Role of network topology based methods in discovering novel gene-phenotype associations

Emre Güney

Department of Experimental and Health Sciences

PhD Thesis

Supervisor: Baldomera Oliva Miguel

Barcelona, 2012

For those who strive for the self-conscious selflessness.

Acknowledgements

I would like to start by thanking to my family –my father, mother and elder brother– for their everlasting support. As I grow older, the influence of the family on shaping one's character has come to my notice more and more.

Of course, one of the main actors who made this thesis possible is my advisor, *Baldo Oliva*, to whom I am truly indebted. Not only he provided the scientific guidance whenever I needed, but also he never hesitated helping me in any non-phd-related topic. The doctoral process could be challenging and less engaging from time to time and without his encouragement it would have been a lot harder. Cheers, Baldo!

I am grateful to the members of the thesis tribunal, *Ben Lehner*, *Nuria López Bigas* and *Patrick Aloy* for their consideration and careful evaluation.

I can not appreciate enough the support of my colleagues in SBI lab, in particular, *Aggeliki, Alessandra, Danielino, David, Jascha, Jaume, Javi, Joan* and *Ori*. I would like to also thank to *Trey Ideker* and his post-doc *Janusz Dutkowski* for giving me the opportunity to experience a different research perspective during my research stay in San Diego. Moreover, I would like to mention my gratitude to friends I met there including *David, Caro(s), Eley, Gjergji, Hyunchul, Mustafa* and *Qi*.

I owe sincere thankfulness to my beloved friends, with whom I shared most of my Barcelonian life; *Alba(s), Ana, Anna, Andrey, Billur, Besray, Gio, Güneş, Ingo, Kashif, Jelena, Jonas, Kader, Michi, Michael, Milica, Nils, Onuralp, Özgen, Sinan, Sonja(s), Walid, Zina* as well as my good-old friends; *Aslı(s), Ayça, Erhan, Seçil, Ufuk, Yiğitcan*. A special thanks goes to *Alice* who did a great job in proof-reading the text. The list of people I am indebted to for offering their friendship goes a long way and for those I fail to include here, I thank you all!

Finally, I would like to acknowledge "Generalitat de Catalunya, L'Agència de Gestió d'Ajuts Universitaris i de Recerca" (AGAUR) for supporting me through FI and BE fellowships.

Abstract

The cell is governed by the complex interactions among various types of biomolecules. Coupled with environmental factors, variations in DNA can cause alterations in normal gene function and lead to a disease condition. Often, such disease phenotypes involve coordinated dysregulation of multiple genes that implicate inter-connected pathways. Towards a better understanding and characterization of mechanisms underlying human diseases, here, I present GUILD, a network-based disease-gene prioritization framework. GUILD associates genes with diseases using the global topology of the protein-protein interaction network and an initial set of genes known to be implicated in the disease. Furthermore, I investigate the mechanistic relationships between disease-genes and explain the robustness emerging from these relationships. I also introduce GUILDify, an online and user-friendly tool which prioritizes genes for their association to any user-provided phenotype. Finally, I describe current state-of-the-art systems-biology approaches where network modeling has helped extending our view on diseases such as cancer.

Resum

La cèl·lula es regeix per interaccions complexes entre diferents tipus de biomolècules. Juntament amb factors ambientals, variacions en el DNA poden causar alteracions en la funció normal dels gens i provocar malalties. Sovint, aquests fenotips de malaltia involucren una desregulació coordinada de múltiples gens implicats en vies interconnectades. Per tal de comprendre i caracteritzar millor els mecanismes subjacents en malalties humanes, en aquesta tesis presento el programa GUILD, una plataforma que prioritza gens relacionats amb una malaltia en concret fent ús de la topologia de xarxe. A partir d'un conjunt conegut de gens implicats en una malaltia, GUILD associa altres gens amb la malaltia mitjançant la topologia global de la xarxa d'interaccions de proteïnes. A més a més, analitzo les relacions mecanístiques entre gens associats a malalties i explico la robustesa es desprèn d'aquesta anàlisi. També presento GUILDify, un servidor web de fácil ús per la priorització de gens i la seva associació a un determinat fenotip. Finalment, descric els mètodes més recents en què el model·latge de xarxes ha ajudat a extendre el coneixement sobre malalties complexes, com per exemple a càncer.

Preface

Science and technology have been continuously pushing the limits of our understanding of our habitat and our lives. Not only they change the way we perceive the universe but undoubtedly they also redefine how we "survive" in it. Consequently, we inform, express, entertain or even nurture ourselves in different ways than we used to do[1].

Rather ironically, science itself has taken its share from the change. Multi-disciplinary fields (such as computational biology) have emerged and large scale experimental and computational techniques have become increasingly available. Due to the accumulative nature of science, often one may feel that the problem he/she tries to address is a highly specific one, as opposed to key inventions we have witnessed over the past centuries. Furthermore, there are many more researchers, articles and journals today then there has ever been, making the criteria for the evaluation of the global impact of the research conducted harder to set than it was done before[2].

I can not say that I was fully aware of what was expecting me when I decided to embark a carrier on computational biology research several years ago, after I graduated from the university with a degree in computer engineering. Biology has always attracted me and I thought it was a brilliant idea to combine informatics with biology when I first heard about bioinformatics. Then, all of a sudden, I found myself in the middle of a huge jungle, the jungle of computational biology. The boat I have taken on the river passing through this jungle brought me to the laboratory of Baldo Oliva, where the preliminary ideas of this thesis was born on the top of recent findings suggesting that the genes encoding proteins that interact with each other tend to share similar functions or tend to be involved in similar diseases. There started another journey within the journey. Though initially was a "small-scale" project, developing algorithms to assess the relatedness of genes within the context of protein-protein interaction network and its implications on the characterization of diseases ended up being the main commitment of this thesis.

During the fruitful years of graduate studies, at least I have learnt one thing for sure: *as long as you do something you like to do and you spend enough effort to rationalize the concepts you are working on, failure is not an option*. And this is probably the most important thing I have learnt. Another lesson I have taken home with me is that Parkinson's law[3] also applies in the case of scientific works. That is, apart from procrastination, new questions to be addressed are added into the project, to fill the time allocated for the project.

All in all, it has been a great pleasure to have worked on a subject that I was interested in and I am glad that I have the opportunity to report what I have done during the past four or so years in this thesis. I hope you find it handy.

---

[1] Though, these changes are often coined with the word "improvement", –taking into account the overall effect they cause on the (dis)order in the nature– I argue that there are actually few cases where the world has unambiguously became a better place for everyone.

[2] Still, "thanks" to the overwhelming consumption capacity of mankind, relatively small scientific and technological innovations such as the ones in the fields of hardware development, search engine optimization, online marketing and social networks have had a huge impact on our daily lives.

[3] Proposed by Cyril Northcote Parkinson in 1955, Parkinson's law reads as follows: *work expands to fill the time allocated to it*.

Contents

List of figures **xi**

List of tables **xiii**

I Overview **1**

1 Introduction **3**

1.1 From genotype to phenotype 3

1.1.1 Variations in human genome 4

1.1.2 Human genetic variants and diseases 5

1.1.3 Disease-gene association databases 7

1.1.4 Disease-gene prioritization 9

1.2 Systems biology 13

1.2.1 Protein-protein interaction (PPI) networks 14

1.2.2 Understanding disease mechanisms through PPI networks . . . 14

1.2.3 Robustness as an emergent property of biological networks . . . 15

1.2.4 Cancer studies in the era of systems biology 16

1.3 Motivation of this thesis 16

II Objectives **19**

III Results **23**

2 Exploiting Protein-Protein Interaction Networks for Genome-wide Disease-Gene Prioritization **25**

3 Robustness of network-based disease gene prioritization methods points to pathophenotypic plasticity **57**

4 GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms 83

5 Understanding Cancer Progression Using Protein Interaction Networks 93

IV Discussion 119

V Conclusions 129

VI Appendix 133

A BIANA: A software framework for compiling biological interactions and analyzing networks 135

B Extending signalling pathways with protein-interaction networks. Application to apoptosis 137

C Networks of Protein-Protein Interactions: from uncertainty to molecular details 139

D Identifying genes involved in human cell fate determination 143

List of Figures

1.1 Examples of genetic variation in human 4
1.2 Overview of a genome-wide association study 8

List of Tables

1.1 Methods for detecting human genetic variants 6
1.2 Available data repositories for genetic variants and disease-gene associations 10
1.3 Available disease-gene prioritization tools 12
1.4 PPI detection methods 14

Part I

Overview

Chapter 1

Introduction

No two humans, including monozygotic twins, are genetically identical. The reason for this is the variation in genetic sequence (i.e. mutations during development in the case of monozygotic twins). Although such variations flourish genetic diversity, not all of them are beneficial for the organism. Some mutations, in combination with environmental factors, can disrupt the complex machinery of the cell and cause functional abnormalities. The past decade has witnessed dramatic advances in genome sequencing and a substantial shift in the number of genetic association studies, therefore strengthening our understanding behind disease phenotypes. Nevertheless, the genetic elements and mechanisms underlying diseases are still not very well characterized. Pathophenotypic characterization remains as a complex problem, mostly due to the pathophenotype being the outcome of the perturbations in the inter-connected pathways, where products of typically more than one gene co-operate through various mechanisms.

This thesis aims to extend our understanding of disease phenotypes incorporating information encoded in the protein-protein interaction network. In this chapter, I will briefly go over basic concepts related to the subject. In the following chapter, I will introduce a network-based phenotypic characterization method and demonstrate that it outperforms similar methods using protein-protein interactions (PPIs) when applied to the problem of disease-gene prioritization in human. Then, I will investigate the behaviour of the network-based prioritization methods against random perturbations and show that for several diseases, disease-gene annotations are easier to be recovered with network-based prioritization methods even when as much as half of the interaction network is perturbed. Next, I will present a web-server version of the prioritization method that can be run by entering only some keywords defining the phenotype. In the last chapter, I will explain how protein-protein interaction networks have helped understanding cancer progression. Finally, I will conclude with a discussion of the findings presented in this thesis.

1.1 From genotype to phenotype

Genotype of the organism gives rise to the phenotype. *Genotype* is the inheritable information coded within the organism. *Phenotype* is the physical manifestation –anything that is part of the observable structure, function or behavior– of a living organism.

In genetics, *allele* is one of the possible forms of a gene (or a genetic locus). Humans, like most multicellular organisms are *diploid* which means that there are two sets of chromosomes in human somatic cells. In each of these chromosomes resides one allele (a copy of each gene). If the same, the two alleles (in the two chromosomes) are *homozygotes*, if different they are *heterozygotes*. Individuals of a population typically have multiple alleles at each locus in their chromosomes yielding in different observable phenotypes (*polymorphism*). For example, blood type in humans is a three-allele phenotype (A, B, 0) where the alleles in two chromosomes in one individual defines the three phenotype (A, B, 0) through one of the six possible genotypes (AA, A0, BB, B0, AB, 00).

Predicting the phenotype of an organism (or the phenotype of the cells taken from it) is a complex problem. To begin with, our knowledge on the genetic elements mediating the phenotype is fairly limited. Furthermore, due to synergistic nature of genes involved in biological processes, the same mutation may even lead to different phenotypes for phenotypically identical individuals of the same organism given two different genetic backgrounds [Dowell et al., 2010]. Yet, understanding the variations in DNA sequence is the first step towards characterizing phenotypes, in particular disease phenotypes.

1.1.1 Variations in human genome

Depending on their frequency of occurrence in the population, human genetic variants can be either common or rare. *Common variants*, are the genetic variants where the minor allele (less common allele) frequency (MAF) is 1% or higher in the population. *Rare variants*, on the other hand, have a MAF of less than 1%. Polymorphisms are considered as common variants rather than rare variants since they are mediated by inheritance and thus their prevalence in the population is higher (equal or higher than 1%) than one would expect as a result of random mutations.

Alternatively, the human genetic variations can also be grouped in two main categories with respect to the change in the nucleotide composition they incur (Figure 1.1).

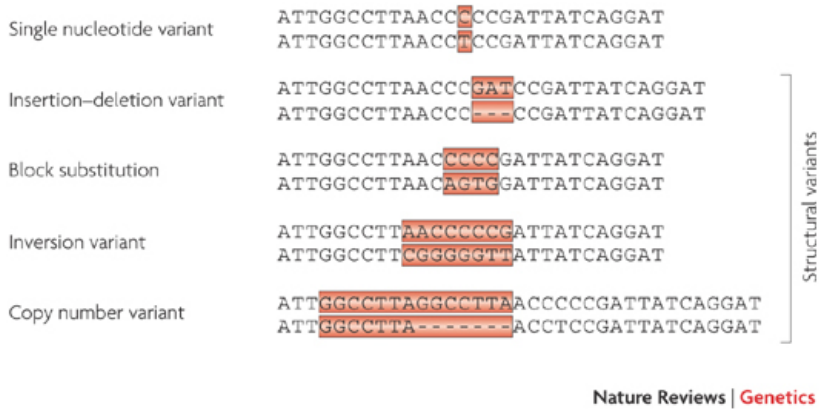


Figure 1.1: Examples of genetic variation in human (reproduced from [Frazer et al., 2009])

Single nucleotide variants Single nucleotide variants consist of single point mutations, that is alteration of a single nucleotide (A,T,G or C) in the DNA sequence. Single nucleotide substitution, also known as *single nucleotide polymorphism (SNP)*, are among the most common human genetic variants. The human genome consists of 3 billion nucleotide base pairs and it is estimated to contain at least 11 million SNPs, all occurring with a MAF of over 1% and some 7 million of them occurring with a MAF of over 5% [Frazer et al., 2009]. Nonetheless, most of the SNPs are neutral (has no effect on the phenotype) and only a small portion of SNPs (around 5%) are functional. The alleles of SNPs located in the close genomic intervals tend to be inherited together yielding in linkage disequilibrium (non-random association of alleles).

Structural variants All DNA sequence differences between variations individuals of a population, that are not single nucleotide variants are broadly defined as structural variants. Typically, these variants affect a sequence length of 1kb (=kilo base) to 3Mb (=mega base). Structural variants are estimated to account for more than 20% of all genetic variants in humans covering more than 70% of the variant bases [Frazer et al., 2009]. Structural variants are mainly categorized as follows:

- Copy number variation (CNV): A segment of DNA that is present at different number of copies in comparison to a reference genome. CNVs occur due to large (>1kb) insertions, deletions or duplications in the DNA sequence. CNVs are estimated to account for roughly 13% of human genomic DNA [Stankiewicz and Lupski, 2010].
- Insertion-deletion (indels): Insertion or deletion of several (less than 1kb) nucleotides.
- Inversion of DNA sequence: A segment of DNA that is reversed in orientation with respect to the rest of the chromosome.
- Translocation: A change in position of a chromosomal segment within a genome. Translocations do not change the total DNA content.

It is worth mentioning that other schemes for the classification of genetic variations exist. These include classification by effect on function (loss- or gain-of-function mutations), by effect on fitness (harmful, beneficial or neutral mutations), by contribution to the phenotypic variation (neutral, near-neutral or non-neutral) or by impact on protein sequence (frameshift, nonsense, missense, neutral or silent mutations).

A wide variety of methods has been developed to detect human genetic variations experimentally [Feuk et al., 2006]. These methods can be broadly categorized with respect to their coverage: genome-wide scans (covering all the DNA sequence) and targeted scans (directed to a certain subsection of the DNA sequence). See Table 1.1 for an overview of the experimental methods used to identify human genetic variants.

1.1.2 Human genetic variants and diseases

Most human diseases are complex genetic traits that are influenced by multiple genetic and environmental factors. During the past decades, a substantial amount of effort has

Table 1.1: Methods for detecting human genetic variants (adapted from [Feuk et al., 2006])*.

	Method	Small sequence variants (<1 kb)	CNV	Translocation & Inversion
Genome-wide scans	Karyotyping	No	Yes (>3 Mb)	Yes (>3 Mb)
	Clone-based array-CGH	No	Yes (>50 kb)	No
	Oligonucleotide-based array-CGH	No	Yes (>35 kb)	No
	SNP array	Yes (SNPs)	Yes	No
	Sequence-assembly comparison	Yes	Yes	Yes
	Clone paired-end sequencing	No	Yes **	Yes
Targeted scans	Microsatellite genotyping	Yes	Yes (deletions)	No
	MAPH	Yes	Yes	No
	MLPA	Yes	Yes	No
	QMPSF	Yes	Yes	No
	Real-time qPCR	Yes	Yes	No
	FISH	No	Yes	Yes
	Southern blotting	Yes	Yes	Yes

* CGH, comparative genome hybridization; FISH, fluorescence in situ hybridization (including metaphase, interphase and fibre); MAPH, multiplex amplifiable probe hybridization; MLPA, multiplex ligation-dependent probe amplification; QMPSF, quantitative multiplex PCR of short fluorescent fragments; qPCR, quantitative PCR

** >8 kb of deletions; <40 kb of insertions

been exerted to delineate sequential variations in human DNA and their consequences on human biology [Altshuler et al., 2008]. These studies have established ties between genetic variations and diseases in human where the genetic variant contributes to the susceptibility for a diverse set of diseases. For example, a number of SNPs are known to affect susceptibility to many Mendelian and complex diseases including cancers, diabetes, asthma, obesity, schizophrenia, Parkinson disease, Alzheimer's Disease (see [Shastry, 2002, Hirschhorn et al., 2002] for two reviews on the topic). Furthermore, inversions in DNA have been identified to be involved in haemophilia, Hunter syndrome and muscular dystrophy [Feuk et al., 2006]. Copy number variations have also been found to be associated with different types of cancer, Crohn's disease, Alzheimer's Disease and neuropsychiatric conditions such as autism, schizophrenia [Stankiewicz and Lupski, 2010, Eichler et al., 2007, Jr and Scherer, 2008].

Recombination mapping is among the earliest methods to identify disease causing genes and mutations. Also known as linkage analysis, recombination mapping consists of genotyping the alleles of individuals within a family at particular genomic polymorphic loci (sites on each chromosome) and then checking whether a certain variant shows correlated segregation with a particular trait or disease. The genomic vicinity of the identified chromosome region is typically further genotyped (e.g., by positional cloning) to identify the exact casual locus for the trait or disease. The experimental design has an important role in defining the effectiveness of recombination mapping. In particular, *i)* the phenotype under consideration should be clearly definable (as opposed to intermediate phenotypes), *ii)* the phenotype should bear obvious genetic basis (i.e. supporting evidence from studies on twins) and *iii)* the population should be selected such that the genetic homogeneity is maximized and environmental factors are minimized (e.g., isolated populations) [Broeckel and Schork, 2004].

Recombination mapping has been used to identify genetic factors contributing significantly to several Mendelian or monogenic diseases such as cystic fibrosis and Huntington's disease and neurofibromatosis [Broeckel and Schork, 2004, Carlson et al., 2004]. However, its success in identifying casual genes in complex traits is shuttered due to the polygenic nature of these phenotypes. Complex disease phenotypes involve many genetic variants and contribution of a genetic variant to the phenotypes might be confounded by other genetic variants. Association studies offer a promising alternative to detect casual loci in human diseases by comparing the frequencies of genetic variants between affected and unaffected individuals [Hirschhorn et al., 2002]. Recently, the advancement in the sequencing techniques gave rise to genome-wide association studies (GWAS) where the association between the disease group (individuals bearing the phenotype) and the control group (healthy individuals) is investigated at genome level [Hirschhorn and Daly, 2005] (see Figure 1.2 for the overview of a typical GWAS). GWAS have been reported numerous novel associations for a wide range of common diseases and clinical conditions such as age-related macular degeneration, diabetes, obesity, inflammatory bowel disease, prostate cancer, breast cancer, colorectal cancer, rheumatoid arthritis, systemic lupus erythematosus, celiac disease, multiple sclerosis, glaucoma, gallstones, asthma, coronary heart disease, atrial fibrillation and restless leg syndrome as well as continuous traits such as lipid levels, height, fat mass, hair color, eye color, freckles and HIV viral set point [Altshuler et al., 2008, McCarthy et al., 2008].

Genetic association studies pose intrinsic limitations affecting the interpretation and applicability of the findings from these studies [Hirschhorn et al., 2002, Wang et al., 2005, Hirschhorn and Daly, 2005, Frazer et al., 2009]. Probably one of the most striking concerns attributed to genetic association studies is the dependence on the sample population and thus reproducibility. Sample size also plays as an important factor in the reliability of the outcome of the study. Furthermore, these studies often fail to identify rare high risk variants and modest risk variants with small effect sizes for which the contribution to the disease prevalence is more than rare high risk variants. The analysis of the findings is further complicated by the fact that a large portion of the identified loci does not correspond to part of the DNA that is transcribed. Even for the sequence variants that correspond to genes, most of these variants do not induce a change in the protein sequence. Such changes are predicted to rather have implications on the transcriptional or translational efficiency and on the gene expression [Hardy and Singleton, 2009, Frazer et al., 2009]. Additionally, so far, the association studies have not been able to take the effects of gene-gene and gene-environment interactions into account. Consequently, for most of the disease-gene associations annotated using genetic association studies explain much less than 10% of the genetic variance [Frazer et al., 2009]. Therefore, improved experimental design strategies and more comprehensive functional annotation methods have taken priority in the research community in order to explain the links between genomic intervals and complex traits [Freedman et al., 2011, Clarke et al., 2011].

