325	0.84	1,002	0.41
2.	180	0.70	47	0.67	653	1.58

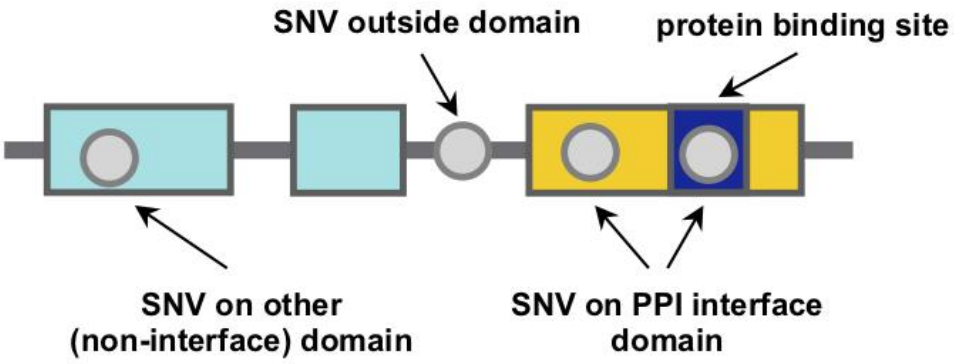


Figure 3-5 Basic principles of the SNV structural analysis with respect to the protein domain architecture. Grey circles represent SNVs and their location on the protein, while rectangles correspond to the protein domains.

3.2.3 Functional annotation of disease SNVs indicates widespread disruption of interactome and synergistic edgetic effects of SNVs

While the enrichment of the pathogenic nsSNVs on the interaction interfaces suggest that pathogenic nsSNVs play an important role in the protein interactions, their mere presence on the PPI interface does not guarantee that each such mutation would have a functional effect on the interaction. Our recently developed SNP-IN tool was designed to differentiate between the interaction-neutral nsSNVs and those ones affecting the interaction [24]. SNP-IN tool accurately predicts the effects of non-synonymous SNVs on the existing wild-type PPIs, given the interaction's structure. It is designed as a set of classifiers leveraging a new Random Forest self-learning protocol. A 3-class classification problem was considered in our study (Fig. 3-6), where the classes corresponded to the three functional effects of SNVs on the protein interaction assigned based on the difference between the binding free energies of the mutant and wild-type complexes: detrimental, neutral, and beneficial (see Methods for the definitions).

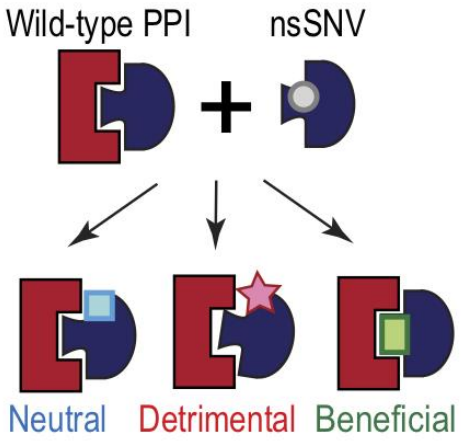


Figure 3-6 Three basic classes of SNVs annotated by SNP-IN tool: Neutral, Detrimental and Beneficial.

SNP-IN tool requires the structure of the PPI complex, which comes from either an experimentally resolved structure or an accurate model from the homology modelling approach. In total, we were able to provide at least partial structural characterization of 1,491 PPIs. To meet this requirement, on one hand, we performed a comprehensive search in the PDB database [191] (see Methods), obtaining 499 experimental structures with structurally resolved PPI interface. Furthermore, we have obtained the full-length homology models for 818 PPIs and partial domain-domain interaction homology models for 174 PPIs. There has not been a common agreement on to what extent the disease associated mutations could affect the PPIs. One study concluded that only a small number of disease-associated mutations were expected to specifically affect PPIs. However, it has been suggested that perturbations of PPIs (disruptions or enhancements) played an important role in the pathogenesis of many disease genes, more than previously expected [21]. Our results showed that, among all 3,401 SNVs annotated by SNP-IN tool, which accounted for about 1/3 of the total disease-associated SNVs we collected, 2,592 SNVs (76.2%) were predicted as detrimental to at least one PPI that the corresponding disease protein was involved in, and 48 SNVs (1.4%) were labelled as beneficial.

Further, we explored whether these pathogenic mutations tend to work synergistically or antagonistically. We grouped the beneficial and neutral mutations into a new class, labelled as interaction preserving, and named the detrimental mutations as the second class, interaction disrupting (Fig. 3-7). We then defined a synergistic genetic interaction as a mutation pair that had the same effect for the corresponding PPI, either preserving or disruptive. Similarly, we defined the antagonistic interaction as a mutation pair that had the opposite PPI rewiring effect according to the SNP-IN tool annotation. Based on this definition, we focused on the PPIs with at least two annotated nsSNVs. These mutations could be on the same protein or located on the two separate interacting partners. However, they should target the same interaction. In total, we collected 1,491 PPIs with at least two nsSNVs. We found that 24,922 mutation pairs have the same disruptive effect on the same protein-protein interaction, while 9,334 mutation pairs had the same interaction preserving effect, which accounts for 55.1% and 20.6% respectively in the all the 45,205 possible pairwise mutation combinations within the same interaction. At the same time, we had 10,949 antagonistic mutation pairs, accounting for

24.2%. Further, we excluded the mutations with the neutral effects in this analysis and focused the PPIs containing beneficial mutations. We found that only 506 pairs of mutations tend to work antagonistically, a small percentage of all possible mutation combination (11,859). These results suggest that genetic mutations tend to work synergistically and can be explained by the fact that an individual mutation might not be "disruptive" enough to cause a major dysfunction of the corresponding protein-protein interaction, while a group of two or more mutations with the same rewiring effect could be sufficient.

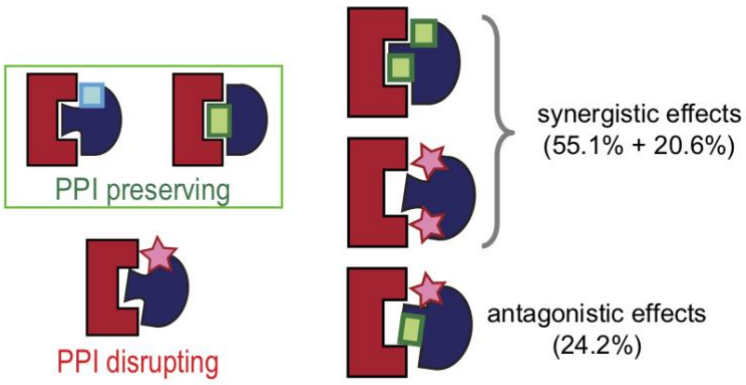


Figure 3-7 Using the basic classes of SNVs from Fig. 3-6, two basic classes of network perturbing mutations are defined: interaction preserving and interaction disrupting. Shown on the right-hand side is the basic principle of the antagonistic and synergistic effects of mutations.

3.2.4 Comparison with experimental edgetic profiling shows greater prediction coverage of the in-silico approach

Recently, the first large-scale edgetic profiling of the missense mutations associated with disease phenotypes has been done by using interaction assays [22]. Specifically, 2,449 mutant proteins and their 1,072 corresponding WT proteins were screened against all partners found in the human interactome HI-II-14 [103]. In total, the interaction profiles for 460 mutant proteins and their 220 WT counterparts were obtained resulting in 521 perturbed interactions found in 1,316 PPIs. The work also provided systematic measurements of the PPI profile changes caused by mutations using a strategy referred to as "edgotyping" [123]. The effects of missense disease mutations on PPIs were grouped into several major categories (Fig 3-8): mutations causing no apparent detectable change in interactions ("quasi-WT"), mutations causing specific loss of one or several interaction ("edgetic"), and mutations causing a complete loss of all interactions ("quasi-null").

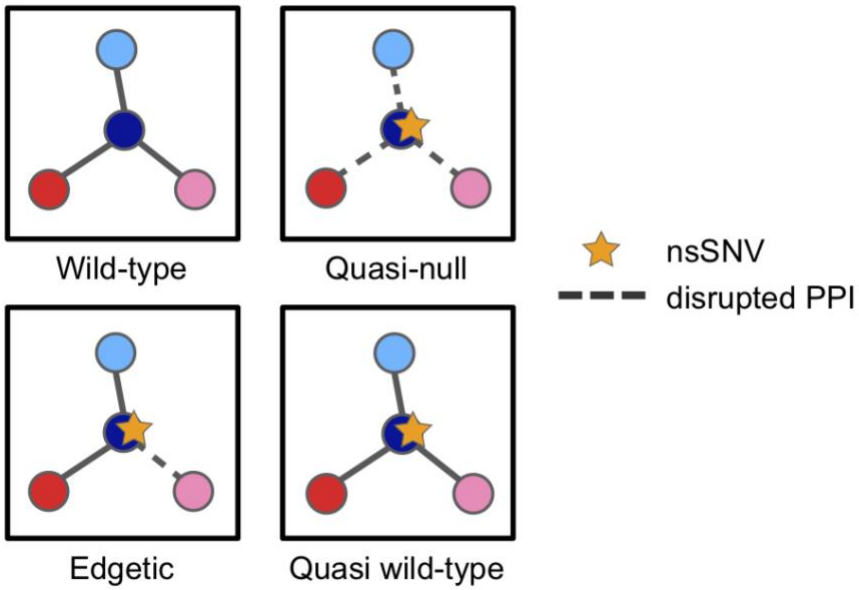


Figure 3-8 Basic edgotypes used in this work. The first one is the wild type interactions. The other three are showing different effects of SNVs on PPI: quasi-null, edgetic, and quasi-wildtype.

Here, we wanted to compare our in-silico edgetic profiling method with the experimental profiling approach in order to find the advantages or disadvantages of the former. We found that our computational approach has a significantly higher coverage in terms of number of genes, mutations, and PPIs being profiled (Table 3-2): in total, we have systematically characterized the effects of 3,401 mutations carried by 669 proteins on 1,491 PPIs. Next, focusing on a missense mutation set for which the corresponding protein has two or more interaction partners, the experimental edgetic profiling identified 26% of the mutations as quasi-null, 31% as edgetic and 43% as quasi-WT. In our case, for a mutation set meeting the same criteria, we determined 32% of them to be quasi-null, 38% as edgetic and 30% as quasi-WT. The distributions of quasi-null, edgetic, and quasi-WT alleles were statistically indistinguishable between the experimental and computationally predicted datasets. Interestingly, in spite of the highly similar distributions, the overlap between the mutation sets from the experimental analysis and our work was minimal: only 56 mutations carried by 33 genes, which is ~4% of a total of 889 genes and 1% of 4,862 mutations considered in both edgetic profiling studies, were shared among the two mutation sets, demonstrating great complementarity between the two approaches.

Table 3-2 Comparison between the recent large-scale experimental edgetic profiling study and the current study performed using an in silico approach

Edgetic profiling approach	N of Genes	N of PPIs involved	N of Mutations profiled	Edgotyping characterization		
				Quasi-null	Edgetic	Quasi-WT
Experimental	220	1,316	460	26%	31%	43%
In silico	669	1,491	3,401	32%	38%	30%

3.2.5 Cumulative damage analysis of PPI network reveals network rewiring behavior caused by genetic mutations

Next, we wanted to estimate the synergistic rewiring effect of pathogenic mutations on the whole interactome. To do so, we tested several measures that quantify the cumulative damage caused by a group of mutations. The idea of cumulative damage is exactly opposite to the idea of the network robustness that is commonly used in the network theory to describe the ability of a network to withstand malicious attacks damaging it. First, we performed the node-based cumulative damage analysis on each of the two individual interactomes (Fig 3-9A, 3-9B). As a result, we observed similar trends in both interactomes, in spite of the fact that these two interactomes were distinct, overlapping only over a small subset of PPIs and proteins. In particular, we found that both HINT and HI-II-14 interactomes were robust to the "random failure" (i.e., random removal of protein nodes) but vulnerable to "malicious attack" (i.e., the removal of nodes based on the size of the largest network component; nodes whose removal reduces the largest component the most are removed first), which was consistent with the previous findings [17].

We also found that when we used the node-based cumulative damage measure to estimate the amount of changes in the interactome due to edgetic effects of the pathogenic SNVs, the network damage caused by the pathogenic SNVs was similar to the random failure (Fig 3-9A, 3-9B). However, such insignificant cumulative damage of the network caused by a random failure might not be sufficient to exhibit the disease phenotype, and was contradictory to the idea that disease phenotype could be caused by the network perturbations [123]. Considering the widespread perturbations of PPIs caused by the pathogenic SNVs across a broad spectrum of genetic diseases, it raised a question if the node-based definition of cumulative damage was an accurate measure to characterize the network damage. To substantiate our contention, we also performed the same analysis for the frameshift mutations for both HI-II-14 and HINT interactomes. As discussed earlier, a frameshift mutation often results in an incomplete protein fragment that is typically degraded, which corresponds to a node removal in the interactome. On the other hand, nsSNVs are likely to produce a full-length protein with a potential defect in the

corresponding PPI(s), which would correspond to some missing edges but the node associated with the protein is likely to remain intact. Therefore, based on the edgotyping classification, most frameshift mutations would fall into the quasi-null category and are supposed to cause very different rewiring effect on the network. However, our results using the node-based cumulative damage measure (Fig 3-9C, Fig 3-9D) suggested that the frameshift mutations have the similar behaviour as the pathogenic nsSNVs. This further confirmed the fact that the node-based cumulative damage measure definition could not differentiate the rewiring behaviour of nsSNVs and frameshift mutations.

Following the above analysis, we next adopted a more reasonable edge-based definition of cumulative damage. Specifically, we leveraged an edge removal scheme proposed in a recent paper that studied edge-based robustness [184]. We found that in both HINT and HI-II-14 interactomes, the pathogenic mutations could cause more severe damage than a random failure (Figs. 3-9E, 3-9F), and the network damage was more similar as the one during malicious attacks (i.e., the removal of edges based on the on the size of the largest network component; edges whose removal reduces the largest component the most are removed first). These results suggested that the edge-based definition was more suitable to characterize the network damage caused by the pathogenic SNVs and might be helpful in assessing phenotypic changes driven by the PPI-rewiring mutations.

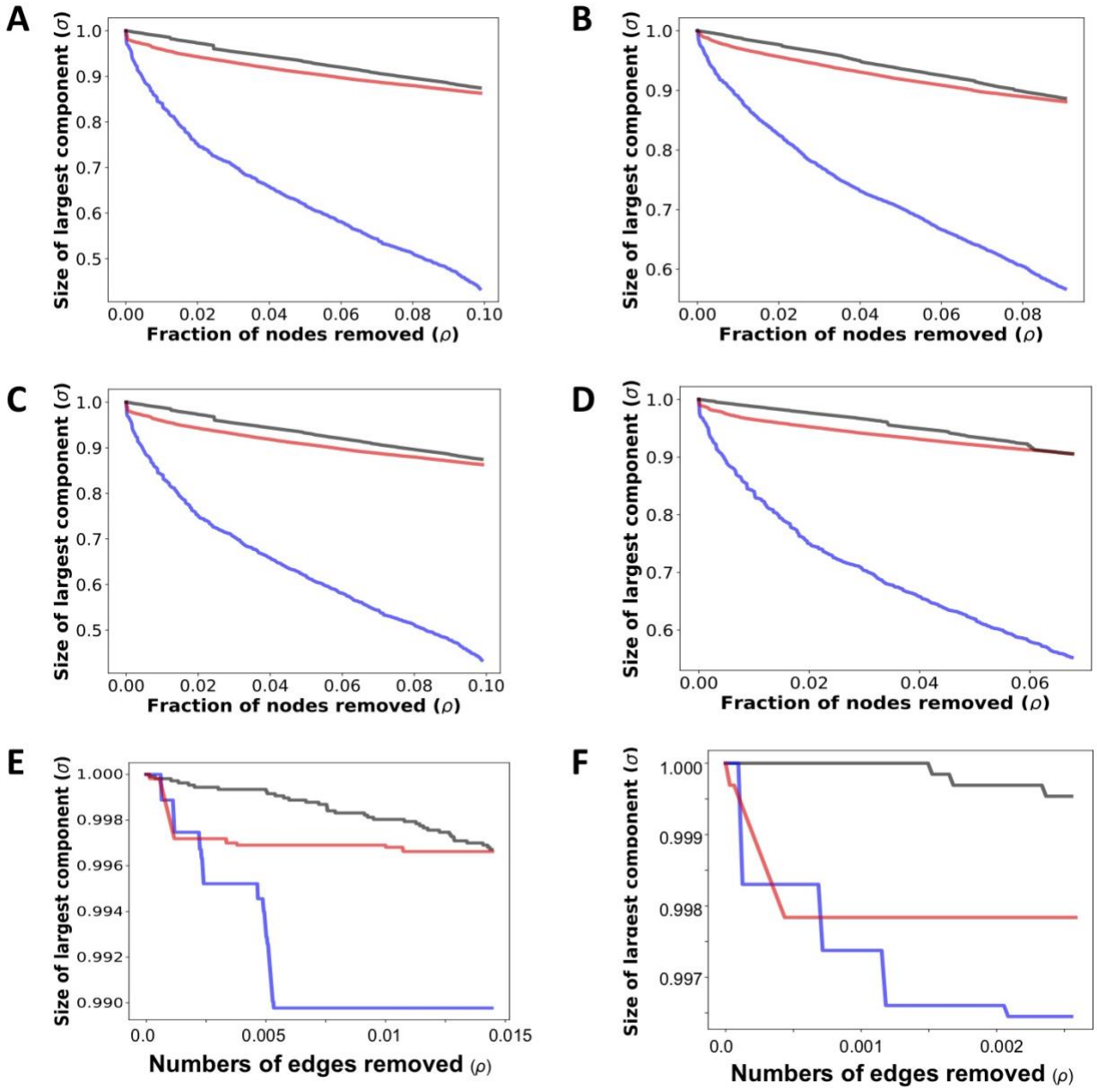


Figure 3-9 Cumulative damage calculated for both interactomes, HINT (panels A, C, and E) and HI-II-14 (panels B, D, and F). Grey lines correspond to the random attack strategy, where the nodes are removed randomly. Blue lines correspond to the removal of nodes based on the size of the largest network component; nodes whose removal reduces the largest component the most are removed first. Red lines correspond to the cumulative damage done exclusively by either pathogenic SNVs or pathogenic frameshift mutations. Three robustness strategies were used and compared: (A), (B) correspond to the cumulative damage calculated using node-based cumulative damage measure, with red lines corresponding to the cumulative damage done exclusively by the pathogenic SNVs that rewire the interactome; (C), (D) correspond to the cumulative damage calculated using node-based cumulative damage measure, with red lines corresponding to the cumulative damage done exclusively by the pathogenic frameshift mutations; (E), (F) correspond to the cumulative damage calculated using edge-based cumulative damage measure, with red lines corresponding to the cumulative damage done exclusively by the pathogenic SNVs disrupting the PPIs.

3.2.6 Network analysis identifies a disrupted network clique of proteins associated with type 2 diabetes mellitus

As a first case-study, we carried out the analysis of an interaction network centered around proteins associated with type 2 diabetes mellitus (T2DM). The important role of protein-protein interactions in T2DM was recently proposed [192-194]. Most of these works focused on integrating different sources of data to discover novel candidate genes for T2DM. Here, we aimed at studying the mutation-induced rewiring of the T2DM-centered PPI network. First, we curated 131 T2DM genes from ClinVar database, together with 346 pathogenic SNVs on those genes. To define a T2DM-centered PPI network, we extracted the interaction partners for the T2DM proteins from both HINT and HI-2014 interactomes. Together, we curated 655 interactions. Based on the edgetic profiling done by SNP-IN tool, we were able to annotate 185 T2DM-related mutations, and 139 of them were labelled by SNP-IN tool as disruptive, suggesting global rewiring of the PPI network in T2DM (Supplementary Table S3-3).

The analysis provided us with several interesting findings. First, we found that T2DM-related genes formed a clique (Fig. 3-10). To find how tightly the genes associated with T2DM are connected with each other, we used the MCODE [195] in Cytoscape [196] to perform graph clustering. The highest-scoring cluster consists of 12 genes, which are all associated with T2DM, and 62 PPIs. Based on the SNP-IN tool annotation, we observed 3 disrupted interactions inside a clique (Fig. 3-10). Such an inter-connected cluster could be a central functional hub of the T2DM network, thus the PPI perturbations inside the clique could play a central role in the disease phenotype.

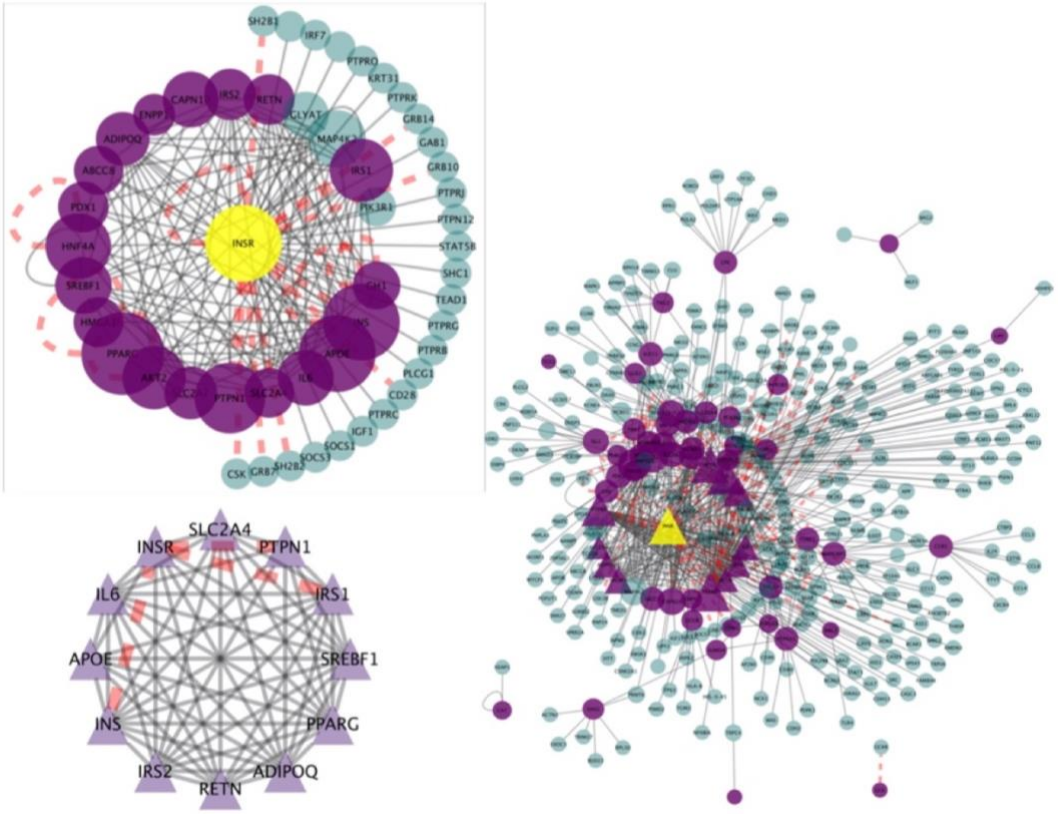


Figure 3-10 Case study of the T2DM-centered network. Shown on the right-hand side is the visualization of the entire T2DM-centered network. The central yellow triangle corresponds to INSR gene, which is surrounded by its interaction partners. The purple nodes correspond to the diabetes genes. The triangle nodes correspond to the genes involved in the clique subnetwork. The size of the node corresponds to the node degree. The red dash lines are the disrupted interactions. The left top network is a subnetwork that focuses on INSR gene and its interacting partners. The left bottom subnetwork corresponds to a clique of diabetes genes identified in the network.

Further analysis revealed that among the genes associated with T2DM, INSR carried the highest number of disruptive mutations. The INSR gene encodes a transmembrane insulin receptor (UniProt ID: P06213), a member of the receptor tyrosine kinase family that mediates the pleiotropic actions of insulin and plays a key role in the regulation of glucose homeostasis [197]. Binding of insulin leads to the phosphorylation of several intracellular substrates, including insulin receptor substrates (IRSs). Two main signalling pathways, PI3K-AKT/PKB and the Ras-MAPK, are activated following the phosphorylation of IRSs. The PI3K-AKT/PKB pathway is responsible for most of the metabolic actions of insulin, and the Ras-MAPK pathway regulates specific gene expressions and cooperates with the PI3K-AKT/PKB pathway. INSR was found to be

involved in the interactions with the majority of known T2DM-associated genes (Fig 3-10); it is also a member of the clique subnetwork described above. We annotated 21 mutations associated with INSR, and 17 of them were annotated as disruptive. Strikingly, each of the 11 protein-protein interactions that INSR participated in, was disrupted by at least one pathogenic SNV, suggesting that cumulatively, the mutations could substantially limit the functioning of this gene.

3.2.7 Interaction enhancing mutations provide new insights into transient interactions and their roles in diseases

In our second case study we investigated potential roles of a small number of pathogenic nsSNVs that were determined beneficial, i.e. causing a significant increase in binding affinity of the existing wild-type PPIs. The network topology does not get affected by such mutations, at least when the network dynamics is not considered. Therefore, we hypothesized that the beneficial mutations are mainly involved and affected the transient PPI. A permanent interaction is typically long-term, stable, and irreversible. On the other hand, transient protein complexes form and break down recurrently, with the involved proteins often interacting in a brief period of time and in a reversible manner [198].

We first would like to check how many interactions that the beneficial mutations affected were the transient interactions. Determining the permanent of transitive state of a PPI is an extremely challenging task. Experimental methods that can characterize the transient or permanent state of a PPI are laborious and costly, and a high-throughput characterization of all interactions in the interactome has yet to be carried out. Here, we chose a computational approach, NOXclass [199], which determines the protein-protein interaction type as either obligate or non-obligate. Non-obligate PPIs constitute of proteins that can form stable well-folded structures alone, obligate protein complexes are mainly involved with those proteins that are unstable on their own and become stabilized only through an interaction. Typically, the obligate interactions are permanent, whereas non-obligate interactions are more likely to be transient [200]. Using NOXclass, among all the 55 interactions, we were able to obtain prediction results for 45 interactions. And

about 50% are predicted to be non-obligate, suggesting a possible mechanism of action for beneficial mutations—stabilizing the transient complexes.

We further studied an interesting case of one nsSNV, rs104894227, located on HRAS gene (protein HRas; Uniprot ID: P01112) that potentially enhanced three independent PPIs: HRas and Raf-1, HRas and SOS-1, as well as HRas and SOS-2 (Fig. 3-11). HRAS gene is a member of the Ras oncogene family encodes a protein located at the inner surface of cell membrane. Mutations in HRAS were found to associate with conventional follicular carcinoma [201] and Costello syndrome [202]. The enhancement of PPIs caused by rs104894227 could explain possible malfunction of the molecular mechanisms underlying the interactions. For instance, the transient interaction between HRas and SOS-1 has been known to maintain a delicate balance between being a strong enough connection to facilitate the nucleotide release and being ready to "unzip" for acceptance of the new nucleotides [203]. The PPI-enhancing mutation is likely to break this balance, potentially affecting intracellular signaling pathways that control cell proliferation and differentiation. Another interesting functional impact is in strengthening HRAS-Raf-1 interaction. A counterintuitive phenomenon was described where RAF inhibitors were found to enhance ERK signaling, facilitating tumor cell proliferation, an adverse effect that was seen with RAF inhibitors in melanoma patients [204]. This phenomenon was explained only recently by the fact that the inhibitors promote RAS–RAF association by disrupting RAF kinase domain autoinhibition [205]. The above missense mutation is also likely to play a role in promoting RAS-RAF association, resulting in a similar phenotype. Development of inhibitors for HRas-RAF and HRas-SOS interactions have been discussed as promising directions in cancer therapeutics [203, 205]. The identified edgotypes may be useful in modifying these therapeutic strategies to account for the presence of PPI-enhancing mutants.

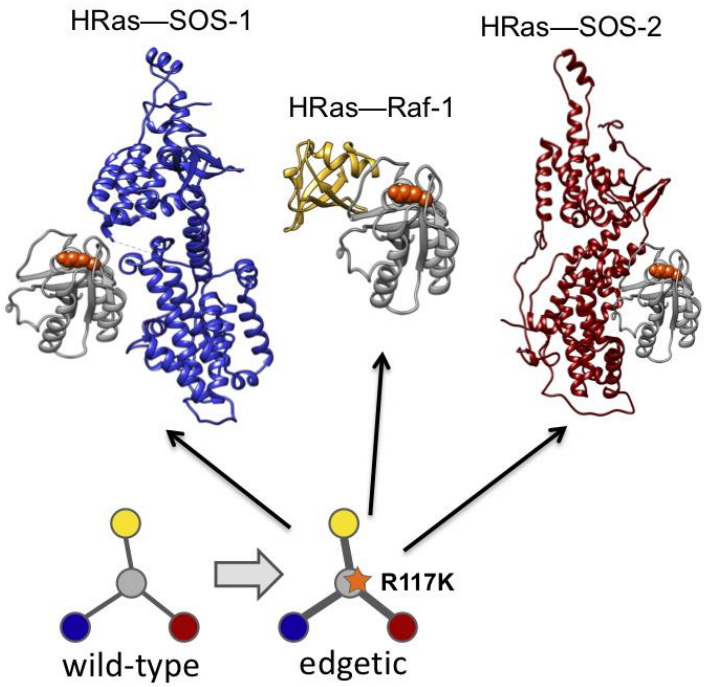


Figure 3-11 Case study of HRAS gene and the beneficial mutation on it. An nsSNV, rs104894227, located on HRAS gene is predicted to enhance three independent PPIs: HRas and Raf-1, HRas and SOS-1, as well as HRas and SOS-2 The enhancement of PPIs caused by rs104894227 could explain possible malfunction of the molecular mechanisms underlying the interactions.

3.2.8 Interaction disrupting mutations on cancer drivers correlate with decreased survival

Lastly, to explore clinical significance of the disruptive mutations, we studied the survival rates and relapsing times of the corresponding cancer patients. Specifically, we found that the acute myeloid leukemia and liver cancer patients with the disruptive somatic mutations carried by the cancer drivers would suffer decreased survival time and relapse time. We first curated a list of 869 high-confidence cancer genes from the Cancer Gene Census [186] and recent literature [187] (see Methods). Among these 869 genes, 227 genes had pathogenic SNVs previously found to be implicated in the cancer progress (Fig 3-12). By applying SNP-IN tool, we were able to provide annotation about the effects of these SNVs on PPIs for half of the set, 107 genes. We note that SNP-IN tool did not cover all the mutations for these genes, since some of the mutations could not be mapped to the structures of PPI complexes.

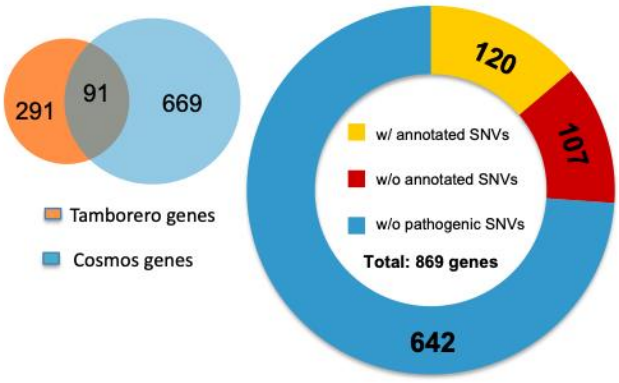


Figure 3-12 Basic statistics of cancer driver genes used in the studies (left) and the annotated SNVs (right). The left part of the panel is showing two main sources of cancer drivers, COSMOS gene consensus and literature collection (Tamborero)

In summary, we annotated 784 mutations for these 107 genes, and 58.3% of them showed to be disruptive to PPIs. The top five cancer genes with the highest numbers of disruptive SNVs include MLH1, MSH2, STAT3, SOS1, and VHL (Fig 3-13). On average, a cancer gene carries 18 pathogenic SNVs, and more than 1/3 of them are annotated as disruptive (Fig 3-14, Supplementary Table S3-4). This suggests that mutations in cancers might target protein-protein interactions, and the corresponding rewiring could be a key factor in driving the cancer progress.

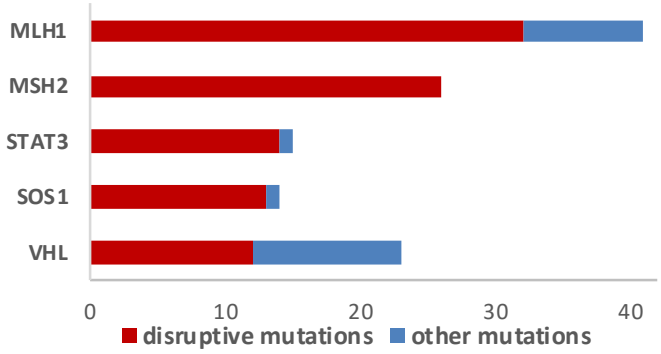


Figure 3-13 Functional annotation of the pathogenic SNVs on cancer drivers. Shown are the top 5 cancer driver genes with the highest numbers of disruptive mutations. The red bar corresponds to disruptive mutations, the blue bar corresponds to other mutations.

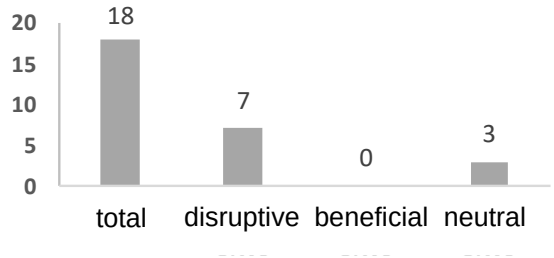


Figure 3-14 The average number of mutations of each type on a cancer driver.

We next collected the somatic mutations from several cancer sequencing projects and linked the clinical results with the genotyping information of cancer patients with the goal to understand the predictive power of the mutations associated with protein-protein interactions in cancer. The initial idea was to compare the survival statistics between the two groups of cancer patients, one carrying disruptive mutations on the cancer driver genes and another carrying any other mutation (i.e., primarily neutral). The genetic variation data were processed using several bioinformatics tools [59] to curate the corresponding gene information and the mutation position on the protein sequence (see Methods). We required that the selected cancer driver genes had the higher mutation rates than the background mutation rate, focusing on nsSNVs from these highly mutated cancer driver genes, and discarding all non-coding variations, nonsense mutations, indels, and other variants. The analysis of the resulting datasets suggested that while subsets of mutations were annotated as neutral and disruptive, some the patient groups were not big enough to carry out the unbiased statistical tests. As a result, we studied a more general question by selecting the groups of patients with mutations located on the PPI interface versus patients with mutations located outside the interface. In this case, the patients were divided into two groups (87/74 for acute myeloid leukemia; 121/113 for liver cancer). Kaplan–Meier statistics was calculated for both groups, and the corresponding statistical significance was calculated using log rank test (see Methods).

We found that the mutations located on the PPI interfaces of the known cancer drivers significantly correlate with the decreased survival and relapse time (Fig 3-15, Fig 3-16). For acute myeloid leukemia patients, the survival correlation is evident in comparison with those patients with mutations on the corresponding cancer drivers that lie outside

the interaction interface (log-rank test p<0.0001 for patient survival time, p = 0.04 for patient relapse time), suggesting that mutations directly associated with protein-protein interactions could be treated as a survival indicator for acute myeloid leukaemia patients. A similar correlation between the interaction interface associated mutations and decreased prognosis is seen among the French liver cancer patients (log rank test p<0.0001 for both, patient survival time and relapse time). In sum, the results above demonstrate that mutations directly associated with the protein-protein interactions in the cancer patients present a strong indicator of the patient's prognosis and the edgetic perturbation of the interactome might play a key role in cancer progress and lead to decreased survival.

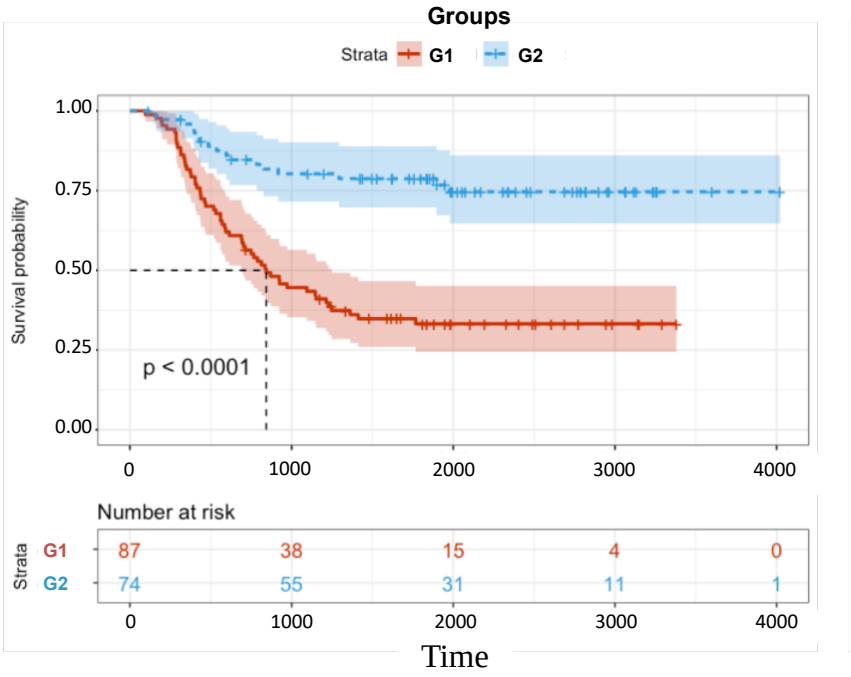


Figure 3-15 The survival analysis for the survival time in cancer patients. Based on the edgetic profiling of their mutations, the patients are divided into two groups: G1 (red) includes the patients with the disrupted cancer-centered network, and G2 (blue) corresponds to the patients with the cancer-centered network undisrupted. The lower bar shows the number of patients at risk at the different time points for these two groups.

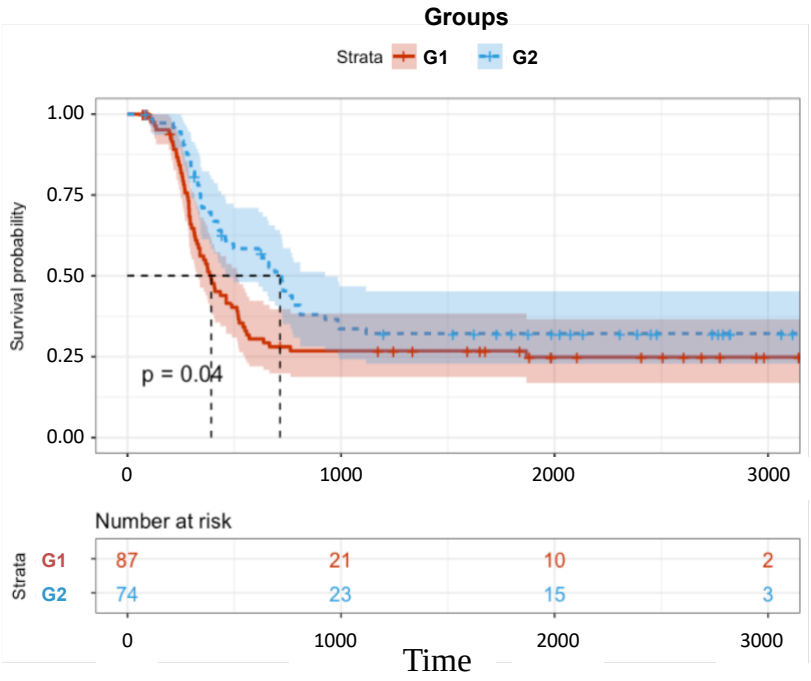


Figure 3-16 Similar survival analysis for the relapse time. The groups are defined in the same way as in Fig 3-15.

3.3 Discussion

In this work, we have presented a systematic multi-layered analysis of the disease interactome using in silico edgetic profiling of mutations and leveraging two independent experimentally validated PPI networks. Our high-throughput approach allows drawing the difference between the system-wide distributions and functional effects by the truly damaging pathogenic mutations and mutations that are expected to have minimal or no effect on the protein-protein interactions, while being located on or near the interaction interface. Our analysis of the three basic groups of mutations associated with the diseases, non-pathogenic nsSNVs, pathogenic SNVs, and frameshift mutations, has shown that both pathogenic groups are distinct from the non-pathogenic nsSNVs in their topological distribution on the network. In fact, both pathogenic types of mutations show remarkable similarity in all three network centrality measures, which is surprising given the different nature of the structural, and therefore functional, impact of these two types of mutations. A frameshift mutation typically results in a significant truncation or nonsense-mediated

decay, which leads to the loss of all PPIs mediated by this protein [206], while a missense mutation often has a small, localized effect on the protein's structure, resulting in a loss of a few, and often just one, PPIs [123]. These results suggest that that the "topological" role of a protein carrying the interaction-rewiring pathogenic mutations in the network is a key contributing factor to our understanding the overall system-wide damaging effect that leads to the disease phenotype.

In spite of their similarity with respect to the network topology, the two types of the pathogenic mutations differ drastically in their capacity to be linked to the pleiotropic effects. A group of detrimental nsSNVs on the same gene is more likely to cause different disease phenotypes than a group of frameshift mutations, which can be explained by the fact that two frameshift mutations are likely to lead to the same deleterious effect, from the point of view of the protein function, with all interactions mediated by this protein being lost. Mutations in a single pleiotropic gene are known to cause different diseases or a wide range of symptoms. However, it is still unknown which type of genetic variation is the main contributing factor to the pleiotropy. Here, based on the edgetic effects analyzed in this work, one could conclude that the disease phenotypes related to the frameshift mutations in the same gene are likely to be more conservative than the disease phenotypes related to the pathogenic nsSNV, and therefore nsSNVs are more likely to be the source of gene pleiotropy. Another distinctive feature of the pathogenic nsSNVs is their enrichment in the PPI interfaces, which is not observed in either non-pathogenic nsSNVs or frameshift mutations, and suggests that many mutations that are associated with the disease phenotypes target the mechanisms behind the macro-molecular interactions

The edgetic profiling results show the surprisingly widespread perturbations of the human interactome: 76% of disease-associated SNVs are predicted to rewire PPIs. The complex patterns of mutation-induced network rewiring in different diseases lead us to a question if all these mutations incur the comparable amount of damage, and whether the cumulative effect of damaging mutations is amplified through the synergy of their individual effects. We answer this question by developing a cumulative damage analysis that quantifies the mutation-induced network damage using the basic principles of the network robustness theory. Perhaps, the most important finding from this analysis is the

fact the traditional node-based measures used in the robustness theory are not well-suited to capture the damage from edgetic effects, and an edge-based alternative should be used instead. We note that the original network robustness concept characterizes how the network withstand failures and perturbation, while our goal is to quantify the amount of damage made by the pathogenic mutations perturbing the network. There are other critical attributes of complex networks that could potentially be included into the definition of cumulative network damage. Future study could consider network efficiency and network navigability. Network efficiency quantifies the exchange of information across the network. Network navigability studies the structural characteristics of many complex networks that support the efficient communication without the global knowledge on their structure. Routing information in networks is a common phenomenon in many complex systems, including biological networks [207]. In the future, it may be helpful to integrate the interactome's topological structure, edgetic annotation, and navigation strategies to understand how the genetic variations can influence the network efficiency and network navigability.

We considered several case studies that provide important insights into the mechanisms of complex genetic disorders, such as cancer and type 2 diabetes mellitus (T2DM). While the role of PPIs in diabetes mellitus has been recently recognized, the previous studies could not distinguish between the neutral and network-damaging mutations. Our edgetic profiling of the genes associated with T2DM suggests protein-protein interaction to be the key molecular mechanism that gets malfunctioned in T2DM because of several reasons. First, the majority of the known mutations in T2DM-associated genes are predicted to disrupt the PPIs. Second, PPIs drive the formation of clusters of tightly interconnected T2DM genes. Last, INSR, the gene that encodes a transmembrane insulin receptor has been found to carry the highest number of PPI-rewiring nsSNVs, which cumulatively disrupt all twelve interactions in which this protein is involved, with a potential to affect the important metabolic pathways.

Of special interest is the analysis of beneficial mutations, which strengthen the PPI instead of disrupting it. While the number of annotated beneficial mutations is much smaller compared to the disruptive mutations, the former group is likely to play an

important role in disease mechanisms targeting transient interactions. The example of the cancer-related HRAS gene and its interactions that was considered in this work lead to an important conclusion: the knowledge about the PPI enhancing mutations may affect the therapeutic strategies designed to inhibit specific interactions.

When studying genes related to cancer, we find a significant proportion of their mutations (more than 58%) to be disruptive. This number could be potentially higher should the coverage of the SNP-IN tool increase. The last question one can investigate is the importance of PPI-associated mutation in the analysis of the clinical data. Specifically, we ask if the knowledge of the fact that the pathogenic nsSNVs are located in the interaction interface can be helpful in predicting the survival rates and relapsing times. Our analysis has shown that the pathogenic mutations on the cancer drivers correlate well with the decreased survival in cancer patients. Our observations are consistent with some recent publications. For example, Ruffalo et. al. [208] developed a method that extend standard network smoothing techniques for predicting the disruption of specific interactions in cancer patients using somatic mutation data and protein interaction networks. They further related patient survival to each edge's mutation scores across cancer patients and found a significantly high association value with survival. Similar to our findings, this indicates that mutations related to these interactions indeed impact disease progression. However, we note that interaction disruptions caused by somatic mutations in cancer are worth further interrogation. Some fundamental questions still need to be addressed. For example, it is unclear whether these somatic mutations on the same interface share the same edgetic properties and whether they tend to co-exist in the cancer patients. We believe the use of an edgetic profiling method with the higher coverage is expected to provide further evidence to the determined correlation in the future.

In conclusion, our in-silico edgetic profiling approach aims to provide mechanistic insights into genotype-phenotype relationship. The role of a fast and inexpensive computational edgotyping approach is becoming increasingly important with the rapid growth of the personalized genomics data and ever-increasing catalogue of disease-associated variants. Such an approach can also reduce the cost of the experimental

interaction assays by prioritizing the genes and mutations according to the predicted edgetic effects.

Chapter 4 Edgotype Based Analysis of Population-specific Mutations

Since the completion of the Human Genome Project, researchers have made tremendous advancements in high-throughput genotyping technology, especially next-generation sequencing technology (NGS)[209-211]. Together with large consortium sequencing efforts, such as the International HapMap project[212] and the 1000 Genomes Project[213], genotyping hundreds of thousands of individuals has become common practice. Such practice has not only identified mutations with clinical relevance to common or rare diseases previously unknown, it has also revealed distinct mutation frequency pattern in different normal populations. These large consortium sequencing projects have made it clear that many genetic variations are population-specific[213, 214]. These genetic differences among available population data sets are critical to interpret the phenotypic differences in between populations and they have important implications in human health and diseases. Despite their potential significance, population-specific SNPs have not been studied extensively, let alone being taken into consideration in clinical practice.

Along with the development of NGS technologies, many genetic variation databases have been developed to help us make sense of the huge amount NGS data, such as 1000 Genomes [47] or Database of short Genetic variations (dbSNP) [48]. Additionally, computational approaches for functional annotation are increasingly important, since many variants are not previously described in the literature [6, 54]. There are many bioinformatics tools for functional annotation of genetic variations. Several recent reviews [55-58] give a comprehensive survey of state-of-art variant annotation tools. Most of the tools focus on the annotation of Single Nucleotide Variants (SNVs), as they are

easier to capture and analyze. And the majority of them are either sequence based or evolutionary conservation based[6]. If a SNV can be mapped on the experimentally determined protein structure or a corresponding homology model, then one can compute a number of properties using the structure information which could improve the accuracy of predicting the functional impact of this mutation [67]. Recently, our group has developed a new computational method [24], called the SNP-IN tool. SNP-IN tool predicts the effects of non-synonymous SNV on PPIs, provided the interaction's structure or structural model. It leverages supervised and semi-supervised feature-based classifiers, including a new Random Forest self-learning protocol. The accurate and balanced performance of SNP-IN tool makes it useful for functional annotation of non-synonymous SNV.

Recently, a concept of “edgotype” has been proposed [123], which is concerned with the functional outcomes of genetic variants on protein-protein interactions (PPIs) and the corresponding rewiring effect on the interactome. We emphasize that edgotype provides alternative molecular explanations for mutation’s impact and why they underlie many complex genotype-to-phenotype relationships [146]. Edgetic perturbation models view mutations as and interaction-specific or edge-specific (‘edgetic’) alterations in the human interactome. An edgetic alteration can cause the removal of one or a few interactions but leaving the rest intact and functioning. It might have subtler impact on the network, and does not necessarily result in disease phenotype[147]. More importantly, edgetic perturbation model can easily explain confounding genetic phenomena, such as genetic heterogeneity[21, 94]. Edgetics is a new approach to interpret genotype-to-phenotype relationships in the context of the biological network. It also shows us a way of studying population genetics. As large consortiums, like 1000 genome project[148] and ENCODE[149], have generated numerous amount of genetic variation data from different populations around the world, it provides us a great opportunity to shift from traditional population genetics to population edgetics and apply the new methodology to study the genetic differences within and between populations. We also expect that distinct edgetic profiles, rather than one or several mutations, harbored by populations or individuals, can better explain why there is different disease frequency patterns and susceptibility across different populations. In sum, network-based analysis of the genetic architectures

may not only shed light on biological mechanisms underlying complex phenotypes, but also yield better ways of measuring the genetic predisposition to a certain disease.

In this work, we create a comprehensive catalog of population-specific mutation's edgetic effects at the whole-interactome level. Our work benefits from our recently developed SNP-IN tool to determines interaction-rewiring effects of non-synonymous single nucleotide variants (nsSNVs). The method was applied to 46,599 nsSNVs collected from the 1000 Genome Project by leveraging the structural information on PPI complexes. We determined that a considerable amount of normal population specific nsSNVs can cause disruptive impact to at least one PPI. We also showed that genes enriched with disruptive mutations obtained from the healthy populations are in fact associated with diverse functions and are implicated in various diseases. Our analysis indicates that some gene edgetic profiles are distinct from each other among 5 major populations and can help explain the population phenotypic variance. Finally, network analysis reveals phenotype-associated modules are enriched with disruptive mutations and the difference of the accumulated damage in such modules may suggest population-specific disease susceptibility. We expect that our approach will provide a more accurate and in-depth characterization of the functional consequences of ethnic-specific alleles, leading to a better understanding of the clinical and phenotypic outcome.

4.1 Methods and Materials

Figure 4-1 gives an overview of this systematic analysis of population-specific mutation in the human interactome. We start with mutation data collection and processing. After that, nsSNVs collected from 1000 Genomes Project are then mapped to protein-protein interaction structures. This structural information is necessary for our SNP-IN tool to predict mutation's effect on the protein-protein interaction. With these function characterizations provided by SNP-IN tool, we evaluate the perturbations in the human interacotme caused by normal mutations. Following is a large-scale edgetic profiling of genes enriched with disruptive mutation. Based on the edgotype strategy, we investigate their potential role related to the phenotypic variance across populations. The last part is

network analysis, including topological analysis and functional module enrichment analysis about disruptive mutations in the interactome.

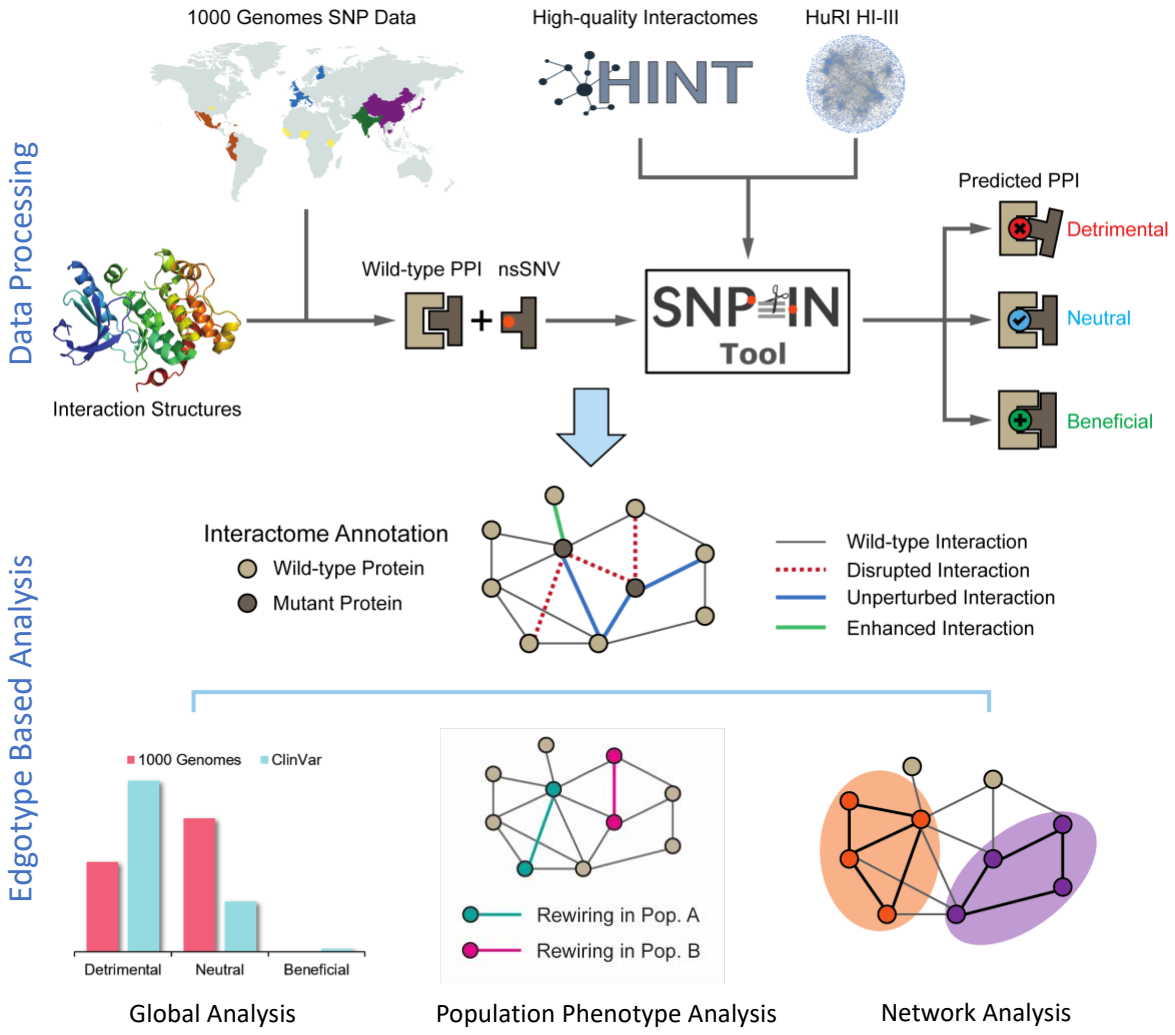


Figure 4-1 Overview of the analytical workflow in Chapter 4. The mutation data is collected from the 1000 Genomes Project. We map the mutations to structurally resolved protein-protein interaction complexes and apply SNP-IN tool for functional annotation. Analysis works includes global analysis of mutation's rewiring effect, edgotype based analysis of population phenotypic variance and network analysis of disruptive mutations.

4.1.1 Genetic mutation data processing and construction of the human interactome

The mutation data are collected from 1,000 Genome Project [213]. 1,000 Genomes Project is one major subsequent international consortium effort after the Human Genome Project to catalogue the most genetic variants. In the functional annotation and later analysis, we mainly focus on only non-synonymous, missense SNVs and excluded non-coding variations, synonymous SNVs, short indels and structural variations. SNV is a single nucleotide substitution occurring in a genome. It is the simplest and yet the most common type of genetic variation among people. Specifically, in this work we focus on non-synonymous SNVs, which occur in the coding region, because they are most likely to make a functional impact on the PPI. Hence, we only deal with nonsynonymous SNVs. The mutation data was first processed with ANNOVAR [59] to retrieve SNV locations on the genes and the corresponding residue change information. All these information is essential to apply our recently developed SNP-IN tool (non-synonymous SNP INteraction effect predictor tool) [24] to predict the mutation's rewiring effect.

To construct a unified human interactome, we resort to two different protein–protein interaction data sources: High-quality INTeractomes database (HINT) [175], and Human Reference Protein Interactome Mapping Project (HuRI) [132]. HINT (http://hint.yulab.org) is organized as a centralized database of high-quality human PPIs integrated from several other databases and annotated using both, an automated protocol and manual curation. Unlike the HINT database, HuRI is a primary source for experimentally validated PPIs using yeast-two-hybrid experiments. The two PPI sources were merged because they provide complementary views of the whole human interactome. The HINT database contains 63,684 interactions. For the HuRI dataset, we collect 76,537 interactions. In total, we generate a human interactome consisting of 105,087 interactions, where 35,134 interactions exist in both data sources.

4.1.2 Functional annotation of the nsSNV

To accurately determine the functional damage, with respect to a protein–protein interaction, caused by a mutation, our approach requires the information about the mutation and protein-protein interaction structure as an input [24, 215]. In this study, we focus on one major mutation types, the Single Nucleotide Variant (SNV). The SNP-IN tool predicts the rewiring effects of nsSNVs on PPIs, given the interaction's experimental structure or accurate comparative model. More specifically, the SNP-IN tool formulates this task as a classification problem. There are three classes of functional effects predicted by the SNP-IN tool: beneficial, neutral, and detrimental. The effects are assigned based on the difference between the binding free energies of the mutant and wild-type complexes ($\Delta\Delta G$). Specifically, $\Delta\Delta G = \Delta G_{mt} - \Delta G_{wt}$, where ΔG_{mt} and ΔG_{wt} are the mutant and wild-type binding-free energies correspondingly. The beneficial, neutral, or detrimental types of mutations are then determined by applying two previously established thresholds to $\Delta\Delta G$ values [178, 216]. We note that SNP-IN tool requires a PPI structure in which a mutated gene is involved. First, if a PPI already has a native structure, we extract it from the Protein Data Bank [217]. If there is no native structure for a protein–protein interaction, we apply homology, or comparative, modelling [218] to get a structural model, either for the full protein length interaction or at least for a pair of protein domains that form the interaction interface [24, 215].

4.1.3 Evolutionary rate calculation and comparison

The high-confidence collection of cancer genes was a union of the Cancer Gene Census[219] and a recently published computationally predicted cancer gene set, MutPanning [220]. The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) is an ongoing effort to catalogue those genes which contain mutations that have been causally implicated in cancer. Each gene in Cancer Gene Census comes with an expert-curated description of the genes driving human cancer. Such manual annotation explains how dysfunction of these genes drives cancer, and CGC is used as a standard in cancer genetics across basic research, medical reporting and pharmaceutical development[219]. MutPanning is a computational method for driver-gene identification

that combines the characteristic contexts around passenger mutations with the signals of mutational recurrence[220]. It's another resource of driver gene across 28 tumor types. Those two cancer driver gene sources contain 723 and 460 genes, respectively, with an overlap of 196. Thus, in total, we collected 987 cancer genes. Housekeeping genes are genes that are essential for the existence of a cell and the maintenance of basic cellular functions[221]. They expected to express in all cells of an organism under normal and pathophysiological conditions[222]. The list of housekeeping genes we retrieved is compiled by Eli Eisenberg and Erez Lavanon[223]. It consists of 3804 genes. Lastly, we define disruptive gene as a gene carrying at least one mutation, which causes disruptive effect on the related protein-protein interactions based on SNP-IN tool output.

The evolutionary rate (ER) of a gene is represented by the ratio between the number of non-synonymous substitutions (dN) to the number of synonymous substitutions, dN/dS[224]. Specifically, we calculate the dN/dS ratio by comparing human gene sequences with the orthologous groups of Homo sapiens (humans). The orthologues genes of Macaque, Gorilla, Orangutan, Chimpanzee, and Gibbon corresponding with human genes (GRCh38.p13) were queried from Ensembl Biomart platform[225], as well as the dN and dS data. Evolution rate with individual organism was calculated by dN/dS, while missing values and infinite values were removed. For sets of homologs that did not include exactly one representative in each organism, the group mean of ER was taken as the representative. The eventual ER for a human gene is the average value of ERs with individual organism, and we only include genes with homologue in more than three organisms. The comparison of ER between different sets of genes are done with Wilcoxon test since no prior information about the underlying distribution is known.

4.1.4 Calculation of disruptive mutation rate in proteome and GO enrichment analysis

For a human gene, the disruptive mutation rate is simply the ratio of the number of mutated residues causing disruptive impact on PPIs to the protein sequence length. In other words, the total number of disruptive mutations on a gene is normalized by the protein sequence length. We first collected all the disruptive mutations occurring on the

gene based on our SNP-IN tool annotation. The protein sequence length information was retrieved from Uniprot database[226]. The average disruptive mutation rate across the proteome is considered as background rate, and we define genes with their disruptive mutation rate larger than the average plus the standard deviation as genes enriched with disruptive mutation.

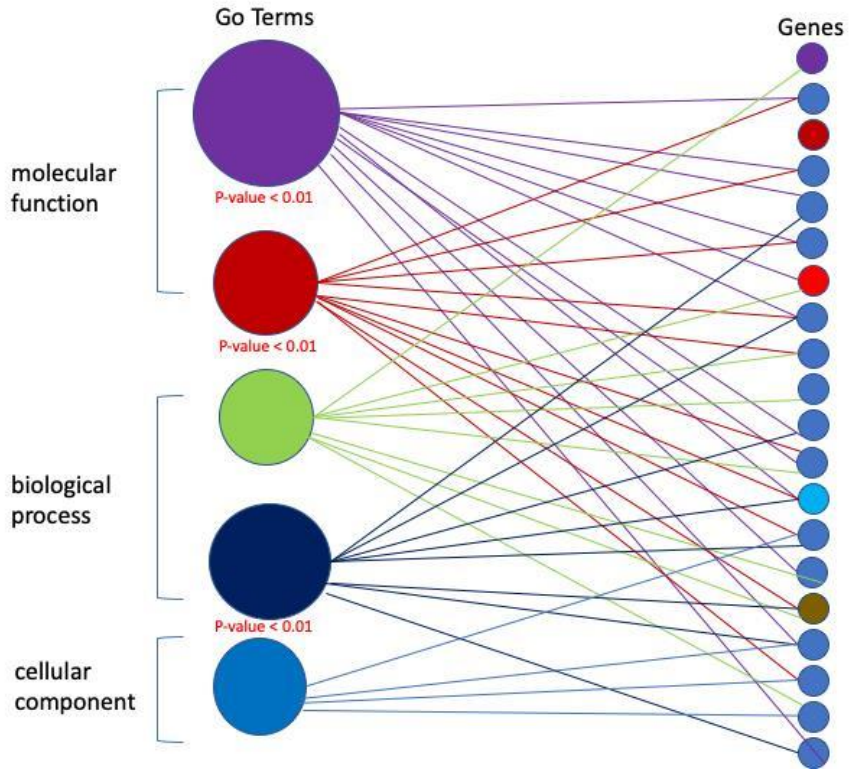


Figure 4-2 Illustration of GO enrichment analysis. GO term enrichment analysis is done by testing the input gene set for each term to see if it is enriched compared against the background.

We investigate the biological implication of genes enriched with disruptive mutations based on their agreement with the available biological knowledge, such as Gene Ontology [227]. The Gene Ontologies are structured as a Directed Acyclic Graph (DAG), with nodes in the graph representing GO terms and edges in the graph representing the relations between GO terms. Genes are associated to Gene Ontology terms via GO annotations. Each gene can be linked with multiple GO terms, and some of these GO terms can be the same GO type. One of the main uses of the GO is to perform enrichment analysis on gene

sets (See Fig 4-2). Basically, we are given a set of genes of interest (e.g., genes with significantly high number of disruptive mutations in this work); we name this gene set as study set. And all genes in the human genome are considered as background, named as population set. For each GO term, we first count the number of genes (k) in the study set that are associated to the term, and the number of genes (n) in the population set that are associated to the same term. Then we compute the likelihood that we obtain at least k genes associated to the term if n genes would be randomly sampled from the population set. For the gene set enriched with disruptive mutations, we perform gene enrichment analysis and obtain the corresponding list of enriched GO terms. In the GO enrichment analysis, we use the third level of the GO hierarchy and kept the GO terms with P-value ≤ 0.01. The third level represents a trade-off between having too general, but well-populated GO terms from the second level (e.g., GO:0050789 regulation of biological process) and more specific but not well-populated terms from the fourth level, which cannot be used for the enrichment analysis. The GO enrichment was performed using DAVID [228], and multiple testing correction was done via false discovery rate estimation [229].

4.1.5 Population specific edgetic profiles of genes enriched with disruptive mutations

For a gene enriched with disruptive mutations, we introduce the population gene edgetic profile concept to describe the diverse network rewiring effects caused by disruptive mutations centering around the mutated gene across different populations. Simply put, the edgetic profile is represented as a sequence of vectors with different length. Each vector consists of a list of disruptive mutations targeting the same interactions. The element takes only binary values, where 1 stands for a disruptive mutation with non-zero allele frequency, and 0 means a disruptive mutation is not present in this population. For example, *ep* = [[1, 0, 0, 1], [0, 1, 1], [1, 0, 0, 1, 1]]. In a population, for gene/protein A, it has three interaction partners B, C, D. For interaction A-B, there are four mutations predicted as disruptive occurring in this interaction. For each in mutation, 1 indicates such disruptive mutation is present in this population, 0 otherwise.

Thus, for genes enriched with disruptive mutations, we collect all the annotation information of mutations, and build the edgetic profile for each of them. To measure the gene edgetic profiles in different populations, we calculate the Manhattan distance between two vectors after flattening the original list of vectors into a homogeneous vector:

$$d\left(ep^{(i)}, ep^{(j)}\right) = \left\| ep^{(i)} - ep^{(j)} \right\|_1 = \sum_{k=1}^{l} \left| ep_k^{(i)} - ep_k^{(j)} \right|$$

The Manhattan distance between two edgetic profiles is further normalized by taking the total number of disruptive mutations into account:

$$d_{norm}\left(ep^{(i)}, ep^{(j)}\right) = d\left(ep^{(i)}, ep^{(j)}\right)/l$$

To evaluate the average difference of gene edgetic profiles between any two populations, we sum up all pair-wise difference, and divided by the total pair number:

$$d\left(ep^{(i)}, ep^{(j)}\right) = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} d_{norm}\left(ep^{(i)}, ep^{(j)}\right)}{N(N-1)}$$

4.1.6 Examination of topological properties of disruptive genes in human interactome

In our comparative network analysis, we first investigate their topological importance of the rewired edges targeted by pathogenic mutations and normal population mutations separately. Specifically, we investigate two major edge centrality measures in the network science: shortest-path edge betweenness and current-flow betweenness. Betweenness centrality was proposed as a general measure of centrality[230]. it has been applied to a wide range of real-world problems, including problems related to biological networks[231], social networks[232], transportation networks[233]. Typically, an edge with higher betweenness centrality in a complex network would have more control over the network, because more information will flow through that edge. In the context of human interactome, it means these mutations target at interactions that occupy critical positions in the human interactome. The shortest path edge betweenness centrality is a

measure of centrality based on the number of the shortest paths that go through an edge. It is defined as the sum of the fraction of all-pairs shortest paths that pass through an edge [163]. Formally, the shortest path edge betweenness centrality of an edge *e* is given by the expression:

$$c_B(e) = \sum_{u,v \in V} \frac{\sigma(u, v|e)}{\sigma(u, v)}$$

where V is the set of nodes, $\sigma(u, v)$ is the number of shortest between u and v, and $\sigma(u, v|e)$ is the number of those paths passing through edge e. Current-flow betweenness is another global centrality measure, which is based on an electrical current model for information spreading. It is also known as random-walk betweenness centrality[234].