1.1.3 Disease-gene association databases

The major data sources providing information on human genetic variants are given in Table 1.2. Among these, Online Mendelian Inheritance in Man (OMIM) [Amberger

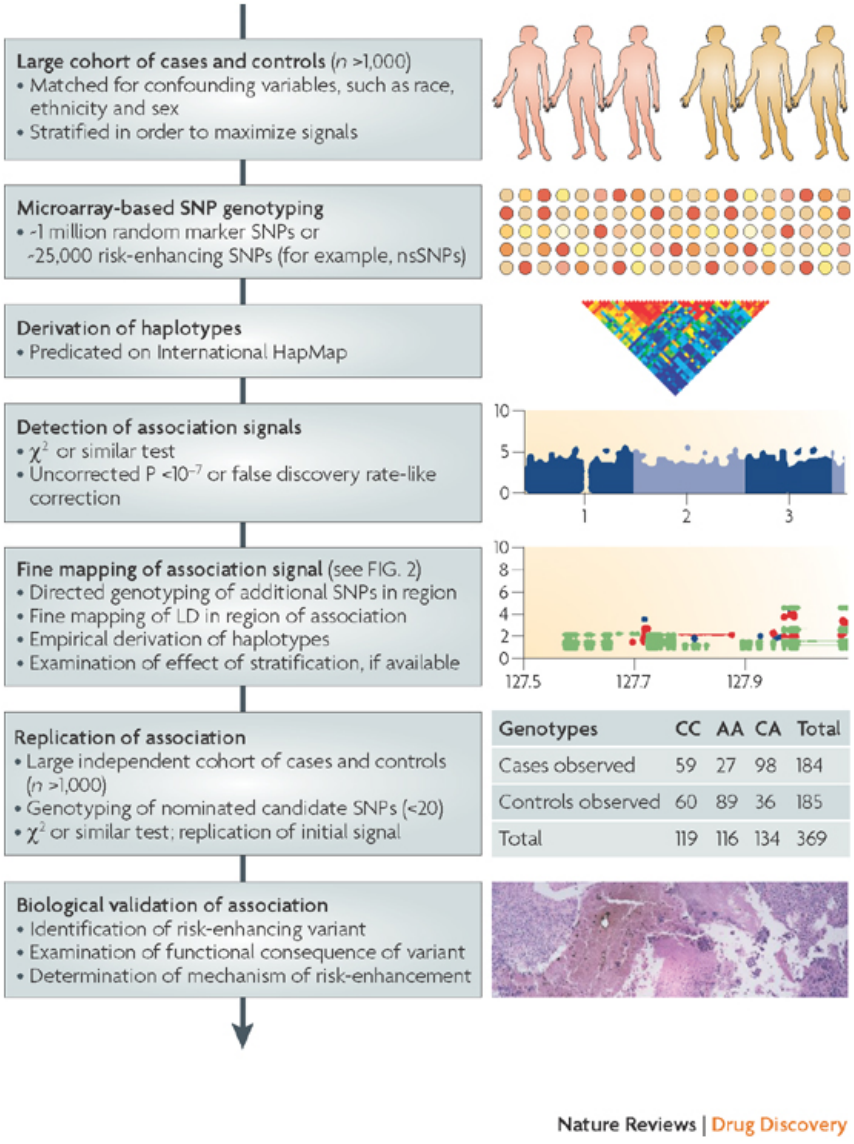


Figure 1.2: Overview of a genome-wide association study (GWAS) (taken from [Kingsmore et al., 2008])

et al., 2009], SwissVar [Mottaz et al., 2010], Human Gene Mutation Database (HGMD) [Stenson et al., 2009] and Genetic Association Database (GAD) [Becker et al., 2004] annotate the genetic variants and genes associated with the human disorders and diseases. These databases are primary source of curated disease-gene annotations reporting associations identified by experimental studies.

In particular, most studies rely on the data in OMIM [Amberger et al., 2009] since it is one of the most comprehensive, authoritative and up-to-date repositories on human genes and genetic disorders. The information in OMIM is expert curated and lists a reviewed overview of the mutations associated with the diseases along with the publications reported them. Phenotypic associations for genes can be downloaded from OMIM morbid map (`omim.org/downloads`), a file containing entries in the following format: disorder name, disorder MIM id (MIM is a six digit number describing phenotypes and genes in OMIM), known genes whose mutations are associated with the phenotype, MIM id of the gene whose association is biologically and historically most significant and the associated linkage interval. Similarly, GAD [Becker et al., 2004] provides valuable information on disease-gene associations cataloging disease-gene associations curated from genetic association studies. GAD collects findings of low significance in addition to those with high significance.

Several pharmacogenomic projects such as PharmaKGB [Hernandez-Boussard et al., 2007], The Comparative Toxicogenomics Database (CTD) [Davis et al., 2010] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al., 2012] compile information on disease-gene associations as well. Furthermore, computationally derived "meta databases" such as GeneCards [Safran et al., 2010], PhenoGO [Sam et al., 2009], PhenomicDB [Groth et al., 2007] and DisGeNet [Bauer-Mehren et al., 2010] integrate data from mutliple data resources. In addition to these sources, a number of large-scale projects sequencing various types of cancer catalogs somatic mutations in oncogenesis (see cancer genes and variants part in the Table 1.2).

1.1.4 Disease-gene prioritization

Experimental studies provide valuable sources of information on the genetic variants and the associated disorders, however often these studies identify a large genomic interval containing as many as thousands of genes. Pinpointing the casual disease-gene requires further experimental testing which increases the costs associated with identifying disease-genes substantially. During the past decade, numerous computational methods have been developed to find the most promising candidates among a list of possible genes. Named as disease-gene prediction/prioritization methods, these methods employ "guilt-by-association" principle. That is, the genes linked to known disease-genes somehow will be more likely to be implicated in a disease. Disease-gene prioritization methods narrow down the list of the candidates using various types of evidences are given below (Table 1.3). Different types of evidences used to associate genes with each other are described below (see [Kann, 2010, Tranchevent et al., 2011, Capriotti et al., 2012] for three extensive reviews).

Sequence / structure Common sequence and structural similarity give rise to homology and thus functional similarity. Disease phenotypes are more likely to involve genes with similar function, thus similarity at sequence and structural level

Table 1.2: Available data repositories for genetic variants and disease-gene associations (reproduced from [Capriotti et al., 2012]).

Database	**URL**
Short variations	
1000 Genomes	`www.1000genomes.org`
dbSNP	`www.ncbi.nlm.nih.gov/projects/SNP`
HapMap	`www.hapmap.org`
Structural variations	
dbVar	`www.ncbi.nlm.nih.gov/dbvar`
DGV	`projects.tcag.ca/variation`
DGVa	`www.ebi.ac.uk/dgva`
General variants associated with phenotypes	
HGMD	`www.hgmd.org`
OMIM	`www.omim.org`
SwissVar	`swissvar.expasy.org`
GWAS and other association studies	
dbGaP	`www.ncbi.nlm.nih.gov/gap`
EGA	`www.ebi.ac.uk/ega`
GAD	`geneticassociationdb.nih.gov`
NHGRI GWAS Catalog	`www.genome.gov/gwastudies`
Cancer genes and variants	
ICGC	`www.icgc.org`
COSMIC	`sanger.ac.uk/genetics/CGP/cosmic`
Cancer Gene Census	`sanger.ac.uk/genetics/CGP/Census`
Cancer Gene Index	`ncicb.nci.nih.gov/NCICB/projects/cgdcp`
TCGA	`cancergenome.nih.gov`
Pharmacogenomic genes and variants	
DrugBank	`drugbank.ca`
PharmGKB	`www.pharmgkb.org`
CTD	`ctdbase.org`
KEGG	`www.genome.jp/kegg`
Crowdsourced genes and variants	
Gene Wiki	`en.wikipedia.org/wiki/Portal:Gene\_Wiki`
SNPedia	`www.snpedia.com`
WikiGenes	`www.wikigenes.org`
Computationally-derived / meta databases	
GeneCards	`www.genecards.org`
PhenoGO	`www.phenogo.org`
PhenomicDB	`www.phenomicdb.de`
DisGeNet	`ibi.imim.es/DisGeNET/DisGeNETweb.html`

implies involvement in similar disease phenotypes. In particular, disease-genes could be distinguished from non-disease genes by their coding region length, sequence conservation, exon number, structural domains, sequence motifs, chromosomal location, (proteins') subcellular location.

Pathway involvement Disease genes affect various biological pathways consisting of genes that are interconnected with each other. The genes in the same pathway perform similar functions, therefore, genes pertained in a pathway implicated in a disease have an increased likelihood of being disease-related. In addition to pathways annotated in the literature (such as KEGG [Kanehisa et al., 2012], Reactome [Croft et al., 2011], BioCyc [Caspi et al., 2011], GenMAPP [Salomonis et al., 2007] and MSigDB [Liberzon et al., 2011]), regulatory networks (e.g., functional links such as gene co-expression) and PPI networks are widely used to define the biological context of the genes [Barabasi and Oltvai, 2004].

Non-human data Transferring functional information from non-human species (orthology) can shed light on disease-genes in human.

Ontologies Common ontological annotation (e.g., from GO [Ashburner et al., 2000]) provides clues on functional and phenotypic similarity making it possible to associate genes with diseases.

Literature Extracting disease-gene relationships through (text-)mining the literature can also reveal links between genes and diseases (e.g., by checking co-occurrence of relevant terms).

Mutations Several methods use existing knowledge / predictions on functional and structural effects of the mutations. Note that, a related set of tools are available for the functional annotation of SNPs. These tools are reviewed elsewhere [Kann, 2010, Karchin, 2009, Capriotti et al., 2012].

The evidence types explained above are used to describe "similarity" between known disease-genes and candidate genes in the guilt-by-association scheme, such that if a gene is similar to known disease-genes, it is also predicted to be implicated in the pathology of the disease. Over the past decade, the amount of pathway information (coordinated gene-gene relationships) such as gene co-expression patterns and PPIs has accumulated rapidly (e.g., high-throughput techniques to detect gene expression and interactions). Considering this continuous growth in the information, among these different evidence types, pathway involvement based approaches to describe the similarity have particularly attracted interest for their unmatched potential in identifying novel associations.

The disease-gene prioritization methods using PPIs to describe similarity to genes already known to be associated with a disease (seeds) can be distinguished by the way they define proximity between the gene products in the network of PPIs. Early attempts to identify novel disease genes check whether the protein encoded by the gene of interest interacts with the products of seeds (direct neighborhood) [Xu and Li, 2006, Oti et al., 2006, Lage et al., 2007, Pujana et al., 2007, Wu et al., 2008, Aragues et al., 2008]. Though the direct neighborhood (direct interaction partners) of genes offers substantial benefits to describe similarity between genes, this approach can be extended indirect

Table 1.3: Available disease-gene prioritization tools (adapted from [Capriotti et al., 2012][*]).

Method	URL	Description
aGeneApart	www.esat.kuleuven.be/ageneapart	L
BITOLA	ibmi.mf.uni-lj.si/bitola	L
CAESAR	polaris.med.unc.edu/projects/caesar	ESPNOML
CANDID	dsgweb.wustl.edu/hutz/candid.html	ESPNL
DADA	compbio.case.edu/dada	PL
DomainRBF	bioinfo.au.tsinghua.edu.cn/domainRBF/gene	SOML
ENDEAVOR	www.esat.kuleuven.be/endeavour	ESPNOL
G2D	www.ogic.ca/projects/g2d_2	ESPOL
GeneDistiller	www.genedistiller.org	ESPNOL
GeneProspector	www.hugenavigator.net	SNML
GeneSeeker	www.cmbi.kun.nl/GeneSeeker	NL
GeneWanderer	compbio.charite.de/genewanderer/GeneWanderer	PNML
Genie	cbdm.mdc-berlin.de/tools/genie	ESPNL
Gentrepid	www.gentrepid.org	ESPL
MedSim	www.funsimmat.de	SPNOL
MimMiner	www.cmbi.ru.nl/MimMiner	SL
PGMapper	www.genediscovery.org/pgmapper	ESPL
PhenoPred	www.phenopred.org	SPO
PINTA	www.esat.kuleuven.be/pinta	EP
PRINCE	www.cs.tau.ac.il/~bnet/software/PrincePlugin	EP
PolySearch	wishart.biology.ualberta.ca/polysearch	L
PosMed	omicspace.riken.jp/PosMed	L
PROSPECTR	www.genetics.med.ed.ac.uk/prospectr	SNML
SNPs3D	www.snps3d.org	SPNOML
SUSPECTS	www.genetics.med.ed.ac.uk/suspects	ESPNML
ToppGene	toppgene.cchmc.org	ESPNOL
TOM[**]	www-micrel.deis.unibo.it/~tom	EP
VAAST	www.yandell-lab.org/software/vaast.html	EM

[*] (E) Experimental observation (S) Sequence, structure, tissue specificity (P) Pathway involvement (N) Non-human data (O) Ontologies (M) Mutations (L) Literature

[**] Resource not available at the time of redaction

connections between genes (e.g., neighbors of neighbors in the network). For this purpose, several studies utilize clustering based methods [Milenkovic et al., 2010, Navlakha and Kingsford, 2010]. More recently, in order to fully exploit the network topology, global topology based approaches have been proposed. Some of these works rank the genes in the network with respect to the shortest distance between disease genes [Franke et al., 2006, Wu et al., 2008, Kohler et al., 2008, Dezso et al., 2009]. Some other works use kernel based diffusion over the links of the network (where the furtner nodes have less influence) [Ma et al., 2007, Nitsch et al., 2010, Qiu et al., 2010] while some others simulate a random walk model (where each node is assigned with the probability of a random surfer ending up in the node while surfing through the links of the network) [Kohler et al., 2008, Chen et al., 2009, Vanunu et al., 2010]. Methods based in global topology, especially the ones based on random walks were demonstrated to outperform methods based on local topology [Kohler et al., 2008, Navlakha and Kingsford, 2010, Vanunu et al., 2010]

Taking into account incompleteness (false negatives) and noisiness (false positives) of available PPI data, statistical adjustment methods to remove the bias towards highly connected known disease nodes in PPI networks has also been proposed [Erten et al., 2011] where the scores computed by prioritization algorithms are normalized using random networks. Furthermore, several approaches incorporate gene expression and data on functional similarity in addition to physical PPIs [Franke et al., 2006, Aerts et al., 2006, Aragues et al., 2008, Ala et al., 2008, Linghu et al., 2009] to increase the quality of the network underlying the methods mentioned above.

1.2 Systems biology

The cell is governed by the complex interactions of various types of biomolecules such as DNA, RNA, proteins and small molecules. With the recent advances in biological data collection and bioinformatics techniques, the gene-centric approaches for phenotypic characterization are being replaced by more systematic approaches that account for the interactions between biomolecules. In order to characterize the dynamics of the biological system, systems biolgoy studies interactions between the components of biological systems within the framework of the biological system as a whole.

Through integration of different data sources such as protein sequence, gene expression and protein-protein interactions, our understanding on the functioning of the organism has evolved rapidly towards the co-operation of groups of biomolecules that constitute biological networks including metabolic networks, PPI networks, regulatory networks (i.e. gene-protein interactions) and genetic interaction networks (i.e. gene-gene interactions). Furthermore, these networks are interconnected with the biomolecules shared between each other.

Network biology is the study of these coupled biological networks giving rise to the behaviour of the cell. It aims to "map out, understand and model in quantifiable terms the topological and dynamic properties of the biological networks" [Barabasi and Oltvai, 2004]. Considering that the diseases arise from perturbations in the biological networks underlying the cellular processes, network biology plays a key role in providing insights to disease mechanisms.

1.2.1 Protein-protein interaction (PPI) networks

The proteins encoded by the genotype of the organism determines the observed phenotype (e.g., proteins giving the color of the eye). Although the components involved in the translation from the genetic trait to the phenotypic outcome remain elusive engaging all types of biological networks mentioned above, the interactions between proteins are among the most important.

A variety of experimental techniques has been developed to detect whether a protein interacts with other proteins. These techniques are given in Table 1.4 along with method-specific information such as the ability to detect binary interactions and/or complexes, the applicability at large scale and the capacity to provide structural information.

Table 1.4: PPI detection methods (adapted from [Garcia-Garcia et al., 2012])

Method	**Binary**	**Complex**	**High-throughput**	**Structure**
Yeast Two Hybrid (Y2H)	✓		✓	
Mammalian PPI trap (MAPPIT)	✓			
Tox-r dimerization assay (TOXCAT)	✓			
Bimolecular Fluorescence Complementation (BiFC)	✓			
Proximity Ligation Assay (PLA)	✓			
Förster/fluorescence Resonance Energy Transfer (FRET)	✓			
Bioluminescence Resonance Energy Transfer (BRET)	✓			
Protein microarrays	✓	✓	✓	
Surface Plasmon Resonance Array (SPR)	✓	✓		
Tandem Affinity Purification (TAP)		✓	✓	
Cryo-electron microscopy	✓	✓		✓
X-ray crystallography	✓	✓		✓
Nuclear Magnetic Resonance (NMR)	✓	✓		✓

The PPIs identified using experimental techniques are deposited in various databases such as BIND [Isserlin et al., 2011], BioGRID [Stark et al., 2010], DIP [Salwinski et al., 2004], HPRD [Keshava Prasad et al., 2009], IntAct [Kerrien et al., 2011], MINT [Licata et al., 2011], MIPS [Mewes et al., 2010], and MPact [Guldener et al., 2006]. In order to compile interaction data spread in such databases, several computational tools including PIANA [Aragues et al., 2006], ONDEX [Kohler et al., 2006] and BIANA [Garcia-Garcia et al., 2010] have recently been developed. Yet, due to the interaction data not being complete (false negatives), containing noisy interactions (false positives) and not being able to capture the time and location dependent aspects of the cellular events, the networks created using available interaction data serve as a snapshot of the PPI networks.

1.2.2 Understanding disease mechanisms through PPI networks

Aberrations in normal gene function lead to a disease condition. However, complex disease phenotypes rarely result from a single disease-gene. Thus, human genetic disorders

often involve coordinated dysregulation of multiple genes that implicate interdependent pathways. PPI network modeling comes into play in order to spot such disease-genes engaged in interrelated biological processes.

The topological information encoded in the PPI network provides valuable insights to complex diseases [Ideker and Sharan, 2008, Vidal et al., 2011]. For instance, recent studies demonstrated that proteins encoded by disease-genes tend to interact with each other compared to the rest of the proteins [Gandhi et al., 2006, Goh et al., 2007]. Under the light of these findings, exploiting aforementioned guilt-by-association principle, network-based disease-gene prioritization methods associate genes using the links in-between genes such as the topology of the PPI network [Barabasi et al., 2011]. Furthermore, methods using global topology of the network have been demonstrated to outperform methods than methods taking only the local or no topology information into account [Navlakha and Kingsford, 2010].

The benefits of network modeling can be further improved by integrating different types of biological data. Among various types of key molecular information such as data on post-translational, transcriptional (e.g., gene expression profiles), metabolic and epigenetic events help to define a finer-grained biological context. Consequently, towards characterizing disease states, several studies have overlaid gene-expression patterns over the PPI network and identified groups of genes that account for the dysregulation observed in various cancer types (active subnetworks) [Nibbe et al., 2010a]. These subnetworks were also shown to mediate the survival patterns and distinguish between good and poor prognosis patient groups better than conventional (single gene-based) biomarkers [Chuang et al., 2007].

1.2.3 Robustness as an emergent property of biological networks

Complex systems are intrinsically tolerant to internal mechanistic failures and changes in environmental conditions. Biological organisms are not an exception [Kitano, 2004]. In biological systems, robustness is mainly achieved by controlling the system through feedback [Morohashi et al., 2002], decoupling the parts of the system as functional modules [Hartwell et al., 1999] and redundancy [Agrawal, 2001].

Network topology has an important role in generating robustness [Albert et al., 2000, Barabasi and Oltvai, 2004]. Particularly, PPI networks are shown to be scale-free, that is the degree distribution follows a inverse power law ($P(k) \sim k^{-\gamma}$ where k is the degree of a node and $2 < \gamma < 3$). The scale-free property of PPI networks implies that most of the proteins in the network have few connections and there are only a small number of hub proteins that connect a large number of proteins [Barabasi and Oltvai, 2004]. Scale-free networks bear so called "ultra small world" property such that the paths connecting nodes are much shorter than what would be expected in a random network [Barabasi and Albert, 1999]. Moreover, PPI networks have a high clustering coefficient, bringing genes together in relatively more densely connected modules.

Though groups of genes in a cell are organized to minimize the effects of perturbations, biological systems are prone to be disrupted by certain types of rare but specialized perturbations leveraging the fragility of the system [Carlson and Doyle, 2002, Kitano, 2004]. For example, seizing the hub proteins in PPI network can predominantly break up the network [Albert et al., 2000].

1.2.4 Cancer studies in the era of systems biology

Cancer is the disturbance of the regulation circuitry controlling normal cell growth emerging from genetic and epigenetic perturbations mediated by environmental situmuli. A systems level understanding of the pathways underlying the progression of cancer is key to delineate the mechanisms and counteract the consequences of cancer (i.e. uncontrolled cell growth and invasion of other tissues) [Laubenbacher et al., 2009].

PPI networks provide a framework to study the functional relationships among the biological molecules involved in cancer [Jonsson and Bates, 2006, Vogelstein and Kinzler, 2004]. However, PPI networks by themselves provide only a partial view of the complex biological processes. A comprehensive understanding of cancer lies beneath the integration of biological data at different levels such as genomic, transcriptomic, proteomic and metabolomic and the analysis of the relationships between biomolecules in a dynamic context [Hanahan and Weinberg, 2011]. For instance, complementary to PPI networks, network models describing relationships between co-expressed genes can capture the expression changes mediated by the disease [Ergun et al., 2007, Mani et al., 2008].

Stem cell research has attracted substantial attention during the past years, due to the regenerative capacity these cells bear. Embryonic stem cells (ESC) can produce more stem cells (self-renewal) and can differentiate into diverse cell types (differentiation potency). Studying the regulatory circuits governing ESC has provided insights into mechanisms in several human diseases [Young, 2011], particularly in cancer [Ben-Porath et al., 2008] where the tumor cell show a parallel proliferative and plastic behaviour to the one of the stem cell [Reya et al., 2001]. Furthermore, stem cells provide precise disease models enabling valuable opportunities for drug discovery [Park et al., 2008]. Though still at its infancy, stem cell based novel therapeutical applications are also expected to drastically change the treatment of many diseases such as neurodegenerative disorders (e.g., through cellular replacement) [Lindvall et al., 2012]. ESC are dynamically regulated by transcription factors and epigenetic modifiers such as miRNAs and methylation events [MacArthur et al., 2009]. Understanding the mechanisms involved in the control of the stem cell behaviour and its relationship with the diseases requires a system-wide modeling of the underlying network of the genetic and epigenetic components.

1.3 Motivation of this thesis

The experimental data on the genetic variants and the disorders associated with these variants shed light on our understanding of human genetic disorders as such data continue to accumulate. However, pinpointing genomic variants causing diseases is still hindered by the expenses such experimental techniques incur. Computational techniques such as network-based disease-gene prioritization methods play an indispensable role in the characterization and eventually prevention of the consequences of such genomic variants on human health.

Typically, network-based approaches tackle the problem of prioritization of genes within a linkage interval. Recently, these methods have also been applied to prioritize

genes from GWAS [Lee et al., 2011, Akula et al., 2011]. In fact, using the whole genome to prioritize disease-gene variants is expected to produce more robust results in identifying modest-risk disease-gene variants than using high-risk alleles [Carlson et al., 2004]. Nonetheless, existing prioritization methods substantially suffer from a lack of linkage interval information [Navlakha and Kingsford, 2010] and depend on the quality of the interaction network [Erten et al., 2011]. Furthermore, available network-based prioritization methods treat all the paths between nodes equally relevant for a particular disease.

Thus, network-based disease-gene prioritization can be improved. Such improvement can then empower discovery of novel disease-genes and disease-related pathways and help developing therapeutics for complex diseases such as cancer which involves many interconnected pathways.

Part II

Objectives

This thesis aims to fulfill the following objectives:

- To develop an algorithm that effectively exploits guilt-by-association principle to transfer phenotypic annotation through the network topology.
- To evaluate the prediction of accuracy of the developed prioritization method (algorithm) on human genetic diseases and compare it with existing methods.
- To assess the dependence of the developed prioritization method on the underlying network and known disease-gene associations.
- To provide an online and user-friendly framework where the genes can be prioritized for their association to any user-provided phenotype.
- To investigate further uses of the developed method, particularly in complex diseases such as cancer.

The first two objectives are addressed in the next chapter (Chapter 2) where I introduce a network-based phenotypic characterization method and demonstrate that it outperforms similar methods using protein-protein interactions (PPIs) when applied to the problem of disease-gene prioritization in human. The dependence of the developed prioritization method on the underlying network and known disease-gene associations is investigated in Chapter 3. In accordance with the fourth objective, an online web-server is presented in Chapter 4. The following chapter (Chapter 5), complies with the final objective and provides an overview of network modeling approaches that have helped understanding cancer progression.

In the appendix section, I include some of the works that I was involved during my doctoral studies. These works are not directly related to the main objectives of the studies but are either used towards achieving or extending the goals of the thesis. BIANA, a biological data integration tool, in the development of which I actively participated, is introduced in Appendix A. BIANA compiles information on the biological molecules and the relationships between them (such as PPIs) in a single database and provides an easy way to access and analyze proteomics and interactomics data spread over various databases. In Appendix B, a method to extend functional annotation of genes using PPIs is described. In this work, starting from known elements of apoptosis signaling pathway, PPIs are used to discover genes that are potentially involved in the same pathway. I was involved in the implementation of network-based methods for functional extension and cross-validation analysis. I also contributed to a review on the acquisition, integration, analysis and application of PPI information (Appendix C). Finally, in Appendix D, I present the results of an on-going collaboration with Trey Ideker's group in University California, San Diego in which I am responsible of analyzing genome-wide gene-expression from various human cell samples at different points of differentiation and developing a framework to identify genes mediating human cell fate.

Part III

Results

Chapter 2

Exploiting Protein-Protein Interaction Networks for Genome-wide Disease-Gene Prioritization

Guney E, Oliva B. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. PloS One. 2012;7(9):e43557.

Guney E, Oliva B. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization1. Supplementary material. PLoS One. 2012;7(9):e43557.

Chapter 3

Robustness of network-based disease gene prioritization methods points to pathophenotypic plasticity

Robustness of Network-based Disease Gene Prioritization Methods Points to Pathophenotypic Plasticity

Emre Guney and Baldo Oliva*

Structural Bioinformatics Group (GRIB). Universitat Pompeu Fabra. Barcelona Research Park of Biomedicine (PRBB). Catalonia, Spain.

Keywords: network based disease gene prioritization, protein-protein interactions, disease pathways, robustness, phenotypic plasticity

Abbreviations: protein-protein interaction (PPI), Receiver Operating Characteristics (ROC), area under ROC curve (AUC), Gene Ontology (GO), tandem affinity purification (TAP), Alzheimer's Disease (AD), Online Mendelian Inheritance in Man (OMIM).

* Corresponding author, baldo.oliva@upf.edu

Abstract

Complex biological systems usually pose a trade-off between robustness and fragility where a small number of perturbations can substantially disrupt the system. In genetic diseases, a mutation perturbs the system and exploits its fragility. Recent advances in identifying and analyzing the sequential variations beneath human disorders help to comprehend a systemic view of the mechanisms underlying various disease phenotypes. In this study, we have hypothesized that disease-gene prioritization methods based on the network of protein-protein interactions can be employed to investigate the mechanistic relationships between disease-genes and explain the robustness emerging from these relationships. Network-based disease-gene prioritization methods rank the relevance of genes in a disease under the hypothesis that genes whose proteins interact with each other tend to exhibit similar phenotypes. We have tested the robustness of several network-based disease-gene prioritization methods with respect to the perturbations of the system using various disease phenotypes from the Online Mendelian Inheritance in Man database. These perturbations have been introduced either in the protein-protein interaction network or in the set of known disease-gene associations. As the network-based disease-gene prioritization methods are based on the connectivity between known disease-genes, we have further used these methods to understand the plasticity of the pathophenotypes. Our results have suggested that pathophenotypes such as breast cancer, diabetes and obesity bear more plasticity compared to the rest of the compared pathophenotypes.

Introduction

A fundamental characteristic of biological systems is tolerance to noise. The ability to counteract both internal mechanistic failures and changes in environmental conditions plays a central role in the survival of the organism. The main components of robustness are controlling the system through negative and positive feedback (Morohashi et al., 2002), splitting the parts of the system as functional units (Hartwell et al., 1999) (modularity and decoupling), and phenotypic plasticity (Agrawal, 2001) (typically achieved by redundancy). In a biological system, groups of genes are optimized in functional decoupling, redundancy and diversity such that the effects of perturbations are minimized (H. Kitano, 2004a). However, complex biological systems have to balance between robustness and fragility which implies that a small number of rare perturbations can substantially disrupt the system (Carlson and Doyle, 2002). In particular, some mutations are the main cause of diseases by exploiting the fragility of the biological system.

Similar to the underlying biological system, the disease phenotypes themselves are likely to be robust against external changes (Pujol et al., 2010). For instance, the robustness of disease phenotypes has been recently discussed for HIV (H. Kitano, 2004a), diabetes (H. Kitano et al., 2004) and cancer (H. Kitano, 2004b). Complex genetic disorders involve cooperation of multiple genes engaged in various biochemical pathways associated with a disease (Goh et al., 2007; Barrenas et al., 2009). Employing several genes often induces resilience against the mechanisms of defense of the organism. Thus, understanding the strengths and weaknesses of pathophenotypes against perturbation can play a crucial role for developing effective therapeutic strategies (Pujol et al., 2010).

During the past decade, genome-wide efforts such as linkage analysis and association studies have successfully associated numerous causal loci with human disorders (Altshuler et al., 2008). Still, much effort needs to be taken to fully understand the complex implications on the whole system and its protein-protein interaction (PPI) network. Hence, several methods have been developed to amplify available disease-gene associations using the principle of “guilt-by-association”. These methods typically exploit relationships of the disease causing genes with other candidate genes, initially using the neighborhood of known associations in the physical (Oti et al., 2006; Xu and Li, 2006) or functional (Pujana et al., 2007) interaction network and more recently extending the approach to account for the global topology of the underlying network (Kohler et al., 2008; Chen et al., 2009). Although considering neighborhood to known disease causing genes is a simple approach to identify novel candidate genes, global topology based methods have proved useful in associating genes with the diseases they mediate (network-based disease gene prioritization) (Navlakha and Kingsford, 2010; Akula et al., 2011; Lee et al., 2011).

Following the emergence of high-throughput experimental techniques that produce large amount of biological data, several studies have investigated robustness of a complex system in respect to the underlying network topology. Different types of networks have been studied with this purpose, such as metabolic networks (Jeong et al., 2000), protein-protein interaction networks (Jeong et al., 2001; Huang et al., 2006) and regulatory interaction networks (Demongeot et al., 2009). However, to our knowledge, robustness of network-based disease-gene prioritization methods, where the underlying network itself is perturbed, has not been extensively investigated. Due to the fact that network-based disease-gene prioritization methods use the connectivity between genes associated with the disease, we hypothesize that they may serve to evaluate robustness of diseases against perturbations in the underlying interaction network. The definition of robustness is problem specific (Rizk et al., 2009). We define the robustness as the observed change in the prediction capacity of the prioritization methods when perturbing the underlying interaction network.

In this paper, our main goal has been to test the quality and robustness of several network-based

disease-gene prioritization methods against perturbations introduced either to the underlying protein-protein interaction network or to known disease-gene associations. Next, to investigate the relationship between robustness and modularity, we have examined the capability of these methods to identify disease modules (groups of genes that are enriched with the functions relevant to the disease). We have used these results to decide the best network-based prioritization approach. Then, we have proven that for some diseases, the robustness of the outcome of the network-based prioritization was a result of the connectivity of disease-associated genes in the underlying network. Therefore, through the analysis of the prediction performance of the network-based prioritization method under *in silico* perturbations, we have categorized various disease phenotypes in Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al., 2002) with respect to the pathophenotypic plasticity. Our results have suggested that several pathophenotypes, particularly the ones with high prevalence in the society such as breast cancer and diabetes, bore more plasticity than the rest of the compared pathophenotypes. We have found that robustness of the pathophenotype was independent of the number of initial genes associated with the disease and rather mediated by how well these genes are connected in the interaction network. Furthermore, we have argued that robust diseases achieved pathophenotypic plasticity by pathologically aiming many more pathways compared to non-robust diseases.

Results and Discussion

Network-based disease-gene prioritization methods as a framework to investigate biological robustness of pathophenotypes

Network-based disease gene prioritization methods rank the relevance of genes in a disease using known disease-gene associations and the network topology. Any perturbation in the network topology induces a change on the ranking of the genes for the disease in concern. Systematically introducing perturbations at different levels and analyzing the changes in the ranking of the disease-genes provide a way to measure the robustness of the prioritization. We hypothesize that if any, such robustness emerges either from the prioritization method applied or it is an intrinsic characteristic of the pathophenotype in concern. Then, if we prove that this robustness is not due to the method itself but caused by the connectivity between the gene products of the genes associated with a disease, we can use this robustness as a measure of the plasticity of the pathophenotype. We adopt the plasticity concept for disease phenotypes following the original definition of phenotypic plasticity. Phenotypic plasticity is defined as the ability of a single genotype to alternate its phenotype in response to environmental conditions (West-Eberhard, 1989, -). Accordingly, we define phenotypic plasticity of a pathophenotype as its ability to adapt the phenotype in response to changes in the environment while preserving its pathogenicity. In other words, pathophenotypic plasticity is our incapacity to affect the pathophenotype with the use of drugs or chemical intervention (Hiroaki Kitano, 2007) and we model such possible interventions as perturbations introduced either in the interaction network or in the known disease-gene associations (seeds).

To assess the tolerance of a given phenotype to the noise in the underlying interaction network or in the seeds, we used five network-based prioritization methods. The prioritization methods rank the nodes of the network according to their implication in the pathopheontype. The network-based prioritization approaches obtain this rank by disseminating the information of seeds through the protein interaction network. In this work, we applied recently proposed topology-based ranking algorithms available in GUILD software package (Guney and Oliva, 2011, [CSL STYLE ERROR: reference with no printed form.]) using an integrated human interactome and 23 disease phenotypes curated from OMIM database (Hamosh et al., 2002). The human protein interaction network was obtained by extracting human protein-protein interactions from several publicly available repositories and integrating them with BIANA (Garcia-Garcia et al., 2010). The network consists of 11250 nodes (gene products) and 59220 edges (physical interactions) connecting

them. Initial disease-gene associations were retrieved from OMIM (Hamosh et al., 2002) database for phenotypes that have at least 25 causative genes associated with them (see Methods for details).

First, we mapped the disease-gene associations to their corresponding nodes (these are named seeds) in the interaction network. Then, we introduced several perturbations on known disease gene associations and on the interaction network. Next, we used the network-based prioritization methods on the original and perturbed data sets. Finally, we evaluated the prediction performance of the prioritization methods (see methods for details). We compared the robustness of five prioritization methods at different percentages of perturbation (see next section). To explain the robustness, we compared the number of seeds, its location in the protein-interaction network and their connectivity. Interestingly, we proved that some disorders showed higher robustness than others due to a different degree of connectivity between the seeds independent of the number of seeds. This argument was used to hypothesize that the degree of robustness could also rank the degree of resilience, adaptation and plasticity of some pathogenic phenotypes. Finally we categorized the disorders and analyzed them by means of the results of the disease-gene prioritization methods.

In silico analysis points out to alternative routes connecting the genes involved in disorders

We first questioned whether the prioritization methods depend on the number of genes associated with a disorder. To address the dependence on the number of seeds, we replaced the seeds with non-seeds in the network at varying percentages (10% to 100%). That is, we disturbed the initial disease-gene associations at different levels by introducing wrong associations between genes and pathophenotypes. Then we calculated the area under ROC curve (AUC) of network-based prioritization methods using a five-fold cross-validation setting on the perturbed disease-gene annotations. An increased percentage of mis-annotated seeds reduced the reliability of predictions for all methods (Figure 1a). If more than 70% of seeds were false, the AUC reduced to less than 50% for all methods. Only NetShort resulted in an AUC higher than 50% with 40% false seeds. In conclusion, all methods were dependent on the quality of initial associations while NetShort was less affected, compared to the rest, in predicting new genes associated with the disease. In order to ensure that this was not an artifact of the interaction network, we repeated the analysis on the interaction network used by Goh et al. (Goh et al., 2007). This network contained a set of high confident protein-protein interactions in human (we refer to this network as Goh network hereafter, see Methods). The AUCs for the methods at different perturbation percentages using Goh network are given in Supplementary Figure 1a.

To prove the relevance of the quality of the interactions, we randomly swapped the edges of the network. Also, to investigate the relevance of the number of interactions, we randomly deleted edges of the network. The variation in the edges of the network ranged between 10% and 100%. We applied the prioritization methods to these perturbed networks and calculated the average AUC over all diseases. Figure 1b shows the decrease in AUC produced using false interactions (randomly swapped edges) for all methods. It is noteworthy that PageRank was the most robust method, the prediction performance of which was less affected from the perturbation of edges than the rest. On the other hand, edge deletion decreased the AUC for NetScore, Functional Flow, and PageRank, but NetZcore and NetShort improved the prediction quality by increasing the AUC, and it only began to drop after more than 60% of the interactions were removed (Figure 1c). Repeating the same analysis with the Goh network revealed that the prioritization approaches exhibit a similar behavior (Supplementary Figures 1b and 1c), indicating that these features were independent of the underlying network. Although this behavior was unexpected, it could be explained by the way the prioritization algorithms work. These algorithms used the seed nodes to disseminate information through the network. For each disease, there were very few number of seeds compared to non-seeds in the interaction network. Therefore, random deletion of edges disconnected fewer seeds because the number of edges connecting two seeds was

much less than the number of remaining edges (that connected a seed and a non-seed or two non-seeds). Consequently, Functional Flow and NetScore were more affected than others because of their dependence on the number of paths that connected seeds with each other. However, the effect of deletion diminished in the case of NetZcore since it normalized the score using random networks and the scores of the nodes connected with seeds improved. To understand this with an example, let's take a node *n* that is relatively more connected to seeds in comparison to any of the random networks. The random deletion of an edge would be more likely to remove a link connecting a non-seed. Hence, it would be more likely that node *n* would remain relatively more connected to seeds in comparison to random networks. NetShort, which used the shortest paths leading to seeds, also improved the quality of the predictions, probably due to seeds having secondary routes unaffected by the deletion of links. Such backup circuits constitute a fail-safe mechanism and explain the resilient nature of cells (H. Kitano, 2004a).

The effect of perturbing interactions on the pathophenotype: the case of Alzheimer's Disease

In Figure 1 we observed that the perturbations in the interaction network improved the prediction performance of several prioritization algorithms. In order to gain an insight on the consequences of the changes in the interaction network, we analyzed the connectedness of the genes associated with the Alzheimer's Disease (AD), a relatively well studied pathophenotype. We used the human interaction network and genes associated with AD in OMIM (Hamosh et al., 2002) (AD-seeds). We took the neighbors of seeds in the interaction network and checked how many of their neighbors are implicated in AD using an independent set of genes taken from literature (validation set, AD-related genes) (Krauthammer et al., 2004). We repeated this procedure on perturbed networks, interactions of which were either randomly swapped (permuted) or deleted (pruned) at different percentages (see Methods). Figure 2 shows the total number of genes and AD-related genes in the neighborhood of seeds. Not surprisingly, as the percentage of perturbation increases, the number of AD-related genes in the neighborhood of AD-seeds decreased. However, in the case of interaction pruning, the ratio of AD-related genes versus the total number of genes in the neighborhood of AD-seeds increased. This suggested that AD-related genes tended to remain connected with at least one AD-seed in the interaction network.

Categorizing pathophenotypic robustness

The analysis of the prediction performance implied that disease-gene associations could still be recovered considerably using network-based prioritization methods even when half of the interactions were perturbed (see Figures 1b and 1c). We further investigated whether this was an intrinsic property of the disease. We first defined the robustness of a disease as the amount of perturbations required to cause a "critical" AUC change when we applied network-based prioritization to predict its associated genes (see Methods for details). Among the different approaches to prioritize candidate genes we used NetScore for this analysis, since it had good overall prediction accuracy (Figure 1) and produced the smallest number of clusters coherently enriched in the functions related to the corresponding disease (see Supplementary Results and Supplementary Figure 2). Note that NetZcore had also similar properties, however the perturbation analysis above proved that NetZcore was robust against perturbations whereas the AUC for NetScore fell down linearly as the percentage of perturbations increased (Figure 1). In order to minimize the bias on the prediction performance due to using a robust prioritization method and focus on the robustness of the disease, we selected NetScore over NetZcore. Thus, we grouped diseases into two categories based on the differences of the prioritization performance (assessed by the AUC) between the original network and perturbed networks using NetScore. For each pathophenotype, we checked the amount of perturbation (both for edge swapping and removal) required in the interaction network that caused the prioritization performance fall below the critical AUC. If after perturbing more than 50% of the interactions, the prioritization of a disease still achieved a performance higher than the critical AUC, we labeled this disease as being robust. On the other hand, if the critical AUC was reached with

perturbations affecting less than 50% of the interactions in the network, we labeled the disease as being non-robust. See Supplementary Figure 3 for the changes in the NetScore prioritization AUC upon perturbations in the interaction network for robust diseases. According to our criteria, robust pathophenotypes such as *breast cancer, cardiomyopathy, diabetes, insulin, leukemia, obesity* and *parkinson disease* had larger plasticity and capability of adaptation against perturbations in the underlying network than the rest of the pathophenotypes. It is worth mentioning that pathophenotypes with more plasticity or resiliency corresponded to diseases of high prevalence in developed countries.

Comparison between robust and non-robust pathophenotypes

We further checked whether the diseases in different categories bore similar properties in terms of known disease annotations associated with disease (seeds), connectivity of these seeds, modularity and functional plasticity. In principle one may think that the definition of robustness seems to be dependent on the number of seeds of the phenotype. This would also suggest an explanation to the fact that diseases of high prevalence were categorized more robust. Nevertheless, there was not a significant difference between the number of seeds of robust and non-robust diseases (Figure 3a). In contrast, the seeds of robust pathophenotypes were significantly connected to each other with shorter paths compared to non-robust pathophenotypes (Figure 3b, associated p-value with the two-sided Wilcoxon rank sum test was *0.03*). This feature, the average length of shortest paths connecting any two seeds in a disease, was independent of the number of seeds.

Next, to investigate the relationship between robustness and modularity, we have checked the number of modules associated with diseases (see Supplementary Methods and Supplementary Results). Robust diseases showed slightly increased modularity compared to non-robust diseases (see the section on modules associated with diseases above) yet the difference in the number of clusters grouping similar number of genes in both categories was not significant (Figure 3c). Furthermore, we counted the number of GO terms enriched in the seeds of the diseases (seed GO terms) for both categories. In order to avoid possible bias towards the number of seeds, we normalized the number of seed GO terms dividing them by the total number of seeds for each pathophenotype. Robust diseases contained more seed GO terms per seed, proving that a larger number of biological functions associated with seeds were involved in robust diseases than in non-robust pathophenotypes (Figure 3d, associated p-value with the difference was *0.01*). Similarly, the clusters of top ranked genes in robust diseases covered a larger number of seed GO terms than the clusters of non-robust diseases (Figure 3e). When we consider all GO terms enriched in the clusters, the clusters of robust pathophenotypes covered more GO terms than the clusters of non-robust pathophenotypes (Figure 3f). Although the difference between robust and non-robust pathophenotypes was only significant for the number of seed GO terms per seed, these findings suggested that robust phenotypes tended to group genes involved in a larger number of functions compared to non-robust phenotypes. Therefore, functional diversity might be more important for robustness than a particular biological function. We also calculated the Jaccard index for quantifying GO terms shared between each pair of diseases in the modules of top ranking genes (see Supplementary Figure 4). Several diseases like breast cancer, lymphoma and ataxia had a large number of common functions compared to rest. Such common features at the functional level might be useful for understanding the etiology of complex genetic disorders (Cotterchio et al., 2002). Nonetheless, robust and non-robust pathophenotypes did not share a substantial amount of biological functions.

Supplementary Figure 5 shows the GO terms shared by at least three diseases within robust and non-robust pathophenotypes. Robust pathophenotypes such as *leukemia, obesity* and *diabetes* were observed to share GO terms of biosynthetic and metabolic processes while non-robust diseases shared GO terms associated with membrane activity. Remarkably, biological processes associated with regulation of cellular metabolism were more often implicated in robust diseases

than in non-robust diseases, which supported the role of positive and negative feedback mechanisms in defining robustness.

Conclusions

In this study, we have analyzed robustness of several disease-gene prioritization algorithms that were based on known disease associations and protein-protein interaction networks. Our analysis on randomly perturbed interaction networks pointed to the existence of backup circuits within the interaction network constituting a fail-safe mechanism. Strikingly, the interactome might tolerate an interaction removal rate as high as 50%, the point at which methods using alternative paths start suffering from low prediction performance. This suggested that only half of the interactions of the network were sufficient to explain a given pathophenotype. Our study has also shown the robustness of some prioritization methods upon perturbations. NetZcore and NetShort have shown to be consistently effective when reducing the number of interactions, while PageRank was instead more robust against the introduction of false interactions.

We also applied the prioritization methods to identify disease modules and to study the implication of several biological processes in various diseases. Our findings confirmed that the prioritization methods were able to distinguish between groups of connected genes with functions identical to those of the known disease-associated genes (disease modules).

We categorized diseases into robust and non-robust pathophenotypes using a network-based prioritization method. Interestingly, robust pathophenotypes included many diseases with high prevalence in society. We further investigated whether there were characteristic differences between diseases involved in these categories. The connectedness of proteins encoded by seed genes (known disease-genes) stoke out as the most important factor in defining robustness of a pathphenotype. That is, a disease was more likely to be robust if the proteins encoded by the genes involved in that pathophenotype were connected with shorter paths to each other. Interestingly, our results suggest that network-based robustness of pathophenotypes may also be defined based on the connectivity of seeds. We also observed that genes involved in robust pathophenotypes were implicated in more biological processes than the genes of non-robust pathophenotypes. The clusters of top prioritized genes in the robust pathophenotypes tended to be involved in a larger number of biological functions than the top ranking genes in the clusters of non-robust pathophenotypes.