Another graph-based metrics to characterize the rewiring effect caused by disruptive mutations is network efficiency[235]. The concept of efficiency measures how efficiently the network propagates and exchanges information. To compute the efficiency of the subnetworks targeted by pathogenic mutations and normal mutations, we first constructed two separate subnetworks by grouping the corresponding disrupted interactions together. For a pair of nodes in the network, the efficiency is the multiplicative inverse of the shortest path distance between the pair. The global efficiency of a graph is defined the average efficiency of all pairs of nodes. Mathematically, it is expressed as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d(i,j)}$$

where N is the number of nodes in a network and *d(i,j)* is the length of the shortest path between a node pair of i and j.

4.1.7 Phenotype associated community detection in human interactome and their enrichment with disruptive mutation

Discovering biologically relevant modules is a challenging task [167, 168]. These methods primarily come in two different flavours. The first group of methods identify the modules

in a biological network by relying exclusively on the network's topology. This is a challenging task due to the lack of information about specific genes/proteins contributing to biological functionality. Methods from the second group start with the "seed genes", and gradually extract additional genes in the network to grow the module.

For the first one, we adopt a strategy based on the idea of "Diffusion State Distance" (DSD)[236], Such strategy has been proven to be the best performer in the DREAM challenge[237], as it yields most phenotype associated modules in the final evaluation round. This approach has two main steps: computing the DSD matrix and applying the spectral clustering on the DSD matrix to identify phenotype associated modules. Diffusion State Distance (DSD) is an alternative proximity measure to the traditional shortest-path. Typical shortest-path measure favors the hub-like nodes in the network and did not incorporate a lot of informative structure information in the network. Intuitively, a protein pair connected by paths through low-degree nodes share more functional similarity than other protein pairs connected by paths that goes through the hubs. (See Supplementary Figure S5-1). So, DSD is a more fine-grained measure of similarity that downweighs the hubs in the human interactome. Formally, given an undirected graph *G(V, E)* consisting of a node set : V ={v1, v2, … , vn} and |V|=n, define a vector $He^k(A)$:

$$He^k(A) = (He^k(A, v_1), He^k(A, v_2), \dots, He^k(A, v_n))$$

where $He^k(A, v_i)$ is expected number of times that a simple symmetric random walk starting at node A, and proceeding for k steps, will visit node v_i. So $He^k(A)$ is the global distance measure from node A to all the other nodes of the network. And the DSD between node A and B is defined as following:

$$DSD(A, B) = \|He(A) - He(B)\|_1$$

The DSD matrix was calculated using the "cDSD" method[238] from software available at: http://dsd.cs.tufts.edu/capdsd. The follow-up spectral clustering on the DSD matrix is performed with scikit-learn package[239].

For the second seed base module detection scenario, we resort to a recently published method:, DIseAse MOdule Detection (DIAMOnD) [166]. DIAMOnD is a disease module detection algorithm that utilizes known seed genes to identify disease modules according to the connection significance to the seed proteins. It exploits the fact that disease associated proteins do not reside within locally dense communities. And the significance of their connections is a more predive measure than the local network density. Also, the use of the significance of the number of connections reduces the spurious detection of high-degree proteins compared against using their absolute number of connections. The algorithm outputs a connected disease module with a list of candidate disease-associated proteins ranked by their connectivity significance.

4.2 Results

4.2.1 A substantial amount of nsSNVs among normal populations can disrupt PPIs

Genetic variations are first collected from the 1000 Genomes Project data portal. 1000 Genomes Projects have catalogued more than 88 million genetic variants. The majority of them are SNVs (84.7 millions); it also included 3.6 million short indels and 60,000 structure variants[213]. As this work mainly focuses on SNVs, we filter out non-synonymous SNVs and extracted the residue change information on the protein sequence with ANNOVAR[59]. To apply SNP-IN for functional annotation on SNVs, we first the mapped them to native protein complexes in PDB databases. If there was is no native structure for a PPIs, we resort to homology modelling to get a comparative structural model, either for this the full-length interaction or only at least for a partial pair of protein domains that form the interaction interface. In summary, we were able to map nsSNVs to 5,324 native protein-protein interaction structures. And 4,258 full-length protein-protein interaction models and 983 domain-domain interaction models are built to meet the requirement of SNP-IN tool. (See Fig 4-3)

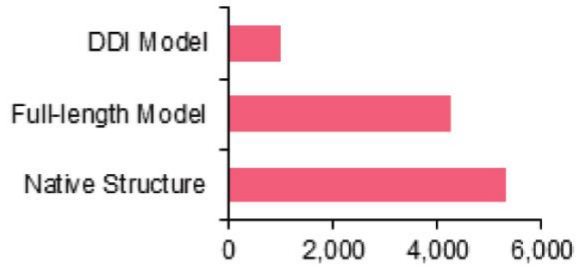


Figure 4-3 Statistics about protein complex data collection from three sources: native PPI structure, full length PPI model and domain-domain interaction model.

Our results showed that, among all 46,599 SNVs annotated by SNPIN tool, which accounted for about 5% of the total SNVs we collected, 25,185 SNVs (54%) were predicted as detrimental to at least one PPI that the corresponding disease protein was involved in, and only 313 SNVs were labelled as beneficial (See Fig 4-4). In previous work, we applied the same strategy to annotate all the pathogenic mutation curated from ClinVar database. We retrieved a list of human disease genes from ClinVar databases, as well as 3,401 pathogenic non-synonymous mutations. Applying the same functional annotation procedure, we found that about 1/3 of the total disease-associated SNVs we collected, 2,592 SNVs (76.2%) were predicted as detrimental to at least one PPI, and 48 SNVs (1.4%) were labelled as beneficial (See Fig 4-5). This shows that a significantly large proportion of disease-associated SNVs could cause specific alteration of PPIs. In other words, there is a global and wide-spread rewiring of PPI network, and it would play a key role in the disease pathogenesis. On the contrary, the SNVs observed in healthy people have less severe effects. But there is still a significant amount of them can cause detrimental impact to protein-protein interactions, even not as widespread as pathogenic ones. The difference could justify the functional utility of our tool. Also, this also confirms our previous observation that genetic variations enhancing the protein-protein interaction or causing beneficial effects are rare events in the human genome.

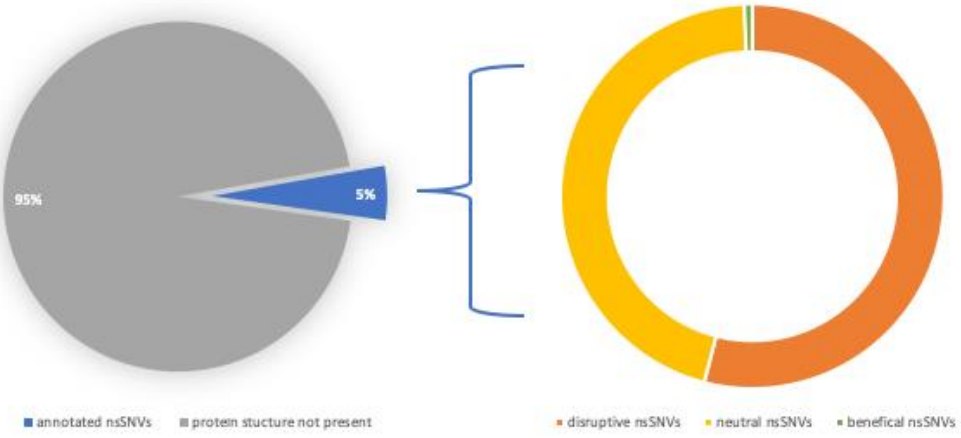


Figure 4-4 Results of mutation data collection from 1000 Genomes Project and SNP-IN tool annotation. Our annotation covers about 5% of total non-synonymous SNVs in the 1000 Genomes Project. And we predict about 54% of them can cause disruptive impact on the protein-protein interactions.

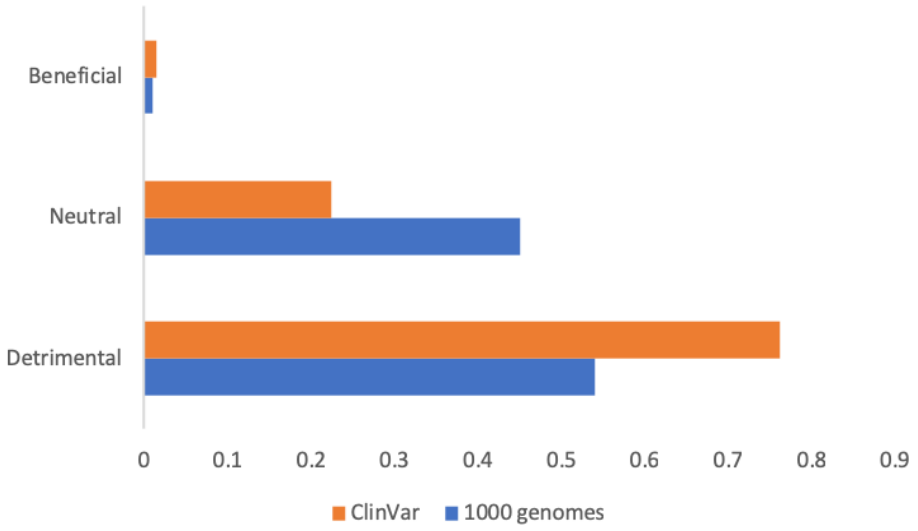


Figure 4-5 Comparison of SNP-IN tool annotation results from pathogenic mutations from ClinVar database and normal mutation from 1000 Genomes Project.

It is somewhat unexpected that such a significantly large number of normal SNVs could disrupt protein-protein interactions in healthy populations. This reveals the abundance of deleterious nsSNVs in human genome. We note some previous research[240] have also shown up to 48% of nsSNVs specific to a single genome are deleterious in nature.

This intrigued us to study the evolutionary features of genes carrying disruptive mutations, especially whether they evolve at slower or faster rate than the gene sets, such as cancer genes and housekeeping genes. Thus, we integrated cancer genes from two sources: COSMOS and MutPanning. This cancer gene set contains 987 cancer genes with high confidence (Fig 4-6). We also curated a housekeeping gene set, which consists of 3804 genes (Fig 4-6) (See methods for details). After calculating the evolutionary rate for each group, we found that the differences of evolutionary rates between three gene sets are all significant. Cancer genes are most evolutionary conserved. Both housekeeping genes and disruptive genes have a faster evolutionary rate than cancer genes (P = 1.443e-6 and P =0.003 separately) (See Fig 4-7). At the same time, disruptive genes are evolutionary conserved than housekeeping genes (P = 0.004). This indicates that the disruption imposes unique functional constraints on the human genes, and it could have a wide implication to understand the genetic basis of human genome evolution.

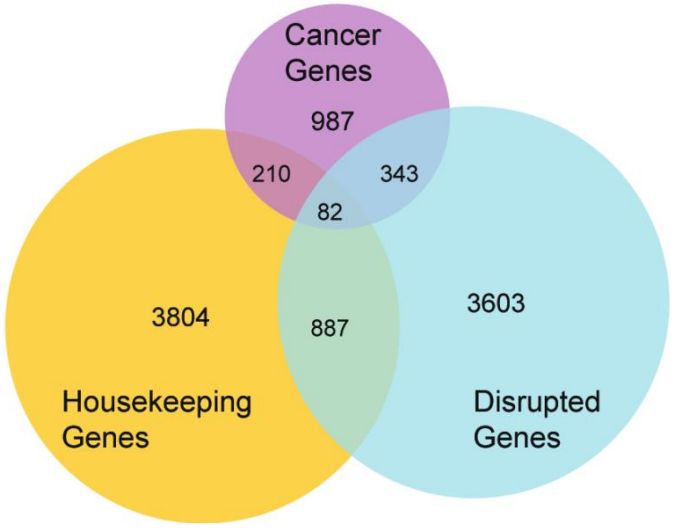


Figure 4-6 Statistics for three sets of genes in evolutionary rate calculation: disruptive genes, cancer genes and housekeeping genes.

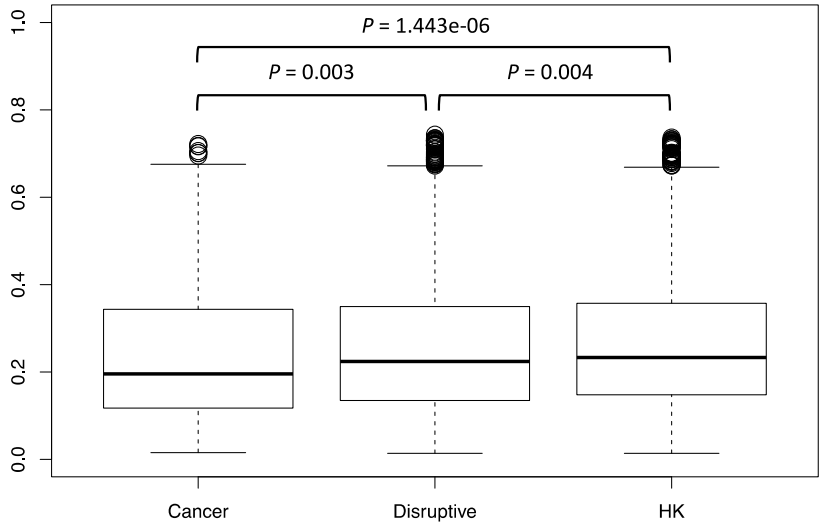


Figure 4-7 Comparison of evolutionary rate of three gene sets. Among them cancer genes are most evolutionary conserved. Disruptive genes evolve slower than housekeeping genes.

4.2.2 Genes enriched with disruptive mutations are associated with diverse molecular functions and are implicated in various diseases

Following the observation that there is widespread disruption of in the human interactome caused by common variations in normal populations, we next focus on genes enriched with disruptive mutations. We define genes enriched with disruptive mutations as genes whose mutation rate is larger than average rate plus standard deviation (See methods for details). Among all the 3603 genes, which carry at least one disruptive nsSNPs based on SNP-IN tool output, we got 461 genes with a significantly higher rate of disruptive mutations.

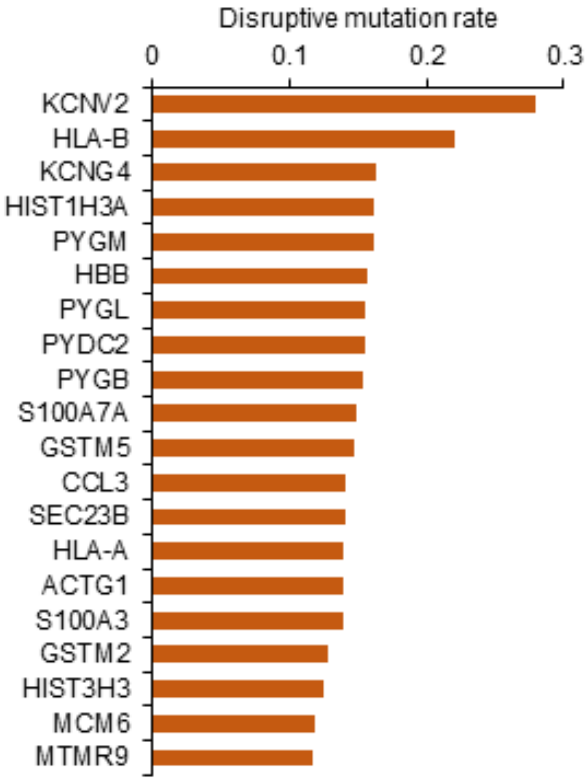


Figure 4-8 Top 20 genes with highest disruptive mutation rate; disruptive mutation rate is normalized by protein sequence length.

Next, we performed GO enrichment analysis on these 461 genes with a high disruptive mutation rate and investigate what molecular functions and biological processes these genes are involved in[226]. We only selected the third level GO terms in the GO hierarchy tree. In total, we have 458 GO terms in total with significant p-value after multiple hypothesis test correction. After examining these significant GO terms, we have some interesting findings. First, in the molecular function category, we obtained 62 GO terms enriched in the highly disrupted gene set. Interestingly, among the top 20 molecular function GO terms (See Supplementary Fig 4-2), the majority of them are related to other kinds of "binding" activities, such as GO:0000166: nucleotide binding, GO:0046977: TAP binding, GO:0030170: Pyridoxal phosphate binding. This indicates that these mutations can potentially exert their effect through a different mechanism. As to the biological process GO term category, more than one third are related to immune system response. Immune system consists of many biological structures and processes. It protects an

organism against a wide variety of pathogens and diseases. We point out that, besides the occurrence of significantly high number of disruptive mutations on these genes, there is also notably uneven distribution of these mutation across different populations. Hence, we suspect that difference of the resulted network rewiring effect could be linked to the population phenotypic variance, such as different disease susceptibility among different populations.

Generally speaking, such disruptive effect on the protein-protein is expected to cause damage to regular biological processes or molecular functions. Thus, we further investigated any potential links between these genes enriched with disruptive mutations and complex diseases. We examined these genes in OMIM [241] and HGMD [49] databases. Indeed, the 461 genes enriched with disrupted mutations are associated with various complex diseases. (See Supplementary Table S4-1) We also overserved that the higher the disruptive mutation rate are, the more disease phenotypes the genes are related to. However, as these mutations are carried by normal population, their disruptive effects are not likely to be enough to result in the disease phenotype. We suspect that a different group of mutations on the same gene set can cause more severe damage to the protein-protein interaction or they can cause the disease via different mechanisms. Further, these normal disruptive mutations can contribute to the collective damages on the protein-protein interaction, and their disruptive effect might increase certain disease susceptibility. Indeed, such hypothesises got more supporting evidence from our analysis in Section 4.2.5.

4.2.3 Edgotype analysis reveals distinct gene edgetic profiles in major populations

Genes and their products interact with each other within the human interactome, rather than functioning individually. The human interactome, or the human protein-protein interaction network is the set of protein–protein interactions that occur in human cells. The interactome is depicted as nodes and edges representing individual proteins and their mutual interactions, respectively. Edgetic network perturbation model, or edgotype, emphasizes the disruption of specific edges (See Fig 4-9). However, unlike the traditional

gene-centric model, which mainly focuses on the overactivity or silencing of gene expression. Edgotype can easily incorporate the influence of genetic variation. It is an extension and complement of the classic gene-centric paradigm.

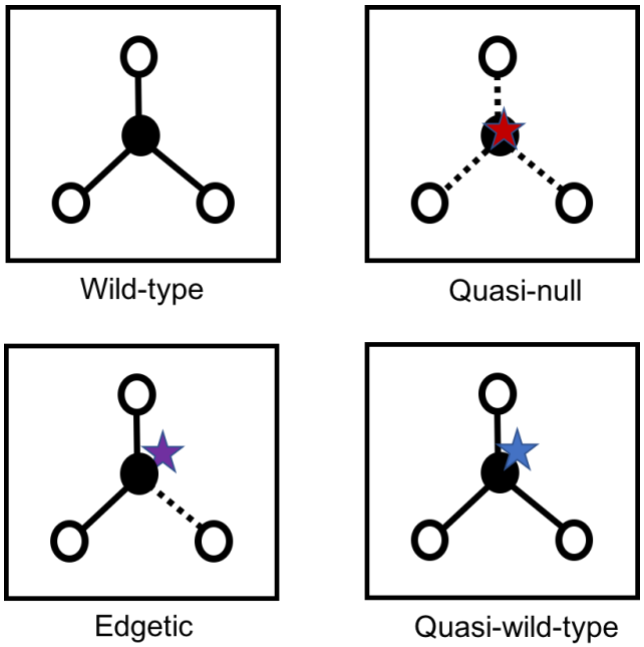


Figure 4-9 Illustration of edgotype concept. Based on the edgotype idea, mutations can be categoried into four groups: wild-type, quasi-null, edgetic and quasi-wild-type.

Inspired by the "edgotype" idea, we systematically characterize the edgetic profile for genes enriched with disruptive mutations in the human genome across different populations. To simplify the analysis, we treat the neutral and beneficial ones as the same, namely mutations have preserving effects on the protein-protein interactions. We exclude the preserving mutations from constructing the edgetic profile. Thus, the edgetic profile of a gene in a specific population is represented as a list of binary vectors including just 1s and 0s, where 1 stands for a disruptive mutation with non-zero allele frequency, and 0 means a disruptive mutation is not present in this population. The difference between the gene edgetic profiles in two different populations is quantified using the Manhattan distance between the extended vectors (See methods for details.). For a single gene, we calculate the average pair-wise edgetic profile difference between any two populations. This average difference is further normalized by the total number of disruptive mutations. On average, the normalized difference between any two major populations for a gene in

the human genome is 43%. In other words, on average, 43% of the total disruptive mutations occurring on one gene have distinct prevalence between two populations. We believe that, such distinct prevalence of disruptive mutations in different population, combined with their unique interactions they target, can help better explain the genetic diversity and phenotypic variance across populations.

4.2.4 Edgetic properties of population-specific SNPs could help explain the phenotypic variance across different populations

Our computational predictions enable the characterization of edegtic property of genetic mutations. More importantly, these functional characterizations of mutations, together with population-specific genetic architecture, enable us to generate a concrete molecular-mechanism hypothesis for certain complex phenotypes and help explain the phenotypic variance across different populations. This approach to analyze genotype-phenotype relationship can be well exemplified by the one edgetic mutation rs671 on the gene ALDH2. ALDH2 gene encodes aldehyde dehydrogenase 2, a member of a family of enzymes that metabolize alcohol. ALDH2 plays a major role in ethanol catabolism[242]; it catalyzes conversion of acetaldehyde to acetic acid. Similarly, gene ALDH1A1 encodes retinal dehydrogenase 1, it mainly oxidizes retinaldehyde to retinoic acid[243]. In fact, it has a broader specificity and oxidize other aldehydes[243]. The missense rs671 mutation of ALDH2 gene is well known as the culprit of the phenomena of "Asian Flush" [244, 245], in which a person, often of Asian descent, develops flushes and turns red on the face, neck and shoulders after drinking alcohol[246, 247]. The frequency of individuals carrying rs671 is highest in eastern Asia (MAF = 17.6%) but is almost absent among other major populations[246]. ALDH2 is involved in two protein-protein interactions: one self-interaction (ALDH2-ALDH2), and one hetero-interaction (ALDH1A1-ALDH2). According to SNP-IN tool prediction, rs671 could disrupt both of them (Fig 4-10). Naturally, ALDH2 is a randomized tetramer. The rs671 allele results in a mutant ALDH2*2 protein. And it causes critical disruption to ALDH2-ALDH2 interaction. Such disruption destabilizes the tetramer, interferes with catalytic activity, and increases protein turnover[242]. Eventually, it makes ALDH2*2 protein defective at metabolizing alcohol. It is worth noting that rs671 also cause damaging impact on ALDH1A1-ALDH2

interaction, which further reduce the ethanol catabolism. Interestingly, there is another mutation rs8187929 on gene ALDH1A1 can exert similar impact to ALDH1A1-ALDH2 interaction (Fig 4-10). A recent study reported that rs8187929 is related to alcohol consumption and drinking behaviours in Japanese population[248]. Similar to rs671, rs8187929 also predominantly exists in East Asian population, but has a much lower allele frequency (MAF = 4.56% in East Asians). More importantly, according to the edgetic profiling, rs8187929 has limited impact to the enzymatic activities involved in ethanol metabolism. So, compared against rs8187929, rs671 has a wider impact to corresponding protein-protein interactions and a higher prevalence in East Asian population. In other words, rs671 has a distinct edgetic profile from rs8187929. This can explain why rs671 a better indicator of drinking behaviour and impose lower risk of alcoholism than rs8187929 in East Asian population.

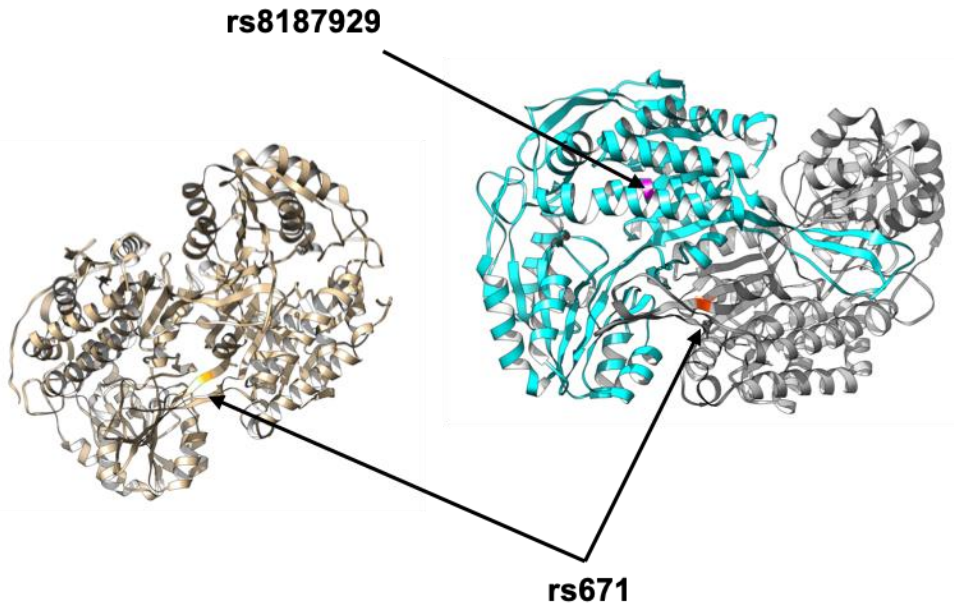


Figure 4-10 Case study about "Asian Flush" and relevant genes: rs671 and rs8187929. rs671 is a known culprit for “Asian Flush”; it disrupts two important interactions. rs8187929 is also related to human drinking behaviour, but less effective than rs671.

Another case study is about gene HLA-B. HLA-B caught our eyes, as it is among genes with the highest disruptive mutation rate. HLA-B is an interesting one, with a disruptive mutation rate of 22.1% (Fig 4-11). HLA-B (major histocompatibility complex, class I, B) is a human gene encoding a protein that plays a critical role in the immune system. HLA-B is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria[249]. It is well known that HLA-B gene has many different normal variations. Many studies have shown that such phenomena can enable each individual's immune system to react to a wide range of foreign invaders[250-253]. In fact, HLA is the human version of the major histocompatibility complex (MHC). The HLA complex is one of the most complicated complexes in human body. In humans, the HLA-B gene and two related genes, HLA-A and HLA-C, are the major genes in MHC class I. According to our characterization, HLA-A is also among the top 461 genes enriched with disruptive mutations (Supplementary Figure S4-3). Given the high number of normal variations can disrupt protein-protein interactions, this might suggest that the wide-range ways of immune response to numerous pathogens can be partially attributed to many distinct HLA gene edgetic profiles and the different rewiring of the subnetwork centering around the HLA family genes. We further conjure that, together with the unequal distribution of mutation frequency among different population, the rewiring effect of HLA complex rewiring could contribute to the different disease susceptibility and how the immune system responds to the pathogens across different populations.

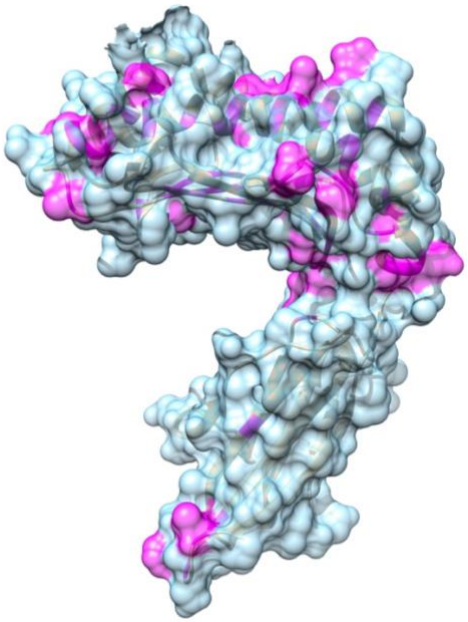


Figure 4-11 Case study of HLA-B gene. HLA-B gene is one of top genes with highest disruption mutation rate. Shown is the protein structure of the HLA-B gene. Disruptive mutations are widespread on the structure.

4.2.5 Comparative network analysis shows disease mutations target at less efficient subnetworks and normal mutations might contribute to disease susceptibility

As a comparison analysis, we further examined the topological properties, as well as the relevant rewiring, of both pathogenic mutations curated from ClinVar database and normal mutations from the 1000 Genomes Project in the human interactome. We first constructed the human interactome by utilizing two main protein-protein interaction data sources: HINT and HuRI. HINT (http://hint.yulab.org) is organized as a centralized database of high-quality human PPIs integrated from several other databases. Unlike the HINT database, HuRI is a primary source for experimentally validated PPIs using yeast-two-hybrid experiments. In short, we think these two interactome complements each; hence we merged them as a unified interactome. The final resulted human interactome consist of 105,087 interactions, with the largest components consisting of 104,563 interactions. (See methods for more details.)

For the list of human diseases genes retrieved from ClinVar databases, 576 genes carrying at least one disruptive mutation; and the disruptive mutations occurring on these pathogenic genes rewires 1,162 interactions. On the other sides, the number of genes carrying at least one disruptive mutation from 1000 Genome Project is 3,603, and number of rewired interactions is 4,529. Interestingly, the overlap between two gene sets and the two interaction sets caused by the corresponding nsSNVs are 449 and 736 respectively. The overlap is relatively high, and it showed that both pathogenic mutations and normal variations can cause detrimental effects on the same interaction subset of the interactome (See Fig 4-12). This suggests that pathogenic mutations should cause the disease phenotype in a more complicated manner, rather than solely rewiring the protein-protein interaction. We further hypothesize that disruptive mutations in normal population are not enough to cause the disease phenotype and but could contribute to the disease susceptibility.

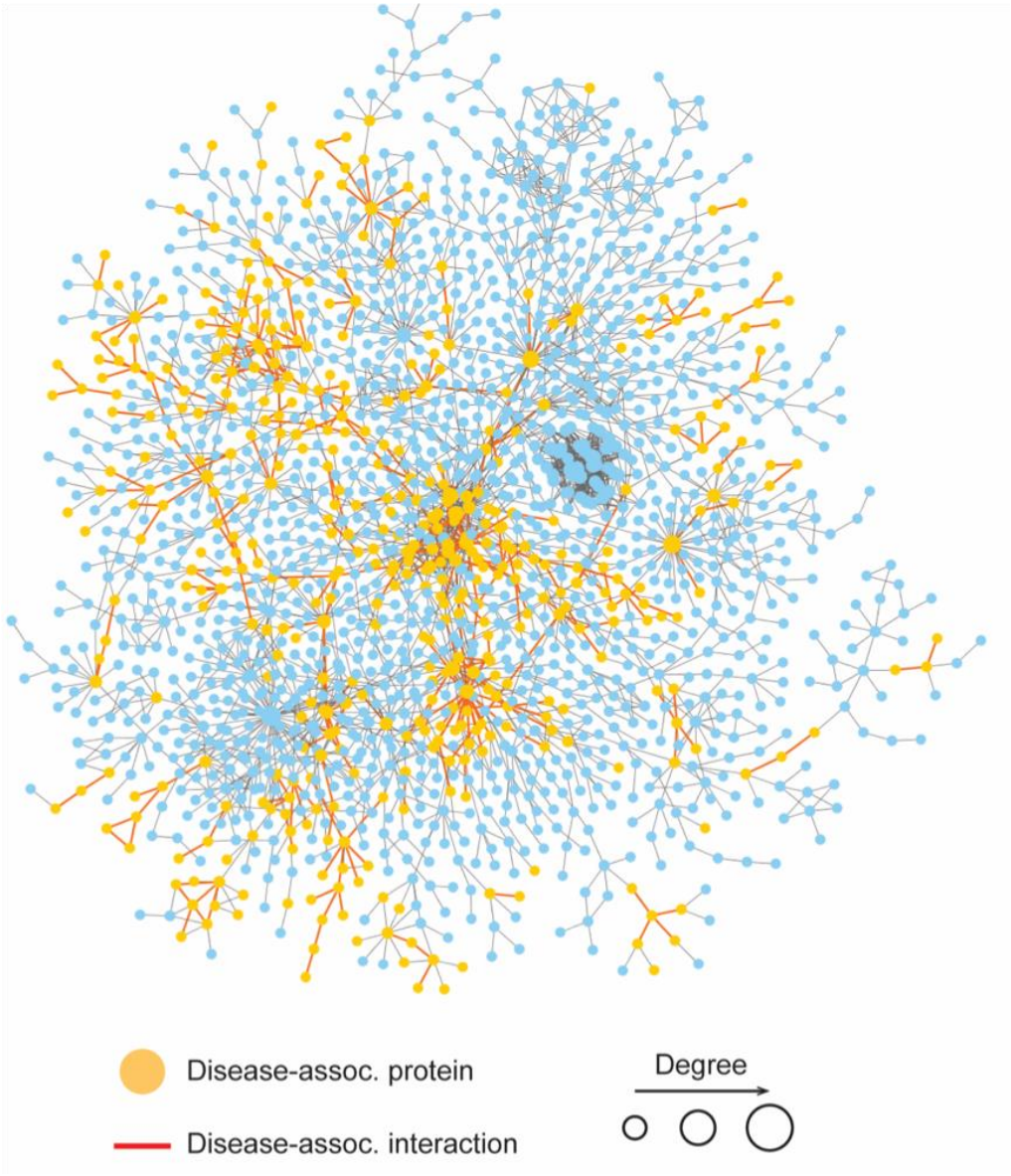


Figure 4-12 Network visualization of two largest connected component of subnetworks targeted by pathogenic mutations and normal mutations. The yellow node is gene carrying pathogenic mutations, and red edge is the interactions target by both pathogenic mutations and normal mutations.

Then we put two groups of disrupted interactions caused by nsSNPs in the context of human interactome and examine their topological properties. We first measure the centrality of the interactions rewired by two groups of mutations. We first exclude the 736 shared interactions. We consider two main network edge centrality measures: shortest-path edge betweenness and flow edge betweenness. In short, shortest-path edge betweenness is the sum of the fraction of all-pairs shortest paths that pass through the

edge. In contrast, flow edge betweenness centrality uses an electrical current model for information spreading. (See methods for more details) Our results showed there is no significant difference between two disrupted interaction sets in terms of both edge centrality measures. To further investigate the network rewiring effect, we first collapsed the two disrupted interaction sets into unified subnetworks. For these two subnetworks, we examined the network efficiency. Briefly, network efficiency is a quantitative measure describing how efficiently the network exchanges information. The results show that the subnetwork disrupted by pathogenic mutations have lower network efficiency with a nearly two-fold difference (0.0117 v.s. 0.0223). Together with our observation about the disrupted interaction edge centrality, we conclude that, compared against normal mutations, pathogenic mutations don't have a strong tendency to disrupt interactions having very high centrality. This can avoid some consequential breakdowns of the interactome. On the other hands, they target at subnetworks with less efficiency. Such decrease of network efficiency might play a role in the physiological processes that cause the disease.

4.2.6 Phenotype associated modules in the human interatome are enriched with disruptive mutations

Modular structure is one of the essential characteristics of biological networks [161]. The identification of these modules is a crucial step in network analysis towards elucidating the biological mechanisms underlying complex phenotypes[15, 16]. To identify biologically relevant modules in the human interactome, we applied two distinct module detection methods: one only relies on the network topological information, with no prior knowledge needed; the other one is seeds based method. For the first one, the method we applied is based on the idea of "Diffusion State Distance"[236], which has been proven as the top performing module detection method in the Module Identification DREAM challenge. Given its best performance in DREAM challenge[237], we expect to get the largest number of discovered modules that are significantly associated with complex traits, including both normal phenotypes and complex diseases. For the second one, we adopt a method named DIAMOnD[166]. DIAMOnD aims to identify the full disease

module around a set of known disease proteins. For more technical details about two module detection methods, see the Methods section.

To discover phenotype associated modules with no prior seed information, we first computed the DSD matrix. The final modules resulted from performing spectral clustering on the DSD matrix. The module size is pre-set to be at least 3 genes and at most 100 genes, as it is much easier to gain biological infight from a medium size module. In total, we collected 1055 possible phenotype associated modules with high confidence. We further investigate whether the PPI disruptions and rewiring inside these phenotypes associated modules caused by disruptive mutations have any role in developing the complex trait. To do so, we randomly generate a set of modules with same number. We first compared the number of rewired edges between the two module sets. Then, the number of the rewired edges in the module is normalized by the total number of nodes in the module. The results showed that phenotype associated modules have significantly more disruptive mutation than random ones (See Fig 4-13). This lends support to our hypothesis that the rewiring caused by disruptive mutations inside phenotype associated modules contributing to the development of certain phenotypes. We went through the module list and found some very interesting cases. For instance, shown in Fig 4-14 is a detected module with 24 genes inside it. After literature search, we confirm that this module is related to pre-mRNA splicing. Some genes in the module, like *PRPF8* (Uniprot: Q6P2Q9), *GEMIN2* (Uniprot: O14893) and *SNRNP200* (Uniprot: O75643), are key components of the spliceosome. A spliceosome is a large and complex molecular machine, which consists of small nuclear RNAs (snRNAs) and approximately 80 proteins. It removes introns from a transcribed pre-mRNA and plays a crucial role plays a crucial role in the central dogma of molecular biology. In this module, two hub-like mutant proteins with a disrupted interaction are *PRPF8* and *SNRNP200*. *PRPF8* forms a scaffold to help in the assembly of the snRNAs and proteins of the complex, and *SNRNP200* mediates the interaction of snRNAs and catalytic activity. Mutations are these genes have been linked to various traits. These traits include both normal phenotypes and disease phenotypes, such as leukocyte count, body mass index, balding measurement and coronary artery disease, etc. Although the full-length nature of these variants has not been determined, our results suggest that the disruptions and rewiring inside the module

caused by disruptive mutations could have significant impacts on complex phenotype development. Moreover, it indicate that the interaction and coordination of a set of phenotypes associated genes within the localized module matter more than their individual functionality.

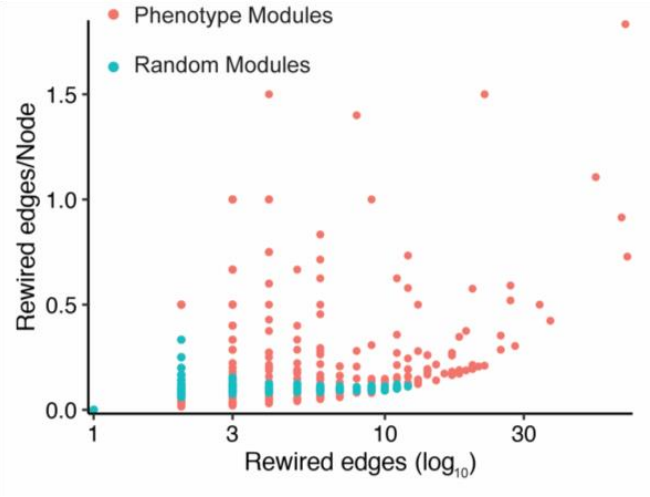


Figure 4-13 Comparison of interaction disruptions accumulated in the phenotype associated modules and random modules. The figure clearly shows that phenotype associated module carries more disruptive mutations than a random one.

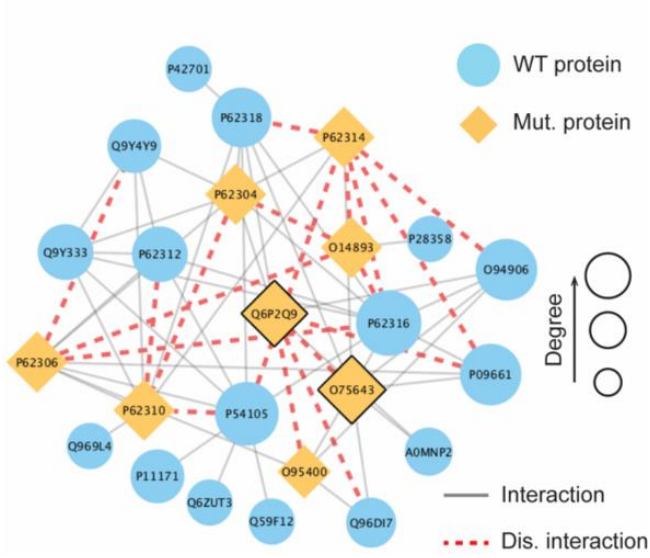


Figure 4-14 An example phenotype associated module in the human interactome detected based on the DSD idea. Shown is a dense PPI module with mutant proteins carrying interaction disrupting variations, indicated by the orange nodes and red dashed edges respectively

We also carried out seed-based module detection on the human interactome. To do so, we first collected disease associated genes as seeds for 70 complex diseases. We shall note that these mutations are from normal populations, and they are most likely benign. So, their disruptions are not expected to cause severe damage to the interactome and further resulting in the disease. However, we suspect that there is some difference in these diseases associated modules caused by disruptive mutations between populations, and such difference can help explain the ethnic disease susceptibility to certain diseases across different populations.

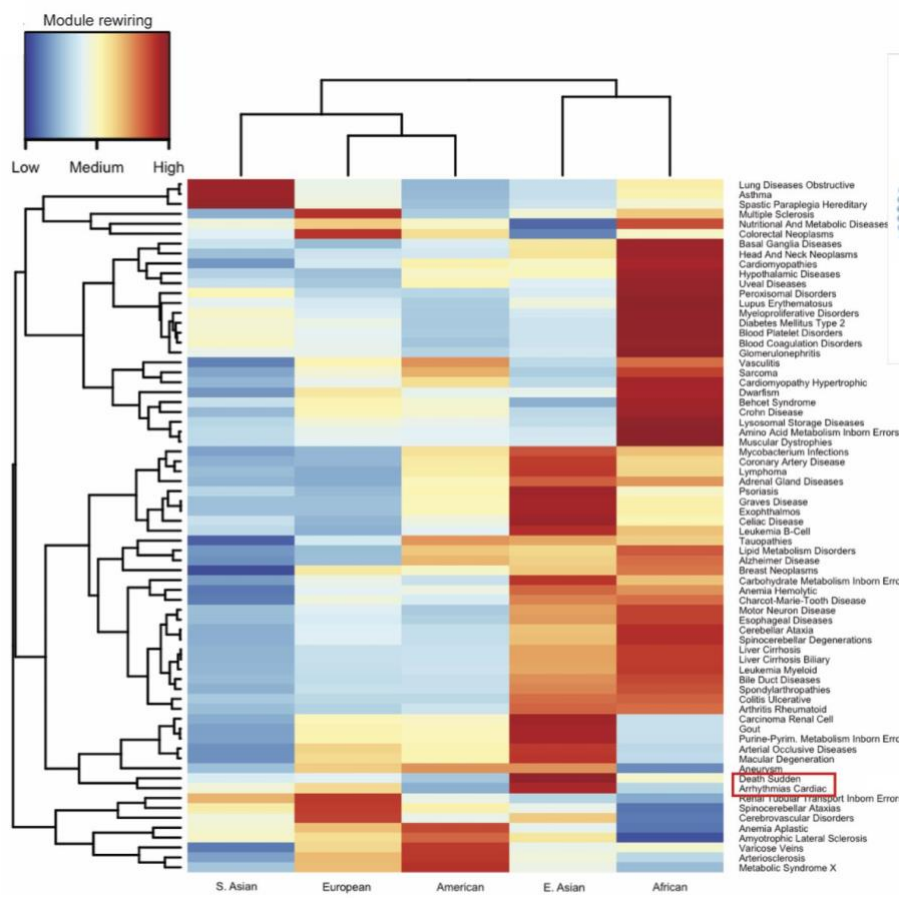


Figure 4-15 A heatmap showing the different prevalence and disruption level caused by the population specific mutation across different populations. The disease-associated modules are represented by rows and the populations are represented by columns. The color of each cell in the heatmap represents the prevalence of rewiring in a specific disease module and a population

We then run the DIAMOND algorithm with the disease associated genes as partial input. We chose default parameters for DIAMOND algorithm. For the 70 resulted disease modules collected from DIAMOND output, we summarize the network aggregated results in the form of a heatmap (Fig. 4-15). In the heatmap, the disease-associated modules are represented by rows and the populations are represented by columns. The colour of each cell in the heatmap represents the prevalence of rewiring in a specific disease module and a population, where red represents a higher occurrence of disrupted interactions in the module. To uncover patterns, we also applied hierarchical clustering to group diseases and populations by the similarity of rewiring prevalence. As shown in the heatmap (Fig 4-15), it clearly suggests that, among five major populations, disruptive mutation and the resulted module rewiring are most prevalent in East Asians and Africans across 70 disease modules.

We further investigate an interesting disease module about cardiac arrhythmia. Cardiac arrhythmia is a group of conditions that cause the heart to beat irregular, too slowly, or too quickly [254]. Arrhythmia affects millions of people; it happens in the normal healthy population as well. There may be no symptoms. Or, symptoms may include a fluttering in the chest, chest pain, or dizziness. Many people don't requirement medical attention or treatment for cardiac arrhythmia. But it can cause severe outcome, such as sudden cardiac death. Based on the results of the heatmap, we see a large difference in rewiring prevalence between the East Asian and American population. Even though cardiac genetic studies in Asian populations are far fewer than in Western populations, a recent review on arrhythmias[255] found that Japanese individuals carry a higher prevalence of long QT syndrome and brugada syndrome that can both increase the risk for sudden cardiac death. Then, we select the disrupted interactions of each of these populations and map them to the module as seen in Fig. 4-16. It shows distinct rewiring patterns in interactome modules associated with arrhythmias in East Asians and Americans. We further focus on a few specific interactions shown in Fig. 4-17. A protein-protein interaction that is important to brugada syndrome and to sudden death, is that between FGF12 (Fibroblast growth factor 12) and SCN5A (sodium voltage-gated channel alpha subunit 5). Fibroblast growth factor homologous factors (FGF11 - FGF14) perform many intracellular functions and are well known to bind to sodium and calcium channels to

modulate cardiac currents, and FGF12 is the major fibroblast growth factor expressed in the human cardiac ventricle. We have predicted the interaction to be disrupted due to mutations carried by FGF12, in which the disruptive variant is more frequent in East Asians than in Americans, denoted by the thick and thin edges respectively in Fig. 4-17. This interaction disruption has been experimentally studied in mice and humans [256]; it showed that such disruption can cause a sodium channel loss-of-function phenotype. More specifically, experiments revealed reduced binding of the mutant FGF12 to SCN5A and resulting reduction in sodium channel current which affects cardiac ventricular action potential. Additionally, SCN5A plays a key role in modulating electrical impulses and their conduction in the heart. SCN5A mutations [257] are implicated in many arrhythmias such as long QT syndrome, brugada syndrome, atrial fibrillation, progressive cardiac conduction defect and sick sinus syndrome. The main function of SCN5A is in the cardiac sodium channels, but also interacts with calmodulin gene products (CALM1, CALM2 and CALM3) that are also associated with predisposition to arrhythmias [258]. Our analysis showed that the disruption (Fig. 4-17) between SCN5A and CALM1 that is more frequent in the East Asian population. Additionally, we predict another interaction of SCN5A to be disrupted, due to a detrimental variation. This interaction is with FGF13 and is also implicated with arrhythmias [259].

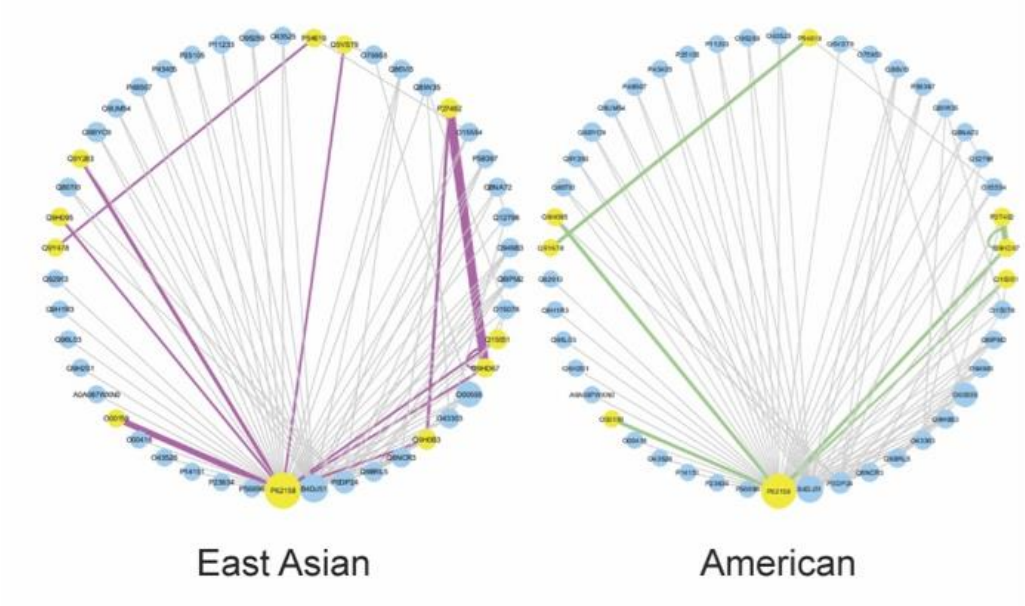


Figure 4-16 Distinct rewiring patterns in interactome modules associated with arrhythmias in East Asians and Americans. E. Asian population has 11 proteins that carry mutations (yellow nodes) resulting in 11 rewired interactions (magenta edges), whereas the American population has 6 mutant proteins (yellow nodes) with 6 rewired interactions (green edges). Additionally, the higher allele frequency of a mutation and the corresponding rewired interaction is represented by the increasing thickness of edges, where we see that E. Asian population has higher frequency of these mutations.

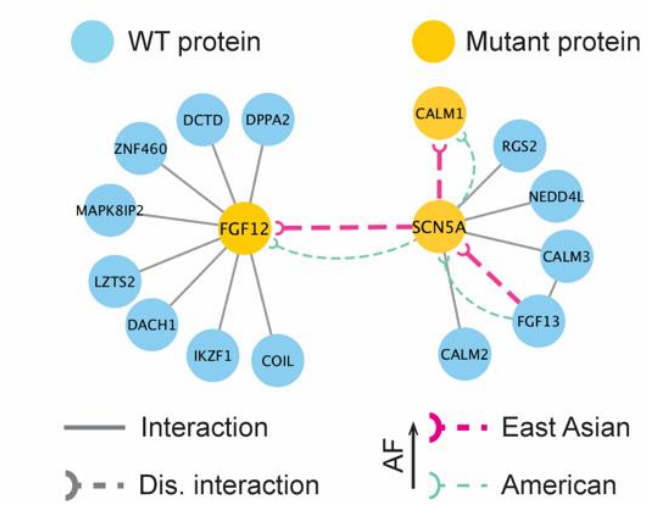


Figure 4-17 Key interactions damaged in the module associated arrhythmias carries disruptive mutations with different prevelence in East Asian and Americans. FGF12 and SCN5A are two important genes implicated in arrhythmias, and disruptions associated with FGF12 and SCN5A are more frequent in East Asians than in Americans

4.3 Discussion

In this work, we create a comprehensive catalog of population-specific edgetic effects at the whole-interactome level. Our work leverages a recently developed novel machine learning approach, SNP-IN tool, that determines interaction-rewiring effects of non-synonymous single nucleotide variants (nsSNVs). The method has been applied to variants associated with diseases in Chapter 3, showing feasibility of a large-scale in-silico edgetic study. To compare the functional impact of population-specific variations, we have applied our approach to annotate an unprecedented set of 46,599 nsSNVs collected from the 1,000 Genomes Project by leveraging the structural information on PPI complexes. The functional impact of a variant on a PPI is annotated as neutral (no change to PPI), disruptive (loss of PPI), or beneficial (increased binding affinity of PPI). We determined that 25,185 SNVs (54%) were predicted as disruptive to at least one PPI, indicating that a significant number of nsSNVs from the normal populations can alter protein function. In fact, this percentage of disruptive mutation is relatively high. In a recent study, Fragoza et al[260] applied an experimental approach (Yeast Two Hybrid) to investigate the impact of 2,009 missense single nucleotide variants (SNVs) from ExAC database[261] across 2,185 protein-protein interactions. They found that about 20% of them can be disruptive to at least one protein-protein. This is a considerable difference between the two approaches. However, we think several things could justify such difference.

First, our computational approach has a much bigger interaction space compared to the experimental protocol. Essentially, the more interactions you include in the experiment to test whether a mutation have impact on them, the more likely that the disruptive impact of such mutation is observed in at least one of them. Second, Yeast Two Hybrid experiment is notoriously known for its high false positive rate [262-264]. Given the nature of their experimental protocol, we suspect such experimental approach will greatly underestimate the disruptive mutation rate. Lastly, we think our SNP-IN can be also biased, because of the stringent requirement of interaction structure. A gene/protein with detrimental impact gets more attention from scientist and carries more significance. On

the other hand, a gene with more significance is more likely to be structurally resolved experimentally [265-267]. Thus, it is very possible that the inputs to the SNP-IN tool is biased towards including more disruptive mutations. What's more, when we prepare the interaction complex structure, our protocol involves partially modelling the interface structure; this excludes mutations outside of the interface in the SNP-IN prediction process. It is also well known that mutations on the interaction interface is more likely to cause disruptions to the interaction [21, 215, 268]. Thus, our SNP-IN tool is likely to overestimate the disruptive mutation rate.

The 'node-centric' gene removal approach is convenient and useful to approximate the disruption of mutated genes. Edgotype focus on specific alterations in distinct molecular interactions and is a more powerful and flexible approach. First, many mutations have been shown to be edgetic [22, 215]. Besides, for a protein, different interactions may not occur independently. So, mutations might accumulate on the same gene at individual genome level; edgotype can better explain the individual phenotypic difference, even though they share might similar genetic makeup. Edgetic network perturbation models have also been proposed and adopted to study complex genetic diseases [21, 22, 146, 215, 269]. Sahni et. al. further suggested a relation between edgetic perturbations and disease severity [22]. More importantly, edgotype provides a plausible explanation for some complicated genetic phenomena, such as locus heterogeneity and gene pleiotropy[21, 270]. In addition, the edgetic perturbation model offered a network-based hypothesis to explain modes of inheritance [146]. We further argue that edgotype can be extended and incorporated with different mutation frequency pattern across populations to study the population genetics. Case studies in this work have shown that this edgotype based approach can better explain the phenotypic variance across populations.

Many results in our work imply that the disruptive mutations and their network rewiring effects can be linked to disease phenotype and disease susceptibility. We summarize our findings here. The implication about disease susceptibility in this study has two-fold meanings: 1) mutations disrupting interactions can potentially contribute to disease onset or progress, and 2) different population-specific rewiring pattern of disease module suggest different disease susceptibility. Two main findings in our work support the former

claim. First, genes enriched with disruptive mutations are associated with various complex genetic diseases. Also, in our comparative network analysis, these normal mutations can target at the same group interactions and causing disruptive impact as the known pathogenic mutations. As to the latter claim, it was implied by two main observations in our network disease module detection and analysis. The first one is that phenotype associated modules are enriched with disruptive mutations compared to random ones. And among different populations, disruptive mutations cause distinct rewiring patterns in disease associated network modules, as shown in Fig 4-15, 4-16.

Chapter 5 DIMSUM: Discovering most IMpacted SUbnetworks in interactoMe

Rapid progress in high-throughput -omics technologies moves us one step closer to the datacalypse in life sciences. In spite of the already generated volumes of data, our knowledge of the molecular mechanisms underlying complex genetic diseases remains limited. Increasing evidence shows that biological networks are essential, albeit not sufficient, for the better understanding of these mechanisms. The identification of disease-specific functional modules in the human interactome can provide a more focused insight into the mechanistic nature of the disease. However, carving a disease network module from the whole interactome is a difficult task.

The need of integrating Genome-wide Association Studies (GWAS) and the functional impact of the disease-associated mutations with the systems data is supported by the increasing body of evidence that large-scale biological systems and cellular networks underlie the majority, if not all, of complex genotype–phenotype relationships in diseases [15, 21, 215, 268]. Understanding the biological network is essential in studying a genetic disease, because such a disease is likely a result of the disruption and rewiring of the complex intracellular molecular network, rather than a dysfunction of a single gene [16]. Along with the increasing availability of the high-throughput human protein interactomics data [103, 132], new computational approaches have been developed. In particular, network propagation has recently emerged as a prominent approach in network biology [133].

In this work, we developed a novel algorithmic framework named DIMSUM (Discovering most IMpacted SUbnetworks in interactoMe) to identify functional disease module in the human interactome. The DIMSUM framework includes three major steps: (1) network annotation, (2) network propagation, and (3) subnetwork extraction. Our approach benefits from integrating GWAS data, functional annotation information, and the protein–protein interaction network. We evaluated our approach using a set of eight complex diseases against two state-of-the-art seed-based module detection methods: DIAMOnD and Seed Connector Algorithm (SCA). From the set of complex diseases, we also carried out two case studies: the first study centered around genes associated with coronary artery disease, and the second one focusing on joint analysis of Schizophrenia and Bipolar Disorder. The evaluations results show that DIMSUM outperforms both DIAMOnD and SCA, because the discovered modules have stronger association with the disease and are more biologically relevant with the seed-gene pool.

5.1 Methods and Materials

At the preprocessing stage, we carried out the human interactome construction, GWAS data collection, and data processing (Figure 5-1). At the same time, we mapped the mutations to the structures of affected PPIs, and applied our SNP-IN tool [24] to characterize mutation-induced rewiring effects on the PPIs [215]. The computational workflow of DIMSUM consists of 3 main stages. In the first stage, we updated the node and edge weights of a fully annotated human PPI network: the node weight reflects the association with a disease, while the edge weight reflects the cumulative damage made to the corresponding interaction. Second, we applied a network propagation strategy with a goal to boost the signal for the genes from the GWAS study with weak association, thus increasing the pool of candidate genes. The last stage includes sub-network extraction: we proposed an iterative procedure to find the disease module with the greatest impact on the disease.

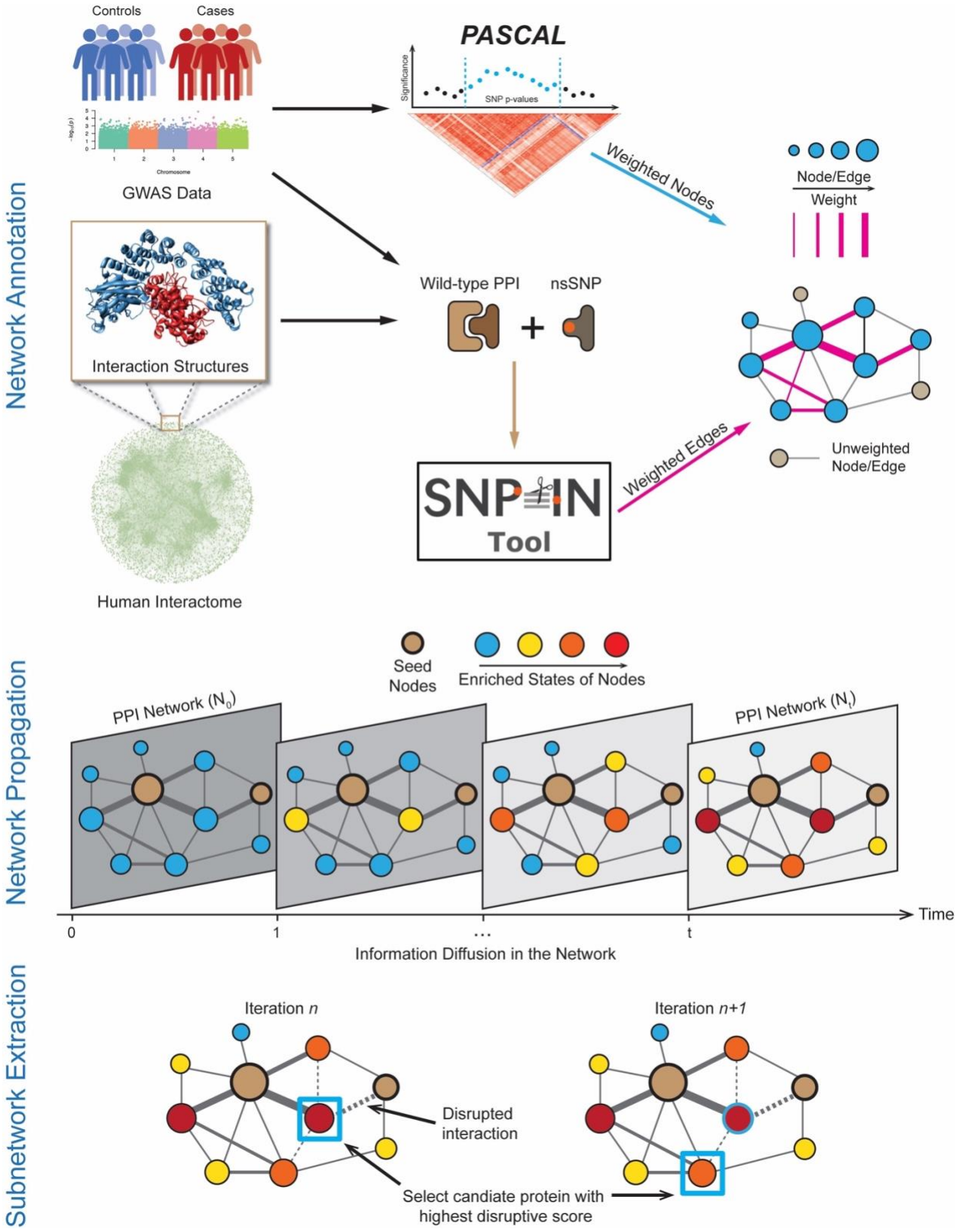

Network Annotation
Controls
Cases
GWAS Data
PASCAL
Significance
SNP p-values
Weighted Nodes
Node/Edge
Weight
Interaction Structures
Wild-type PPI
nsSNP
SNP-IN Tool
Weighted Edges
Unweighted Node/Edge
Human Interactome
Network Propagation
Seed Nodes
Enriched States of Nodes
PPI Network (N_0)
PPI Network (N_t)
Time
0
1
...
t
Information Diffusion in the Network
Subnetwork Extraction
Iteration n
Iteration n+1
Disrupted interaction
Select candiate protein with highest disruptive score

Figure 5-1 Basic workflow of Discovering most IMpacted SUbnetworks in interactoMe (DIMSUM) computational framework. The module detection framework contains three major steps: network annotation; network propagation and subnetworks extraction. In Network Annotation stage, we collect GWAS data and use Pascal tool to aggregate single nucleotide polymorphisms (SNP) summary statistics. Then we apply the non-synonymous SNP INteraction effect predictor tool (SNP-IN) tool to properly assess SNP's impact on the protein–protein interactions. Finally, we integrate this information with the human interactome to generate a fully weighted network. In Network Propagation stage, we apply a network propagation procedure to implicate genes most likely to be influenced by disruptive SNPs. Finally, in Subnetwork Extraction stage, we apply an iterative strategy to identify the most impacted module.

5.1.1 Human Interactome Construction

In this work, we utilized two different protein–protein interaction data sources to construct the human interactome. The first data source is the High-quality INTeractomes (HINT) database [175]. It integrates several databases and filters out low-quality and erroneous interactions. The other source is the Human Reference Protein Interactome Mapping Project (HuRI) [132]. HINT is a manually curated repository of PPIs mainly from the literature, whereas HuRI is a primary source of experimentally validated PPIs. We considered these two PPI sources as they complement each other, hence we constructed the interactome by combining both. The current release of HINT interactome (Version 4) contains 63,684 interactions. Combining all the three proteome-scale human PPI datasets released from HuRI at different stages of the project, we obtained 76,537 interactions. In total, we generated a human interactome consisting of 105,087 interactions, where 35,134 interactions existed in both data sources.

5.1.2 GWAS Data Collection and Processing

A distinctive feature of our method was that it integrated GWAS data into the interactome to improve the disease module detection. To do so, we compiled eight publicly available GWAS datasets. The datasets spanned a broad range of diseases, including neurodegenerative, metabolic, and psychiatric disorders. For the GWAS integration, the dataset was only required to have pre-calculated summary statistics, no individual level information was needed. To generate the seeds for the later stage of network propagation, we computed gene scores by aggregating SNP p-values from GWAS studies using the

Pascal tool (Pathway scoring algorithm) [271]. Integrating SNP p-values from GWAS studies has proven itself as a powerful method to improve statistical power.

Pascal is a fast and rigorous computational tool developed to aggregate SNP summary statistics into gene scores with the high power, while absolving the need to access the original, individual-level, genotypic data. It can be considered as an alternative to the traditional p-value estimation approaches, including chi-squared statistics (SOCS) and the maximum of chi-squared statistics (MOCS) [272], which measure the average and the strongest associations of signals per gene, respectively. Pascal relies on the assumption that a pairwise correlation matrix of the contributing genotypes underlying the null distributions of the MOCS and SOCS statistics can be estimated from the ethnicity-matched, publicly available genotypic data. To calculate gene scores, we selected the sum of chi-squared statistics (SOCS) of all SNPs from genes of interest. To properly correct for linkage disequilibrium (LD) correlation structure in GWAS data, we used the European population of the 1000 Genomes Project [273], because GWAS studies in this work are predominantly the European cohorts. The disease-associated genes with significant p-values that Pascal produces were then defined as the seed genes for our approach. The final list of seed genes was acquired after correcting for multiple testing using Bonferroni correction.

5.1.3 Functional Annotation of nsSNV with the SNP-IN Tool

To properly assess the functional damage caused by mutations with respect to protein–protein interactions, we applied our recently developed SNP-IN tool (non-synonymous SNP INteraction effect predictor tool) [24]. The SNP-IN tool predicts the rewiring effects of nsSNVs on PPIs, given the interaction's experimental structure or accurate comparative model. More specifically, the SNP-IN tool formulates this task as a classification problem. There are three classes of edgetic effects predicted by the SNP-IN tool: beneficial, neutral, and detrimental. The effects are assigned based on the difference between the binding free energies of the mutant and wild-type complexes ($\Delta\Delta G$). Specifically, $\Delta\Delta G = \Delta\Delta G_{mt} - \Delta\Delta G_{wt}$, where $\Delta\Delta G_{mt}$ and $\Delta\Delta G_{wt}$ are the mutant and wild-type binding free energies correspondingly. The beneficial, neutral, or detrimental types

of mutations are then determined by applying two previously established thresholds to ΔΔG [179, 274]. The annotation workflow begins with processing the GWAS data. Most mutations in GWAS datasets come with only dbSNP RefSNP cluster ID's (rs#). The variant data is preprocessed using ANNOVAR [59] to retrieve SNV locations on the genes and the corresponding residue change information. For each mutated gene and the corresponding SNP, we collected all their interacting partners in the merged interactome (see Section 5.2.1). Because the SNP-IN tool is a structure-based classifier, we needed both the mutation information and interaction structure as an input. There are several different cases when generating the PPI structure in which a mutated gene is involved (Supplementary Figure S5-1). First, if a PPI already has a native structure, it is extracted from a protein data bank (PDB) [217]. In the corresponding PDB file, we first identified the interacting subunit pair for each PPI using the 3did database [275]. The 3did database maintains information regarding the two interacting domains with physical interfaces. If there was no native structure for a protein–protein interaction, two options were explored. First, if a structural template for such interaction (i.e., a homologous protein interaction complex) existed, a comparative model of this interaction could be obtained [276]. When a full-length PPI could not be modeled, we only modeled the domain–domain interaction that included the domain containing the mutation. Homology modeling was done through Interactome3D [181], a web service for structural modeling of PPI network.