Overall, robust diseases bore higher connectedness between their seeds and an increased functional diversity, suggesting plasticity to some extend at pathophenotype level. The findings presented here may help developing novel network-medicine approaches that try to characterize the interconnected pathways implicated in diseases and possibly suggest points of action to compensate these changes induced by the disease (Zanzoni, 2009). Seed connectivity and functional diversity can explain why polypharmacological approaches, which typically target many gene products simultaneously via administration of multiple drugs (Hopkins, 2008), may work better on some diseases such as AIDS or cancer (Yang et al., 2008; Vazquez, 2009; Motter, 2010).

Methods

Gene–phenotype associations. Genes and their associated disorders were taken from Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al., 2002). OMIM is one of the most comprehensive repositories of genes with Mendelian mutations and the disorders associated with them. Phenotypic associations for genes were extracted from the OMIM Morbid Map (ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap retrieved August 27, 2009) by searching for

keyword entries associated with the disorders given in Supplementary Table 3. A disorder was considered if at least 25 genes were associated with it in the Morbid Map after merging several diseases under the same keyword (e.g., for Alzheimer's disease we collected Alzheimer's disease types 1, 2, etc... using the keyword "alzheimer"). Supplementary Table 3 summarizes all diseases used under the context of this study and the number of genes associated with them.

Genes associated with a disorder were mapped to their products (proteins) in the protein-protein interaction network and assigned an initial score for their phenotypic relevance. Thus, proteins translated by genes known to be involved in a particular pathology were termed *seeds* and have the highest score (*1.0*) in the network. All other proteins in the network were assigned a *non-seed* score (lowest score in the network: *0.01*). The correspondence between genes and their products (proteins) was determined using the data integration protocol of BIANA (Garcia-Garcia et al., 2010).

Protein–protein interaction data sets. We compiled a human protein-protein interaction network from publicly available major interaction data repositories using BIANA integration tool (see Supplementary Table 4). First, protein-protein interactions from different sources were integrated with BIANA (Garcia-Garcia et al., 2010) to obtain a human interactome. High throughput pull down interaction detection methods introduce many indirect relationships (such as being involved in the same complex) in addition to direct physical interactions. Thus, we removed the subset of interactions obtained by Tandem Affinity Purification (TAP) and called this network as the bPPI (binary PPI) network. We used the largest connected component of the network since we apply a set of methods based on the transfer of information through the edges of the network (i.e. unconnected nodes can not receive information from other nodes). This component will be referred as network hereafter. The disease-gene annotations from OMIM were mapped to the corresponding proteins in the network using gene symbols provided in UniProt database. Note that some of the gene-disease associations were excluded since not all gene products have an interaction in the largest connected component of human interactome. We also used the human interactome from Goh et al. (Goh et al., 2007) (referred to as the Goh network), which combined two high quality yeast two-hybrid experiments (Rual et al., 2005; Stelzl et al., 2005) and PPIs obtained from the literature.

Network-based prioritization of disease-genes. *GUILD* (Genes Underlying Inheritance Linked Disorders) is a network-based disease gene prioritization framework (Guney and Oliva, 2011, [CSL STYLE ERROR: reference with no printed form.]). The framework contains a set of methods which use known disease genes and their interactions to rank the relevance of genes in a disease or disorder based on the hypothesis that genes whose proteins interact with each other tend to exhibit similar features, such as function and/or phenotype. These methods require an initial set of genes associated with a particular phenotype (e.g., a Mendelian disorder) and interactions between the products of these genes. *GUILD* proposes three topology-based ranking algorithms: NetShort, NetZcore, and NetScore (see Supplementary Methods for details). Additionally, two state-of-the-art algorithms have been included in GUILD: PageRank with priors (White and Smyth, 2003) (as used in ToppNet (Chen et al., 2009)) and Functional Flow (Nabieva et al., 2005). PageRank with priors has recently been proven to be superior to available topology-based prioritization methods (Chen et al., 2009; Navlakha and Kingsford, 2010; Lee et al., 2011). We also apply Functional Flow, a method originally addressed functional annotation problem with success (Nabieva et al., 2005).

Evaluation. To evaluate the prioritization methods, we used a five-fold cross validation. Proteins known to be associated with a phenotype (seeds) were split into five groups; four of them were used to apply the prioritization functions and one set to evaluate the prediction. This process was repeated five times, changing the set for evaluation. *The AUC* were averaged over the five evaluations. These averages and their standard deviations were used to assess the quality of the

predictions. ROCR package was used to calculate these values (Sing, 2005).

We have to note that the algorithms being studied depend solely on the topology of the network, implying that unconnected nodes and very small components cannot transfer the score information along the network. Consequently, only the largest connected component of the network was used for the evaluation (see Supplementary Table 3 for the number of proteins in the largest component of each network).

In the context of gene-phenotype association studies, obtaining negative data (proteins/genes that have no effect on a diseasee) is a challenge. We considered all proteins not associated with a particular disease as potential negative instances. Then, we used a random sampling of the potential negatives to calculate an average score. This score was defined as the score of a *negative* instance. We calculated as many scores of *negative* instances as seeds (positive instances) in the evaluation set such that we had the same number of positive and negative scores.

Dependency of prioritization methods on network features and gene associations. We evaluated the dependency of the five prioritization methods implemented in *GUILD* on the interaction and seed sets. This dependency was studied by modifying the data with three tests: 1) permuting interactions at random, 2) randomly removing interactions of the network, and 3) permuting the seeds at random. We tested the effect of the modifications on the bPPI and Goh networks using OMIM gene-phenotype associations. The degree of modifications ranged from 10% to 100% for each network and seed set. Ensembles of 100 random networks and random seed sets were used to assess average prediction performance for each perturbation level. Thus, in each test, we used nine groups of networks or seed sets (corresponding to the percentage of perturbation, 10 to 100 with the increments of 10) and each group contained 100 randomly perturbed networks or seed sets. For the first test, nine groups of 100 networks were generated by swapping the edges of the original network (randomly creating new edges and removing old ones), and each group contained a different number of random permutations corresponding to the 10% to 100% of the number of edges. For the second test, the edges were randomly deleted to create nine groups of 100 networks in which the number of deletions varied between 10% and 100% of the number of edges. For the third test, a varying percentage of seed nodes (10% to 100%) was replaced with non-seed nodes 100 times, yielding nine groups of seed sets and the percentage of non-seeds in each group ranged between 10 and 100. The prioritization methods were applied to these modified data sets and for each group the AUC averaged over the 100 randomly modified networks or seed sets.

Functional enrichment analysis. GO terms enriched among genes corresponding to high scoring proteins in a network were identified using the FuncAssociate2.0 (Berriz et al., 2009) web service. Proteins in the network were mapped to the genes using the gene symbols provided by UniProt, and these symbols were fed to the web service. All genes in the network were used as the background gene list. A GO term was associated with the gene set, if and only if, the adjusted p-value associated with the term was less or equal to *0.05*.

Defining robustness of a pathophenotype. As mentioned above, we applied network-based prioritization methods on two types of perturbed networks (random interaction swapping and interaction removal) and analyzed the prioritization accuracy (AUC) using a five-fold cross validation scheme. For the phenotypic robustness analysis, we chose NetScore method since *i)* it had the highest prediction performance *ii)* it produced clusters that are functionally more relevant to the disease compared to the rest of the prioritization methods and *iii)* the method itself was not robust against perturbations (Figure 1 and Supplementary Figure 2). We defined robustness based on the amount of interaction perturbation required to cause a “critical” AUC change in network-based prioritization of a disease. For each disease, the critical AUC change was set as half of the AUC difference from expected AUC that would be obtained by random predictions.

That is, for each pathophenotype p, critical AUC was calculated using the following formula:

$$AUC_{crit}(p) = \frac{AUC_{init}(p) - 0.5}{2}$$

A disease phenotype was called robust if the amount of interaction perturbation (both interaction swapping and removal) required to cause a critical AUC change was lower than 50%. Similarly we called a disease phenotype non-robust if the amount of interaction perturbation for critical AUC change was higher than 50%.

For testing significance of differences in distribution of values between robust and non-robust diseases we used Wilcoxon rank-sum test. Alpha values were set to 0.05. R software (http://www.r-project.org) was used to compute statistics.

Acknowledgements

EG is supported through FI fellowship granted by “Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu”. GRIB is the Biomedical Informatics node of the Spanish Institute of Bioinformatics (INB). This work was also supported by grants from the Spanish Ministry of Science and Innovation (MICINN) FEDER BIO2011-22568; and by EU grant EraSysbio+ (SHIPREC) Euroinvestigación (EUI2009-04018).

Author Contributions

Designed the experiments: EG BO. Performed the experiments: EG. Analyzed the data: EG BO. Wrote the paper: EG BO.

References

Agrawal,A.A. (2001) Phenotypic plasticity in the interactions and evolution of species. Science, 294, 321–6.

Akula,N. et al. (2011) A network-based approach to prioritize results from genome-wide association studies. PLoS One, 6, e24220.

Altshuler,D. et al. (2008) Genetic mapping in human disease. Science, 322, 881–8.

Barrenas,F. et al. (2009) Network properties of complex human disease genes identified through genome-wide association studies. PLoS One, 4, e8090.

Berriz,G.F. et al. (2009) Next generation software for functional trend analysis. Bioinformatics, 25, 3043–4.

Carlson,J.M. and Doyle,J. (2002) Complexity and robustness. Proc Natl Acad Sci U S A, 99 Suppl 1, 2538–45.

Chen,J. et al. (2009) Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 10, 73.

Cotterchio,M. et al. (2002) Human breast cancer and lymphomas may share a common aetiology involving Mouse Mammary Tumour Virus (MMTV). Medical hypotheses, 59, 492–494.

Demongeot,J. et al. (2009) Robustness in regulatory interaction networks. A generic approach with applications at different levels: physiologic, metabolic and genetic. Int J Mol Sci, 10, 4437–73.

Garcia-Garcia,J. et al. (2010) Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics, 11, 56.

Goh,K.I. et al. (2007) The human disease network. Proc Natl Acad Sci U S A, 104, 8685.

Guney,E. and Oliva,B. Exploiting Protein-Protein Interaction Networks for Genome-wide Disease-Gene Prioritization. submitted.

Guney,E. and Oliva,B. (2011) Toward PWAS: discovering pathways associated with human disorders. BMC Bioinformatics, 12, A12.

Hamosh,A. et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 30, 52–55.

Hartwell,L.H. et al. (1999) From molecular to modular cell biology. Nature, 402, C47–52.

Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology, 4, 682–690.

Huang,C.-H. et al. (2006) Topological Robustness of the Protein-Protein Interaction Networks Systems Biology and Regulatory Genomics. In, Eskin,E. et al. (eds), Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 166–177.

Jeong,H. et al. (2001) Lethality and centrality in protein networks. Nature, 411, 41–2.

Jeong,H. et al. (2000) The large-scale organization of metabolic networks. Nature, 407, 651–4.

Kitano,H. (2004a) Biological robustness. Nat Rev Genet, 5, 826–37.

Kitano,H. (2004b) Cancer as a robust system: implications for anticancer therapy. Nature Reviews Cancer, 4, 227–235.

Kitano,H. et al. (2004) Metabolic syndrome and robustness tradeoffs. Diabetes, 53, S6–S15.

Kitano,Hiroaki (2007) A robustness-based approach to systems-oriented drug design. Nature Reviews Drug Discovery, 5, 202–210.

Kohler,S. et al. (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 82, 949–58.

Krauthammer,M. et al. (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease. Proc Natl Acad Sci U S A, 101, 15148–15153.

Lee,I. et al. (2011) Prioritizing candidate disease genes by network-based boosting of genome-

wide association data. Genome Res, advance online article.
Morohashi,M. et al. (2002) Robustness as a measure of plausibility in models of biochemical networks. Journal of theoretical biology, 216, 19–30.
Motter,A.E. (2010) Improved network performance via antagonism: From synthetic rescues to multi-drug combinations. BioEssays, 32, 236–245.
Nabieva,E. et al. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 21, i302–i310.
Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics, 26, 1057–1063.
Oti,M. et al. (2006) Predicting disease genes using protein-protein interactions. J Med Genet, 43, 691–698.
Pujana,M.A. et al. (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet, 39, 1338–1349.
Pujol,A. et al. (2010) Unveiling the role of network and systems biology in drug discovery. Trends in pharmacological sciences, 31, 115–123.
Rizk,A. et al. (2009) A General Computational Method for Robustness Analysis with Applications to Synthetic Gene Networks. Bioinformatics, 25, i169–i178.
Rual,J.F. et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature, 437, 1173–1178.
Sing,T.S. (2005) ROCR: visualizing classifier performance in R. Bioinformatics, 21, 3940–3941.
Stelzl,U. et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell, 122, 957–968.
Vazquez,A. (2009) Optimal drug combinations and minimal hitting sets. BMC systems biology, 3, 81.
West-Eberhard,M.J. (1989) Phenotypic plasticity and the origins of diversity. Annual Review of Ecology and Systematics, 20, 249–278.
White,S. and Smyth,P. (2003) Algorithms for estimating relative importance in networks. In, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, Washington, D.C., pp. 266–275.
Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 22, 2800–2805.
Yang,K. et al. (2008) Finding multiple target optimal intervention in disease-related molecular network. Molecular Systems Biology, 4.
Zanzoni,A.S.-L. p. (2009) A network medicine approach to human disease. FEBS Lett, 583, 1759–1765.

Figure Legends

Figure 1 Robustness of the methods with the perturbation of the edges of the bPPI network and initial disease-gene associations. The interactions of the bPPI network were perturbed **(a)** by swapping the links in order to make false interactions or **(b)** by removing links. Plots show the average AUC and confidence intervals calculated for the prediction of gene-disease associations for 23 diseases using NetScore (red), NetZcore (yellow), NetShort (green), Functional Flow (blue), and ToppGene (purple). The percentage of interactions swapped or removed varied between 0 and 80%. **(c)** Dependence on the number and quality of seeds. The average AUCs are given as the percentage of mis-annotated seeds goes from 0% to 80%.

Figure 2 Change in the number and ratio of AD-related genes in the neighborhood of AD-seeds with respect to the amount of interaction permutation. The interactions of bPPI network were perturbed (**a**) by swapping the links in order to make false interactions or (**b**) by removing interactions. The percentage of interactions swapped or removed varied between 0 and 80%. The bars correspond to the number of genes whereas the line shows the ratio of the average number of AD-related genes in the neighborhood of AD-seeds to the number of all genes in the neighborhood.

Figure 3 Comparison of robust and non-robust pathophenotypes with respect to **(a)** known disease annotations associated with disease, **(b)** the connectivity of these seeds (assessed by the average shortest path length between seeds), **(c)** the modularity, **(d)** the number of seed GO terms associated with the diseases, **(e)** the number of seed GO terms in the modules and **(f)** the number of all GO terms in the modules.

Tables

Table 1. Network-based prioritization performance on the original and perturbed networks*.

Pathophenotype	AUC (%)				Pathophenotype	AUC (%)			
	org.	crit.	perm.	del.		org.	crit.	perm.	del.
alzheimer	78.3	64.2	62.5	62.8	lung cancer	85.0	67.5	65.8	68.4
anemia	70.3	60.2	56.4	57.9	lymphoma	79.7	64.9	62.3	71.8
asthma	31.7	n/a	34.8	28.4	mental retardation	56.3	53.2	46.6	45.3
ataxia	62.6	56.3	54.2	53.8	myopathy	86.0	68.0	67.3	72.0
breast cancer	76.7	63.4	70.7	75.8	neuropathy	42.5	n/a	38.1	33.2
cardiomyopathy	69.5	59.8	65.0	70.5	obesity	72.0	61.0	67.4	70.4
cataract	72.0	61.0	53.9	52.8	parkinson disease	80.0	65.0	70.9	78.5
diabetes	61.4	55.7	58.4	63.4	prostate cancer	68.0	59.0	52.7	62.7
epilepsy	62.1	56.1	47.4	47.4	schizophrenia	53.3	51.7	40.9	42.1
hypertension	70.0	60.0	47.7	51.8	spastic paraplegia	31.7	n/a	31.7	31.1
insulin	80.0	65.0	71.6	73.6	systemic lupus erythematosus	86.3	68.2	64.2	72.7
leukemia	84.6	67.3	75.8	81.6					

* Table shows the AUC values using the original network (org.), the critical AUC values (crit.) and the AUC values using perturbed networks (perm.; 50% of interactions permuted, del.; 50% of interactions deleted) for each pathophenotype.

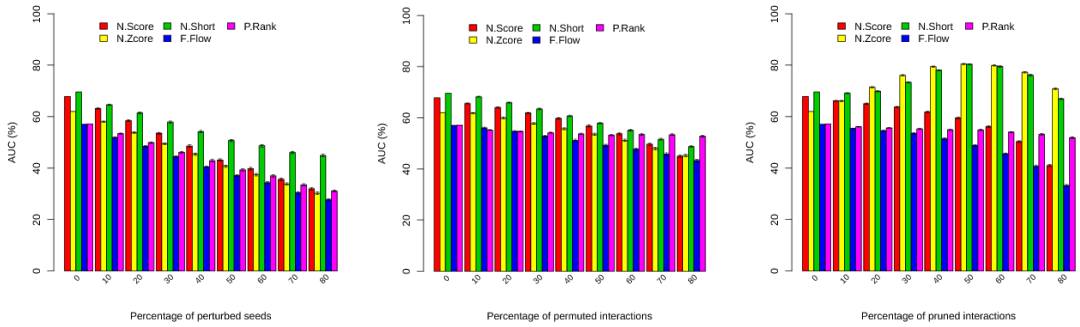


Figure 1

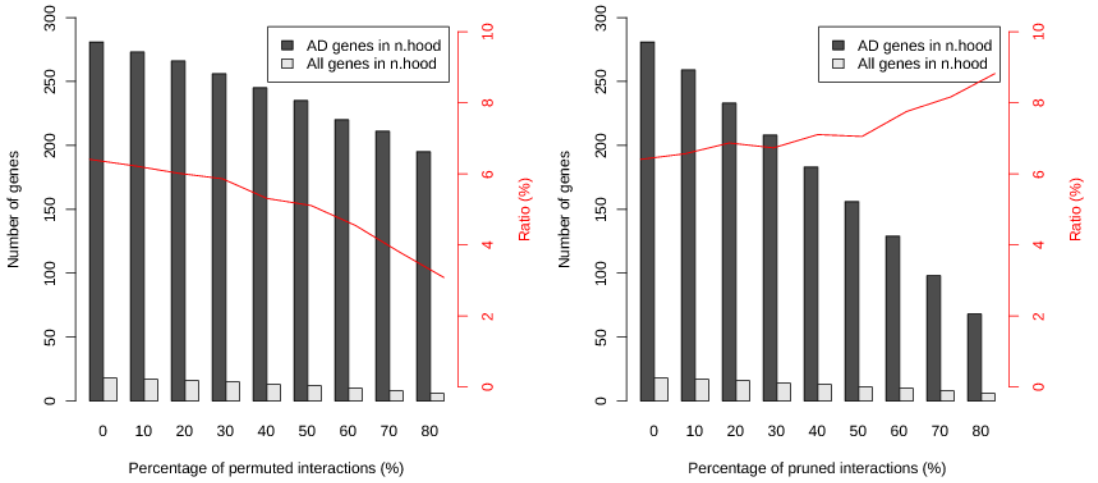


Figure 2

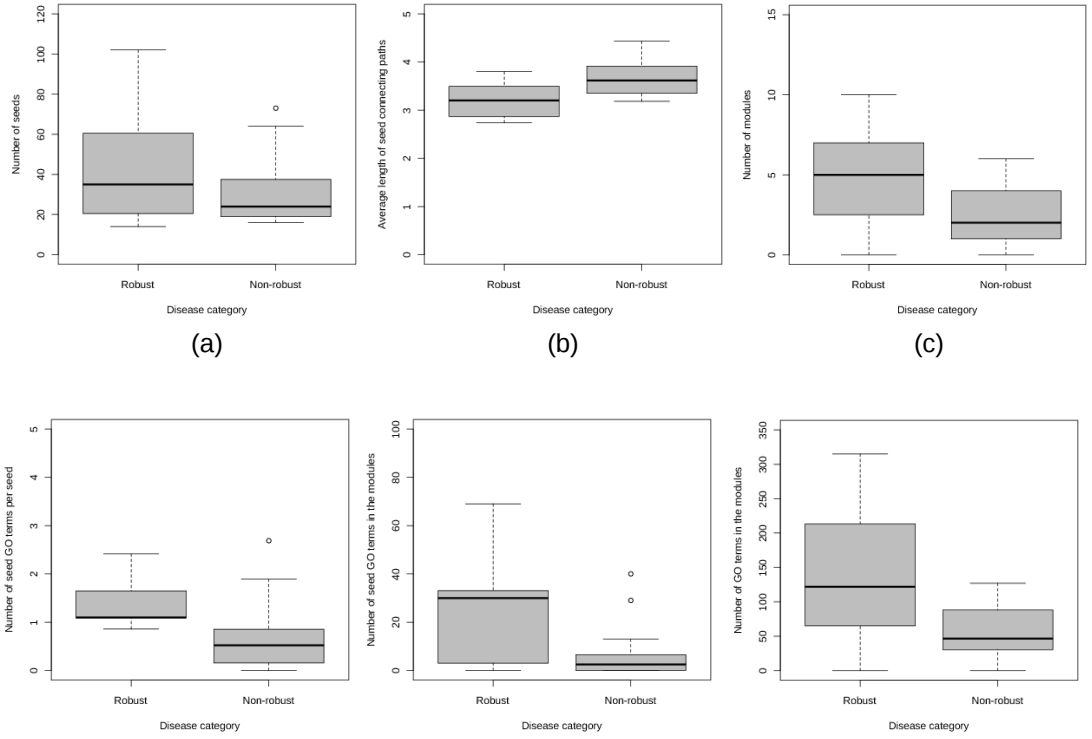


Figure 3

Supplementary material for Guney E. and Oliva B., Robustness of Network-based Disease Gene Prioritization Methods Points to Pathophenotypic Plasticity

SUPPLEMENTARY RESULTS

Investigating functional modularity of the top-ranking genes identified by the prioritization methods

Cells exhibit a modular nature in which genes are organized into various interconnected pathways and such organization of genes is often attributed as an essential component of robustness [1]. In this work, we questioned whether the prioritization methods identified groups of genes that were functionally associated with the disease phenotype. First, we carried out a module-based functional enrichment analysis, using top scored nodes associated with a disease identified by network-based prioritization approaches. Then, we examined the extent to which these prioritization algorithms were able to discover functions implicated in complex genetic disorders. We acquired GO terms that are significantly enriched in the set of genes associated with the disease (seed GO terms). We used five network-based prioritization approaches to score proteins associated with diseases from the OMIM database. We also used a trivial prioritization approach where all neighbors of the seeds were prioritized for the same pathophenotype (direct neighborhood). Next, we applied a clustering protocol based on MCL [2], a random-walk based graph clustering algorithm, to the sets of highly scored nodes or the neighbors of the seeds. For each disease, this procedure produced a set of clusters of proteins potentially implicated in the pathology of the disease. A cluster was potentially implicated in the disease, if and only if it contained two or more seeds (see Supplementary Methods for details). We checked the percentage of seed GO terms among all GO terms significantly enriched in these clusters, based on the GO terms enriched among the genes in the clusters. The number of clusters was smaller when prioritizing with NetScore and NetZcore than with the rest of methods (Supplementary Figure 2a). Moreover, the genes in the clusters prioritized by NetScore and NetZcore were slightly more enriched in seed GO terms, suggesting that they could identify genes more relevant to the disease, these being either seeds or genes involved in similar functions (Supplementary Figure 2b). In particular, the GO term enrichment in the clusters when prioritizing with the neighborhood approach was very low, while the number of clusters was large. We wish to note that if the number of clusters is large, the seeds may be functionally spread among several sets. In contrast, if the number of clusters is small, the genes (or proteins) in these clusters have similar functions that may characterize the main functional features of the disease. Our results showed that NetScore and NetZcore tended to form a small set of functionally coherent groups. These functionally coherent groups of genes are likely to entail pathways selectively altered throughout disease progression (Supplementary Tables 3 and 4).