5.1.4 Network Annotation and Network Propagation

The human interactome is next represented as the graph G = (V, E), where V is the set of nodes representing the genes and E is the set of edges representing the protein–protein interactions. After applying the Pascal tool to the GWAS studies, we obtained p-values of disease associated genes. These values, together with the functional annotations of the corresponding mutations from the SNP-IN tool, were used to weight the nodes and edges in the network. Thus, G is a graph with weighted nodes and edges. Specifically, for a disease-associated gene i, the corresponding node was weighted with $-\log(P_i)$, where P_i is the p-value obtained from Pascal tool. A node corresponding to a gene that is not listed in the GWAS study was assigned a zero weight. The edge was weighted according to the

damage accumulated on the corresponding PPI by the disruptive SNVs. Specifically, a PPI between genes i and j was weighted with the total number of disruptive SNVs on both genes targeting the same interaction. In other words, the node weight reflected the "relevance" of the gene to the disease, and the edge weight reflected the accumulated "damage" on the interaction.

Next, we applied the network propagation strategy to implicate other core disease genes affected by the perturbations of disruptive SNVs. Let F: $V \rightarrow \Re$ represent a function reflecting the relevance of gene i to a specific complex disease. The goal of the network propagation was to prioritize the genes that were not showing significant association based on the GWAS study, but were expected to have possible relevance to the disease. We imposed two constraints on the prioritization function F: the computed function should be (1) smooth and (2) compliant with the prior knowledge. Smoothness of the function was defined by the assignment of similar values to the interaction partners (nodes) of disease gene i. The compliance with the prior knowledge implies that the difference between the final computed value for a disease gene F(i) and the initial value is minor. The values of F can be iteratively obtained as follows:

$$F_{t+1} = \alpha W' F_t + (1 - \alpha) Y$$

where Y is the prior knowledge defined as the node weight, α is a parameter reflecting the importance of the two constraints mentioned above with the default value α = 0.5, W' is a |V|*|V| matrix whose values are determined by the edge weights and that is defined as a normalized form of network edge weight matrix W. Formally, we introduced a diagonal matrix D, such that D(i, i) was the sum of row i of W. We then set $W' = D^{-1/2}WD^{-1/2}$. F_t was initialized as Y. The equation could be solved iteratively and guaranteed to converge to the system's solution [134]. This iterative algorithm could be considered as propagating the prior information from some nodes through the network. The disease genes first sent the signal to their neighbors, and every node then propagated the received signal to its neighbors (Figure 5-1). There were also additional constraints on the weighted network. First, when we propagated the information through the network, both the node and edge

weights should have been in (0,1) range [134]. In this work, the seed genes in the network carried the largest weights. Thus, we normalized the node weight:

$$p_i' = \frac{p_i - p_{min}}{p_{max} - p_{min}},$$

where p_{max} and p_{min} are the maximum and minimum of the initial node weights, i.e., $-\log(P_i)$. After network propagation, we de-normalized the node weight at the subnetwork extraction step (see Section 5.2.5). For the edge weight, the higher weight meant that the information flow was more likely to go through that edge. Thus, the weight was converted to (0,1) range using a sigmoid transformation:

$$w_{i,j}{}' = \frac{1}{1 + e^{w_{i,j}}},$$

where the $w_{i,j}$ is the original weight of the edge. After convergence, the disease association information from seed genes was diffused into the interactome, and all the nodes were weighted with their relevance to the disease.

5.1.5 Sub-Network Extraction

Finally, we extracted the sub-network with the greatest "impact" on the disease progress. To do so we defined disease-associated genes with significant p-values obtained from Pascal as the seed genes. The goal was to extract a sub-network containing all the seed genes while maintaining the greatest impact. Intuitively, the impact was defined based on the "severity" of the network damage caused by the disruptive SNVs located on the genes with high relevance to the disease. The relevance was reflected by the node weight after the network propagation procedure, while the network damage was determined based on the edge weights reflecting the total number of disruptive mutations occurring in each interaction. The subnetwork extraction was then formulated as an iterative procedure:

I. Assume that a seed gene set $g_1, g_2, \ldots, g_k$ induces a subnetwork with initial size N_0. For all immediate interacting partners of the seed gene set that are not in the subnetwork, define an impact score:

$$S(i) = p * \sum_j w_{i,j}$$

where p is the updated p-value after the network propagation and $\sum_j w_{i,j}$ is the total number of mutations that disrupt the PPIs between the candidate gene i and every gene j in the module gene set. Thus, the impact score combines the disease relevance and potential disruption to existing subnetwork caused by the candidate gene.

II. All the immediate interaction neighbors of the seed gene set, not included in the subnetwork, are ranked according to their impact score.

III. Select the gene with the largest impact score. If there are multiple candidates with the same impact score, randomly pick one gene to break the tie. Add the gene with the biggest impact to the set of seed genes and increase the size of the induced subnetwork by 1: $N_{t+1} = N_t + 1$.

Given a number of maximum iterations, steps I–III were repeated in a loop until the number of added genes equals maximum iteration number.

5.1.6 Validation and GO Analysis

To validate the performance of our module detection method, we compared it against two seed-based module detection methods, a widely used method called DIAMOnD [166], and a recently published method called SCA [277] For both methods, we used the same seed genes that were used for DIMSUM. For each method, we also limited the number of candidate genes forming a disease module to 100. To validate the disease association of the predicted candidate genes, we first compiled a list of known disease-related genes from two databases, Human Gene Mutation Database (HGMD) [49] and Online

Mendelian Inheritance in Man (OMIM) [241]. The candidate genes supported by the literature were considered as true positives.

Furthermore, to compare the biological relevance of candidate genes to the seed genes we performed Gene Ontology (GO) enrichment analysis [278] on the two gene sets. In the GO enrichment analysis, we used the third level of the GO hierarchy as a trade-off between the too general and well-populated GO terms at the second level, and specific but not well-populated terms at the fourth level. The GO enrichment was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [37], and GO terms with a p-value ≤ 0.01 were selected. During the analysis, we first identified significantly enriched GO terms within the seed genes. We then checked how many of the GO terms significantly enriched in the pool of candidate genes were identical to the enriched GO terms in the pool of seed genes. The higher number of the identical GO terms suggests the stronger relevance between the candidate genes and the seed genes.

5.2 Results

5.2.1. Seed Genes Generated from GWAS Datasets

We collected eight GWAS datasets from various public sources. The collected GWAS data cover a diverse range of eight complex diseases, including Alzheimer's, bipolar disorder, coronary artery disease, macular degeneration, osteoporosis, rheumatoid arthritis, schizophrenia, and type 2 diabetes mellitus (Supplementary Table S5-1). The GWAS datasets were only required to contain SNP-phenotype association summary statistics, no individual level genotype information was used. These studies were predominantly from European cohorts, with the total number of SNPs reported in each study varying from 2 million to 12 million (Supplementary Table S5-1).

Our subnetwork detection strategy benefited from integrating the GWAS dataset and network data. Specifically, we derived the seed genes for network propagation from the GWAS dataset (see Section 5.2.2 in Methods). The length of the gene list generated from

Pascal varied for each disease due to different sizes of GWAS datasets. The number of seeds for each disease ranged from 26 to 301 (Supplementary Table S5-2).

5.2.2. Functional Predictions from the SNP-IN Tool

The lack of functional knowledge for SNPs obtained from GWAS studies limits our understanding of the mechanistic processes that underlie diseases. Although there is a plethora of functional annotation tools for SNPs, most of them provide with annotations of generic putative deleterious effects of SNPs [6]. In particular, they do not provide the means to determine how SNPs disrupt protein–protein interactions, while such information could lead to a better understanding of how SNPs rewire the human interactome and help identify the impacted subnetworks responsible for the disease. Our recently developed SNP-IN tool accurately predicts how mutations affect the PPIs, given the interaction's structure (see Section 5.2.3 in Methods). Given that the sizes of GWAS datasets considered in this work varied, the prediction coverage of non-synonymous SNPs also varied for different diseases, ranging from 547 to 8323 (Supplementary Table S5-3). Previous studies reported high percentage of disease-associated mutations that affected PPIs [22, 215, 268]. Specifically, our latest study [215] showed that out of all pathogenic mutations collected from the ClinVar database, 76.2% were predicted to have a disruptive effect on PPIs. In the present work, on average, 51.1% of the annotated mutations from all eight GWAS studies were predicted to have detrimental effects on PPIs. The percentages of disruptive mutations in this study were lower than our previous work, which could be attributed to the fact that some of mutations detected in the GWAS studies were random mutations or passenger mutations without significant functional impact. Nevertheless, the current study showed that a considerable amount of mutations occurring in a disease could rewire the human interactome. In addition, these results reaffirmed our previous findings that the beneficial mutations, strengthening the PPIs, were rare in the human genome. The reported beneficial mutations are less than 1% percent in all cases (Supplementary Table S5-3). Given such a low percentage, we discard this beneficial mutation during network propagation.

5.2.3. Network Annotation and Network Propagation

A key idea of our approach is in integrating GWAS data and functional annotation data with the human interactome data to improve the network module detection. Specifically, we utilized the GWAS study results and functional annotations from the SNP-IN tool to properly weight the network. The node weight reflected the relevance of the gene to the disease, while the edge weight represented the cumulative damage imposed on the corresponding interaction (see Section 5.2.4 in Methods). Once the fully weighted network was generated, we examined the topological properties of the genes carrying disruptive mutations and the damaged interactions in the human interactome. In particular, between the eight datasets we compared the distribution of the node degrees in the human interactome for disease-associated genes. The average node degree for the genes carrying disruptive mutations among the eight diseases ranged from 14 to 24 (Figure 5-2). Compared to the average node degree of the human interactome, all disease networks showed increased average degree suggesting that disease-associated mutations tend to disrupt genes occupying a central spot in the human interactome, rather than lying on the periphery. The results also suggested that mutations captured in complex diseases were more likely to cause the network rewiring than random mutations.

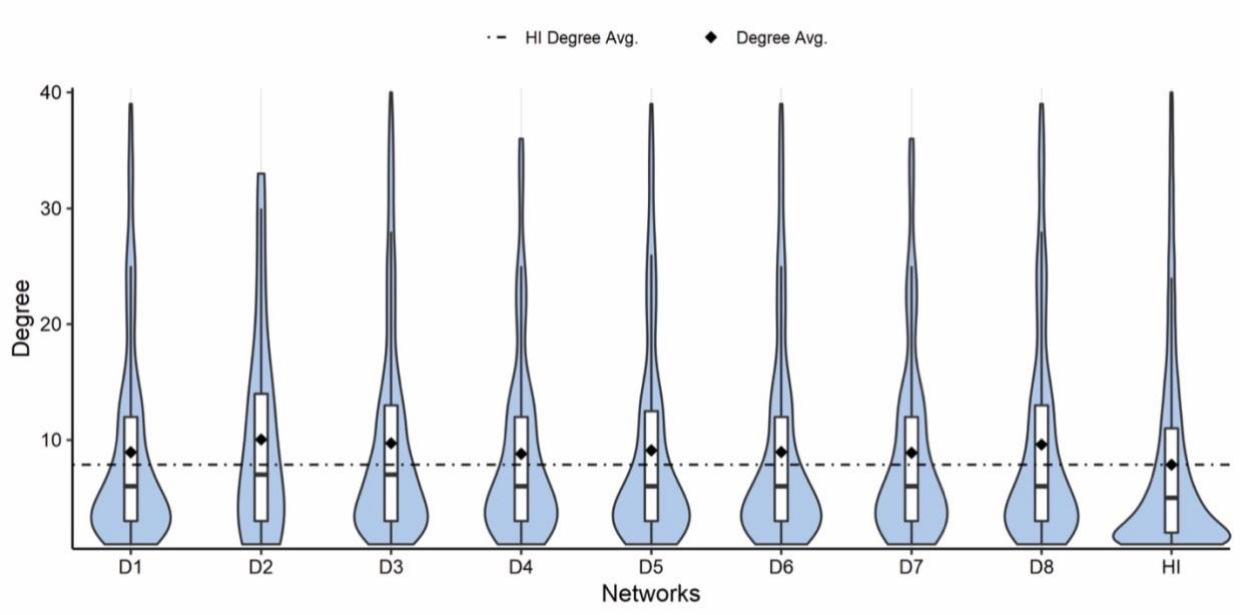


Figure 5-2 Comparison of the node degree of the disrupted genes with the avg. degree of HI. The first eight violin plots represent the node degree distributions of disruptively mutated genes for eight complex diseases; the last violin plot is the node degree distribution of all genes in the human interactome (HI); the avg. degree of the disrupted genes is much greater than the avg. degree of HI, showing that highly connected genes are disrupted.

Following annotation of the interactome with GWAS and functional data, we adopted a network propagation strategy to implicate protein interactions most likely to be influenced by disruptive SNVs and proposed a novel subnetwork extraction algorithm to find the mutation-specific module with the most "impact". To determine if our protocol benefited from integrating GWAS data and functional annotation into the interactome, we compared our protocol against a naïve network propagation solution on the basic human interactome, i.e., without integration GWAS or functional annotation data [134, 279]. We also compared it against another network propagation strategy with only GWAS data integrated. We used three selection ratios, 25%, 50%, and 75%, to randomly pick the number of seeds and compared the numbers of remaining seeds that could be rediscovered after applying either DIMSUM or naïve propagation. The edges were assigned with the same weight value 1 for each edge for both the naïve and GWAS based propagation approaches. After propagation, we selected genes with the highest node score to add to the module (Figure 5-3 and Supplementary Figure S5-2). The results demonstrated that our protocol had a substantially higher fraction of discovered seed genes compared to the naïve and the GWAS based network propagation strategy. As an

additional experiment, we used the entire set of seeds for the network propagation in both DIMSUM and naïve approach, and then examined the top 100 discovered genes. When checking the overlap between these two gene sets, we found that the set of overlapping genes consisted of six genes on average across eight diseases. The results suggested that our method emphasized the genes with the greatest functional impact on the interactome and were not driven exclusively by the information propagated from the seeds. In other words, the genes extracted in the last step of our method indicated strong association with the disease and also reflected the severe damage caused by the mutations.

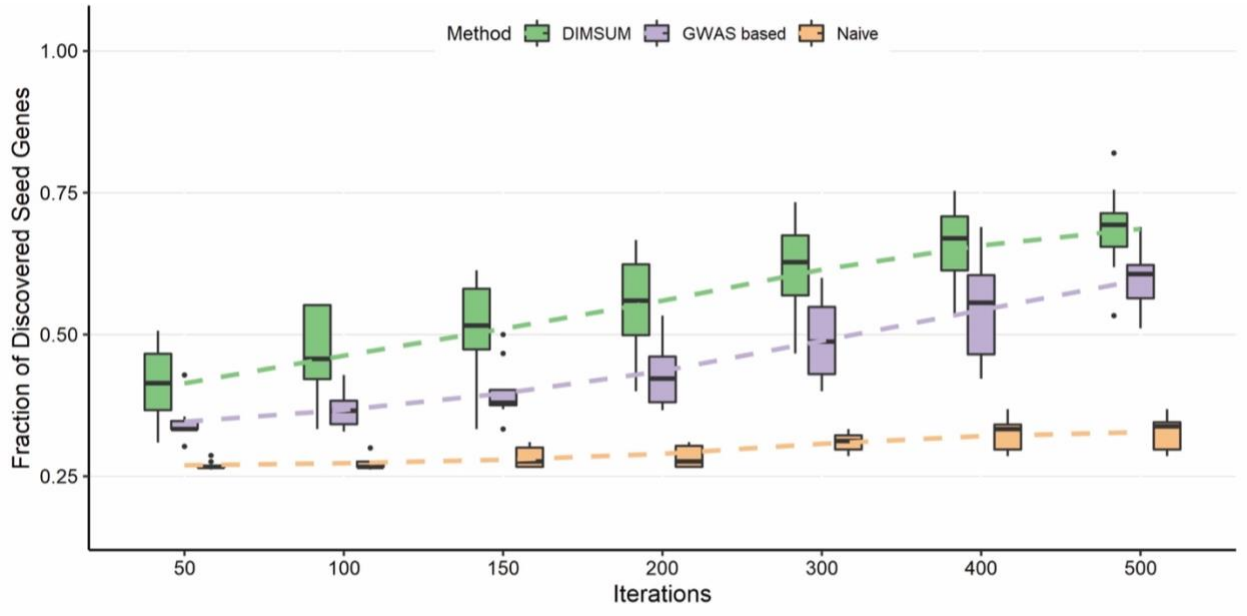


Figure 5-3 Comparison of the seed genes discovered when randomly selecting 25% of the seed gene pool as seeds. Each box plot represents the fraction of discovered seed genes across all eight diseases from DIMSUM, a GWAS based and a naïve network propagation at different iterations. DIMSUM performs significantly better than the other two approaches at the initial 50 iterations and improves drastically with increasing iterations.

5.2.4. Comparison Against DIAMOnD and SCA

To validate the performance of our methods, we compared our method against two seeds-based module detection methods, DIAMOnD [166] and SCA [277]. DIseAse MOdule Detection (DIAMOnD) is one of the most popular methods for module detections. It was developed based on the observation that the connectivity significance is a more predictive quantity characterizing the module's interaction patterns, rather than connection density. The core idea of the algorithm was that, given a set of disease genes as seeds, it ranked all the candidates connected to the seeds based on their connectivity significance and added them to the existing seed set. Seed Connector Algorithm (SCA) is a recently developed seeds-based module detection method. SCA was built on the idea of seed connectors, which served as "bridges" of different network branches that were induced by seed genes. It selected a gene that maximally increased the size of the largest connected component of the subnetworks as the seed connector, and added them to the existing module.

We compiled a list of known disease-related genes for eight GWAS datasets as our benchmark (Supplementary Table S5-4). We then manually checked which of the added genes were supported with the literature evidence. We found that the lists of seed genes across eight diseases had between five and 29 of the disease genes supported by literature (Table 5-1). We next checked whether DIMSUM outperformed the other two methods in terms of the prediction accuracy in discovering the genes with supporting evidence of known disease association from literature in the added gene set. We observed that our method outperformed both DIAMOnD and SCA in all eight cases but one (osteoporosis), where DIMSUM did not find a match while both DIAMOnD and SCA found a single gene match. In fact, the predictions from DIAMOnD and SCA barely found any literature-supported disease genes (Table 5-1). The results also showed that except for the genes with very strong statistical signals from GWAS studies, implicating other disease genes through a network-based approach is a challenging task. As to the agreement between the methods, the overlap of the resulted modules between every two methods is very low (Supplementary Table S5-5). This suggests that the algorithm design determines the way the disease module grows, and different algorithm favor different genes.

Table 5-1 Comparison of the number of disease associated genes in the detected modules with literature evidence between three methods: DIAMOnD, SCA and DIMSUM.

Disease ID	Disease Name	Seeds Match	DIAMOnD Match	SCA Match	DIMSUM Match
D1	Coronary artery disease	8	1	0	10
D2	Diabetes mellitus, Type 2	14	0	1	2
D3	Macular degeneration	17	0	0	3
D4	Osteoporosis	5	1	1	0
D5	Alzheimer's disease	8	0	0	3
D6	Rheumatoid arthritis	19	0	0	4
D7	Bipolar disorder	12	0	0	8
D8	Schizophrenia	29	0	0	14

We next carried out GO enrichment analysis on each gene set using all three categories of GO terms at the third level of GO hierarchy. GO annotation was then used to find how many genes from the newly obtained module shared the same GO terms with the seed genes. The results (Figure 5-4) showed that DIMSUM on average yielded a higher number of GO terms compared to both DIAMOnD and SCA. The identical GO term number is further normalized with the total number of enriched terms (Supplementary Table S5-6). Although DIMSUM has the most enriched GO terms above the threshold, the normalized ratio is still higher than the rest two. So, taking the total number of enriched terms into account does not negate our conclusion; on the contrary, it shows that DIMSUM extracts groups of genes that are more functionally coherent. However, our method did not always have the highest number of shared GO terms for an individual disease. Interestingly, our analysis also showed that the dominant GO terms for genes extracted by all three methods fell in the Biological Process category.

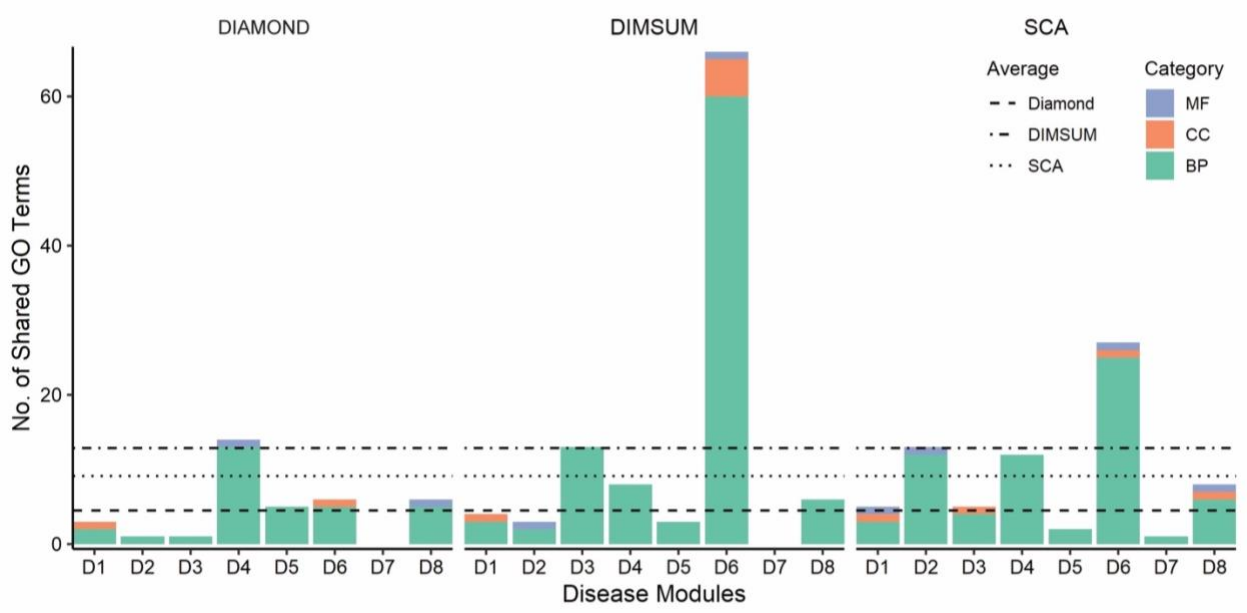


Figure 5-4 Comparison of biological relevance of disease modules detected using three methods. Bar graphs representing the number of the GO terms enriched in the added genes overlapped with the seed genes for DIAMOnD, SCA, and DIMSUM for the eight disease modules. The three GO term categories are Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). The average of significant and identical GO terms from DIMSUM is higher than from DIAMOnD or SCA.

Next, we examined the topological properties of the modules generated from the three methods. To quantify the structural difference of the modules generated from three methods, we focused on two topological properties. First, we calculated the connection density of the disease modules. Previous studies [166] showed that the connection density was not the primary quantity to characterize the connection patterns among disease proteins. It was further argued that, in biological networks, the paths through low-degree nodes bore stronger indications of functional similarity than the paths that went through the high-degree nodes, or hubs [280]. These findings suggested a good strategy for a module detection method should reduce the density of the detected modules and mitigate the influence of the hubs in the human interactome. When comparing these three methods, we found that DIAMOnD had the highest connection density in the detected networks, whereas modules generated from SCA and DIMSUM had much lower density (Figure 5-5 and Supplementary Figure S5-3). We also observed that SCA favored the genes with extremely high degrees, as indicated by the genes in the long tails (Fig 5-5). The low density of the modules and node degree distributions from DIMSUM suggested

that our method was not biased towards interaction hubs, and thus was expected to extract genes with similar functions from the rest of the interactome more efficiently than DIAMOnD.

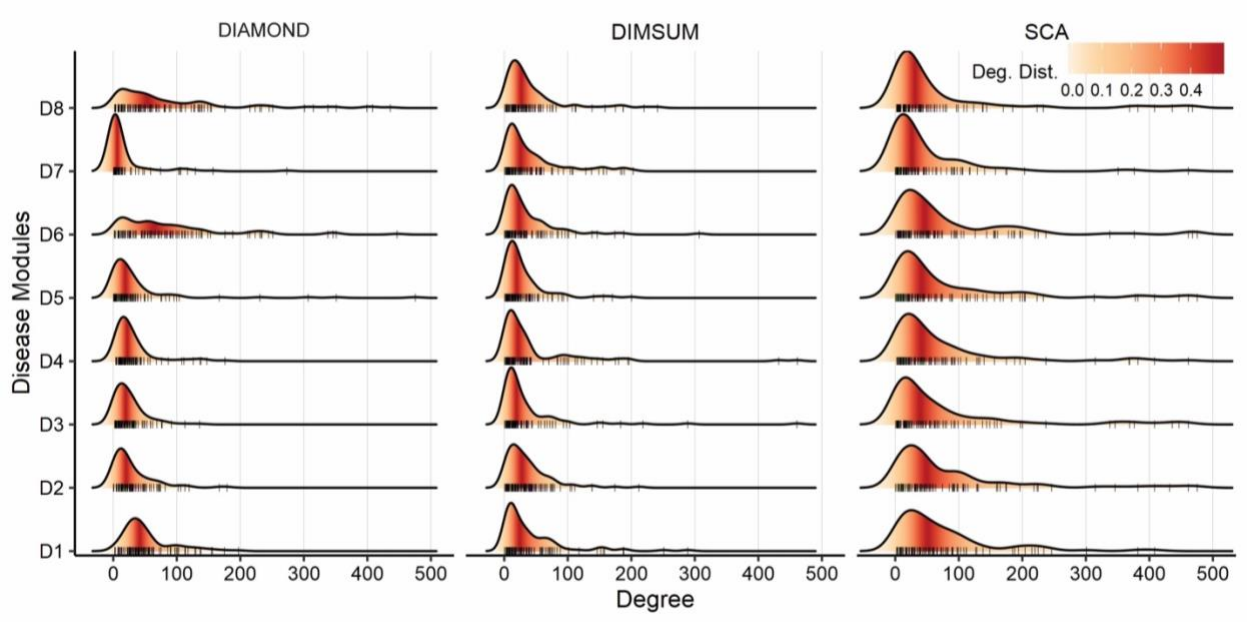


Figure 5-5 Comparison of topological properties of disease modules detected using three methods. Degree distribution of the added genes during module detection for the eight diseases by each method. When building a module, DIMSUM avoids bias towards always including the nodes of high degree and has lower node degrees for all eight modules.

Another interesting distinction between the methods was the fact that DIAMOnD and SCA tended to grow a single giant globular component, whereas our method typically built a major component accompanied with a set of smaller, "satellite", components. In particular, we found some of the disease-associated seed genes occurring in the small satellite subnetworks from the modules obtained by DIMSUM, but not DIAMOnD or SCA. This observation suggests that the smaller subnetworks play equally important role in defining the disease phenotype.

5.2.5. Case Study 1: Coronary Artery Disease

As the first case study, we considered an application of DIMSUM to extract a network module centered around coronary artery disease (CAD) and compared the module to those ones derived by SCA and DIAMOnD using the same set of seed genes (Figure 5-6,

5-7, and Supplementary Figure S5-3). The CAD GWAS dataset was obtained from the CARDIoGRAMplusC4D Consortium [276]. After pre-processing and applying the Pascal tool, we were able to curate a seed gene pool consisting of 37 genes, 24 of which could be mapped to the human interactome. The seed genes were spread across the entire interactome, with very few direct interactions between each other. For each of the three methods, we extracted 100 genes in addition to the original seed gene set to form a functional module. We first validated the obtained new genes from each of the three modules against known CAD-associated genes that were collected from literature. DIMSUM outperformed both DIAMOnD and SCA (Table 5-1): DIAMOnD had only one and SCA reported no known CAD-related genes.

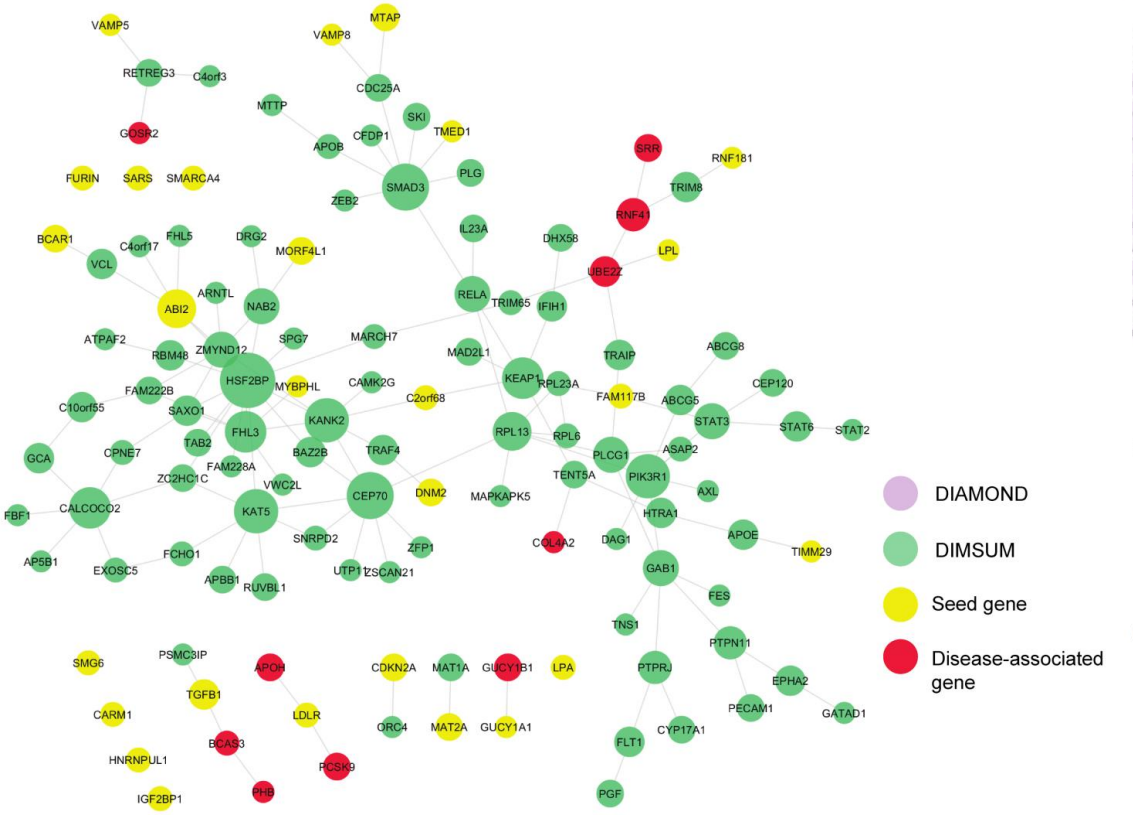


Figure 5-6 Largest connected component and satellite components detected by DIMSUM. Yellow nodes represent the seed genes, and red nodes represent disease-associated genes that are supported by literature

The CAD disease module from DIAMOnD formed a clique-like structure (Figure 5-7). There were dense connections inside the largest connected component, the property

typically not observed in a functional module [166]. Besides, the largest component originated from only seven seed genes. During the later stage of the DIAMOnD algorithm, extraction of additional genes to form the disease module was determined by several genes, including FAM209A, STX1A, and CREB3L1, which were not seeds and which were added in the early steps of the method run, rather than from the initial seed gene pool. We surveyed the literature and did not find a strong link between these non-seed genes and CAD. We also observed that the rest of the seed genes became isolated and separated from the largest component. On the contrary, SCA generated a globular structure for the disease module, in which the largest connected component (LCC) includes most of the seed genes. This is not surprising, as SCA is specifically designed to add a "seeds connector" to grow the LCC maximally, rather than the genes with functional importance. The addition of the seeds connector, therefore, was biased towards the hubs in the interactome. SCA tended to add many genes with high values of node degrees, as indicated by the long tails of the degree distributions. This phenomenon was not only demonstrated in the case of CAD module, but was also evident for the other diseases we studied (Figure 5-5). However, recent work revealed that proteins connected to the high-degree hubs were less likely to have similar functions, compared to the proteins that interacted with a protein of significantly lower node degree [165].

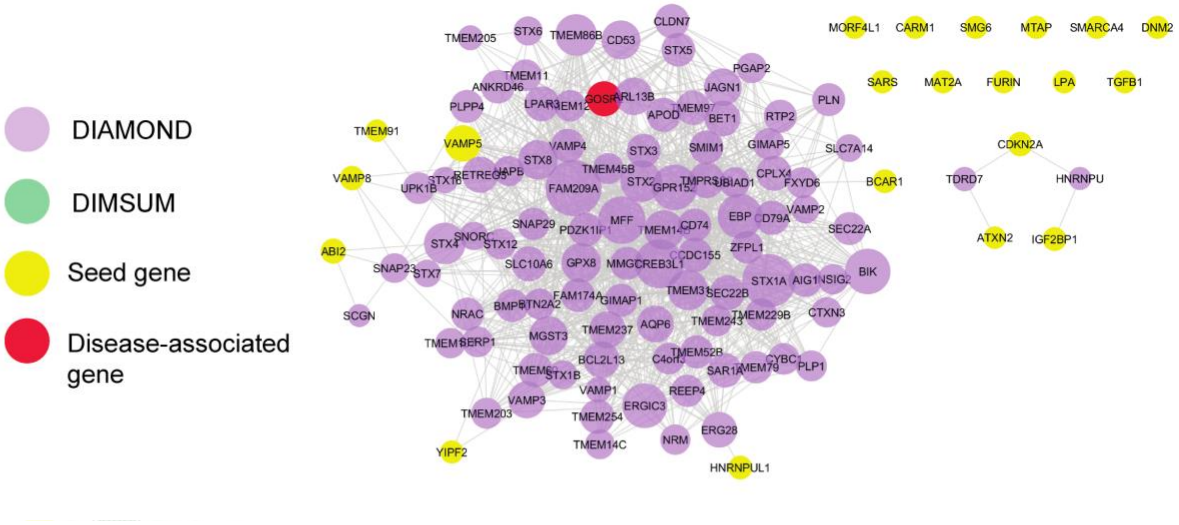


Figure 5-7 Large and high-density module detected by DIAMOnD. DIMSUM identifies ten CAD associated genes, whereas DIAMOnD identifies only one.