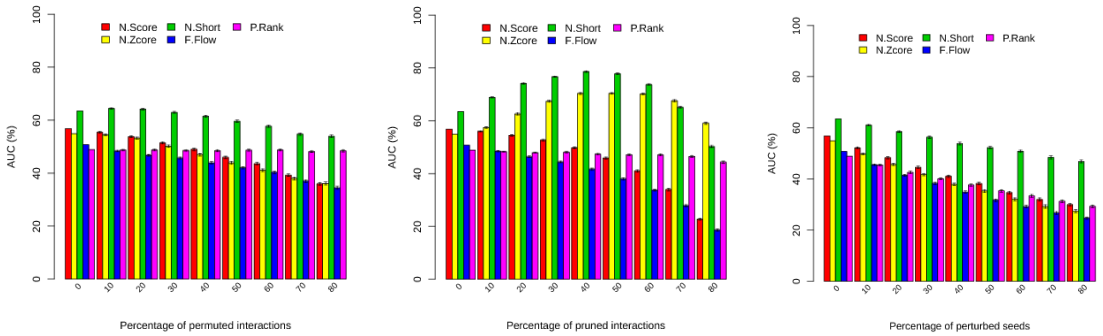
SUPPLEMENTARY METHODS

Network-based disease-gene prioritization methods

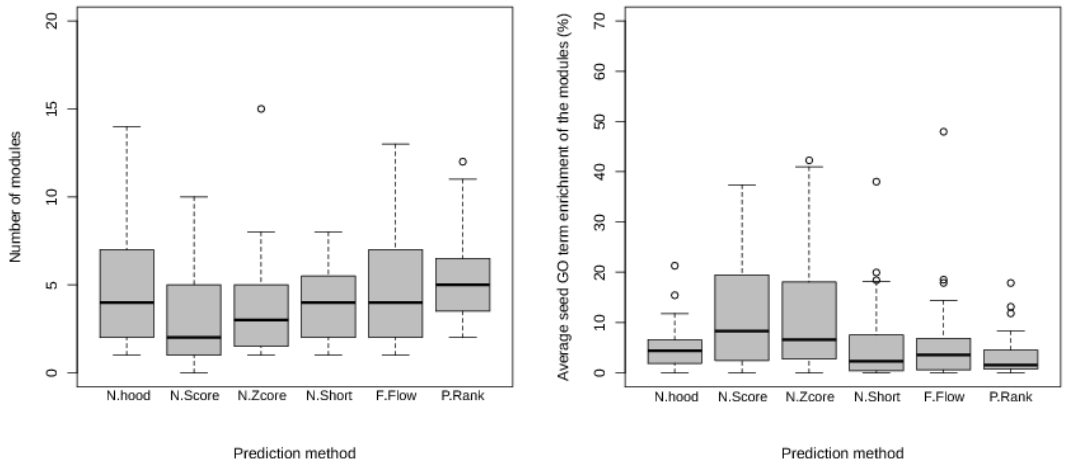
We recently proposed three disease-gene prioritization methods that use known disease-gene annotations and protein-protein interaction network to rank genes (or their proteins) with respect to a given phenotype [3,4]. The first method, named *NetShort,* considers not only the number of links that reach to the phenotype-associated node but also the number of phenoype-associated nodes that are included in the path. This is achieved by modifying the weight of the links in shortest path algorithm such that the links connecting seed nodes are shorter than the links connecting non-seed nodes. The second method, *NetZcore,* iteratively assesses the relevance of a node for a given phenotype by averaging the normalized scores of the neighboring nodes in a network. The normalized scores for each node is calculated using observed scores of the same node in a catalog of random networks that have the same topology as the original network. Finally, the third method, *NetScore* considers multiple shortest paths (if exist) from the source of information to the target for each node and ignores all other paths between them. All of these algorithms were implemented in C++ (requires GNU C++ compiler and GNU make) and freely available at http://sbi.imim.es/GUILD.php .

Identification of clusters associated with the disease. We defined two types of sub-networks to compare the enrichment of GO terms and specific functions in putative pathways: 1) a trivial approach also named “neighboring” used the nodes of a network formed by seeds of a disease and their interaction partners; and 2) a more sophisticated approach, using the highly scored nodes (proteins) of the network identified by a prioritization method. The edges of these sub-networks were taken from the original network, using all edges that connected any of the selected nodes. For assigning GO terms we used the genes that produce these selected proteins. We identified clusters in these sub-networks using the Markov cluster (MCL) algorithm [2]. The inflation parameter of MCL was chosen to be *1.7*, the optimum value obtained by Brohee and van Helden [5]. Only the clusters that contain two or more proteins translated by disease-associated genes were taken into consideration. We then determined the percentage of GO terms significantly enriched in the set of genes associated with the disease (seed GO terms) among all GO terms significantly enriched in identified clusters.

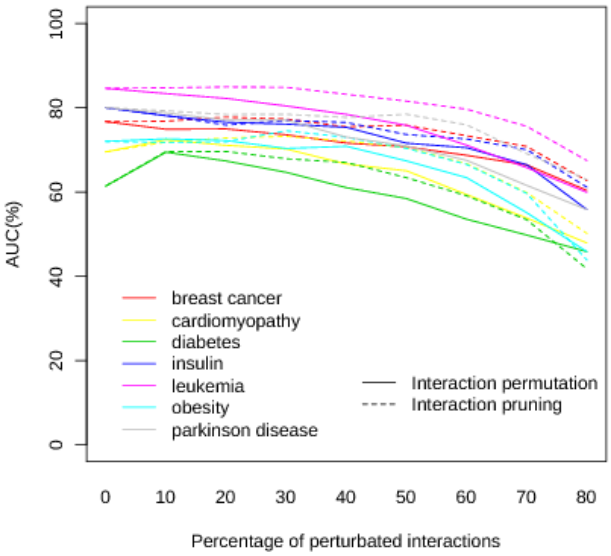
SUPPLEMENTARY FIGURES



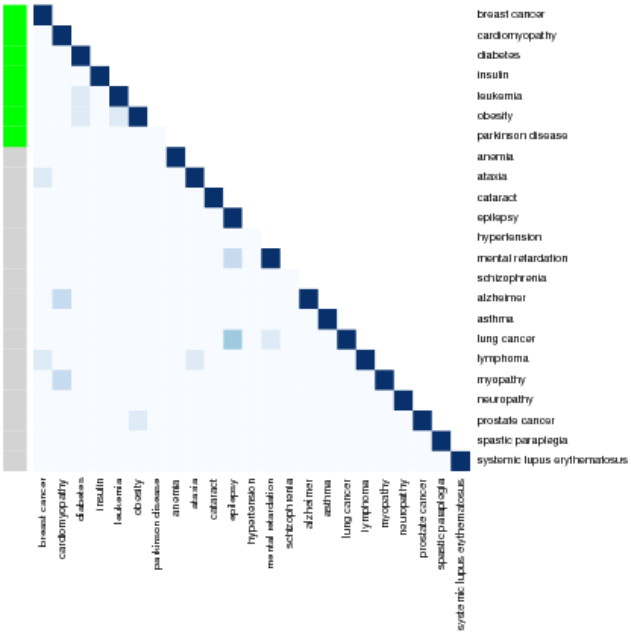
Supplementary Figure 1 Robustness of the methods with the perturbation of the edges of the Goh network. Plot of the average AUC (shown in bars) and confidence interval (shown with error lines) calculated for the prediction of gene-disease associations by NetScore (red), NetZcore (yellow), NetShort (green), Functional Flow (blue), and ToppGene (purple). The interactions of Goh network were perturbed (**a**) by swapping the links in order to make false interactions or (**b**) by removing interactions. The percentage of interactions swapped or removed varied between 0 and 80%. (**c**) Plot of the average AUC and confidence intervals calculated for the prediction of gene-disease associations as the percentage of mis-annotated seeds goes from 0% to 80%.



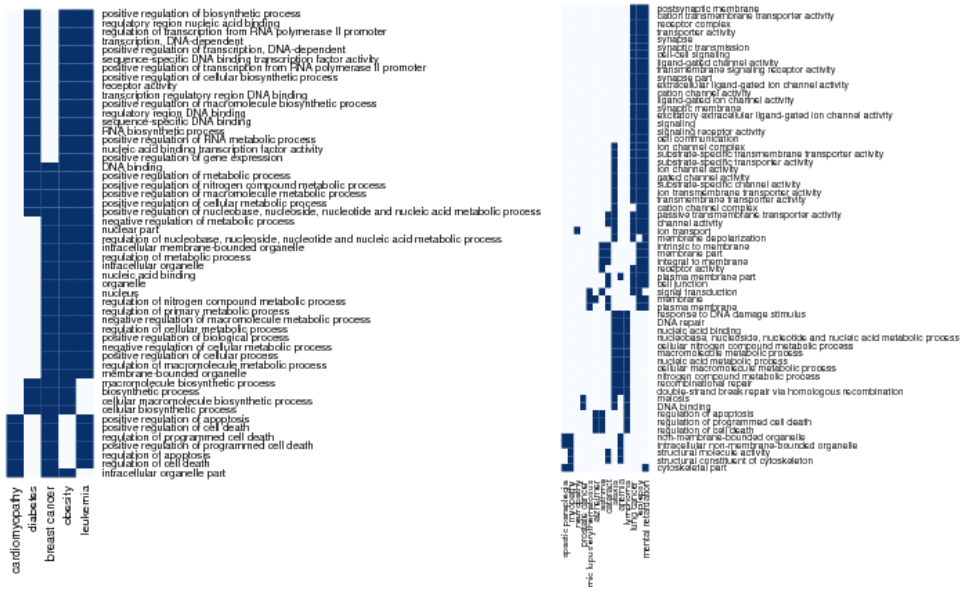
Supplementary Figure 2 Module-based functional enrichment analysis of prioritized subnetworks. **(a)** Number of modules identified in the neighborhood of known disease associations (N.hood) and in high scoring subnetworks identified by Functional Flow (Func.Flow), NetScore, NetZcore, NetShort, or ToppGene (T.Gene) prioritization methods. **(b)** Percentage of seed GO terms (GO terms significantly enriched in the set of genes associated with the disease) among all GO terms significantly enriched in the identified modules.



Supplementary Figure 3 AUC change at different levels of perturbations in the network for robust pathophenotypes. Solid lines correspond to AUC of the disease when NetScore prioritization method is applied to the interaction network whose edges are randomly swapped and dashed lines correspond to AUC of the disease when the prioritization method is applied to the interaction network whose edges are randomly deleted.



Supplementary Figure 4 Heatmap representation of Jaccard indices calculated for every pair of pathophenotypes using GO terms in the high scoring (top 5%) modules identified by NetScore prioritization method.



Supplementary Figure 5 GO terms shared in the high scoring modules of at least 3 diseases **(a)** in the robust diseases category **(b)** in the non-robust diseases category.

SUPPLEMENTARY TABLES

Supplementary Table 1 The list of the diseases used in this study, the number of genes associated with them and the number of gene products corresponding to these genes in the interaction network.

	# of genes	**# of gene products covered by the largest component (LCC) of the interaction network**	
Phenotype		bPPI	Goh
alzheimer	48	18	17
anemia	129	73	54
asthma	47	16	12
ataxia	131	44	30
breast cancer	70	35	29
cardiomyopathy	138	51	43
cataract	85	25	20
diabetes	182	70	63
epilepsy	139	33	26
hypertension	51	19	15
insulin	54	14	13
leukemia	162	102	75
lung cancer	43	22	16
lymphoma	39	23	22
mental retardation	257	64	46
myopathy	102	38	30
neuropathy	100	37	26
obesity	59	25	22
parkinson disease	52	16	13
prostate cancer	63	25	23
schizophrenia	45	19	13
spastic paraplegia	66	16	14
systemic lupus erythematosus	43	21	14

Supplementary Table 2 The interaction sources used to create the human protein-protein interaction network.

Network	Sources	Date of retrieval or version	# of nodes in LCC	# of edges in LCC
Goh [6]	Interactions collected from literature (i.e. large-scale Y2H experiments by Rual et al. [7] and Stelzl et al. [8])	n/a	7279	21911
bPPI network (BIANA [9])	DIP [10]	January 2009	11250	59220
	HPRD [11]	September 2007		
	IntAct [12]	January 2009		
	MIPS (MPACT) [13]	October 2008		
	BioGRID [14]	2.0.49		

Supplementary Table 3 Module-based functional enrichment analysis of prioritized subnetworks.

Provided as a separate XLS file.

Supplementary Table 4 Modules and GO terms enriched in the prioritized subnetworks.

Provided as a separate XLS file.

SUPPLEMENTARY REFERENCES

1. Kitano H (2004) Biological robustness. Nat Rev Genet 5: 826–837.

2. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.

3. Guney E, Oliva B (n.d.) Exploiting Protein-Protein Interaction Networks for Genome-wide Disease-Gene Prioritization. submitted.

4. Guney E, Oliva B (2011) Toward PWAS: discovering pathways associated with human disorders. BMC Bioinformatics 12: A12. doi:10.1186/1471-2105-12-S11-A12.

5. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

6. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685.

7. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–1178.

8. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.

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

10. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449–51.

11. Prasad TS, Kandasamy K, Pandey A (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol 577: 67–79.

12. Hermjakob HM-P (2004) IntAct: an open source molecular interaction database. Nucleic Acids Research 32: D452.

13. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, et al. (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436–41.

14. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res 36: D637–40.

Chapter 4

GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms

GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms

Emre Guney, Javier García-García, Baldo Oliva*
Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), Barcelona, 08003, Spain
* To whom correspondence should be addressed. Tel: +34 933160509; Fax: +34 933160550; Email: baldo.oliva@upf.edu

ABSTRACT

Determining genetic factors underlying various phenotypes is hindered by the involvement of multiple genes acting cooperatively. Over the past years, disease-gene prioritization has been central to identify genes implicated in human disorders. Special attention has been paid on using physical interactions between the proteins encoded by the genes to relate them with diseases. Such methods exploit the guilt-by-association principle over the protein interaction network to uncover novel disease-gene associations. These methods rely on the proximity of a gene in the network to other genes known to be involved in the same phenotype and typically require a known set of initial associations. Here, we present GUILDify, an easy to use web server for the phenotypic characterization of genes. Available at http://sbi.imim.es/GUILDify.php, GUILDify offers a whole genome protein-protein interaction network-based prioritization where the initial phenotype-gene associations are retrieved via free text search on biological databases. GUILDify web server does not restrict the prioritization to any predefined phenotype and supports multiple species.

INTRODUCTION

During the past decade, disease-gene prioritization has been central to research efforts in the field of human genetics. The promise of suggesting novel associations for genetic disorders with implications to therapeutical improvements has yielded a broad spectrum of computational tools (1, 2). These tools take into account a diverse set of data types such as functional annotation, gene expression, sequence properties, orthology, text mining and protein-protein interactions to extend known disease-gene associations. Given that the genetic factors deriving a phenotypic trait usually consist of more than a single gene, a special attention has been paid on using physical interactions between the products of these genes to relate them with diseases (3, 4).

Following the emergence of high throughput interaction detection experiments, protein interaction network-based approaches have been employed to prioritize genes involved in various diseases (5–10). Network-based methods exploit the “guilt-by-association” principle over the network topology to uncover new disease-gene associations. The guilt-by-association principle suggests that the genes whose products (proteins) interact with the products of known disease genes are more likely to be disease genes (11, 12). The methods using global topology of the network has been demonstrated to predict disease genes more accurately compared to methods that use only local neighborhood of genes (8). Recently, we proposed three novel algorithms for genome-wide prioritization of disease-genes using protein-protein interaction (PPI) networks (7) and showed that a consensus method combining these algorithms, outperformed the state-of-the-art global network topology-based prioritization methods (Guney and Baldo, *manuscript in preperation*) using the disease-gene associations in OMIM database (13).

All of the network-based methods rely on the proximity of a gene in the network to other genes known to be involved in the same phenotype, thus require a known set of initial associations (seed genes). Among available global topology network-based tools, GeneWanderer uses known disease-gene annotations in OMIM database (13) and prioritizes the genes lying under a given genomic interval by applying random walk with restart algorithm (14). Similarly, PRINCIPLE retrieves disease-gene annotations from OMIM for a disease term provided by the user and ranks the genes by network propagation algorithm (15). In ToppNet a set of genes provided by the user are fed to PageRank with priors (or to a similar) algorithm to rank another set of genes provided by the user (16). PINTA considers differentially expressed genes in disease-specific expression data for the phenotype of interest as seed genes and employs either diffusion based or random walk based algorithm for whole genome or candidate genes provided by the user (17).

Publicly available biological data repositories can also be utilized for mining initial phenotype-gene associations required by network-based prioritization methods without restricting to any predefined phenotype. Considering the limited availability

of convenient interfaces that bridge network-based prioritization algorithms to end users, we present GUILDify, a whole genome PPI network-based prioritization server for phenotypic characterization. GUILDify ranks the genes for their relevance to a given phenotype. It supports multiple species and can retrieve initial phenotype-gene associations (seeds) via free text search on biological databases for any given phenotype. The genes extracted from biological databases and/or provided by the user are fed to the network-based prioritization methods we have developed recently. GUILDify is available at http://sbi.imim.es/GUILDify.php (free and open to all users and there is no login requirement).

GUILDIFY WORKFLOW

The workflow of GUILDify web server consists of three simple steps (see Supplementary Figure 1 for a schematic overview). In the first step, the user enters some text describing the phenotype of interest and chooses the organism for which she wants to run whole genome prioritization based on protein-protein interaction networks. In the next step, among all the proteins encoded by the genes that have been found to be associated with the phenotype, the user chooses which proteins to designate as seeds (source of information) in network-based prioritization. At this step, the user may add additional genes (and their products) that are not listed by the text-based search at the first step. In the final step, the network-based prioritization algorithm is run and the ranking of the products of genes is displayed to the user. The ranking represents the predicted level of association with the phenotype of interest.

RETRIEVING PHENOTYPE-GENE ASSOCIATIONS

When a query is made through GUILDify web interface (any text describing a phenotype such as disease, biological function or pathway), the query is tokenized into keywords. The descriptive fields of UniProt (both Swissprot and Trembl) database --“description”, “disease”, “function” and “keyword”-- are searched for matching keywords. In addition to these descriptive fields in UniProt, OMIM disorder names and GO term names are searched for matching keywords. GUILDify also supports quoted phrases to query the occurrence of a group of words in biological databases. Therefore to describe a phenotype that consists of multiple words, the words should be quoted.

GUILDify keeps a local copy of these databases for fast search and retrieval of information. Once the genes matching to the query are identified, GUILDify uses BIANA (18), a tool that integrates biological data spread over various publicly available proteomics and interaction databases, to list proteins encoded by the matching genes. The BIANA knowledge base integrated in GUILDify consists of the following biological data repositories (all downloaded after May, 2011): UniProt (19), HGNC (20), DIP (21), HPRD (22), IntAct (23), MINT (24), MPact (25), BioGRID (26), BIND (27). The entries in these databases are unified using common identifiers such as *UniProt Accession*, protein sequence and *Entrez Gene Identifier*.

Though literature provides an initial catalog of phenotype-gene annotations, these annotations may not always be complete. Furthermore, in some cases the annotation provided in the literature could be spurious or the user simply may not want to include these annotations in the whole-genome prioritization. Thus, the users may provide any number of genes that are not listed by GUILDify and/or may choose to use a subset of the listed genes.

WHOLE GENOME PRIORITIZATION USING PPI NETWORK

Given a set of initial phenotype-gene associations (seeds), GUILDify maps these associations onto a genome-wide PPI network and runs a global topology-based prioritization algorithm developed recently in our group under the context of GUILD framework (7). The species-specific PPI network is generated using the interaction databases integrated in BIANA (listed above). GUILD (Genes Underlying Inheritance Linked Disorders) implements three algorithms (see Supplementary Material) to prioritize genes potentially involved in diseases using *a priori* gene-disease associations and protein-protein interactions. GUILDify uses a consensus method combining the scores of the former three algorithms, for scoring the relevance of all protein in the interaction network (see Supplementary Material for details).

OUTPUT

GUILDify outputs a likelihood score (GUILD Score) associating the gene product with the phenotype provided by the user

for each gene product in the PPI network. In addition to GUILD Score, descriptive information of the gene products such as *UniProt Accession*, gene symbol, *Entrez gene id* and description is included in the output. The files containing all of this information as well as the seed proteins used in the prioritization method can be downloaded via the links provided on the result page.

IMPLEMENTATION

The interface of the web server is built using Pyramid, a Python platform for web programming, with an emphasis on simplicity and ease-of-use. MySQL is used to store and retrieve biological data. The prioritization algorithms are implemented in C++ and a two-node cluster is used on the web server to schedule job requests. See BIANA http://sbi.imim.es/web/BIANA.php and GUILD http://sbi.imim.es/web/GUILD.php for the details on their implementation.

GUILDify is designed to be as simple as possible. Many algorithmic details such as internal parameters used by the prioritization algorithms are hidden from the user. These parameters are chosen the values that are shown to be optimum on a large data set of disease phenotypes under the context of GUILD project. The users that are interested in using user-defined parameters are advised to refer to download stand-alone software provided in the GUILD project page.

VALIDATION

We confirmed the validity of our predictions using Alzheimer's disease phenotype in human. For Alzheimer's disease, we used differentially expressed genes in Alzheimer's disease identified in a previous study (28) as true positives. True negatives were randomly selected among genes whose association to the phenotype is unknown (such that there was an equal number of true negatives as there were true positives). The Kratuhammer data set contained 60 genes and only one of them was not in the PPI network. After removing 11 seeds listed by GUILDify, the remaining 48 genes used for the validation. On this data set, the precision obtained by using “alzheimer” keyword on the web server was 83% at 71% sensitivity level.

COMPARISON WITH EXISTING WEB SERVICES

In order to compare GUILDify with existing network-based prioritization web services, we checked the coverage of the genes in the aforementioned data set among the top ranking genes provided by these services (Table 1). The web services included in the comparison were ToppNet and PRINCIPLE. GeneWanderer is excluded from the comparison since it prioritizes the genes only under a given linkage interval. PINTA is also not taken into consideration since it requires phenotype-specific expression data.

We used the seeds identified by GUILDify as training genes and all the genes in the PPI network as test genes in ToppNet. We ranked the genes in the test set using PageRank with priors algorithm with the default parameters and the built-in interaction network of ToppNet. Since ToppNet does not provide a ranking for the training genes, we assumed the best case scenario and considered all the seed genes contained in the benchmarking set (11 genes) as having the highest rank. In the Cytoscape plugin of PRINCIPLE, we chose “Alzheimer Disease (AD)” and only modified the k-cutoff parameter (number of top genes to output) by setting it to its maximum value such. Then we considered all the top ranking genes and their neighbors identified by the plugin using network propagation algorithm and the interaction network integrated into the web service.

Table 1. Comparison of web services with respect to the number of benchmark genes covered among the top ranking genes.

Web Service	**Number of top ranking genes considered**			
	100	**250**	**500**	**1000**
GUILDify	11	17	26	34
ToppNet	16	25	31	37
PRINCIPLE	13	21	27	35

CONCLUSION

Phenotypic characterization of genes plays a crucial role in explaining the mechanisms behind biological processes. We have developed GUILDify, a free and easy to use web server for whole genome prioritization of genes using PPI networks. For a given phenotype, GUILDify uses descriptive fields in several proteomics and genomics databases in combination with network-based prioritization methods and provides a genome-wide ranking. The ranking represents the relevance to the phenotype of interest and can be used to shortlist the set of candidate genes that needs to be further validated. Though the prediction performance is far from perfect, GUILDify serves to generate a meaningful ranking among genes just by applying PPI network-based whole genome prioritization over the data extracted from biological databases.

ACKNOWLEDGEMENTS

E.G. would like to acknowledge the technical support from GRIB IT team, in particular Alfons Gonzalez Pauner and Miguel Sánchez Gómez.

FUNDING

This work was supported by “Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu” through an FI fellowship granted to E.G.; by the grants from the Spanish Ministry of Science and Innovation (MICINN), FEDER BIO2008-0205, BIO2011-22568, PSE-0100000-2007, and PSE-0100000-2009; and by EU grant EraSysbio+ (SHIPREC) Euroinvestigación (EUI2009-04018).

Conflict of interest statement: none declared.

REFERENCES

1. Kann,M.G. (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform*, **11**, 96–110.

2. Tranchevent,L.-C., Capdevila,F.B., Nitsch,D., De Moor,B., De Causmaecker,P. and Moreau,Y. (2011) A guide to web tools to prioritize candidate genes. *Brief Bioinform*, **12**, 22–32.

3. Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

4. Barabasi,A.-L., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**, 56–68.

5. Chen,J., Aronow,B. and Jegga,A. (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, **10**, 73.

6. Franke,L., Bakel,H. van, Fokkens,L., de Jong,E.D., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes. *The American Journal of Human Genetics*, **78**, 1011–1025.

7. Guney,E. and Oliva,B. (2011) Toward PWAS: discovering pathways associated with human disorders. *BMC Bioinformatics*, **12**, A12.

8. Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.

9. Vanunu,O., Magger,O., Ruppin,E., Shlomi,T. and Sharan,R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, **6**, e1000641.

10. Wu,X., Jiang,R., Zhang,M.Q. and Li,S. (2008) Network-based global inference of human disease genes. *Molecular Systems Biology*, **4**.

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

12. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N., et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotech*, **25**, 309–316.

13. Hamosh,A. (2004) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**, D514–D517.

14. Köhler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, **82**, 949–958.

15. Gottlieb,A., Magger,O., Berman,I., Ruppin,E. and Sharan,R. (2011) PRINCIPLE: A tool for associating genes with diseases via network propagation. *Bioinformatics*.

16. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, **37**, W305–W311.

17. Nitsch,D., Tranchevent,L.-C., Goncalves,J.P., Vogt,J.K., Madeira,S.C. and Moreau,Y. (2011) PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Research*, **39**, W334–W338.

18. Garcia-Garcia,J., Guney,E., Aragues,R., Planas-Iglesias,J. and Oliva,B. (2010) Biana: a software framework for

compiling biological interactions and analyzing networks. *BMC Bioinformatics*, **11**, 56.

19. Wu,C.H. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, **34**, D187–D191.

20. Seal,R.L., Gordon,S.M., Lush,M.J., Wright,M.W. and Bruford,E.A. (2010) genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, **39**, D514–D519.

21. Salwinski,L. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, **32**, 449D–451.

22. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A., et al. (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Research*, **37**, D767–D772.

23. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U., et al. (2011) The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, **40**, D841–D846.

24. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E., et al. (2011) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, **40**, D857–D861.

25. Mewes,H.W., Ruepp,A., Theis,F., Rattei,T., Walter,M., Frishman,D., Suhre,K., Spannagl,M., Mayer,K.F.X., Stumpflen,V., et al. (2010) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Research*, **39**, D220–D224.

26. Stark,C., Breitkreutz,B.-J., Chatr-aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X., et al. (2010) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, **39**, D698–D704.

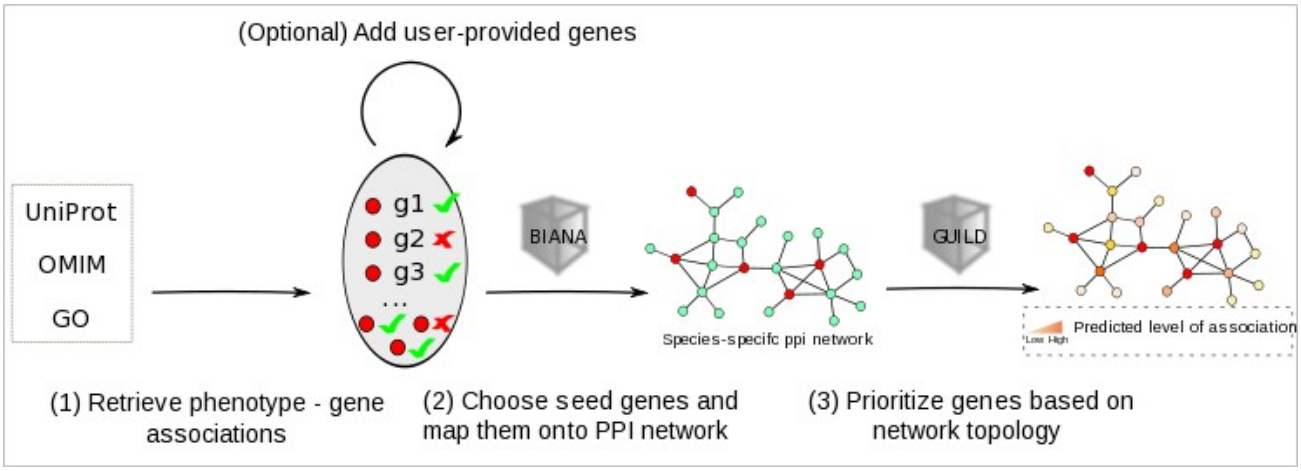
27. Isserlin,R., El-Badrawi,R.A. and Bader,G.D. (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)*, **2011**.

28. Krauthammer,M., Kaufmann,C.A., Gilliam,T.C. and Rzhetsky,A. (2004) Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease. *PNAS*, **101**, 15148–15153.

SUPPLEMENTARY DATA

This text contains the supplementary information for the manuscript entitled "*GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms*" by Emre Guney, Javier García-García and Baldo Oliva.

GUILDIFY WORKFLOW



Supplementary Figure 1. Schematic overview of GUILDify workflow. (1) A text describing the phenotype of interest is queried over several genomics and proteomics databases and the genes (proteins) matching with the input query for the species of interest are listed. (2) The seed genes (proteins) are chosen among these and mapped onto the PPI network. At this step, the user may add additional genes (proteins) that are not listed by the text-based search at the first step. (3) Using the seeds and species-specific PPI, all the genes (proteins) in the network are prioritized and the ranking of the genes (proteins), corresponding to the predicted level of association with the phenotype of interest, is displayed to the user.

PRIORITIZATION METHODS USED IN GUILDIFY

GUILDify uses *NetCombo* algorithm, a consensus method which combines the following three prioritization methods: *NetShort*, *NetZcore* and *NetScore* (1). *NetShort* incorporates "phenotypic-relevance" of the path between a node and the nodes of a given phenotype by considering not only the number of links that reach to the phenotype-associated node but also the number of phenoype-associated nodes that are included in the path. Algorithmically, this is achieved by modifying the weight of the links in shortest path algorithm such that the links connecting seed nodes are shorter than the links connecting non-seed nodes. *NetZcore* iteratively assesses the relevance of a node for a given phenotype by averaging the normalized scores of the neighboring nodes in a network. The normalized scores for each node is calculated using observed scores of the same node in a catalog of random networks that have the same topology as the original network (yielded by swapping nodes in the original network). *NetScore* is based on the propagation of information (i.e. data packets) through the nodes in the network by considering multiple shortest paths (if exist) from the source of information to the target and ignoring all other paths between them. In addition to the scores of the nodes, the data packet contains the weight of the path that the packet has traveled. Iteratively, a score is calculated for each node using the packets it has received from shortest paths (more than once if multiple shortest paths exist) weighted by the edge weights along these paths. *NetCombo* combines NetScore, NetShort and NetZcore in a consensus scheme by averaging the normalized score of each prioritization method.

We tested the prioritization methods we proposed in GUILD using multiple gene-disease association data sets (from OMIM database) and protein-protein interaction networks. In this genome-wide analysis, *NetCombo* outperformed state-of-the-art network-based prioritization methods such as Functional Flow, PageRank with Priors, Random walk with restart and Network propagation (*manuscript under preparation*).

COMPARISON WITH EXISTING WEB SERVICES

We used the seeds identified by GUILDify as training genes and all the genes in the PPI network as test genes in ToppNet (2). We ranked the genes in the test set using PageRank with priors algorithm with the default parameters and the built-in interaction network of ToppNet. Since ToppNet does not provide a ranking for the training genes, we assumed the best case scenario and considered all the seed genes contained in the benchmarking set (11 genes) as having the highest rank. In the Cytoscape plugin of PRINCIPLE (3), we chose “Alzheimer Disease (AD)” and only modified the k-cutoff parameter (number of top genes to output) by setting it to its maximum value. Then, we considered all the top ranking genes and their neighbors identified by the plugin using network propagation algorithm and the interaction network integrated into the web service.

SUPPLEMENTARY REFERENCES

1. Guney,E. and Oliva,B. (2011) Toward PWAS: discovering pathways associated with human disorders. *BMC Bioinformatics*, **12**, A12.

2. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, **37**, W305–W311.

3. Gottlieb,A., Magger,O., Berman,I., Ruppin,E. and Sharan,R. (2011) PRINCIPLE: A tool for associating genes with diseases via network propagation. *Bioinformatics*.

Chapter 5

Understanding Cancer Progression Using Protein Interaction Networks

Springer Book on Systems Biology in Cancer Research and Drug Discovery

Chapter 7 Understanding Cancer Progression Using Protein Interaction Networks

Emre Guney[*], Rebeca Sanz[#], Angels Sierra[+], Baldo Oliva[*]

* Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, PRBB, Dr. Aiguader 88, 08003, Barcelona, Catalunya, Spain

+ Centre d'Oncologia Molecular, Institut de Recerca Oncològica- IDIBELL, Hospital Duran i Reynals, Gran Via, s/n, Km 2.7, L'Hospitalet L1., 08907 Barcelona, Spain

Unit of Biomarkers and Susceptibility, ICO-IDIBELL, Hospital Duran i Reynals, Gran Via, s/n, Km 2.7, L'Hospitalet L1., 08907 Barcelona, Spain

Emails: emre.guney@upf.edu, rsanz@idibell.cat, asierra@idibell.cat, baldo.oliva@upf.edu

Keywords

protein-protein interaction
network biology
network medicine
active subnetwork
metastasis
metastatic breast cancer
guilt-by-association

Abstract

Cancer is produced by perturbations affecting several genes and pathways. Environmental stimuli trigger uncontrolled cell growth and invasion of other tissues. Understanding cancer progression requires a profound knowledge of the pathways involved in the communication between proteins and genes at a systems level. Consequently, protein-protein interaction networks play an important role in delineating cancer related pathways. Our understanding of cancer has evolved towards the co-operation of groups of genes that constitute pathways. In this chapter, we describe the characteristics of genes involved in cancer and the relationships between them in the context of the protein-protein interaction network. We also explain several methods to predict novel candidates that are potentially involved in cancer and its progression using topological information encoded in the protein-protein interaction network. Towards developing effective network-based therapeutics, we give details of identifying dysregulation patterns in cancer using protein-protein interaction networks with an emphasis on the underlying mechanisms of progression in metastatic breast cancer.

Contents

1 Introduction
2 Protein-protein interaction networks underlying cancer
3 Complementary network models based on cancer mediated gene expression changes
4 Network guided prediction of relevant genes involved in cancer progression
5 Predicting disease prognosis using dysregulation patterns in networks
6 Discovery of biological pathways in breast cancer metastases: A network biology perspective
7 Future directions: From networks to systems medicine

Abbreviations

PPI **Protein-protein interaction**
GO **Gene ontology**
OMIM **Online Mendelian Inheritance in Man**
ROC **Receiver-operatoring characteristic**
AUC **Area under ROC curve**
GRP **Glucose Regulated Proteins**

1. Introduction

Cancer is the outcome of perturbations in the orchestral activity of genetic elements. Environmental stimuli disturb the genetic regulation circuitry composed of such genetic elements and trigger uncontrolled cell growth and invasion of other tissues (Hanahan and Weinberg, 2011). These consequences can only be counteracted through profound understanding of the pathways underlying the progression of cancer at a systems level (Laubenbacher, et al., 2009). This argument follows from recent works showing that interactions between gene products mediate pathways involved in cancer (Jonsson and Bates, 2006; Vogelstein and Kinzler, 2004). Therefore, studying dysregulation patterns of the protein-protein interaction network is key to delineate cancer related pathways (Mani, et al., 2008) and to develop effective treatment strategies such as network-based therapeutics (Russell and Aloy, 2008).

Among variety of molecular mechanisms involved in cancer, gene regulation, signaling and cell metabolism pathways as well as the cross-talk between them are the most relevant ones (Hanahan and Weinberg, 2011). We can perfectly assume that cancer originates from a single cell. Due to a change in the expression behavior of one or more genes involved in the regulation of the cell growth and differentiation, the cell starts abnormally replicating itself through mitosis. Effected genes are classified in two categories depending on their functional roles: oncogenes and tumor suppressor genes. Oncogenes promote cell growth and reproduction hindering the cell from going programmed cell death (*apoptosis*), whereas tumor suppressor genes inhibit cell division and survival. The failure in the cell cycle regulation is typically caused by mutations in several oncogenes and/or tumor suppressor genes. These mutations rapidly accumulate over the following generations of cells. This faulty group of cells constitutes the *primary tumor.* Cells of the primary tumor can travel within the body through the lymph and blood and may establish *secondary* (*metastatic*) *tumor* in a location different than the tissue from where it was originated. Metastatic colonization requires adaptation to the microenvironment of the distant organ site and this introduces a constraint on the tissues where metastatic cells can possibly invade (see reviews (Gupta and Massague, 2006; Steeg, 2006; Valastyan and Weinberg, 2011)). Furthermore, the adaptation mechanism varies depending on the tissue of origin of the primary tumor cells and the tissue (or tissues) where these primary tumor cells metastasize (Valastyan and Weinberg, 2011). That is, the genetic and epigenetic programs employed by metastatic breast cancer cells in the bone are different than the ones in metastatic prostate cancer cells or those metastasizing in brain, liver or lungs. Recent studies demonstrated that the invasion capacity of cancer cells are not limited to primary tumors, also metastatic tumors have the ability of infiltrating back to their primary tumors (Comen, et al., 2011; Valastyan and Weinberg, 2011).

Invasion of distant organs through metastatic colonization has especially attracted attention due to almost 90% of causalities being attributed to metastases rather than the primary tumors from which they originated (Gupta and Massague, 2006). However, the cellular processes governing metastases are still far from understood. The two main drawbacks in the study of cancer progression are cellular heterogeneity within tissues and genetic heterogeneity across patients (Chuang, et al., 2010). Cellular heterogeneity between the cells of the same tissue implies that they don't have a well-defined distinct phenotype for a specific cancer type. Genetic heterogeneity refers to the condition where different patients have different expression patterns for the same gene. This clearly implies two different perspectives describing cancer progression: changes of expression of few genes on one side and the pathways (i.e. interaction networks) affected by them on the other. Consequently, the impact of a small perturbation is amplified to the point that the survival of a complex system, such as an organism, is jeopardized. The analysis of gene expression has provided insights into the elements that change their patterns of expression during the progression of various cancer types (Quackenbush, 2006). However, to be able to characterize the associated cellular processes, we need to account for the complex interactions between these genetic elements.

Advances in biological data collection and bioinformatic techniques gave rise to more systematic approaches for the interpretation of the data. Systems biology fundamentally challenges the gene centric view of cancer. This view takes advantage of the present increase and availability of data describing biological molecules and their relationships. Through integration of different data sources such as protein sequence, gene expression and protein-protein interactions, our understanding of cancer has evolved rapidly towards the co-operation of groups of genes that constitute pathways. Consequently, the past decade witnessed a brand new perspective, named network biology. In network biology, genes, their products and the interactions between them are studied within the framework of the biological system as a whole.

In this chapter, we first delineate the characteristics of genes involved in cancer and the relationships between them in the context of protein-protein interaction (PPI) network (Section 2.2). Then, in Section 2.3 we explain several network modeling approaches that incorporate genomics data, complementing the networks created by PPIs. In the following section (Section 2.4), we describe methods to predict novel candidates that are potentially involved in cancer and its progression

using topology information encoded in the PPI network. Next, we give details of several network biology approaches to identify dysregulation patterns in cancer and how to use them to improve our knowledge on the prognosis (Section 2.5). In Section 2.6, we focus on a number of works that shed light on the underlying mechanisms of progression in metastatic breast cancer. Finally, we provide an outlook on the use of network medicine approaches towards developing effective treatment strategies in Section 2.7. The reader, unless familiar with the topics that are mentioned throughout the text, such as the integration of biological data or graph theory concepts, is advised to refer to Boxes 1, 2 and 3 where concise information on these subjects are provided.

2. Protein-protein interaction networks underlying cancer

Coordinated relationships between biological molecules help to describe a particular phenotype. Cancer can be studied under the same approach. PPI networks provide a framework to study the functional relationships among such biological molecules (Barabasi and Oltvai, 2004). For instance, topological features of gene and protein interaction networks helped attributing functions to genes whose functions were not characterized (Schwikowski, et al., 2000). Studying network properties of human genes also provided valuable insights in diseases and cancer in particular (Ideker and Sharan, 2008; Vidal, et al., 2011).

The first step in the systematic study of relationships between genetic elements is the integration of data describing various aspects of the biomolecules as well as the links between them (e.g., physical or functional associations). With the recent emergence of high-throughput interaction detection experiments, a substantial amount of data of physical PPIs in human became publicly available (Dreze, et al., 2010; Venkatesan, et al., 2009). These interaction maps offer an invaluable resource for network-based studies. However, they are incomplete, they content a large amount of noise (false interactions) and occasionally biased towards highly studied proteins (Schwartz, et al., 2009; Venkatesan, et al., 2009). Thus, PPI networks by themselves provide only a partial view of the complex biological processes. A comprehensive understanding of complex diseases, such as cancer, lies beneath the integration of biological data at different levels (genomics, transcriptomic, proteomic and metabolomic) and the analysis of the relationships between biomolecules in a dynamic context (Joyce and Palsson, 2006; Rhodes and Chinnaiyan, 2005, Xia, 2011 #198). Several data repositories widely used in the context of network-based cancer studies are given in Box 1.

10% of human genes are estimated to contribute to oncogenesis (Strausberg, et al., 2003), in fact this seems to be a general trend in all diseases (Amberger, et al., 2009). Cancer genes have shown to be distinguishable from normal genes based on structural, functional and evolutionary properties (Furney, et al., 2008). Nonetheless, cancer is typically seen as a disease where many different perturbations produce a similar phenotype due to the underlying interrelated pathways (Barabasi, et al., 2011; Wood, et al., 2007). Incorporating PPI information is indispensable to characterize such pathways. Topological characteristics of genes implicated in cancer have been thoroughly investigated during the past years. Wachi *et al.* (Wachi, et al., 2005) analyzed cancer tissue samples and normal tissue samples surrounding the tumor from 5 patients with lung squamous cancer (a subtype of lung cancer associated with smoking) in the framework of a protein interaction network. The network contained known physical protein-protein interactions and predicted interactions using interology (via transferring interactions from model organisms). Their analysis showed that genes that were up-regulated in cancer were highly connected in the network (i.e. their interaction partners were also up-regulated genes). Furthermore, these up-regulated genes were identified to be central, where centrality was defined as the enrichment of up-regulated genes in each k-core in the network (refer to Box 2 for the basic concepts in graph theory).

Similarly, Jonsson and Bates (Jonsson and Bates, 2006) used a larger human interactome constructed by interology and a comprehensive census of human cancer genes from the work of Futreal et al. (Futreal, et al., 2004) to analyze network positions of cancer related genes. They revealed that products of genes that were susceptible to mutations leading to cancer had the predisposition to interact with each other twice as much as the products of non-cancer genes. They also identified commonalities among the nodes of the network by clustering the network into overlapping subnetworks (union of k-cliques that share k-1 nodes with each other). The cancer proteins were found to be involved in larger groups than non-cancer proteins. Moreover, they were observed to reside in the overlapping parts of the clusters more often than their non-cancer counterparts.