Finally, when examining the disease module generated from DIMSUM we found that DIMSUM module included most of the seed CAD-related genes (Table 5-1 and Figure 5-6). In addition, there was a core component in the module set which was topologically different compared to the core components generated by DIAMOnD and SCA: it did not form a highly dense clique and it was not biased toward the hubs with very high node degrees (Figure 5-7). In addition to the core component, the DIMSUM functional module included small satellite subnetworks that harbored several functionally important genes known to be associated with CAD but not reported in the seed gene pool. Three of these satellite modules contained five genes associated with CAD, namely: PHB, BCAS3, GOSR2, APOH, and PCSK9.

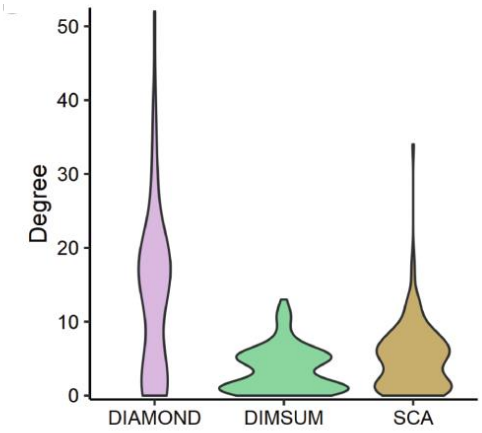


Figure 5-8 Degree distribution of the modules generated by each method shows DIMSUM does not tend to grow a highly dense clique and it is not biased toward the hubs with very high node degrees

5.2.6. Case Study 2: Schizophrenia and Bipolar Disorder

In the second case study, we use DIMSUM to find if two psychiatric disorders, schizophrenia and bipolar disorder, that shared symptoms also shared functional modules. Bipolar disorder (BPD), is a mental disorder, also known as manic-depressive illness, that causes unusual shifts in mood, energy and activity levels, often resulting in periods of depression or mania [281, 282]. Schizophrenia (SCZ) is a chronic and severe mental disorder that is represented by abnormal behavior and an altered notion of reality where the patients hear voices or see objects/persons that are not real [283]. While schizophrenia is not as common as other mental disorders, the symptoms can be very disabling. Schizophrenia and bipolar disorder had many common traits previously documented [284, 285].

The DIMSUM algorithm was supplied with 76 seed genes for Schizophrenia and 15 seed genes with Bipolar Disorder extracted and processed from two GWAS studies [286, 287]. There was no overlap between the seed gene sets for the two diseases. For each disease, additional 100 genes were extracted by DIMSUM to form the disease-centered module. We first queried the genes from the obtained BPD module against a list of BPD-associated genes from another recently published GWAS of the Psychiatric Genomic Consortium Bipolar Disorder Working Group [288]. As a result, we identified four genes from the module that were not among the initial set of seed genes but were found in the above GWAS study by Bipolar Disorder Working Group: RIMS1, ERBB2, STK4, and MAD1L1. These four genes have been previously shown to play functional roles in a number of neurological disorders [289-291]. For example, RIMS1 is a RAS superfamily member, and the encoded protein regulates synaptic vesicle exocytosis [292]. Mutations occurring on RIMS1 genes have been suggested to play a central role in cognition [293]. In addition to BPD, it is associated with autism spectrum disorder, neurodevelopmental disorders, and intellectual disability [289, 294].

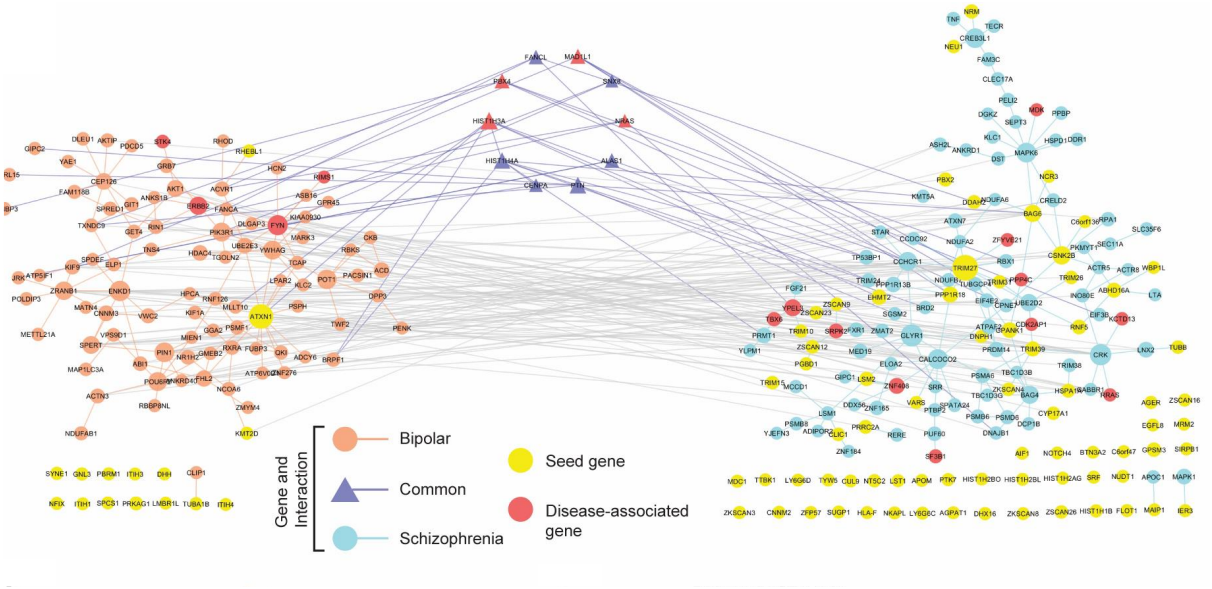


Figure 5-9 Analysis of Bipolar disorder and Schizophrenia modules discovered by DIMSUM. The BPD and SCZ modules are represented by the left and right components respectively, and the genes common to both disorders are in the small central component. DIMSUM discovers a total of nineteen disease-associated genes in both modules.

To determine disease-associated genes in the SCZ module we relied on a recent study that categorized the disease associated genes under three tiers based on diagnosis, polygenic risk scores (PRS) and those reported by the Psychiatric Genomic Consortium (PGC) [295]. In total, we found that 33 genes among the 100 added genes were present in the PGC gene set, of which four were in Tier 1 (CDK2AP1, MDK, ZFYVE21, and RRAS). Genes in this Tier were found to be significantly associated with both diagnosis and PRS. Perhaps the most interesting of these four was MDK, a gene associated with many important neurological processes, e.g., cerebral cortex development, behavioral fear response, short-term memory, and regulation of behavior [296, 297]. Eight more genes belonged to Tier 2, i.e., associated with diagnosis but not PRS.

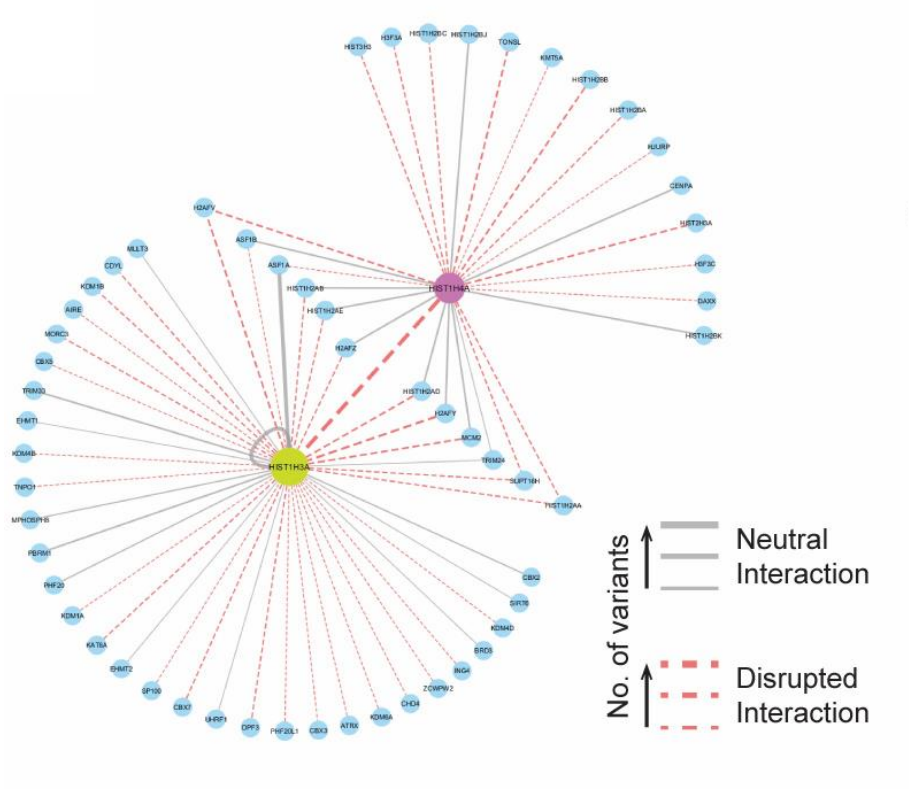


Figure 5-10 The rewiring of the subnetwork centered around the shared histone genes HIST1H3A and HIST1H4A. The dash lines indicate the accumulative damage based on the SNP-IN predictions.

Finally, we determined if the BPD and SCZ modules shared any genes—or more importantly, submodules—in common. It had been established that bipolar disorder and schizophrenia shared a large overlap of genetic risk loci and often exhibited similar symptoms like mania and depression [285]. Thus, in spite of the missing overlap between the two seed genes sets between those disease, the modules enriched with more disease-related genes could share common genes. Intriguingly, the BPD and SCZ modules were found to share a smaller sub-module of 10 genes connected with each other and with other BPD and SCZ genes (Figure 5-9). Out of 10 genes found shared between BPD and SCZ modules, four genes (HIST1H3A, PBX4, MAD1L1, and NRAS) were known to be strongly associated to both disorders (Figure 5-9). We conjectured that the common genes between the BPD and SCZ modules could provide insights into the phenotypic similarities between the two diseases. To support this hypothesis, we revisited the functional predictions from the SNP-IN tool involving these ten genes. We found that while the

mutations occurring in both diseases were quite diverse, a small group of mutations targeted the same subnetwork centering around HIST1H3A and HIST1H4A (Figure 5-10). Both of these genes were the core units of the nucleosome, implicated in a number of neuro-psychiatric disorders [298, 299]. Most of the mutations occurring in this subnetwork were predicted by the SNP-IN tool to disrupt the corresponding PPIs (Figure 5-11, Supplementary Table S5-7). Thus, different mutation frequencies and combinations for BPD and SCZ could give rise to different rewiring, or edgetic, effects of the HIST1H3A/HIST1H4A centric subnetwork. These results lead us to suggest that the different rewiring patterns of the same subnetwork could explain the phenotypic similarity but underlie differences in symptomatic severity between the two diseases.

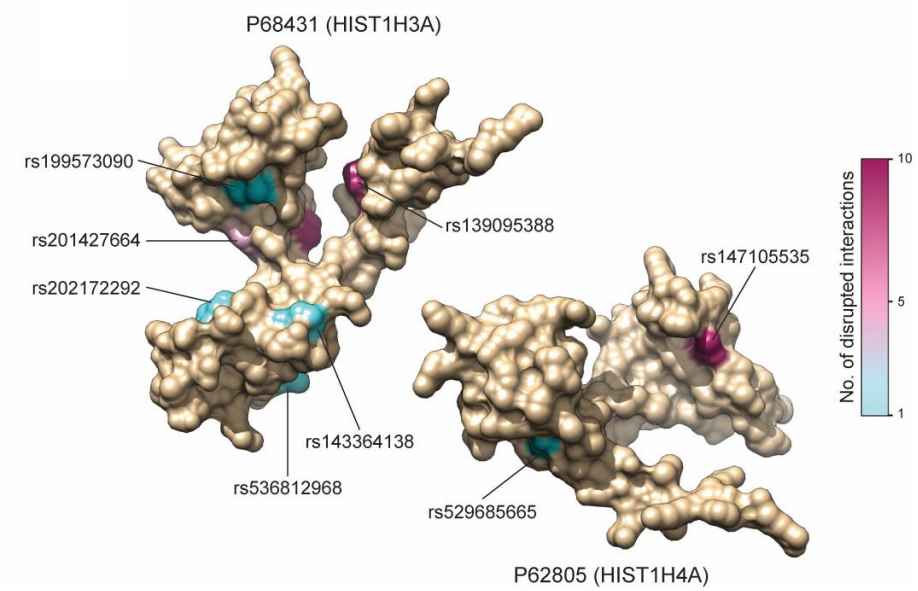


Figure 5-11 Protein structures for the histones HIST1H3A and HIST1H4A on the left and right respectively. Disruptive mutations occurring on these two genes are observed in both BPD and SCZ; HIST1H3A carries six disruptive mutations and HIST1H4A carries two. The color of each mutated residue corresponds to the number of interactions it disrupts

5.3 Discussion

In this work, we proposed a computational framework for functional disease module detection, DIMSUM, which integrates GWAS datasets with the human interactome, propagating the functional impact of nsSNVs. Our module detection approach first annotates the network with the functional information from genes and the associated mutations, followed by the network propagation to determine new genes associated with the same disease and, finally, subnetwork extraction. We assessed our approach using a set of eight complex diseases and comparing the performance of DIMSUM against two state-of-the-art seed-based module detection methods, DIAMOnD and SCA. The integration of multiple data types within a single computational framework allowed us to improve the effectiveness of module detection. In particular, the evaluation results showed that DIMSUM outperformed both DIAMOnD and SCA: our approach was able to yield modules with stronger disease association and greater biological relevance.

Comparison with DIAMOnD and SCA (Table 5-1) methods exhibited poor performances in the experiment carried out as part of this study. However, we would not claim that DIAMOnD and SCA does bear any value. We think they have their own advantages, and they might be suitable for some specific research scenarios. We detailed the reasons here. First, when we carried out the experiment, we set the number of genes added to the initial seed gene pool to 100. The main reason we select 100 as the number of genes to be added to initial gene pool is that it results in a final module with moderate size. Modules comprising hundreds of genes are often too general and almost impossible to gain biological insight and guide follow-up experiments. In DIAMOND, the default number of iteration steps is 200 recommended by the authors. We suspect that if we increase the number limit, both SCA and DIAMOND would include more genes related to the disease. We think this also demonstrate one of DIMSUM's advantage that it can capture the disease associated genes with immediate impact to the interactome in fewer iteration steps. Secondly, there could be some undercount in the cases of Diamond and SCA. We mainly resorted to OMIM and HGMD to check whether the newly added candidate genes are disease related or not. But this does not exhaust all possible sources. There could some scattered evidence in literatures lending support to DIAMOND and SCA's predictions.

The difference of design idea of these algorithms might lead them favor different groups of genes. And this could also contribute the performance difference. All the three methods evaluated in this experiment are seed based, iterative methods for module detection in protein-protein network. But they have distinct nature and design philosophy. For example, SCA is built on the idea of seed connectors, which served as "bridges" of different network branches that were induced by seed genes. It selected a gene that maximally increased the size of the largest connected component of the subnetworks as the seed connector and added them to the existing module. We suspect that the seed connector genes might be more likely to carry pathogenic frameshift mutations. These mutations tend to cause nonfunctional gene product and a complete loss of interactions. On the contrary, the DIMSUM methods might favor genes carrying pathogenic missense mutation, which tend to cause edgetic impact on the protein-protein interactions.

Integrating biomolecular networks across various types of data, including -omics profiles, GWAS, and functional annotation, have proven to be powerful for the detection and interpretation of biological modules [162]. GWAS investigates the entire genome and identifies the genomic loci related to a disease, providing a "macro view" of the underlying genetic architecture. On the other hand, leveraging functional annotation tools like the SNP-IN tool, enables a "micro view" by examining the specific and localized mechanistic effects of mutations and providing insights into disease etiology. Our computational framework facilitates joint interpretation of the biological information originating from those two different perspectives. Furthermore, the network propagation procedure allows to interpret the list of candidate genes into a genome-wide spectrum of gene scores, reflecting the disease association signal. This ability to amplify the signal from the seed gene pool has been previously proven helpful when identifying the genetic modules that underlie human diseases [133].

The case studies of eight complex diseases showed that the final modules obtained by DIMSUM typically include a large connected component containing most of the genes associated with a disease. This capacity of the algorithm to merge the initially isolated seed genes into a connected core component may be useful for elucidating the molecular mechanisms that are often carried out by functioning of molecular complexes and

pathways, rather than isolated proteins. Furthermore, the submodules disconnected from the largest component were also found to harbor a considerable number of genes related to the disease. Examination of the discovered modules and findings from the case studies prompted us with a hypothesis that underlie the importance of the system-wide variation effects for complex genetic diseases. Specifically, we hypothesize that disease phenotypes observed in the complex diseases, such as coronary artery disease and schizophrenia, may be a consequence of rewiring of an orchestrated functional module system rather than the abnormal functioning of the independent genes. Such functional module system would consist of a core module and several smaller satellite modules. The core unit is mainly responsible for the disease, while rewiring of the smaller satellite modules could also contribute to the disease progress and disease phenotype diversity. We further hypothesize that some satellite modules could correspond to a specific symptom node in the recently proposed symptom network for the psychiatric disorders [300]. Thus, the specific rewiring of the functional module system could help explaining the disease subtypes or different symptom combinations among patients.

Chapter 6 Conclusion and Future Work

6.1 Final Conclusion

The significance of this dissertation lies in addressing several important questions about genetic mutations and their relationships with complex phenotypes in the context of human interactome. First, we systematically characterized the network property, rewiring behaviors and accumulative damages caused by the pathogenic SNVs in the human interactome and investigate their link with complex disease. Especially, through translational study with clinical data, our work could contribute to the development of new biomarker or the design of new drug targets. Second, the "population edgetics" study would complement the current population genetic study. The edgetic profile differences between populations can be used to infer evolutionary history and interpret the difference of common phonotypes. Also, combining our module detection approach and population edgetics concept could help explain different disease susceptibility between populations. Lastly, we developed a computational framework incorporating SNP-IN annotation results of SNVs and GWAS study for disease module detection. Identifying functional modules more specific to the disease could help addressing some biological problems about the intrinsic relationship between disease genes and phonotype through following subnetwork and pathway analysis.

The innovation of this dissertation is three-fold: (1) a pioneering work in edgotype based biomarker discovery research; (2) a novel methodology to study genetic disease susceptibility and phenotype variance between different populations; and (3) an

innovative protocol that combines functional annotation, network propagation and GWAS data to identify mutations specific functional modules. The innovations are discussed below.

One innovation is this dissertation is that we combine our edgetic analysis with the clinical data on cancer patients. We demonstrate the disruptive mutations are prevalent in the cancer driver genes. Moreover, our analysis determines the link between the disruptive mutations and the decreased patient relapse time and survival time. To our knowledge, this is one of several translational studies attempting to apply edgetic perturbation model on cancer study and relate the protein-protein interaction to the clinical outcome [126, 269]. We expect this work can help develop a biomarker scoring system with regards to cancer patient survival time or relapse time.

Another innovation in the dissertation's methodology development is for studying the disease susceptibility in populations. A lot of works have identified genetic variations that are common in the general population, but contribute to the disease susceptibility. However, it is very difficult to make sense of how genetic variations in multiple genes, each with a slight impact, could underlie the different susceptibility to many complex diseases between populations. Our methodology is based on the new edgotype concept and combined with state-of-art module identification approaches. This strategy can reveal the distinct rewiring pattern in the disease modules and help explain the different disease susceptibility across populations.

An important advancement of the dissertation is developing a protocol for module detection. With the advancement of high-throughput technologies and enrichment of public databases, more computational approaches have been developed with the aim of integrating network and other biological data source for extracting context-dependent active modules[162]. We note that these methods are unable to detect the disease modules that sense and reflect the perturbations from genetic variations. The modules detected by our protocol could reflect the overall mutation "impact", which combines the "relevance" with the disease and the "damage" caused by genetic perturbations.

In conclusion, our in-silico edgetic profiling approach is a great alternative to costly and laborious experimental approach, such as Yeast Two Hybrid. It can also provide mechanistic insights into genotype-phenotype relationship. The role of a fast and inexpensive computational edgotyping approach is becoming increasingly important with the rapid growth of the personalized genomics data and ever-increasing catalogue of disease-associated variants. Such an approach can also reduce the cost of the experimental interaction assays by prioritizing the genes and mutations according to the predicted edgetic effects. Together with network analysis, we have shown successes in studying both pathogenic mutations and normal mutation. This strategy can help reveal the mechanistic details underlying the complex genotype-phenotype relationship. Further, the integration of multiple data types within a single computational framework allowed us to improve the effectiveness of module detection. In particular, our DIMSUM outperformed some of the state-of-art module detection methods, as our approach was able to yield modules with stronger disease association and greater biological relevance.

6.2 Future Work

As to the future direction, in spite of the achievements, as well as the significance and innovation discussed above, the methodologies developed in this dissertation could be further optimized and extended in several ways.

6.2.1 Leveraging Privileged Structural Information to Increase SNP-IN Tool's Prediction Coverage

Our recently developed SNP-IN tool has played an important role in systematic characterization of non-synonymous SNVs in human intreactome, complementing the recently published large-scale experimental edgetic profiling study [123]. While the accuracy of our, in-silico, approach is expected to be somewhat lower, compared to the experimental interaction assays, the coverage of our method is several times higher than the experimental approach: we were able to profile three time as many genes and more than twice as many pathogenic mutations as demonstrated in Chapter 3. Most

importantly, when comparing the distributions of the main edgotypes for our prediction with the experimental results, we find the distributions to be nearly identical. Given the minimal overlap of ~4% of shared genes and 1% of shared mutations with the experimental study, the results suggest that our approach can be used to guide future interactomics experiments, suggesting the most promising candidate mutations for the experimental edgetic profiling.

Furthermore, SNP-IN tool itself could be improved. While its coverage is currently greater than the experimental approach, our method requires information about the structure of the PPI interface, either from an experimentally resolved structure of macromolecular complex or from an accurate homology model. The requirement of a structurally resolved PPI for the prediction seriously limited the application of SNP-IN tool. First, not all proteins have resolved structures, not to mention much less structure information for the protein-protein interactions[301]. Also, homology modeling cannot help a lot in this case. INStruct is a high-quality protein interactome networks annotated to structural resolution[172]. INStruct contains 11,470 protein-protein interaction entries for Homo Sapiens, while it is expected there are more than 200,000 interactions in the human interactome[93]. This structural information is currently limited and was recently estimated to cover ~15% of the human interactome [302]. As a result, only one-third of all pathogenic non-synonymous SNVs extracted from the ClinVar database could be profiled in this work.

SNP-IN tool relies only on the structural information about the PPIs in both training and prediction stages. On the hand, it has been shown that sequence information regarding the variation and the interaction is also helpful for the prediction of functional impact[303, 304]. Recently, Vapnik introduced Learning Using Privileged Information (LUPI) paradigm[305]. It is a general methodology for utilizing additional (privileged) information about the training samples. The decision boundaries are learned from both standard training data and privileged information, but in the prediction stage, it only requires standard training data sample to make predictions. LUPI can improve the predictive performance and reduce the amount of required training data. Altogether, it inspires us to adopt a new formulation of the classification problem under the LUPI

paradigm by treating the structural information as privileged information. In fact, we can treat the structural information required in SNP-IN tool as the privileged information, and gather sequence information regarding the variation and interaction as standard training data. LUPI has recently attracted a lot of attention in machine learning community and started using in a few research areas[306, 307]. However, there are few successful applications of LUPI to biomedical informatics and computational biology problems. We believe a LUPI based classification method could a powerful alternative to SNP-IN tool and render a genome-wide prediction coverage.

6.2.2 A New Association Test Intergrading Rewiring Effects of SNVs.

Genome-wide association studies (GWAS) is a powerful tool for investigating the underlying genetic architecture of complex disease[308]. The goal of GWAS is to identify genetic factors that may contribute to a person's risk of a certain disease, either common or rare Mendelian diseases. There are many association methods, but the common essential idea is to test the strength of statistical evidence for nonrandom variant distribution in cases and in controls. Such statistical evidence won't be strong or reliable if data sample size is relatively small and/or SNVs are rare[309]. Furthermore, the mechanism of genetic effect is complex, not just a straight line from DNA to disease[91]. A gene may be critical to a disease pathway, but the final disease status is affected by many other factors. So, the association evidence measured solely by genotype and phenotype data could be weak. A promising approach to increasing the statistical power of association studies is to properly integrate the SNV information that reflects the intermediate steps of disease development[310]. PPIs are one important component related to disease development. Several recent genome-wide association studies have reported the value of incorporating PPI information into the pipeline of identifying novel disease genes[169, 311].

The idea of integrating the PPI network data into a traditional statistical framework for disease association has been gaining attention in the study of several complex diseases, such as cancer and diabetes[312-314]. However, their methodology mainly uses generic functional information to filter candidate genes and SNVs for test, while the association

test itself is still without incorporating such information. Such filtering process can loosen the strict genome-wide significance level in favor of relatively weak association factors. Our association test framework could enable implicit incorporation of the prior importance of SNVs regarding their influence on PPI that may involve the intermediate steps of disease development.

6.2.3 Edgotype Based Biomarker Scoring Systems for Translational Research

Edgotype is much better at dissecting the dynamics and complexities of biological systems compared to the traditional the 'node-centric' gene removal approach. It has been proposed and applied to study the molecular alterations observed in human complex genetic diseases. For example, mutations in CBS can cause enzyme deficiency and further gives rise to a metabolic disorder: Homocystinuria [315]. Zhong et. al. [146] tested five mutations (P49L, I278T, P145L, P422L and L539S) on CBS gene for interactions against three interactors of the respective wild-type protein. Two of them (P145L and L539S) have one Y2H interaction retained while two other interactions lost. This interaction-specific perturbation of CBS mutant proteins was further proven to associated with a treatment response of Homocystinuria. Patients carry the edgetic alleles, P145L or L539S, are responsive to pyridoxine, which can alleviate CBS deficiency and reduces the associated disease symptoms [316, 317]. And the rest are not pyridoxine responsive. Furthermore, edgetic perturbation model suggests distinct edgetic perturbations in a protein might cause different disorders and can help explain phenotypic variations among patients, such as incomplete penetrance or variable expressivity [113, 123, 146].

In this dissertation, we have shown that disruptions of protein-protein interactions caused by disease mutation is related decreased survival time. Thus, it is informative for survival time and relapse time prediction and can potentially be used as a biomarker. However, we shall note that there will be a long way to go for this potential edgotype based biomarker discovery research. First, a single edgetic biomarker is likely to lack predictive power [318]. Also, it is not clear whether such edgetic disruption is specific to certain types of cancers. Moreover, ethnicity, life style and other environmental factors, should be

considered, otherwise, such biomarker strategy will not generalize well for other groups of patients. In short, in spite of its potential, developing an edgotype based scoring systems or incorporating the edgetic effect into an existing biomarker scoring system might be a more promising option.

6.2.4 Integration with Other "-omics" Data and Biological Networks

The interactome network considered so far is protein-protein interaction networks. But there are many more other biological networks, such as gene co-expression network, metabolic network and gene regulatory networks etc. Some of these networks consist of physical interactions between macromolecules, while others are composed of functional links. These functional links can be indirect interactions, or even conceptual interactions. Although what the edges represent can be strikingly different, different links between biological entities can complement our knowledge from one specific biological network. For example, Wang et. al. [319] integrated various biological data, including gene expression profiles, genome-wide location data, protein-protein interactions etc. to construct an integrated cellular network of transcription regulations and protein-protein interactions. They further prune all possible interactions and remove those unlikely to exist in a real cellular system. This integrative method was applied to study *S. cerevisiae* stress responses and elucidate the stress response mechanistic details. They demonstrated the predictive power of this integrative approach and identified some genes/proteins which are relevant to the stress responses. In short, such integrated cellular network demands further analysis and experiments in the fields of network biology.

Substantial improvements have been made in the ability of utilizing high-throughput "omics" data for use in clinical environment during the same time period [320]. Scientists have recently started looking at a new important target: diagnostics and treatment of complex genetic diseases by leveraging the omics data of different types, including genomics, proteomics, metabolomics, transcriptomics, glycomics, epigenomics, lipomics, and others [321-323]. However, like most new concepts, there are multiple problems

surrounding attempts to utilize the omics data for treatment and diagnostics, such as a lack of a standardized protocol [324], reproducibility of the results, limited computational resources for data integration. Our works have demonstrated some practical solutions for integrating various biological data sources with the interactome as a platform.

Appendix

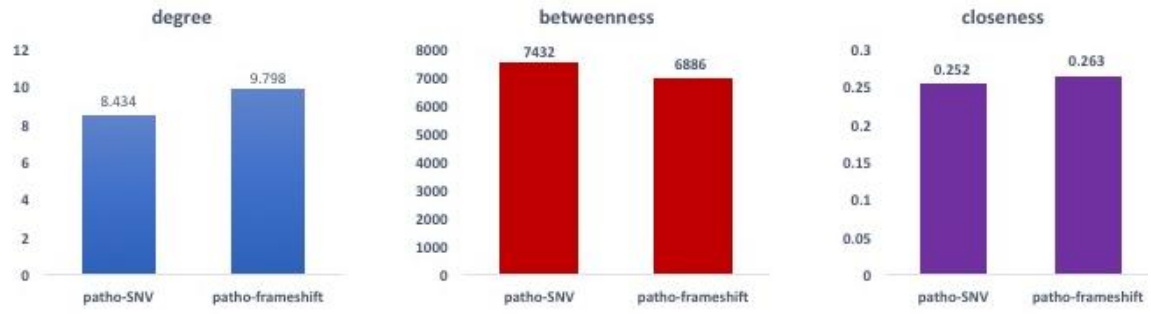
A1. Chapter 3 Supplementary Materials

Supplementary Figure Information

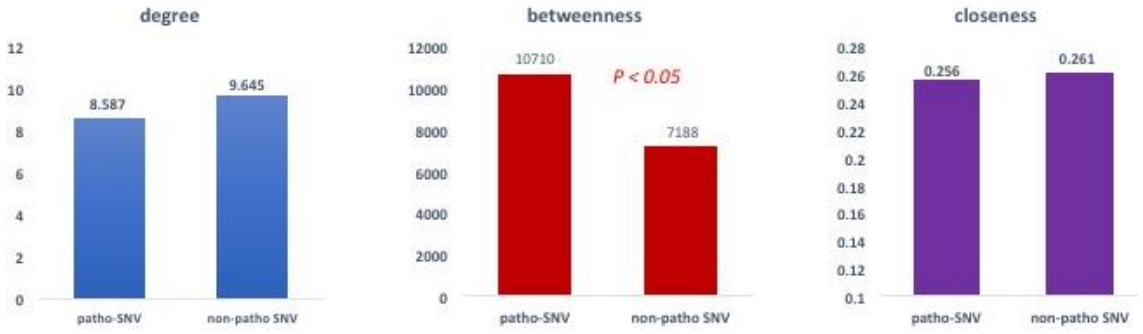
Supplementary Figure S3-1	Comparison of centrality between pathogenic SNVs and frameshift mutations in HI-II-14 interactome
Supplementary Figure S3-2	Comparison of centrality between pathogenic SNVs and non-pathogenic SNVs in HI-II-14 interactome
Supplementary Figure S3-3	Comparison of the survival time of France liver cancer patients with different functional mutations.
Supplementary Figure S3-4	Comparison of the relapse time of France liver cancer patients with different functional mutations.

Supplementary Table Information

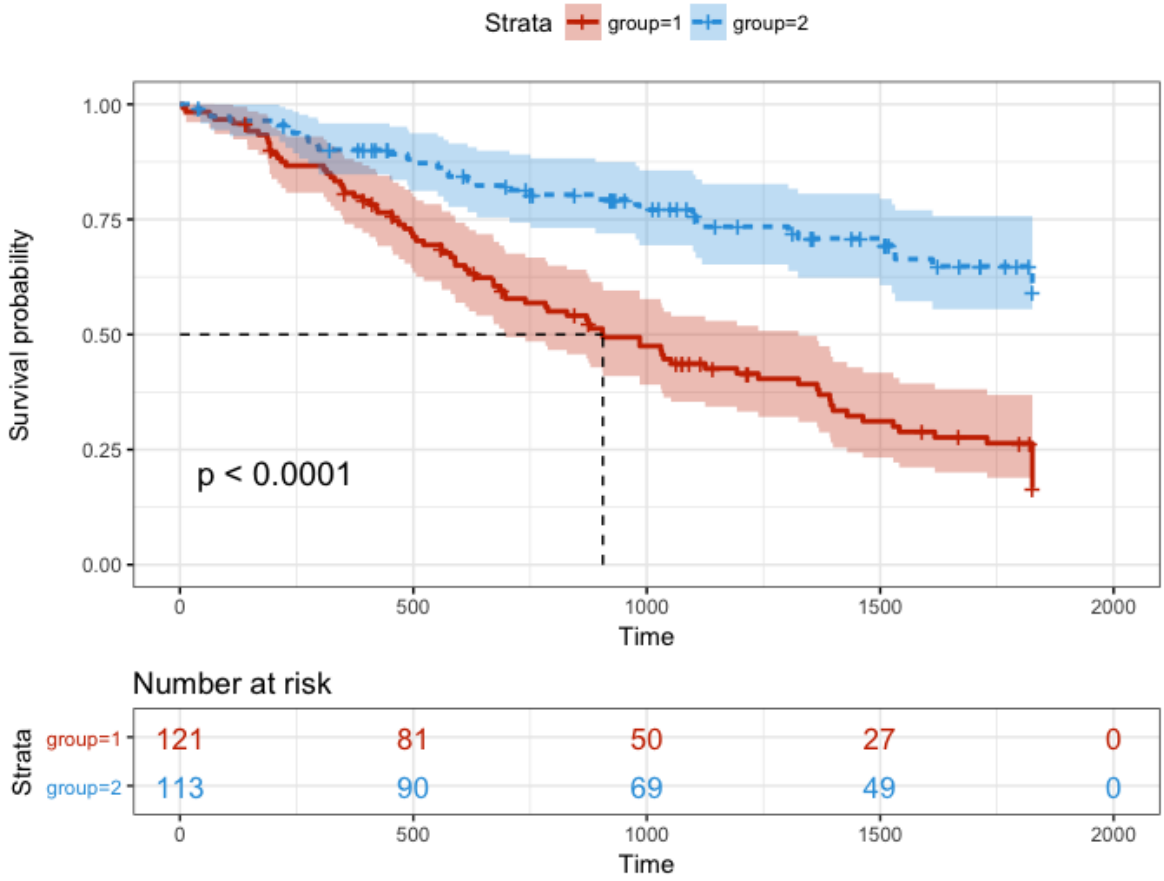
Supplementary Table S3-1	Distribution of functionally annotated pathogenic non-synonymous SNVs in the proteome.
Supplementary Table S3-2	Beneficial mutations with corresponding PPIs and related diseases.
Supplementary Table S3-3	SNP-IN tool annotation results of Type 2 Diabetes Mellitus related pathogenic mutation and corresponding PPIs
Supplementary Table S3-4	Cancer genes and related disruptive mutations.