In a network centric analysis of gene-disease associations for 22 disorder classes from Online Mendelian Inheritance in Man (OMIM) (Hamosh, et al., 2005), Goh et al. (Goh, et al., 2007) showed that products of genes involved in the same disorder class were predisposed to interact with each other. These genes constituted functional modules enriched in Gene Ontology (GO) (Ashburner, et al., 2000) terms. Genes from the same disorder class were also more likely to be homogeneously co-

expressed in tissues related to the pathology. Although, in general, most disease genes were found to be non-essential and had no tendency to encode hubs in the network, their analysis on somatic cancer genes demonstrated that cancer genes likely encoded protein-hubs, pointing out the functional and topological centrality of the cancer genes.

Kar and colleagues (Kar, et al., 2009) overlaid structural information of proteins and their binding sites on top of a protein-protein interaction network. They analyzed the structural properties of binding sites of cancer-related proteins. They started with the human PPI network used by Jonsson and Bates (Jonsson and Bates, 2006) (13,584 nodes and 85,083 edges) and mapped known and predicted protein interfaces (from PDB (Berman, et al., 2000) and PRISM (Ogmen, et al., 2005) web server) on the protein interaction network. The resulting network contained 534 nodes and 549 edges where edges corresponded to structurally characterized interaction sites. Their analysis revealed that cancer-related proteins had smaller, more planar, more charged and less hydrophobic binding sites than non-cancer proteins. Moreover, cancer-related protein-hubs in the interaction network tended to be essential, to interact with each other and to employ distinct interfaces in their interactions with their partners (multi-interface hubs) compared to their non-cancer counterparts. Based on their findings, the authors claimed that cancer proteins usually employ transient interactions and are more likely to be involved in multiple pathways. They further showed that such structural properties of binding interfaces could be used to classify the cancer phenotype in which the protein was implicated (e.g. leukemia, breast cancer or colorectal cancer) with accuracies ranging between 60-70% depending on the cancer type (refer to Box 3 for definitions of metrics used in the evaluation and classifier performance). Their predictions could help efficiency of drug discovery by suggesting potential targets to be used in cancer therapy (Fry and Vassilev, 2005).

3. Complementary network models based on cancer mediated gene expression changes

Network models have been employed to describe and infer relationships between co-expressed genes involved in cancer. Complementary to physical interaction networks, such models capture gene expression changes mediated by the disease. Ergun *et al.* (Ergun, et al., 2007) created a network model of regulatory interactions using 1144 whole-genome expression profiles spanning various cancer types such as adrenal, brain, breast, leukemia, lung, prostate and thyroid. The influence of all transcript concentrations and external factors were taken into consideration to model the transcript synthesis rate and infer the regulatory network using a reverse engineering approach. The regulatory network was then used to attribute genes to the progression of prostate cancer. Using 14 non-recurrent primary and 9 metastatic prostate-cancer samples, they identified the genes in the regulatory network with significant expression changes as the genetic mediators of prostate cancer. Their results demonstrated the role of the androgen receptor pathway in the progression of metastatic prostate cancer.

Based on the hypothesis that genes implicated in cancer initiation and progression show dysregulated interactions with their molecular partners, Mani *et al.* (Mani, et al., 2008) emphasized the identification of molecular interactions that were significantly dysregulated in a particular phenotype. They used a genome-wide cellular interaction network for human B-cells involving several key molecular interaction types such as transcriptional interactions, signaling events, and complex formation in combination with microarray expression profiles from normal, tumor-related, and experimentally manipulated B cells. In this network they used a naïve Bayes integration approach to incorporate protein-protein interactions, protein-DNA interactions and regulatory interactions between two genes modulated by a third gene product. The method distinguished two types of changes in the background regulatory network, namely loss and gain of correlation, which were caused due to the genetic perturbations in the phenotype of interest. If two connected genes in the network were not correlated, according to the samples of the phenotype, the link was classified as loss of correlation. Similarly, if two unconnected genes in the network were correlated in the phenotype, the link between these genes was classified as gain of correlation. The change of correlation between two genes was assessed by the difference of mutual information (calculated with the expression profiles of these genes under all conditions and those with the phenotype of interest). Next, the genes were ranked based on the enrichment of gain and loss of correlation events under the phenotype of interest. They highlighted several highly ranked genes such as *BCL2*, *SMAD1* in Follicular lymphoma; *MYC*, *MTA1* in Burkitt's lymphoma; and *CCND1*, *HDAC1* in mantle cell lymphoma where all of these genes had evidence in the literature for their association to the phenotype. Using an approach solely dependent on the differential expression of the genes could not have identified some of these genes otherwise.

In order to gain insights to cancer progression, several works analyzed enrichment of functionally related groups of genes (gene-sets) under various cancer types such as leukemia, lung cancer, breast cancer and prostate cancer (Segal, et al., 2004; Subramanian, et al., 2005; Tomlins, et al., 2007). Gene-sets contain functionally linked genes. These links were curated using annotations from the literature such as interacting genes, genes belonging to the same pathway, co-expressed genes in

various microarray experiments or genes sharing common regulatory motifs. Several methods were used to identify gene-sets consistently up or down regulated given a set of conditions (gene expression samples). For example, the gene-set signature identified by Chinnaiyan and coworkers (Tomlins, et al., 2007) implicated two new genes (*ZIC2* and *NPAL3*) in the progression to metastatic prostate cancer from benign epithelium.

Instead of relying on predefined gene annotations, Rhodes *et al.* (Rhodes, et al., 2005) created a predicted human interactome (whose links were not necessarily physical connections between proteins) and used this network to identify groups of genes overexpressed in pancreatic adenocarcinoma, myeloma and renal cell carcinoma. They combined the evidence from independent data sources, such as PPIs (Salwinski, et al., 2004), similarity between gene-expression profiles across several human tissue-samples (Rhodes, et al., 2007), domain combinations of known PPIs (Mulder, et al., 2003) or shared functional annotations (Harris, et al., 2004). A PPI network was obtained with the integration of these independent evidences in a probabilistic framework by means of a naive Bayesian classifier. This PPI network contained proteins as nodes and each link had a score with its likelihood. This network helped to implicate *RSU1* as a tumor suppressor in the *integrin* signaling pathway after the experimental validation of the interaction between the genes of *RSU1* and *LIMS1* (an *integrin*-mediated signaling adaptor protein). Network guided implication of genes in various cancer types is explained in the next section (Section 2.4). The subnetworks activated in cancer are further discussed later (see Section 2.5).

4. Network guided prediction of relevant genes involved in cancer progression

Cancer is a complex phenotype that recruits multiple genes whose products (proteins) tend to physically interact with each other(Jonsson and Bates, 2006; Kar, et al., 2009; Wachi, et al., 2005). In the light of the recent findings suggesting that proteins rarely act in isolation, the focus of research has shifted towards identifying the set of genes whose products work in cooperation. To this end, several methods exploiting guilt-by-association principle have been developed. In the context of disease associations, guilt-by-association principle suggests that a gene, whose gene product interacts with the products of known disease genes (seed genes), is likely to be a disease gene (Aerts, et al., 2006; Lage, et al., 2007). PPIs are concise descriptors of relationships between proteins and the genes encoding them. For this reason, the topology of the PPI networks have been extensively used to associate genes with diseases in the past years. An outline of the methodology of associating genes with a phenotype using known associations and PPIs is given in Figure 1.

Early attempts to identify novel cancer-associated genes considered only direct interacting partners of known cancer–associated genes (Aragues, et al., 2008; Ostlund, et al., 2010; Pujana, et al., 2007). Though the local neighborhood (direct interaction partners) of genes offers some clues for associating genes with diseases such as cancer, this approach misses the remaining information in the global network. Towards extending the amount of information extracted using interactions between genes, clustering based methods were also utilized (Milenkovic, et al., 2010; Navlakha and Kingsford, 2010). However, in order to fully exploit network topology, global topology based approaches have been recently proposed. Several works use shortest paths from known disease-associated genes (Dezso, et al., 2009; Guney and Oliva, 2011; Wu, et al., 2008), some use either kernel based diffusion (where more distant nodes have less weight/influence) (Ma, et al., 2007; Nitsch, et al., 2010; Qiu, et al., 2010) or message passing (Guney and Oliva, 2011) over the links of the network and some others simulate a random walk model on the network (where each node is assigned with the probability of a random surfer ending up in the node while surfing through the links of the network) (Chen, et al., 2009; Guney and Oliva, 2011; Vanunu, et al., 2010). Methods based in global topology, especially the ones based on random walk and message passing, were demonstrated to outperform methods based in local topology (Guney and Oliva, 2011; Navlakha and Kingsford, 2010; Vanunu, et al., 2010). Several of these methods have been applied on various cancer types, like prostate and breast cancer, suggesting novel associations for the implication of genes in specific mechanisms of the disease progression (Chen, et al., 2009; Qiu, et al., 2010; Vanunu, et al., 2010). In the following text, we explain in detail some of the works that successfully identified novel associations for various types of cancers.

For example, with the intention of extending a set of known breast cancer oncogenes (*BRCA1*, *BRCA2*, *ATM*, *CHEK2*), Pujana *et al.* (Pujana, et al., 2007) used the functional associations of these four genes with other genes and pinpointed those which probability to be involved in breast cancer was high. First, they found the genes co-expressed with each one of them (assessed by Pearson correlation coefficient). They identified 164 genes commonly co-expressed with all four oncogenes (BRCA1, BRCA2, ATM, and CHEK2). Next, they created a functional association network with these 164 genes, covering protein-protein interactions from literature, complex memberships, phenotypic similarity, co-expression, genetic interactions of orthologous genes and indirect associations (two genes connected by a third one). The genes of this network (consisting of 118 genes and 866 functional links) were scored according to GO term enrichment, conservation of co-expression across species, co-expression in breast tumor-derived cell lines, expression changes in $BRCA1^{mut}$ in breast tumors and functional

similarity with any of the four known oncogenes. This scoring protocol revealed several candidate genes for developing breast cancer, among which *HMMR* was experimentally linked to tumorigenesis and centrosome dysfunction. In a follow-up work, Pujana and colleagues (Bonifaci, et al., 2008) took a similar approach to suggest novel candidate-genes for breast cancer through integration of PPIs with several genomic and proteomic data sources. These data sources included expression changes in tumors relative to normal tissue samples and copy number variations correlating with gene expressions.

In another work, Aragues and coworkers (Aragues, et al., 2008) postulated that integrating PPI and genomics data would improve the prediction of cancer-associated genes. They combined several sources of data, such as gene expression, protein-protein interactions and structural-functional-evolutionary features (Furney, et al., 2006) to predict genes implicated in cancer. A prediction model was built based on the integration of data (Figure 2). Genes were associated with cancer provided that i) their products interacted with known cancer genes (retrieved from literature), ii) they were differentially expressed in a number of different cancer types, and iii) they shared structural, functional and evolutionary properties with known cancer genes. This combined prediction model outperformed each of the single-model predictions and it was used to produce a reliable list of cancer gene candidates.

Similar to the works above, Östlund *et al.* (Ostlund, et al., 2010) adopted the idea of looking at the network neighborhood of known cancer genes to predict novel cancer associations. They curated a set of 812 cancer genes from Cancer Gene Census (Futreal, et al., 2004) and text mining on UniProt (Bairoch, et al., 2005) entries. Then, they ranked the genes in a functional association network (high confidence links of FunCoup (Alexeyenko and Sonnhammer, 2009)) based on the number of cancer-associated genes connected with them. They showed that the higher the ranking of a gene was, the more likely the gene was involved in the biological functions associated with cancer (using GO (Ashburner, et al., 2000)) and it was differentially expressed in cancer tissues (using Human Protein Atlas (Berglund, et al., 2008)). Among 1891 genes connected with at least one known cancer-associated gene, they highlighted 185 genes with 10 or more linked cancer-genes and suggested them as candidates for further confirmation. In a slightly different approach, the comparisons between the PPI networks of a cancer phenotype and normal tissues were employed by Chu and Chen (Chu and Chen, 2008) to address the problem of identification of potential drug targets in human cervical carcinoma. They created condition-dependent PPI networks using a nonlinear stochastic model with microarray data to keep or remove the interactions. The interactions were distinguished as gain-of-function if they existed in the network of human cervical carcinoma cells but were absent in the network of normal primary lung fibroblasts samples. The interactions were distinguished as loss-of-function otherwise. They identified *BCL2*, *caspase-3* and *TP53* as potential drug targets.

Guo *et al.* (Guo, et al., 2007) developed an algorithm to identify the portion of the human interactome that responded to different conditions such as pathophenotype or environmental change and then used this subnetwork to predict novel disease associations. The activity score of an edge connecting two nodes (products of genes) was defined as the covariance of the gene expression between the two nodes. An optimally active subnetwork was gradually constructed such that the addition of an edge increased the overall activity score of the subnetwork. They applied this algorithm to the human interactome retrieved from HPRD (Peri, et al., 2004) using an expression data set of 71 prostate tumors and 41 normal prostate specimens (Lapointe, et al., 2004). The prostate-cancer responsive subnetwork contained 2181 nodes and 3200 edges and covered 74 of 118 prostate-cancer associated genes of the Prostate Gene Database (Li, et al., 2003). Next, a sub-region was defined with the genes of the subnetwork that interact with known prostate-cancer associated genes. 8 out of 17 genes were linked with at least two genes associated with prostate-cancer in the sub-region and they were reported to be involved in the pathology of pancreas cancer too. The detection of this kind of subnetworks (i.e. active subnetworks whose genes change their expression behavior under a certain phenotype) using PPIs and gene expression profiles has been a hot topic of research for the last few years (Ideker, et al., 2002; Ulitsky and Shamir, 2007; Ulitsky and Shamir, 2009). In the next section, we explain how to use these subnetworks for prognosis.

5. Predicting disease prognosis using dysregulation patterns in networks

Revealing alterations in cellular pathways in response to cancer is crucial to determine patient prognosis, where tumor cells manifest disruptions of the normal gene expressions. With this respect, many studies to identify cancer biomarkers by means of the analysis of gene expression patterns exist in the literature (Ludwig and Weinstein, 2005; Sawyers, 2008). However, the combined use of network modeling and gene expression data to discover gene-sets capable of distinguishing different disease states (e.g., good outcome versus poor outcome) is very recent. These gene-sets can be extracted exclusively using regulatory networks (Lim, et al., 2009), but the possibility of using dysregulation patterns in PPI networks to predict a disease outcome or prognosis has motivated many researchers during the past decade because of its potential to improve the

predictions (Nibbe, et al., 2011) (see Figure 3 for a conceptual overview of these methods).

Efroni *et al.* (Efroni, et al., 2007) assessed the activity of pathways described in the literature (Buetow, et al., 2002; Schaefer, 2004) using gene expression data compiled over multiple published studies for various cancer types. They first calculated the probability of a gene being in either up or down state in cancer (i. e. showed higher/lower expression in cancer samples in comparison to normal samples). Next, they calculated an activity score for each interaction in the pathway by incorporating the probabilities of interacting genes. The overall activity score of a pathway was then computed as the average of the activity scores of all interactions. Using a subset of these pathways, selected by a machine-learning approach and a Bayesian classifier, they were able to classify the tumor samples with 98% accuracy. Interestingly, the most discriminative pathways were *Trka* Pathway, DNA Damage pathway, Ceramide Pathway, Telomerase Pathway, *CD40L* Pathway and Calcineurin Pathway.

Similarly, Chuang *et al.* (Chuang, et al., 2007) developed a network-based approach to identify functionally related genes distinguishing post-surgery metastasis in breast cancer patients. The method associated the phenotypic variance among cohorts of patients with clusters of genes. After mapping expression data on the protein interaction network, they identified subnetworks involving products of coherently expressed genes. A score was assigned to a candidate subnetwork by averaging the normalized expression values of its genes using each sample (patient) in two cohorts of metastatic and non-metastatic breast cancer patients (van de Vijver, et al., 2002; Wang, et al., 2005). Starting from a single seed node in the interaction network, a candidate subnetwork was greedily constructed by considering the neighbors of the nodes already included in the subnetwork and within a specified distance from the seed. These candidate subnetworks were scored in terms of their potential to discriminate between the two cohorts (metastatic and non-metastatic patient groups) using mutual information. Their results showed that these subnetworks contained genes playing a central role in connecting differentially expressed genes, even though some of these genes were not differentially expressed. They also proved that the genes of these subnetworks were better predictors of metastasis than markers based on single genes. This improvement of the prediction highlighted the importance of network topology for characterizing genetic elements involved in breast cancer metastasis. In a follow-up study by Lee *et al.* (Lee, et al., 2009) the same subnetwork identification method was used to identify subnetwork dysregulated in acute myeloid leukemia patients. They identified subnetworks that were tightly coupled with key leukemogenic processes such as myeloid differentiation, cell signaling of growth and survival, cell cycle and cell and tissue remodeling.

Nibbe *et al.* (Nibbe, et al., 2009) adapted the subnetwork scoring method of Chuang *et al.* (Chuang, et al., 2007) to identify genes that were discriminative of late stage of human colorectal cancer. Candidate subnetworks were generated by including the partners of 67 seed genes associated with colorectal cancer (according to 2D gel experiments). Gene expression data was also used to score these subnetworks according to their discriminative power between cancer and control. Several evidences were found in the literature supporting the relevant roles of some of the genes of the subnetworks being actively implicated in human colorectal cancer, such as: CNSK2A2, PLK1, IGFBP3 (involved in progression), PDFGRB (with metastatic potential), IFITM1 (as a biomarker). This work presents a clear example of the benefits of integrating proteomics, gene expression and protein-protein interaction data.

Nibbe *et al.* (Nibbe, et al., 2010) further extended their approach by incorporating additional proteomics and gene expression data as well as a guided search of subnetworks using a guilt-by-association score. A random walk algorithm was employed to calculate colorectal cancer association scores of genes in the interaction network, where differentially expressed genes were used as seeds. The subnetworks included the genes interacting with seed genes (as in previous work (Nibbe, et al., 2009)) and those with high association scores. They showed that these subnetworks classified tumor samples better than subnetworks that contained only those genes interacting with seeds.

Taylor *et al.* (Taylor, et al., 2009) investigated rewiring and modularity of the interaction network during tumor progression. First, they analyzed an interaction network containing experimental and predicted protein-protein interactions and including expression data from 79 human tissues (Su, et al., 2004). The hubs in this network were grouped based on the average of co-expression with their partners as either inter-modular hubs (co-expression was restricted to specific tissues) or intra-modular hubs (co-expressed in most of the tissues). According to their study, the interactome was modular with inter-modular hubs connecting modules composed of intra-modular hubs that tend to be functionally more coherent than inter-modular hubs. Furthermore, they observed the predisposition of inter-modular hubs to be associated with cancer phenotypes (according to the mutations listed in OMIM (McKusick, 2007) and the census of cancer genes (Futreal, et al., 2004)) than intra-modular hubs. Using a cohort of sporadic non-familial breast cancer patients (van de Vijver, et al., 2002), they identified 256 hub

genes showing significant changes of co-expression between two patient groups classified by survival time (good and poor outcome patient groups). These findings suggested that altered gene expression in breast cancer affected survival. Accordingly, they defined the subnetworks that define the phenotypic variation (i.e. subnetworks active in cancer patients) as the hubs whose expression is significantly altered plus the genes interacting with them. They developed a classification system using relative expression within these subnetworks and affinity-propagation clustering algorithm. They first determined the hubs for which the relative expression differed significantly between patients who survived versus those who died from disease. Next, they clustered the patients using relative expression of these hubs as features. Then, they predicted the outcome of the prognosis for a patient based on identified clusters. Their classification achieved 71% area under ROC curve (AUC), which increased to 78% when clinical data such as patient age, tumor stage and tumor grade were incorporated. The prediction performance of the proposed classification system compared favorably with the commercially available genetic breast cancer diagnostics.

The works mentioned above calculated the activity of a subnetwork by either aggregating or subtracting the expression of its genes. To capture the effects of complex forms of interactions within subnetworks, such as inhibitory interactions, several studies (Chowdhury, et al., 2011; Dutkowski and Ideker, 2011) focused on identifying subnetworks consisting of genes exhibiting combinatorial expression patterns. These combinatorially dysregulated subnetworks bore collective differential expression of their constituents (e.g. the subnetwork whose genes were not necessarily all up or down regulated but in which the particular combination of genes defined the phenotype better). Combinatorially dysregulated subnetworks were shown to distinguish different stages of cancer with high accuracy(Chowdhury, et al., 2011). During the search for dysregulated subnetworks, Chowdhury *et al.* (Chowdhury, et al., 2011) employed a heuristic to extend the list, increasing its potential to describe a specific phenotype. Gene expression samples were represented in a binary fashion as having either high or low expression (e.g., a binary state). The subnetworks were then referred as state functions where the combination of genes in the subnetwork was informative of the phenotype. Using a neural network model whose inputs were states of the genes in the subnetworks, they classified metastatic colorectal samples from non-metastatic samples with a precision of 88% and sensitivity of 95% on average. In another work, Dutkowski and Ideker (Dutkowski and Ideker, 2011) tackled the problem of identifying combinatorially dysregulated subnetworks that distinguished various classes of samples by adopting a random forest approach. The algorithm generated multiple decision trees using gene expression of genes and the interactions of their products in the PPI network. The rules defined by these decision trees were used to classify metastatic samples from non-metastatic samples in breast cancer as in (Chuang, et al., 2007). Their results also confirmed that combinatorially dysregulated subnetworks distinguished better the phenotype of the samples than using only coherent gene activities. Chen *et al.* (Chen, et al., 2011) tackled the same problem by constraining the search for subnetworks and adopting a Support Vector Machine approach (SVM) to include the interactions of the genes in the feature space. They showed that this approach succesfully classified metastatic and non-metastatic breast cancer samples.

6. Discovery of biological pathways in breast cancer metastases: A network biology perspective

Breast cancer in women is one of the most common forms of cancer in Europe, around 400,000 cases of breast cancer are diagnosed annually (Ferlay, et al., 2010) and the observed incidence of this cancer is expected to continue rising. Although there have been great improvements in early detection and treatment, around 30% of early stage breast cancer patients experience recurrent disease. Following the diagnosis and initial surgery to remove the primary tumor, patients may experience relapse due to the invasion of distant organs by secondary tumors. A major factor affecting survival in these cases is resistance to chemotherapy used to treat primary and secondary tumors.