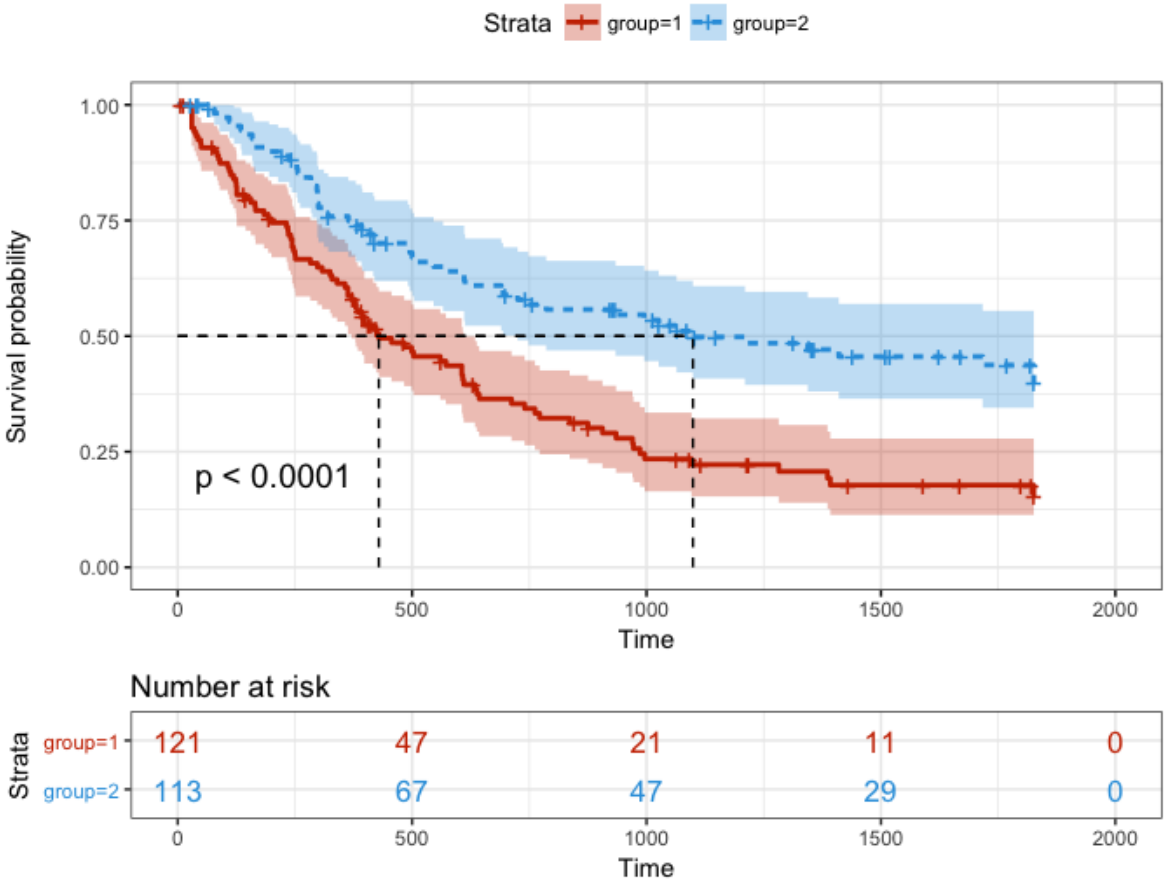
Supplementary Figure S3-1. Comparison of centrality between pathogenic SNVs and frameshift mutations in HI-II-14 interactome.



Supplementary Figure S3-2. Comparison of centrality between pathogenic SNVs and non-pathogenic SNVs in HI-II-14 interactome



Supplementary Figure S3-3. Comparison of the survival time of France liver cancer patients with different functional mutations.



Supplementary Figure S3-4. Comparison of the relapse time of France liver cancer patients with different functional mutations.

	Detrimental	*Neutral*	*Beneficial*	*Total*
Interface Domain	1,324	208	28	1,560
Other Domain	20	16	1	37
Outside	231	103	4	338
Total	1,575	327	33	1,935

Supplementary Table S3-1. Distribution of functionally annotated pathogenic non-synonymous SNVs in the proteome. Three types of functional SNV annotations with respect to a PPI are considered: Detrimental, Neutral, and Beneficial. The structural positioning of an SNV with respect to the protein domain architecture is categorized as belonging to the protein domain containing the PPI interface (Interface Domain), belonging to a protein domain that does not contain the PPI interface, and occurring in the protein termini or interdomain linkers (Outside).

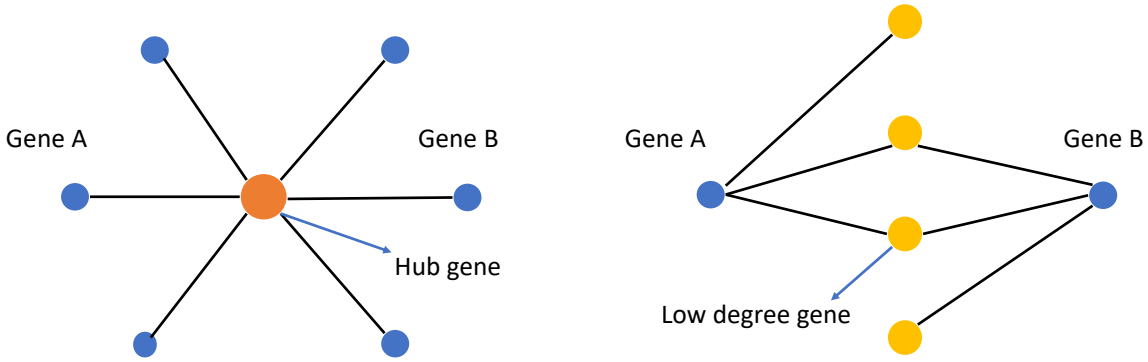
A2. Chapter 4 Supplementary Materials

Supplementary Figure Information

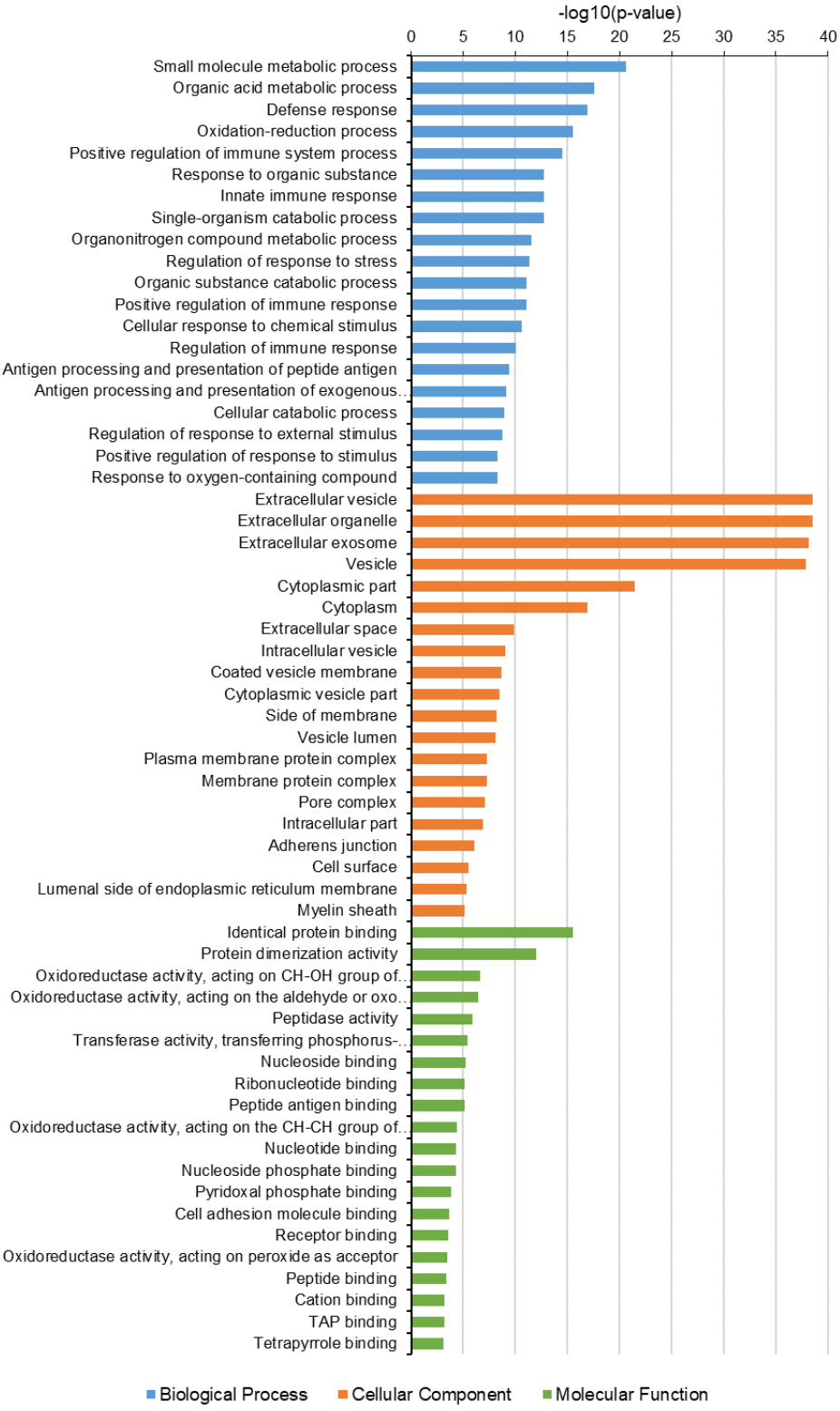
Supplementary Figure S4-1	Illustration of the concept of Diffusion State Distance (DSD)
Supplementary Figure S4-2	Top 20 enriched GO terms in three basic GO categories: biological process, molecular function and cellular component
Supplementary Figure S4-3	HLA-A is one of genes carrying high number of normal disruptive mutation in the HLA gene family

Supplementary Table Information

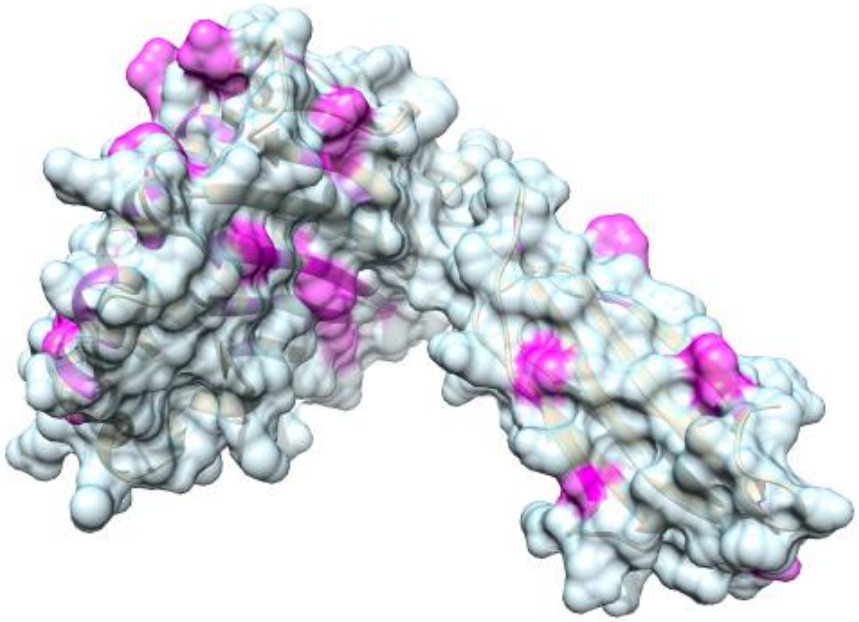
Supplementary Table S4-1	Disease phenotype information associated with genes enriched with disruptive mutation curated from OMIM, HGMD
Supplementary Table S4-2	Disruptive mutations carried by HLA-A and HLA-B gene.



Supplementary Figure S4-1 Illustration of the concept of Diffusion State Distance (DSD). DSD downweighs the influence of the hub gene and favors low degree gene.



Supplementary Figure S4-2 Top 20 significant GO terms for genes enriched with disruptive mutations among three basic GO categories: biological process, molecular function and cellular component.



Supplementary Figure S4-3 HLA-A is one of genes carrying high number of normal disruptive mutation in the HLA gene family

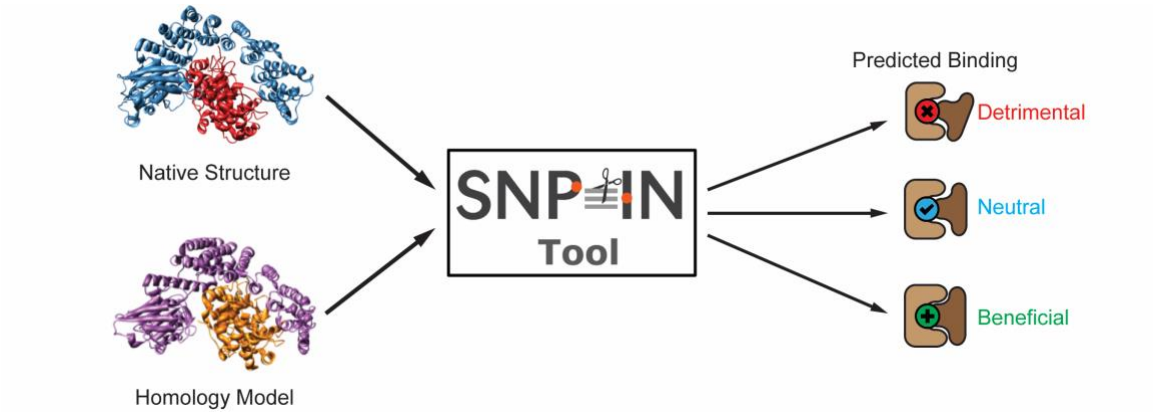
A3. Chapter 5 Supplementary Materials

Supplementary Figure Information

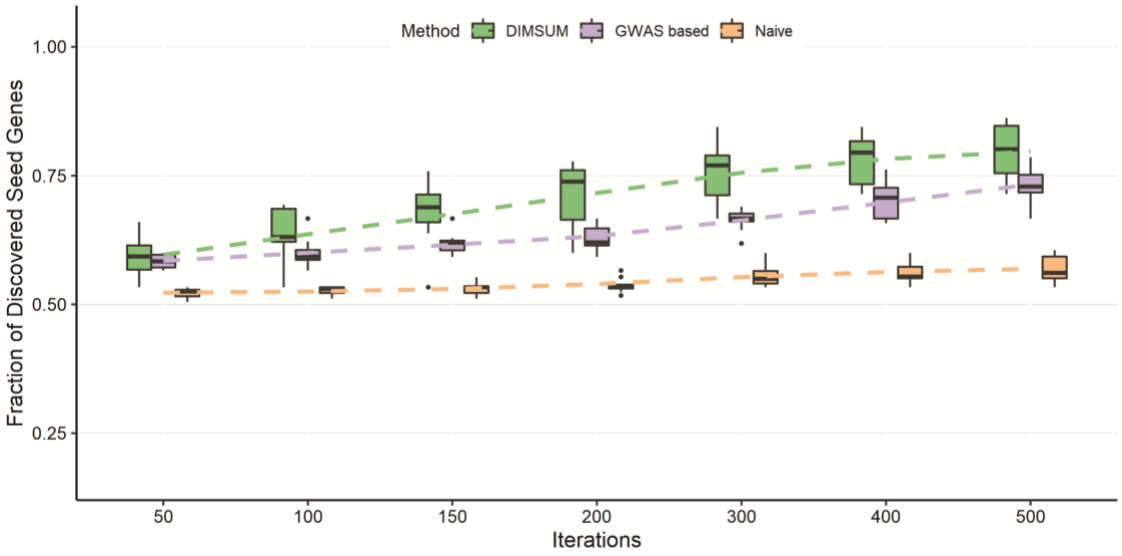
Supplementary Figure S5-1	Structure-based prediction of SNP's effect on PPI when applying SNP-IN tool.
Supplementary Figure S5-2	Comparison of DIMSUM against GWAS based and naïve network propagation procedure with 50% of nodes randomly selected from the seed gene pool.
Supplementary Figure S5-3	Comparison of DIMSUM against GWAS based and naïve network propagation procedure with 75% of nodes randomly selected from the seed gene pool.
Supplementary Figure S5-4	Coronary artery disease (CAD) module discovered by the SCA algorithm.

Supplementary Table Information

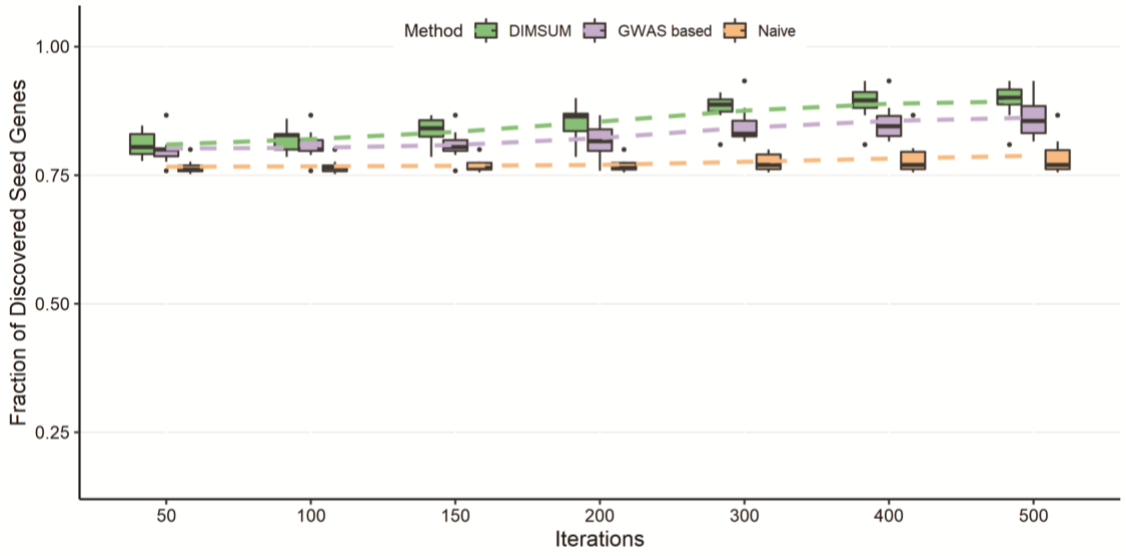
Supplementary Table S5-1	Description of the eight GWAS datasets curated for this work.
Supplementary Table S5-2	Seeds generated by the Pascal tool from eight GWAS datasets of complex diseases.
Supplementary Table S5-3	SNP-IN tool annotation results for eight GWAS datasets.
Supplementary Table S5-4	Disease gene association data curated from OMIM, HGMD
Supplementary Table S5-5	Overlapping genes between the discovered modules from all three methods
Supplementary Table S5-6	Total number of enriched GO terms for eight GWAS datasets from all three methods
Supplementary Table S5-7	HIST1H4A-HIST1H3A centered PPI subnetwork and associated disruptive mutations



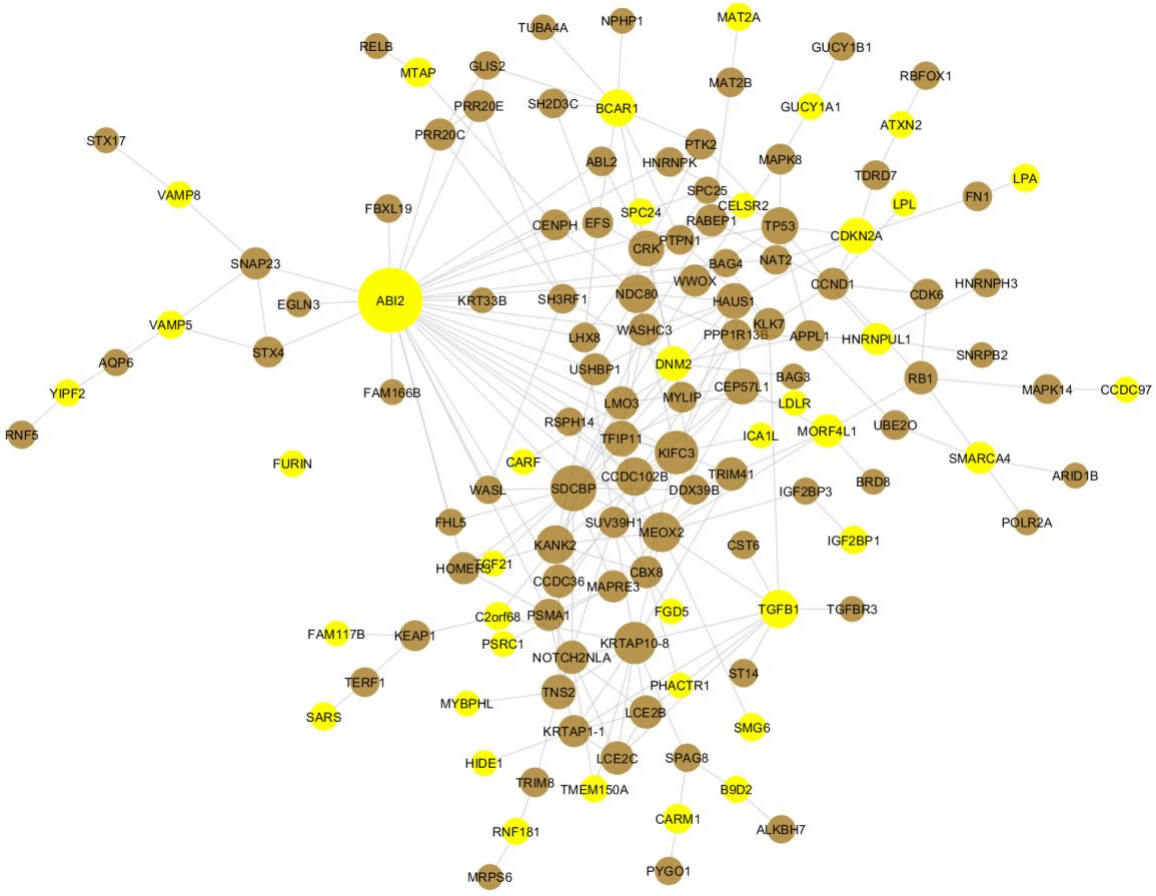
Supplementary Figure 5-1. Structure-based prediction of SNP's effect on PPI when applying SNP-IN tool. The SNP-IN tool uses native structure (when available) or a homology model of the PPI as an input. The output is the predicted PPI-rewiring effect of the SNP with three possible outcomes: (i) Detrimental: where the binding is lost; (ii) Neutral: the binding is preserved and (iii) Beneficial: the binding is strengthened.



Supplementary Figure 5-2. Comparison of DIMSUM against GWAS based and naïve network propagation procedure with 50% of nodes randomly selected from the seed gene pool. DIMSUM has a greater trend in discovering seed genes with respect to the increasing number of iterations reaching an average of 0.75.



Supplementary Figure 5-3. Comparison of DIMSUM against GWAS based and naïve network propagation procedure with 75% of nodes randomly selected from the seed gene pool. DIMSUM again outperforms the GWAS based and naïve method, reaching an average discovery rate of 0.9 with 500 iterations.



Supplementary Figure 5-4. Coronary artery disease (CAD) module discovered by the SCA algorithm. Yellow nodes represent the seed genes for CAD and brown nodes represent the added genes. The SCA algorithm tries to include as many seeds as possible when building the module. No CAD associated genes were discovered by SCA, and the degree distribution was greater (i.e. high degree nodes or hubs were added) than in the module discovered by DIMSUM.

Disease ID	Disease name	Category	No. of SNPs	Reference
D1	Coronary artery disease	Cardiovascular	9,455,778	Nikpay et al., Nat Genet (2015)
D2	Diabetes mellitus - Type 2	Glycemic	7,474,782	Fuchsberger et al., Nat Genet (2016)
D3	Macular degeneration	Other	12,023,830	Fritsche et al., Nat Genet (2015)
D4	Osteoporosis	Other	2,478,338	Estrada et al., Nat Genet (2012)
D5	Alzheimer's disease	Neurodegenerative	7,055,881	Lambert et al., Nat Genet (2013)
D6	Rheumatoid arthritis	Immune	6,446,682	Okada et al., Nature (2014)
D7	Bipolar disorder	Psychiatric	2,427,220	PGC, Nat Genet (2011)
D8	Schizophrenia	Psychiatric	9,898,078	Ripke et al., Nat Genet (2013)

Supplementary Table S5-1 Description of the eight GWAS datasets curated for this work.

Bibliography

1. Anderson, D.E., *Ceinical characteristics of the genetic variety of cutaneous melanoma in man.* Cancer, 1971. **28**(3): p. 721-725.
2. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31.
3. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nature reviews genetics, 2011. **12**(2): p. 87.
4. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Molecular cell, 2015. **58**(4): p. 610-620.
5. Zeggini, E., *Next-generation association studies for complex traits.* Nature genetics, 2011. **43**(4): p. 287.
6. Cui, H., et al., *The variation game: Cracking complex genetic disorders with NGS and omics data.* Methods, 2015. **79**: p. 18-31.
7. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.
8. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in genetics, 2008. **24**(3): p. 133-141.
9. Davey, J.W., et al., *Genome-wide genetic marker discovery and genotyping using next-generation sequencing.* Nature Reviews Genetics, 2011. **12**(7): p. 499-510.
10. Boycott, K.M., et al., *Rare-disease genetics in the era of next-generation sequencing: discovery to translation.* Nature Reviews Genetics, 2013. **14**(10): p. 681-691.
11. Carapito, R., M. Radosavljevic, and S. Bahram, *Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies.* Human immunology, 2016. **77**(11): p. 1016-1023.
12. Consortium, G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061.
13. Nagasaki, M., et al., *Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals.* Nature communications, 2015. **6**: p. 8018.
14. Csermely, P., et al., *Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review.* Pharmacology & therapeutics, 2013. **138**(3): p. 333-408.
15. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome research, 2008. **18**(4): p. 644-652.
16. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nature reviews genetics, 2011. **12**(1): p. 56.
17. Barabasi, A.-L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nature reviews genetics, 2004. **5**(2): p. 101-113.
18. Vidal, M., M.E. Cusick, and A.-L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-998.

19. Cho, D.-Y., Y.-A. Kim, and T.M. Przytycka, *Network biology approach to complex diseases.* PLoS computational biology, 2012. **8**(12): p. e1002820.
20. Barrenas, F., et al., *Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms.* Genome Biol, 2012. **13**(6): p. R46.
21. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nature biotechnology, 2012. **30**(2): p. 159.
22. Sahni, N., et al., *Widespread macromolecular interaction perturbations in human genetic disorders.* Cell, 2015. **161**(3): p. 647-660.
23. Chavez, J.D., et al., *Quantitative interactome analysis reveals a chemoresistant edgotype.* Nature communications, 2015. **6**: p. 7928.
24. Zhao, N., et al., *Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning.* PLoS computational biology, 2014. **10**(5): p. e1003592.
25. Wray, N.R., et al., *Pitfalls of predicting complex traits from SNPs.* Nature Reviews Genetics, 2013. **14**(7): p. 507-515.
26. Prokunina, L. and M.E. Alarcón-Riquelme, *Regulatory SNPs in complex diseases: their identification and functional validation.* Expert reviews in molecular medicine, 2004. **6**(10): p. 1-15.
27. Giacomini, K.M., et al., *The pharmacogenetics research network: from SNP discovery to clinical drug response.* Clinical Pharmacology & Therapeutics, 2007. **81**(3): p. 328-345.
28. Shastry, B.S., *SNPs: impact on gene function and phenotype*, in *Single Nucleotide Polymorphisms*. 2009, Springer. p. 3-22.
29. Ramensky, V., P. Bork, and S. Sunyaev, *Human non-synonymous SNPs: server and survey.* Nucleic acids research, 2002. **30**(17): p. 3894-3900.
30. Kao, S., S. Chong, and C. Lee, *The role of single nucleotide polymorphisms (SNPs) in understanding complex disorders and pharmacogenomics.* Annals of the Academy of Medicine, Singapore, 2000. **29**(3): p. 376-382.
31. Mah, J.T., E.S. Low, and E. Lee, *In silico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery.* Drug discovery today, 2011. **16**(17-18): p. 800-809.
32. Sauna, Z.E. and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human disease.* Nature Reviews Genetics, 2011. **12**(10): p. 683-691.
33. Hayden, E.C., *Technology: the $1,000 genome.* Nature, 2014. **507**(7492): p. 294-295.
34. Hawkins, R.D., G.C. Hon, and B. Ren, *Next-generation genomics: an integrative approach.* Nature Reviews Genetics, 2010. **11**(7): p. 476-486.
35. Kilpinen, H. and J.C. Barrett, *How next-generation sequencing is transforming complex disease genetics.* Trends in Genetics, 2013. **29**(1): p. 23-30.
36. Goldstein, D.B., et al., *Sequencing studies in human genetics: design and interpretation.* Nature Reviews Genetics, 2013. **14**(7): p. 460-470.
37. Eyre-Walker, A., *Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies.* Proceedings of the National Academy of Sciences, 2010. **107**(suppl 1): p. 1752-1756.
38. Metzker, M.L., *Sequencing technologies—the next generation.* Nature Reviews Genetics, 2009. **11**(1): p. 31-46.
39. Bentley, D.R., *Whole-genome re-sequencing.* Current opinion in genetics & development, 2006. **16**(6): p. 545-552.
40. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nature Reviews Genetics, 2011. **12**(11): p. 745-755.

41. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nature reviews genetics, 2010. **12**(2): p. 87-98.
42. Saliba, A.-E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic acids research, 2014: p. gku555.
43. Boguski, M.S., R. Arnaout, and C. Hill, *Customized care 2020: how medical sequencing and network biology will enable personalized medicine.* F1000 biology reports, 2009. **1**.
44. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature reviews genetics, 2008. **9**(5): p. 356.
45. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719.
46. Gonzalez-Perez, A., et al., *Computational approaches to identify functional genetic variants in cancer genomes.* Nature methods, 2013. **10**(8): p. 723.
47. Consortium, G.P., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
48. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic acids research, 2001. **29**(1): p. 308-311.
49. Stenson, P.D., et al., *Human gene mutation database (HGMD®): 2003 update.* Human mutation, 2003. **21**(6): p. 577-581.
50. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic acids research, 2013: p. gkt1113.
51. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic acids research, 2005. **33**(suppl 1): p. D514-D517.
52. *Locus-Specific Mutation Databases*. Available from: http://www.hgvs.org/locus-specific-mutation-databases.
53. Fokkema, I.F., et al., *LOVD v. 2.0: the next generation in gene variant databases.* Human mutation, 2011. **32**(5): p. 557-563.
54. Alexander, R.P., et al., *Annotating non-coding regions of the genome.* Nature Reviews Genetics, 2010. **11**(8): p. 559.
55. Pabinger, S., et al., *A survey of tools for variant analysis of next-generation genome sequencing data.* Briefings in bioinformatics, 2014. **15**(2): p. 256-278.
56. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nature Reviews Genetics, 2011. **12**(9): p. 628-640.
57. Ward, L.D. and M. Kellis, *Interpreting noncoding genetic variation in complex traits and human disease.* Nature biotechnology, 2012. **30**(11): p. 1095-1106.
58. Raphael, B.J., et al., *Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine.* Genome medicine, 2014. **6**(1): p. 5.
59. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic acids research, 2010. **38**(16): p. e164-e164.
60. Hu, H., et al., *VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix.* Genetic epidemiology, 2013. **37**(6): p. 622-634.
61. Dayhoff, M.O. and R.M. Schwartz. *A model of evolutionary change in proteins*. in *In Atlas of protein sequence and structure*. 1978. Citeseer.
62. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein*

blocks. Proceedings of the National Academy of Sciences, 1992. **89**(22): p. 10915-10919.
63. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nature protocols, 2009. **4**(7): p. 1073-1081.
64. Tavtigian, S.V., et al., *Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral.* Journal of medical genetics, 2006. **43**(4): p. 295-305.
65. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nature methods, 2010. **7**(4): p. 248-249.
66. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database, 2011. **2011**: p. bar009.
67. Cline, M.S. and R. Karchin, *Using bioinformatics to predict the functional impact of SNVs.* Bioinformatics, 2011. **27**(4): p. 441-448.
68. Mathe, E., et al., *Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods.* Nucleic acids research, 2006. **34**(5): p. 1317-1325.
69. Makarov, V., et al., *AnnTools: a comprehensive and versatile annotation toolkit for genomic variants.* Bioinformatics, 2012. **28**(5): p. 724-725.
70. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.* Journal of molecular biology, 2002. **320**(2): p. 369-387.
71. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS computational biology, 2010. **6**(12): p. e1001025.
72. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome research, 2003. **13**(9): p. 2129-2141.
73. Bromberg, Y. and B. Rost, *SNAP: predict effect of non-synonymous polymorphisms on function.* Nucleic acids research, 2007. **35**(11): p. 3823-3835.
74. San Lucas, F.A., et al., *Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools.* Bioinformatics, 2012. **28**(3): p. 421-422.
75. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-2070.
76. Jemal, A., et al., *Global cancer statistics.* CA: a cancer journal for clinicians, 2011. **61**(2): p. 69-90.
77. Control, C.f.D. and Prevention, *National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014.* Atlanta, GA: US Department of Health and Human Services, 2014.
78. Perrin, J.M., S.R. Bloom, and S.L. Gortmaker, *The increase of childhood chronic conditions in the United States.* Jama, 2007. **297**(24): p. 2755-2759.
79. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic acids research, 2014. **42**(D1): p. D1001-D1006.
80. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nature Reviews Genetics, 2009. **10**(4): p. 241-251.
81. Oksenberg, J.R., et al., *The genetics of multiple sclerosis: SNPs to pathways to pathogenesis.* Nature Reviews Genetics, 2008. **9**(7): p. 516-526.
82. Didonna, A., et al., *A non-synonymous single-nucleotide polymorphism associated with multiple sclerosis risk affects the EVI5 interactome.* Human molecular

genetics, 2015. **24**(24): p. 7151-7158.
83. Kastritis, P.L. and A.M. Bonvin, *On the binding affinity of macromolecular interactions: daring to ask why proteins interact.* Journal of The Royal Society Interface, 2013. **10**(79): p. 20120835.
84. Mikhail, F.M., *Copy number variations and human genetic disease.* Current opinion in pediatrics, 2014. **26**(6): p. 646-652.
85. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nat Rev Mol Cell Biol, 2009. **10**(11): p. 741-54.
86. Singh, R.K. and T.A. Cooper, *Pre-mRNA splicing in disease and therapeutics.* Trends Mol Med, 2012. **18**(8): p. 472-82.
87. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nature biotechnology, 2010. **28**(10): p. 1057-1068.
88. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.
89. Ramanan, V.K., et al., *Pathway analysis of genomic data: concepts, methods, and prospects for future development.* TRENDS in Genetics, 2012. **28**(7): p. 323-332.
90. Askland, K., C. Read, and J. Moore, *Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission.* Human genetics, 2009. **125**(1): p. 63-79.
91. Carter, H., M. Hofree, and T. Ideker, *Genotype to phenotype via network analysis.* Current opinion in genetics & development, 2013. **23**(6): p. 611-621.
92. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nature Reviews Genetics, 2011. **12**(1): p. 56-68.
93. Stumpf, M.P., et al., *Estimating the size of the human interactome.* Proceedings of the National Academy of Sciences, 2008. **105**(19): p. 6959-6964.
94. Vidal, M., M.E. Cusick, and A.-L. Barabási, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-998.
95. Stark, C., et al., *The BioGRID interaction database: 2011 update.* Nucleic acids research, 2010. **39**(suppl_1): p. D698-D704.
96. Salwinski, L., et al., *The database of interacting proteins: 2004 update.* Nucleic acids research, 2004. **32**(suppl_1): p. D449-D451.
97. Kerrien, S., et al., *The IntAct molecular interaction database in 2012.* Nucleic acids research, 2012. **40**(D1): p. D841-D846.
98. Licata, L., et al., *MINT, the molecular interaction database: 2012 update.* Nucleic acids research, 2012. **40**(D1): p. D857-D861.
99. Hu, Z., et al., *VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology.* Nucleic acids research, 2009. **37**(suppl_2): p. W115-W121.
100. Mewes, H.W., et al., *MIPS: curated databases and comprehensive secondary data resources in 2010.* Nucleic acids research, 2011. **39**(suppl_1): p. D220-D224.
101. Keshava Prasad, T., et al., *Human protein reference database—2009 update.* Nucleic acids research, 2009. **37**(suppl_1): p. D767-D772.
102. Venkatesan, K., et al., *An empirical framework for binary interactome mapping.* Nature methods, 2009. **6**(1): p. 83.
103. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-1226.
104. Yu, H., et al., *Next-generation sequencing to generate interactome datasets.* Nature

methods, 2011. **8**(6): p. 478.
105. Rual, J.-F., et al., *Towards a proteome-scale map of the human protein–protein interaction network.* Nature, 2005. **437**(7062): p. 1173.
106. Wang, Y., et al., *Sequence-based protein-protein interaction prediction via support vector machine.* Journal of Systems Science and Complexity, 2010. **23**(5): p. 1012-1023.
107. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein–protein interaction.* Protein engineering, 2001. **14**(9): p. 609-614.
108. Margolin, A.A., et al. *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. in *BMC bioinformatics*. 2006. Springer.
109. Jaeger, S., et al. *Integrating protein-protein interactions and text mining for protein function prediction*. in *BMC bioinformatics*. 2008. Springer.
110. Botstein, D. and N. Risch, *Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.* Nature genetics, 2003. **33**(3): p. 228-237.
111. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nature Reviews Genetics, 2010. **11**(6): p. 446-450.
112. Zhang, B., Y. Tian, and Z. Zhang, *Network biology in medicine and beyond.* Circulation: Cardiovascular Genetics, 2014. **7**(4): p. 536-547.
113. Furlong, L.I., *Human diseases through the lens of network biology.* Trends in genetics, 2013. **29**(3): p. 150-159.
114. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Molecular systems biology, 2013. **9**(1).
115. Newman, M.E., *The structure and function of complex networks*. SIAM review, 2003. **45**(2): p. 167-256.
116. Mason, O. and M. Verwoerd, *Graph theory and networks in biology*. IET systems biology, 2007. **1**(2): p. 89-119.
117. Albert, R., *Scale-free networks in cell biology.* Journal of cell science, 2005. **118**(21): p. 4947-4957.
118. Wagner, A. and D.A. Fell, *The small world inside large metabolic networks.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 2001. **268**(1478): p. 1803-1810.
119. Santolini, M. and A.-L. Barabási, *Predicting perturbation patterns from the topology of biological networks*. Proceedings of the National Academy of Sciences, 2018. **115**(27): p. E6375-E6383.
120. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nature biotechnology, 2012. **30**(2): p. 159-164.
121. Mosca, R., A. Céol, and P. Aloy, *Interactome3D: adding structural details to protein networks.* Nature methods, 2013. **10**(1): p. 47-53.
122. Duran-Frigola, M., R. Mosca, and P. Aloy, *Structural systems pharmacology: the role of 3D structures in next-generation drug development.* Chemistry & biology, 2013. **20**(5): p. 674-684.
123. Sahni, N., et al., *Edgotype: a fundamental link between genotype and phenotype.* Current opinion in genetics & development, 2013. **23**(6): p. 649-657.
124. Zhong, Q., et al., *Edgetic perturbation models of human inherited disorders.* Molecular systems biology, 2009. **5**(1): p. 321.
125. Dreze, M., et al., *'Edgetic'perturbation of a C. elegans BCL2 ortholog.* Nature

methods, 2009. **6**(11): p. 843.
126. Wang, Y., N. Sahni, and M. Vidal, *Global edgetic rewiring in cancer networks.* Cell systems, 2015. **1**(4): p. 251-253.
127. Zhang, W., et al., *Diagnosing phenotypes of single-sample individuals by edge biomarkers.* Journal of molecular cell biology, 2015. **7**(3): p. 231-241.
128. McGarry, K., et al., *Complex network based computational techniques for 'edgetic'modelling of mutations implicated with cardiovascular disease*, in *Advances in Computational Intelligence Systems*. 2017, Springer. p. 89-106.
129. Lonser, R.R., et al., *von Hippel-Lindau disease.* The Lancet, 2003. **361**(9374): p. 2059-2067.
130. Clarke, N.F., et al., *Mutations in TPM3 are a common cause of congenital fiber type disproportion.* Annals of neurology, 2008. **63**(3): p. 329-337.
131. Laing, N.G., et al., *A mutation in the α tropomyosin gene TPM3 associated with autosomal dominant nemaline myopathy.* Nature genetics, 1995. **9**(1): p. 75-79.
132. Luck, K., et al., *A reference map of the human protein interactome.* bioRxiv, 2019: p. 605451.
133. Cowen, L., et al., *Network propagation: a universal amplifier of genetic associations.* Nature Reviews Genetics, 2017. **18**(9): p. 551.
134. Vanunu, O., et al., *Associating genes and protein complexes with disease via network propagation.* PLoS computational biology, 2010. **6**(1): p. e1000641.
135. Li, H., et al., *Network propagation predicts drug synergy in cancers.* Cancer research, 2018. **78**(18): p. 5446-5457.
136. Menden, M.P., et al., *A cancer pharmacogenomic screen powering crowd-sourced advancement of drug combination prediction.* bioRxiv, 2018: p. 200451.
137. Crow, J.F. and M. Kimura, *An introduction to population genetics theory.* An introduction to population genetics theory., 1970.
138. Weibull, J.W., *Evolutionary game theory*. 1997: MIT press.
139. Hofmann, S., et al., *Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlation with D loop variants and association with disease.* Human molecular genetics, 1997. **6**(11): p. 1835-1846.
140. Falconer, D.S., *Introduction to quantitative genetics.* Introduction to quantitative genetics., 1960.
141. Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez, *Estimating and interpreting heritability for plant breeding: an update.* Plant breeding reviews, 2003. **22**.
142. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era—concepts and misconceptions.* Nature reviews genetics, 2008. **9**(4): p. 255-266.
143. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.
144. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proceedings of the National Academy of Sciences, 2012. **109**(4): p. 1193-1198.
145. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature genetics, 2010. **42**(7): p. 565.
146. Zhong, Q., et al., *Edgetic perturbation models of human inherited disorders.* Molecular systems biology, 2009. **5**(1).
147. Madhani, H.D., C.A. Styles, and G.R. Fink, *MAP kinases with distinct inhibitory functions impart signaling specificity during yeast differentiation.* Cell, 1997. **91**(5):

p. 673-684.
148. Siva, N., *1000 Genomes project*. 2008, Nature Publishing Group.
149. Consortium, E.P., *The ENCODE (ENCyclopedia of DNA elements) project.* Science, 2004. **306**(5696): p. 636-640.
150. Schork, N.J., *Genetics of complex disease: approaches, problems, and solutions.* American journal of respiratory and critical care medicine, 1997. **156**(4): p. S103-S109.
151. Clayton, D.G., *Prediction and interaction in complex disease genetics: experience in type 1 diabetes.* PLoS genetics, 2009. **5**(7).
152. Manolio, T.A., *Cohort studies and the genetics of complex disease.* Nature genetics, 2009. **41**(1): p. 5-6.
153. King, M.-C., J.H. Marks, and J.B. Mandell, *Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2.* Science, 2003. **302**(5645): p. 643-646.
154. Imyanitov, E., K. Hanson, and B. Zhivotovsky, *Polymorphic variations in apoptotic genes and cancer predisposition*. 2005, Nature Publishing Group.
155. Turnbull, C. and N. Rahman, *Genetic predisposition to breast cancer: past, present, and future.* Annu. Rev. Genomics Hum. Genet., 2008. **9**: p. 321-345.
156. Frank, S.A., *Genetic predisposition to cancer—insights from population genetics.* Nature reviews genetics, 2004. **5**(10): p. 764-772.
157. Qi, L., et al., *Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men.* The American journal of clinical nutrition, 2009. **89**(5): p. 1453-1458.
158. Power, R.A., et al., *Genetic predisposition to schizophrenia associated with increased use of cannabis.* Molecular psychiatry, 2014. **19**(11): p. 1201-1204.
159. Postma, D.S., et al., *Asthma and chronic obstructive pulmonary disease: common genes, common environments?* American journal of respiratory and critical care medicine, 2011. **183**(12): p. 1588-1594.
160. Nayak, R.R., et al., *Coexpression network based on natural variation in human gene expression reveals gene interactions and functions.* Genome research, 2009. **19**(11): p. 1953-1962.
161. Barabasi, A.-L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nature reviews genetics, 2004. **5**(2): p. 101.
162. Mitra, K., et al., *Integrative approaches for finding modular structure in biological networks.* Nature Reviews Genetics, 2013. **14**(10): p. 719.
163. Girvan, M. and M.E. Newman, *Community structure in social and biological networks.* Proceedings of the national academy of sciences, 2002. **99**(12): p. 7821-7826.
164. Zhang, Q.C., et al., *Structure-based prediction of protein–protein interactions on a genome-wide scale.* Nature, 2012. **490**(7421): p. 556.
165. Choobdar, S., et al., *Assessment of network module identification across complex diseases.* Nature methods, 2019. **16**(9): p. 843-852.
166. Ghiassian, S.D., J. Menche, and A.-L. Barabási, *A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome.* PLoS computational biology, 2015. **11**(4): p. e1004120.
167. Tripathi, S., et al., *Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules.* BMC bioinformatics, 2016. **17**(1): p. 129.
168. Vlaic, S., et al., *ModuleDiscoverer: Identification of regulatory modules in protein-*

protein interaction networks. Scientific reports, 2018. **8**(1): p. 433.
169. Zhang, D., et al. *Incorporation of protein binding effects into likelihood ratio test for exome sequencing data*. in *BMC proceedings*. 2016. BioMed Central.
170. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-934.
171. Saliba, A.-E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic acids research, 2014. **42**(14): p. 8845-8860.
172. Meyer, M.J., et al., *INstruct: a database of high-quality 3D structurally resolved protein interactome networks.* Bioinformatics, 2013. **29**(12): p. 1577-1579.
173. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic acids research, 2013. **42**(D1): p. D980-D985.
174. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genetics in medicine, 2015. **17**(5): p. 405.
175. Das, J. and H. Yu, *HINT: High-quality protein interactomes and their applications in understanding human disease.* BMC systems biology, 2012. **6**(1): p. 92.
176. Hagberg, A., D. Schult, and P. Swart, *Networkx: Python software for the analysis of networks.* Mathematical Modeling and Analysis, Los Alamos National Laboratory, 2005.
177. Van Driel, M.A., et al., *A text-mining analysis of the human phenome.* European journal of human genetics, 2006. **14**(5): p. 535.
178. Moal, I.H. and J. Fernández-Recio, *SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.* Bioinformatics, 2012. **28**(20): p. 2600-2607.
179. Benedix, A., et al., *Predicting free energy changes using structural ensembles.* Nature methods, 2009. **6**(1): p. 3.
180. Stein, A., R.B. Russell, and P. Aloy, *3did: interacting protein domains of known three-dimensional structure.* Nucleic acids research, 2005. **33**(suppl_1): p. D413-D417.
181. Mosca, R., A. Céol, and P. Aloy, *Interactome3D: adding structural details to protein networks.* Nature methods, 2013. **10**(1): p. 47.
182. Callaway, D.S., et al., *Network robustness and fragility: Percolation on random graphs.* Physical review letters, 2000. **85**(25): p. 5468.
183. Iyer, S., et al., *Attack robustness and centrality of complex networks.* PloS one, 2013. **8**(4): p. e59613.
184. Zeng, A. and W. Liu, *Enhancing network robustness against malicious attacks.* Physical Review E, 2012. **85**(6): p. 066130.
185. Albert, R., H. Jeong, and A.-L. Barabási, *Error and attack tolerance of complex networks.* nature, 2000. **406**(6794): p. 378.
186. Futreal, P.A., et al., *A census of human cancer genes.* Nature Reviews Cancer, 2004. **4**(3): p. 177.
187. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Scientific reports, 2013. **3**: p. 2650.
188. Consortium, I.C.G., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993.
189. Mantel, N., *Evaluation of survival data and two new rank order statistics arising in its consideration.* Cancer Chemother. Rep., 1966. **50**: p. 163-170.

190. Singleton, A.B., et al., *Towards a complete resolution of the genetic architecture of disease.* Trends in genetics, 2010. **26**(10): p. 438-442.
191. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallographica Section D: Biological Crystallography, 1998. **54**(6): p. 1078-1084.
192. Tang, X., et al., *Predicting diabetes mellitus genes via protein-protein interaction and protein subcellular localization information.* BMC genomics, 2016. **17**(4): p. 433.
193. Stevens, A., et al., *Network analysis: a new approach to study endocrine disorders.* Journal of molecular endocrinology, 2014. **52**(1): p. R79-R93.
194. Vyas, R., et al., *Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis.* Computational biology and chemistry, 2016. **65**: p. 37-44.
195. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC bioinformatics, 2003. **4**(1): p. 2.
196. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.
197. Tritos, N.A. and C.S. Mantzoros, *Syndromes of severe insulin resistance.* The Journal of Clinical Endocrinology & Metabolism, 1998. **83**(9): p. 3025-3030.
198. Perkins, J.R., et al., *Transient protein-protein interactions: structural, functional, and network properties.* Structure, 2010. **18**(10): p. 1233-1243.
199. Zhu, H., et al., *NOXclass: prediction of protein-protein interaction types.* BMC bioinformatics, 2006. **7**(1): p. 27.
200. Acuner Ozbabacan, S.E., et al., *Transient protein–protein interactions.* Protein engineering, design and selection, 2011. **24**(9): p. 635-648.
201. Nikiforova, M.N., et al., *RAS point mutations and PAX8-PPARγ rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma.* The Journal of Clinical Endocrinology & Metabolism, 2003. **88**(5): p. 2318-2326.
202. Kerr, B., et al., *Genotype-phenotype correlation in Costello syndrome: HRAS mutation analysis in 43 cases.* Journal of medical genetics, 2006. **43**(5): p. 401-405.
203. Boriack-Sjodin, P.A., et al., *The structural basis of the activation of Ras by Sos.* Nature, 1998. **394**(6691): p. 337.
204. Hatzivassiliou, G., et al., *RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth.* Nature, 2010. **464**(7287): p. 431.
205. Jin, T., et al., *RAF inhibitors promote RAS-RAF interaction by allosterically disrupting RAF autoinhibition.* Nature communications, 2017. **8**(1): p. 1211.
206. Frischmeyer, P.A. and H.C. Dietz, *Nonsense-mediated mRNA decay in health and disease.* Human molecular genetics, 1999. **8**(10): p. 1893-1900.
207. Kirst, C., M. Timme, and D. Battaglia, *Dynamic information routing in complex networks.* Nature communications, 2016. **7**: p. 11061.
208. Ruffalo, M. and Z. Bar-Joseph, *Protein interaction disruption in cancer.* BMC cancer, 2019. **19**(1): p. 370.
209. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nature methods, 2008. **5**(1): p. 16-18.
210. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31-46.
211. Shendure, J., et al., *DNA sequencing at 40: past, present and future.* Nature, 2017. **550**(7676): p. 345-353.
212. Consortium, I.H., *The international HapMap project.* Nature, 2003. **426**(6968): p.

789.
213. Consortium, G.P., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
214. van Rooij, J.G., et al., *Population-specific genetic variation in large sequencing data sets: why more data is still better.* European Journal of Human Genetics, 2017. **25**(10): p. 1173-1175.
215. Cui, H., N. Zhao, and D. Korkin, *Multilayer View of Pathogenic SNVs in Human Interactome through In Silico Edgetic Profiling.* Journal of molecular biology, 2018. **430**(18): p. 2974-2992.
216. Benedix, A., et al., *Predicting free energy changes using structural ensembles.* Nature methods, 2009. **6**(1): p. 3-4.
217. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallographica Section D: Biological Crystallography, 1998. **54**(6): p. 1078-1084.
218. Fiser, A. and A. Šali, *Modeller: generation and refinement of homology-based protein structure models*, in *Methods in enzymology*. 2003, Elsevier. p. 461-491.
219. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nature Reviews Cancer, 2018. **18**(11): p. 696-705.
220. Dietlein, F., et al., *Identification of cancer driver genes based on nucleotide context.* Nature Genetics, 2020. **52**(2): p. 208-218.
221. Zhu, J., et al., *On the nature of human housekeeping genes.* Trends in genetics, 2008. **24**(10): p. 481-484.
222. Butte, A.J., V.J. Dzau, and S.B. Glueck, *Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues".* Physiological genomics, 2001. **7**(2): p. 95-96.
223. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes, revisited.* TRENDS in Genetics, 2013. **29**(10): p. 569-574.
224. Kimura, M., *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): p. 624-626.
225. Kinsella, R.J., et al., *Ensembl BioMarts: a hub for data retrieval across taxonomic space.* Database, 2011. **2011**.
226. *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
227. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
228. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**(1): p. 1-13.
229. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**(1): p. 44.
230. Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, 1977: p. 35-41.
231. Koschützki, D. and F. Schreiber, *Centrality analysis methods for biological networks and their application to gene regulatory networks.* Gene regulation and systems biology, 2008. **2**: p. GRSB. S702.
232. Mizruchi, M.S., et al., *Techniques for disaggregating centrality scores in social networks.* Sociological methodology, 1986. **16**: p. 26-48.
233. Puzis, R., et al., *Augmented betweenness centrality for environmentally aware*

traffic monitoring in transportation networks. Journal of Intelligent Transportation Systems, 2013. **17**(1): p. 91-105.
234. Newman, M.E., *A measure of betweenness centrality based on random walks.* Social networks, 2005. **27**(1): p. 39-54.
235. Latora, V. and M. Marchiori, *Efficient behavior of small-world networks.* Physical review letters, 2001. **87**(19): p. 198701.
236. Cao, M., et al., *Going the distance for protein function prediction: a new distance metric for protein interaction networks.* PloS one, 2013. **8**(10).
237. Choobdar, S., et al., *Assessment of network module identification across complex diseases.* bioRxiv, 2019: p. 265553.
238. Cao, M., et al., *New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence.* Bioinformatics, 2014. **30**(12): p. i219-i227.
239. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
240. Subramanian, S., *The abundance of deleterious polymorphisms in humans.* Genetics, 2012. **190**(4): p. 1579-1583.
241. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic acids research, 2005. **33**(suppl_1): p. D514-D517.
242. Macgregor, S., et al., *Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis.* Human molecular genetics, 2009. **18**(3): p. 580-593.
243. Agarwal, D.P. and H.W. Goedde, *Human aldehyde dehydrogenases: Their role in alcoholism.* Alcohol, 1989. **6**(6): p. 517-523.
244. Wall, T.L., et al., *Hangover symptoms in Asian Americans with variations in the aldehyde dehydrogenase (ALDH2) gene.* Journal of studies on alcohol, 2000. **61**(1): p. 13-17.
245. Cook, T.A., et al., *Associations of ALDH2 and ADH1B genotypes with response to alcohol in Asian Americans.* Journal of Studies on Alcohol, 2005. **66**(2): p. 196-204.
246. Eng, M.Y., S.E. Luczak, and T.L. Wall, *ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review.* Alcohol Research & Health, 2007. **30**(1): p. 22.
247. Ye, L., *Alcohol and the Asian flush reaction.* SURG Journal, 2009. **2**(2): p. 34-39.
248. Matoba, N., et al., *GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits.* Nature human behaviour, 2020. **4**(3): p. 308-316.
249. Shankarkumar, U., *The human leukocyte antigen (HLA) system.* International Journal of Human Genetics, 2004. **4**(2): p. 91-103.
250. Hildebrand, W.H., et al., *HLA-B15: a widespread and diverse family of HLA-B alleles.* Tissue antigens, 1994. **43**(4): p. 209-218.
251. Bihl, F., et al., *Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses.* The Journal of Immunology, 2006. **176**(7): p. 4094-4101.
252. Williams, F., et al., *Analysis of the distribution of HLA-B alleles in populations from five continents.* Human immunology, 2001. **62**(6): p. 645-650.
253. Khan, M.A., *HLA-B27 and its subtypes in world populations.* Current opinion in rheumatology, 1995. **7**(4): p. 263-269.
254. Kong, M.H., et al., *Systematic review of the incidence of sudden cardiac death in the United States.* Journal of the American College of Cardiology, 2011. **57**(7): p. 794-801.

255. Offerhaus, J.A., C.R. Bezzina, and A.A. Wilde, *Epidemiology of inherited arrhythmias.* Nature Reviews Cardiology, 2019: p. 1-11.
256. Hennessey, J.A., et al., *FGF12 is a candidate Brugada syndrome locus.* Heart rhythm, 2013. **10**(12): p. 1886-1894.
257. Ruan, Y., N. Liu, and S.G. Priori, *Sodium channel mutations and arrhythmias.* Nature Reviews Cardiology, 2009. **6**(5): p. 337.
258. Makita, N., et al., *Novel calmodulin mutations associated with congenital arrhythmia susceptibility.* Circulation: Cardiovascular Genetics, 2014. **7**(4): p. 466-474.
259. Musa, H., et al., *SCN5A variant that blocks fibroblast growth factor homologous factor regulation causes human arrhythmia.* Proceedings of the National Academy of Sciences, 2015. **112**(40): p. 12528-12533.
260. Fragoza, R., et al., *Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations.* Nature communications, 2019. **10**(1): p. 1-15.
261. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-291.
262. Parrish, J.R., K.D. Gulyas, and R.L. Finley Jr, *Yeast two-hybrid contributions to interactome mapping.* Current opinion in biotechnology, 2006. **17**(4): p. 387-393.
263. Fields, S., *High-throughput two-hybrid analysis: The promise and the peril.* The FEBS journal, 2005. **272**(21): p. 5391-5399.
264. Hengen, P.H., *Methods and reagents: False positives from the yeast two-hybrid system.* Trends in biochemical sciences, 1997. **22**(1): p. 33-34.
265. Mittl, P.R. and M.G. Grütter, *Structural genomics: opportunities and challenges.* Current opinion in chemical biology, 2001. **5**(4): p. 402-408.
266. Vitkup, D., et al., *Completeness in structural genomics.* nature structural biology, 2001. **8**(6): p. 559-566.
267. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-96.
268. Cui, H. and D. Korkin. *Effect-specific analysis of pathogenic SNVs in human interactome: Leveraging edge-based network robustness*. in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016. IEEE.
269. Kataka, E., et al., *Edgetic perturbation signatures represent known and novel cancer biomarkers.* Scientific reports, 2020. **10**(1): p. 1-16.
270. Mosca, R., et al., *dSysMap: exploring the edgetic role of disease mutations.* Nature methods, 2015. **12**(3): p. 167-168.
271. Lamparter, D., et al., *Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics.* PLoS computational biology, 2016. **12**(1): p. e1004714.
272. Wang, L., et al., *An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies.* Bioinformatics, 2011. **27**(5): p. 686-692.
273. Consortium, G.P., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56.
274. Kamisetty, H., et al., *Accounting for conformational entropy in predicting binding free energies of protein-protein interactions.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(2): p. 444-462.
275. Mosca, R., et al., *3did: a catalog of domain-based interactions of known three-*

dimensional structure. Nucleic acids research, 2013. **42**(D1): p. D374-D379.
276. Russell, R.B., et al., *A structural perspective on protein–protein interactions.* Current opinion in structural biology, 2004. **14**(3): p. 313-324.
277. Wang, R.-S. and J. Loscalzo, *Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications.* Journal of molecular biology, 2018. **430**(18): p. 2939-2950.
278. Rivals, I., et al., *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics, 2006. **23**(4): p. 401-407.
279. Qian, Y., et al. *Identifying disease associated genes by network propagation*. in *BMC systems biology*. 2014. BioMed Central.
280. Cao, M., et al., *Going the distance for protein function prediction: a new distance metric for protein interaction networks.* PloS one, 2013. **8**(10): p. e76339.
281. Craddock, N. and P. Sklar, *Genetics of bipolar disorder.* The Lancet, 2013. **381**(9878): p. 1654-1662.
282. Belmaker, R., *Bipolar disorder.* New England Journal of Medicine, 2004. **351**(5): p. 476-486.
283. Fazel, S., et al., *Schizophrenia and violence: systematic review and meta-analysis.* PLoS medicine, 2009. **6**(8): p. e1000120.
284. Lichtenstein, P., et al., *Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study.* The Lancet, 2009. **373**(9659): p. 234-239.
285. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-752.
286. Ripke, S., et al., *Genome-wide association study identifies five new schizophrenia loci.* Nature genetics, 2011. **43**(10): p. 969.
287. Sklar, P., et al., *Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4.* Nature genetics, 2011. **43**(10): p. 977.
288. Stahl, E.A., et al., *Genome-wide association study identifies 30 loci associated with bipolar disorder.* Nature genetics, 2019. **51**(5): p. 793.
289. Pinto, D., et al., *Convergence of genes and cellular pathways dysregulated in autism spectrum disorders.* The American Journal of Human Genetics, 2014. **94**(5): p. 677-694.
290. Pirooznia, M., et al., *High-throughput sequencing of the synaptome in major depressive disorder.* Molecular psychiatry, 2016. **21**(5): p. 650.
291. Fabbri, C. and A. Serretti, *Pharmacogenetics of major depressive disorder: top genes and pathways toward clinical applications.* Current psychiatry reports, 2015. **17**(7): p. 50.
292. Castillo, P.E., et al., *RIM1α is required for presynaptic long-term potentiation.* Nature, 2002. **415**(6869): p. 327.
293. Sisodiya, S.M., et al., *Genetic enhancement of cognition in a kindred with cone–rod dystrophy due to RIMS1 mutation.* Journal of medical genetics, 2007. **44**(6): p. 373-380.
294. Stessman, H.A., et al., *Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases.* Nature genetics, 2017. **49**(4): p. 515.
295. Radulescu, E., et al., *Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain.* Molecular psychiatry, 2018: p. 1.

296. Winkler, C. and S. Yao, *The midkine family of growth factors: diverse roles in nervous system formation and maintenance.* British journal of pharmacology, 2014. **171**(4): p. 905-912.
297. Muramatsu, T., *Midkine: a promising molecule for drug development to treat diseases of the central nervous system.* Current pharmaceutical design, 2011. **17**(5): p. 410-423.
298. Rao, J., et al., *Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients.* Translational psychiatry, 2012. **2**(7): p. e132.
299. Sharma, R.P., et al., *Valproic acid and chromatin remodeling in schizophrenia and bipolar disorder: preliminary results from a clinical population.* Schizophrenia research, 2006. **88**(1-3): p. 227-231.
300. Borsboom, D., *A network theory of mental disorders.* World psychiatry, 2017. **16**(1): p. 5-13.
301. Chou, K.-C., *Structural bioinformatics and its impact to biomedical science.* Current medicinal chemistry, 2004. **11**(16): p. 2105-2134.
302. Mosca, R., et al., *Towards a detailed atlas of protein–protein interactions.* Current opinion in structural biology, 2013. **23**(6): p. 929-940.
303. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Current protocols in human genetics, 2013: p. 7.20. 1-7.20. 41.
304. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic acids research, 2003. **31**(13): p. 3812-3814.
305. Vapnik, V. and A. Vashist, *A new learning paradigm: Learning using privileged information.* Neural networks, 2009. **22**(5-6): p. 544-557.
306. Shiao, H.-T. and V. Cherkassky. *Learning using privileged information (LUPI) for modeling survival data*. in *Neural Networks (IJCNN), 2014 International Joint Conference on*. 2014. IEEE.
307. Xu, X., W. Li, and D. Xu, *Distance metric learning using privileged information for face verification and person re-identification.* IEEE transactions on neural networks and learning systems, 2015. **26**(12): p. 3150-3162.
308. Korte, A. and A. Farlow, *The advantages and limitations of trait analysis with GWAS: a review.* Plant methods, 2013. **9**(1): p. 29.
309. Bush, W.S. and J.H. Moore, *Genome-wide association studies.* PLoS computational biology, 2012. **8**(12): p. e1002822.
310. Begum, F., et al., *Comprehensive literature review and statistical considerations for GWAS meta-analysis.* Nucleic acids research, 2012. **40**(9): p. 3777-3784.
311. Chen, Y.-C., et al., *A hybrid likelihood model for sequence-based disease association studies.* PLoS genetics, 2013. **9**(1): p. e1003224.
312. Barrett, J.C., et al., *Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes.* Nature genetics, 2009. **41**(6): p. 703.
313. Chasman, D.I., et al., *Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function.* Human molecular genetics, 2012. **21**(24): p. 5329-5343.
314. Consortium, W.T.C.C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661.
315. Kluijtmans, L.A., et al., *The molecular basis of cystathionine β-synthase deficiency in Dutch patients with homocystinuria: effect of CBS genotype on biochemical and clinical phenotype and on response to treatment.* The American Journal of Human Genetics, 1999. **65**(1): p. 59-67.

316. Kozich, V., R. de Franchis, and J.P. Kraus, *Molecular defect in a patient with pyridoxine-responsive homocystinuria.* Human molecular genetics, 1993. **2**(6): p. 815-816.
317. Aral, B., et al., *Two novel mutations (K384E and L539S) in the C-terminal moiety of the cystathionine ²-synthase protein in two French pyridoxine-responsive homocystinuria patients.* Human mutation, 1997. **9**(1): p. 81.
318. Goossens, N., et al., *Cancer biomarker discovery and validation.* Translational cancer research, 2015. **4**(3): p. 256.
319. Wang, Y.-C. and B.-S. Chen, *Integrated cellular network of transcription regulations and protein-protein interactions.* BMC Systems Biology, 2010. **4**(1): p. 20.
320. McShane, L.M., et al., *Criteria for the use of omics-based predictors in clinical trials.* Nature, 2013. **502**(7471): p. 317-320.
321. Mason, C.E., S.G. Porter, and T.M. Smith, *Characterizing Multi-omic Data in Systems Biology*, in *Systems Analysis of Human Multigene Disorders*. 2014, Springer. p. 15-38.
322. Domany, E., *Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives.* Cancer research, 2014. **74**(17): p. 4612-4621.
323. Wang, W.Y., et al., *Genome-wide association studies: theoretical and practical concerns.* Nature Reviews Genetics, 2005. **6**(2): p. 109-118.
324. Parker, L.A., et al., *Methodological deficits in diagnostic research using '-omics' technologies: evaluation of the QUADOMICS tool and quality of recently published studies.* PloS one, 2010. **5**(7): p. e11419.