In patients with controlled local cancer, systemic progression constitutes a major public health problem (Gluck, 2007). Breast tumours show an organ-specific pattern in metastasis formation, in which bone (60%), lung (34 %) liver (20%) and brain (15%) are the most commonly affected organs (Lu and Kang, 2007). The patterns of metastatic spread vary from patient to patient. Some patients may escape relapse entirely. Others develop bone metastases only and may survive for ten years or more following the diagnosis. However, the patients who develop metastases in the various visceral tissues have a much increased mortality rate and shortened life expectancy. There is a need for research to integrate scientific and clinical investigation to understand the basic processes of breast cancer metastasis and translate such insights into clinical care as rapidly as possible.

The seed-and-soil hypothesis proposed by Paget (Paget, 1889) in the 19th century postulated that the development of distant metastases in cancer patients was dependent both on the characteristics of the cancer cells and the cooperation of the cells in the host organ (Fidler, 2003). Breast cancer cell signaling networks are complex systems that integrate information from the cellular environment (Manning and Cantley, 2007). Indeed, metastasis is a complex disease that involves a number of

simultaneous molecular processes (Hortobagyi, 2000; Kaal, et al., 2005; Minn, et al., 2005). The mechanisms that mediate organ-specific pathogenesis of metastases are the combination of modifications that occur in both primary tumor and metastatic cells, during the process of spread and microenvironment adaptation (Waltregny, et al., 2000).

Results of transcriptomic analysis of cell lines with specific organ-tropisms indicate the existence of an organ-specific metastatic phenotype (Kang, et al., 2003). Among the different studies for organ-specific signatures, there is a significant lack of overlap in the selected genes, indicating perhaps a strong platform-dependence or other bias in each study. Also, transcriptomic studies on cell lines do not take account of in situ gene expression, and provide information only from the cancer cell itself, when it is known that interactions with host cells are also critical for the establishment of metastases. Computational approaches are needed to elucidate the regulatory properties of signaling networks of metastasis (Aldridge, et al., 2006; Bhalla, 2003; Justman, et al., 2009). Microarray-based gene studies are difficult to interpret, because of the huge amount of data involved, and it is therefore a challenge to describe biological insights. Maps of complex networks were derived by interconnecting the individual pathways obtained from experimental data (Bhalla and Iyengar, 1999; Weng, et al., 1999). These studies revealed that signaling networks contain numerous features, such as feedback and feed forward loops (Alon, 2007; Ma'ayan, et al., 2005), which render it virtually impossible for the human mind to decipher how signals are integrated within the pathways determining the pathogenic function.

Large-scale computational comparisons of alterations in thousands of genes and proteins in cancer cells documented in inter-laboratory data are essential to identify key genes and/or proteins that are deregulated in metastatic cancer cells (Nguyen and Massague, 2007; Shedden, et al., 2008). Despite the wealth of molecular profiling data that describe breast tumours, our understanding of the fundamental genetic dependences in metastatic progression is relatively poor (Schlabach, et al., 2008). Indeed, molecular classification provides insights into breast cancer taxonomy, but its clinical implementation is hindered by the unreliability of single sample allocations (Weigelt, et al., 2010). To design an appropriate course of treatment, there is a need for comprehensive functional viability profiles to identify the risk of metastasis and develop therapeutic targets.

As mentioned above, a strategy based on mapping expression profiles with protein interactions was described by Chuang et al., 2007 (Chuang, et al., 2007). The authors showed that it was possible to extract relevant biological information about deregulated functions and the relationship between them, and to identify molecules that could be helpful as metastatic markers or therapeutic targets. The use of a PPI network-based approach identified markers not as individual genes but as subnetworks extracted from PPI network, providing a systemic view of the interactome (Grimaldi, et al., 2009). This method served to filter information by picking out key protein functions as metastasis markers. Thus, PPI network-based approach was useful to decipher distinctive phenotypes, since differences between PPI networks revealed characteristic traits of each metastasis (Figure 4).

Coupling microarray data from clinical metastases and immunohistochemistry, Sanz *et al.* (Sanz, et al., 2007) assessed association of proteins in the soft-tissue metastases of breast cancer tumors such as liver and lung metastases. They created protein interaction networks starting from sets of differentially expressed genes for each phenotype. Next, they analyzed the interaction networks to investigate the commonalities between the three soft-tissue human breast cancer metastases and showed that although the studied soft-tissue metastases are phenotypically diverse, several metastatic competency genes are shared among these metastases. These resemblances in the PPI networks reflected redundant phenotypes in metastatic cells that could be useful to colonize several tissues. By this approach, the chaperone GRP75 was found to be only up-regulated in liver metastasis, and this discovery was validated in tissue patients. So, this protein could play an important role in the pathogenesis of liver metastasis. Furthermore, they revealed the link between *HSP60*, a widely recognized mitochondrial chaperone machine, and *BAG2* in both soft tissue metastases.

In a following study by Martin *et al.* (Martin, et al., 2008), 18 proteins identified by protein expression difference in brain metastasis of primary breast cancer tumors were placed into a network context where associations between proteins were defined by protein-protein interactions, functional associations from STRING (von Mering, et al., 2003) and predicted protein-protein interactions using structural similarities or interology. Next, they clustered the extended neighborhood of these initial proteins using functions defined in UniProt (Apweiler, et al., 2004) to characterize functional phenotypes that could enhance brain metastasis in breast cancer cells. Their analysis identified *HSP27*, an ATP-independent molecular chaperone influencing the assembly, transport and folding of other proteins, as a gene implicated in the pathology of metastasis. A similar approach was taken by Sanz-Pamplona *et al.* (Sanz-Pamplona, et al., 2011) to discriminate patients developing brain metastasis from those who didn’t. Based on a functional study in which the PPI network was divided into

modules of interacting proteins sharing a common function, they hypothesized that brain metastasis cells had a characteristic behavior named “endoplasmic reticulum stress resistance phenotype”. They further validated the expression of proteins in primary breast carcinoma, using both samples that developed brain metastasis and the samples that did not, and the search for a multivariate panel of markers revealed the expression of proteins in breast tumors predicting the metastasis in brain. Indeed, *GRP94*, *FN14* and *inhibin* was the best combination to discriminate metastasis samples from non-metastasis samples, achieving 85% of the area under ROC curve.

Moreover, proteins from the family of chaperones and GRP (Glucose Regulated Proteins) act as central hubs in all metastasis networks. These proteins have an active role in the maintenance of networks architecture, acting as a key regulator of cellular systems and working as bridge nodes, binding functional modules of proteins with each other. In the case of environmental stress, the chaperones remodel interactions between these functional modules helping the cell to survive in a hostile environment (Korcsmaros, et al., 2007; Palotai, et al., 2008).

The developments in therapy are now driving a demand for a more precise prognosis, especially with respect to metastasis. The arrival of low toxicity adjuvant chemotherapy has encouraged the identification of breast cancer patients who are at high risk of aggressive cancer. Also, the ongoing development of a range of preventative strategies for metastasis formation has increased the demand for effective classification of patients who are at increased risk of specific metastasis.

7. Future directions: From networks to systems medicine

Macromolecular assemblies carry out most biological processes. The interactions between these macromolecules constitute pathways, which are networks usually involving transient interactions. Since most of these pathways are interconnected, even slightest changes in one pathway can cause abnormal regulation events affecting other biological processes. Taken together with the fact that cancer is a disease of pathways rather than single genes, small perturbations hinder the discovery of novel drugs causing them to fail at the very last (clinical) phases. Therefore, network medicine approaches aim to foresee the outcome of such perturbations in regulation patterns by incorporating protein-protein interactions in addition to the available data, helping to define a dynamic context (i.e. proteomic, genomic, metabolic, physiological and environmental information) and possibly suggesting points of action (see (Pujol, et al., 2010) and (Fliri, et al., 2010) for reviews). These approaches to human disease can have multiple biological and clinical applications: first, they may lead to the identification of disease genes and disease pathways; second, they can be applied in the discovery of new targets and development of new drugs; and third, some of the new targets can be used as more accurate cancer biomarkers or applied for a better classification of cancer, improving personalized therapies and treatment (Barabasi, et al., 2011).

In order to tackle with the complexity of polygenic diseases, recently emerged polypharmacological approaches (Hopkins, 2008) that typically target many proteins simultaneously via the administration of multiple drugs. Such strategies bear a potential to intervene the disease progression mechanism by creating a synergistic (more-than-additive) response and to reduce the likelihood of drug-resistance by eliminating compensatory reactions (Csermely, et al., 2005). The applicability of therapies involving multiple targets was demonstrated for several pathophenotypes such as AIDS or cancer, where optimal drug combinations were proposed (Vazquez, 2009; Yang, et al., 2008). In the near future, however, network based approaches are expected to prove particularly useful in predicting toxicology and repurposing drugs with secondary targets involved in several pathways that are not apparently related to each other.

Another important direction towards effective treatment of cancer is developing DNA-damaging agents that are only toxic for the proliferating cancer cells without affecting normal tissue cells. Genetic interactions provide a theoretical framework for identifying candidate genes that are synthetic lethal (combination of two phenotypes results in lethality) with the mutations causative of cancer (Michod and Widmann, 2007). This kind of “next-generation” approach might replace conventional methods, such as aggressive drugs and chemotherapy, that damage cancer cells as well as normal tissue cells. Although initial studies have reported promising results, where several genes that show synthetic lethality with a handful of oncogenes are identified (Luo, et al., 2009; Scholl, et al., 2009), research in this area is still in its infancy. It can be postulated that integration of genetic interactions (such as in (Bandyopadhyay, et al., 2008) and (Ulitsky, et al., 2008)) will play an essential role in building up clinical applications of such next-generation approaches.

In conclusion, evaluating the genes and their relationships within the context of the network –in particular using PPI networks– has made possible a better understanding of disease states. Still, even if the results extracted from the works mentioned in this chapter are very promising, more research towards delineating network-centric view of cellular processes is required to further develop more effective and possibly more personalized therapeutics.

Acknowledgments

EG is supported through FI fellowship granted by “Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu”. BO acknowledges grants from the Spanish Ministry of Science and Innovation (MICINN), FEDER BIO2011-22568, and PSE-0100000-2009. AS and RS acknowledge MetaBre consortium (LSHC-CT-2004-506049).

2. References

Aerts, S., Lambrechts, D., Maity, S., et al. (2006) Gene prioritization through genomic data fusion, *Nat Biotechnol*, **24**, 537-544.

Aldridge, B.B., Burke, J.M., Lauffenburger, D.A. and Sorger, P.K. (2006) Physicochemical modelling of cell signalling pathways, *Nat Cell Biol*, **8**, 1195-1203.

Alexeyenko, A. and Sonnhammer, E.L. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration, *Genome Res*, **19**, 1107-1116.

Alfarano, C., Andrade, C.E., Anthony, K., et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res*, **33**, D418-424.

Alon, U. (2007) Network motifs: theory and experimental approaches, *Nat Rev Genet*, **8**, 450-461.

Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM), *Nucleic Acids Res*, **37**, D793-796.

Apweiler, R., Bairoch, A., Wu, C.H., et al. (2004) UniProt: the Universal Protein knowledgebase, *Nucleic Acids Res*, **32**, D115-119.

Aragues, R., Sander, C. and Oliva, B. (2008) Predicting cancer involvement of genes from heterogeneous data, *BMC Bioinformatics*, **9**, 172.

Ashburner, M., Ball, C.A., Blake, J.A., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bairoch, A., Apweiler, R., Wu, C.H., et al. (2005) The Universal Protein Resource (UniProt), *Nucleic Acids Res*, **33**, D154-159.

Bandyopadhyay, S., Kelley, R., Krogan, N.J. and Ideker, T. (2008) Functional maps of protein complexes from quantitative genetic interaction data, *PLoS Comput Biol*, **4**, e1000065.

Barabasi, A.L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease, *Nat Rev Genet*, **12**, 56-68.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Barrett, T. and Edgar, R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, *Methods Enzymol*, **411**, 352-369.

Berglund, L., Bjorling, E., Oksvold, P., et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies, *Mol Cell Proteomics*, **7**, 2019-2027.

Berman, H.M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank, *Nucleic Acids Res*, **28**, 235-242.

Bhalla, U.S. (2003) Understanding complex signaling networks through models and metaphors, *Prog Biophys Mol Biol*, **81**, 45-65.

Bhalla, U.S. and Iyengar, R. (1999) Emergent properties of networks of biological signaling pathways, *Science*, **283**, 381-387.

Bonifaci, N., Berenguer, A., Diez, J., et al. (2008) Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes, *BMC Med Genomics*, **1**, 62.

Buetow, K.H., Klausner, R.D., Fine, H., et al. (2002) Cancer Molecular Analysis Project: weaving a rich cancer research tapestry, *Cancer Cell*, **1**, 315-318.

Chen, J., Aronow, B.J. and Jegga, A.G. (2009) Disease candidate gene identification and prioritization using protein interaction networks, *BMC Bioinformatics*, **10**, 73.

Chen, L., Xuan, J., Riggins, R.B., et al. (2011) Identifying cancer biomarkers by network-constrained support vector machines, *BMC Syst Biol*, **5**, 161.

Chowdhury, S.A., Nibbe, R.K., Chance, M.R. and Koyuturk, M. (2011) Subnetwork state functions define dysregulated subnetworks in cancer, *J Comput Biol*, **18**, 263-281.

Chu, L.H. and Chen, B.S. (2008) Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets, *BMC Syst Biol*, **2**, 56.

Chuang, H.Y., Hofree, M. and Ideker, T. (2010) A decade of systems biology, *Annu Rev Cell Dev Biol*, **26**, 721-744.

Chuang, H.Y., Lee, E., Liu, Y.T., et al. (2007) Network-based classification of breast cancer metastasis, *Mol Syst Biol*, **3**, 140.

Collins, F.S. and Barker, A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies, *Sci Am*, **296**, 50-57.

Comen, E., Norton, L. and Massague, J. (2011) Clinical implications of cancer self-seeding, *Nat Rev Clin Oncol*, **8**, 369-377.

Croft, D., O'Kelly, G., Wu, G., et al. (2011) Reactome: a database of reactions, pathways and biological processes, *Nucleic Acids Res*, **39**, D691-697.
Csermely, P., Agoston, V. and Pongor, S. (2005) The efficiency of multi-target drugs: the network approach might help drug design, *Trends Pharmacol Sci*, **26**, 178-182.
Dezso, Z., Nikolsky, Y., Nikolskaya, T., et al. (2009) Identifying disease-specific genes based on their topological significance in protein networks, *BMC Syst Biol*, **3**, 36.
Dickson, D. (1999) Wellcome funds cancer database, *Nature*, **401**, 729.
Dreze, M., Monachello, D., Lurin, C., et al. (2010) High-quality binary interactome mapping, *Methods Enzymol*, **470**, 281-315.
Dutkowski, J. and Ideker, T. (2011) Protein networks as logic functions in development and cancer, *PLoS Comput Biol*, **7**, e1002180.
Efroni, S., Schaefer, C.F. and Buetow, K.H. (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis, *PLoS One*, **2**, e425.
Ergun, A., Lawrence, C.A., Kohanski, M.A., et al. (2007) A network biology approach to prostate cancer, *Mol Syst Biol*, **3**, 82.
Ferlay, J., Parkin, D.M. and Steliarova-Foucher, E. (2010) Estimates of cancer incidence and mortality in Europe in 2008, *Eur J Cancer*, **46**, 765-781.
Fidler, I.J. (2003) The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited, *Nat Rev Cancer*, **3**, 453-458.
Fliri, A.F., Loging, W.T. and Volkmann, R.A. (2010) Cause-effect relationships in medicine: a protein network perspective, *Trends Pharmacol Sci*, **31**, 547-555.
Fry, D.C. and Vassilev, L.T. (2005) Targeting protein-protein interactions for cancer therapy, *J Mol Med (Berl)*, **83**, 955-963.
Furney, S.J., Calvo, B., Larranaga, P., et al. (2008) Prioritization of candidate cancer genes--an aid to oncogenomic studies, *Nucleic Acids Res*, **36**, e115.
Furney, S.J., Higgins, D.G., Ouzounis, C.A. and Lopez-Bigas, N. (2006) Structural and functional properties of genes involved in human cancer, *BMC Genomics*, **7**, 3.
Futreal, P.A., Coin, L., Marshall, M., et al. (2004) A census of human cancer genes, *Nat Rev Cancer*, **4**, 177-183.
Garcia-Garcia, J., Guney, E., Aragues, R., et al. (2010) Biana: a software framework for compiling biological interactions and analyzing networks, *BMC Bioinformatics*, **11**, 56.
Gluck, S. (2007) The prevention and management of distant metastases in women with breast cancer, *Cancer Invest*, **25**, 6-13.
Goh, K.I., Cusick, M.E., Valle, D., et al. (2007) The human disease network, *Proc Natl Acad Sci U S A*, **104**, 8685-8690.
Grimaldi, D., Claessens, Y.E., Mira, J.P. and Chiche, J.D. (2009) Beyond clinical phenotype: the biologic integratome, *Crit Care Med*, **37**, S38-49.
Guldener, U., Munsterkotter, M., Oesterheld, M., et al. (2006) MPact: the MIPS protein interaction resource on yeast, *Nucleic Acids Res*, **34**, D436-441.
Gundem, G., Perez-Llamas, C., Jene-Sanz, A., et al. (2010) IntOGen: integration and data mining of multidimensional oncogenomic data, *Nat Methods*, **7**, 92-93.
Guney, E. and Oliva, B. (2011) Toward PWAS: discovering pathways associated with human disorders, *BMC Bioinformatics*, **12**, A12.
Guo, Z., Wang, L., Li, Y., et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network, *Bioinformatics*, **23**, 2121-2128.
Gupta, G.P. and Massague, J. (2006) Cancer metastasis: building a framework, *Cell*, **127**, 679-695.
Hamosh, A., Scott, A.F., Amberger, J.S., et al. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res*, **33**, D514-517.
Han, K., Park, B., Kim, H., et al. (2004) HPID: the Human Protein Interaction Database, *Bioinformatics*, **20**, 2466-2470.
Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation, *Cell*, **144**, 646-674.
Harris, M.A., Clark, J., Ireland, A., et al. (2004) The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Res*, **32**, D258-261.
Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery, *Nat Chem Biol*, **4**, 682-690.
Hortobagyi, G.N. (2000) Developments in chemotherapy of breast cancer, *Cancer*, **88**, 3073-3079.
Hudson, T.J., Anderson, W., Artez, A., et al. (2010) International network of cancer genome projects, *Nature*, **464**, 993-998.
Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, **18 Suppl 1**, S233-240.
Ideker, T. and Sharan, R. (2008) Protein networks in disease, *Genome Res*, **18**, 644-652.

Jonsson, P.F. and Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome, *Bioinformatics*, **22**, 2291-2297.
Joyce, A.R. and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets, *Nat Rev Mol Cell Biol*, **7**, 198-210.
Justman, Q.A., Serber, Z., Ferrell, J.E., Jr., et al. (2009) Tuning the activation threshold of a kinase network by nested feedback loops, *Science*, **324**, 509-512.
Kaal, E.C., Niel, C.G. and Vecht, C.J. (2005) Therapeutic management of brain metastasis, *Lancet Neurol*, **4**, 289-298.
Kanehisa, M., Goto, S., Sato, Y., et al. (2012) KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res*, **40**, D109-114.
Kang, Y., Siegel, P.M., Shu, W., et al. (2003) A multigenic program mediating breast cancer metastasis to bone, *Cancer Cell*, **3**, 537-549.
Kapushesky, M., Emam, I., Holloway, E., et al. (2010) Gene expression atlas at the European bioinformatics institute, *Nucleic Acids Res*, **38**, D690-698.
Kar, G., Gursoy, A. and Keskin, O. (2009) Human cancer protein-protein interaction network: a structural perspective, *PLoS Comput Biol*, **5**, e1000601.
Kerrien, S., Aranda, B., Breuza, L., et al. (2012) The IntAct molecular interaction database in 2012, *Nucleic Acids Res*, **40**, D841-846.
Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. (2009) Human Protein Reference Database--2009 update, *Nucleic Acids Res*, **37**, D767-772.
Korcsmaros, T., Kovacs, I.A., Szalay, M.S. and Csermely, P. (2007) Molecular chaperones: the modular evolution of cellular networks, *J Biosci*, **32**, 441-446.
Lage, K., Karlberg, E.O., Storling, Z.M., et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders, *Nat Biotechnol*, **25**, 309-316.
Lapointe, J., Li, C., Higgins, J.P., et al. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer, *Proc Natl Acad Sci U S A*, **101**, 811-816.
Laubenbacher, R., Hower, V., Jarrah, A., et al. (2009) A systems biology view of cancer, *Biochim Biophys Acta*, **1796**, 129-139.
Lee, E., Jung, H., Radivojac, P., et al. (2009) Analysis of AML genes in dysregulated molecular networks, *BMC Bioinformatics*, **10 Suppl 9**, S2.
Li, L.C., Zhao, H., Shiina, H., et al. (2003) PGDB: a curated and integrated database of genes related to the prostate, *Nucleic Acids Res*, **31**, 291-293.
Licata, L., Briganti, L., Peluso, D., et al. (2012) MINT, the molecular interaction database: 2012 update, *Nucleic Acids Res*, **40**, D857-861.
Lim, W.K., Lyashenko, E. and Califano, A. (2009) Master regulators used as breast cancer metastasis classifier, *Pac Symp Biocomput*, 504-515.
Lu, X. and Kang, Y. (2007) Organotropism of breast cancer metastasis, *J Mammary Gland Biol Neoplasia*, **12**, 153-162.
Ludwig, J.A. and Weinstein, J.N. (2005) Biomarkers in cancer staging, prognosis and treatment selection, *Nat Rev Cancer*, **5**, 845-856.
Luo, J., Emanuele, M.J., Li, D., et al. (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene, *Cell*, **137**, 835-848.
Ma, X., Lee, H., Wang, L. and Sun, F. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, *Bioinformatics*, **23**, 215-221.
Ma'ayan, A., Jenkins, S.L., Neves, S., et al. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network, *Science*, **309**, 1078-1083.
Mani, K.M., Lefebvre, C., Wang, K., et al. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas, *Mol Syst Biol*, **4**, 169.
Manning, B.D. and Cantley, L.C. (2007) AKT/PKB signaling: navigating downstream, *Cell*, **129**, 1261-1274.
Martin, B., Aragues, R., Sanz, R., et al. (2008) Biological pathways contributing to organ-specific phenotype of brain metastatic cells, *J Proteome Res*, **7**, 908-920.
McKusick, V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM, *Am J Hum Genet*, **80**, 588-604.
Michod, D. and Widmann, C. (2007) DNA-damage sensitizers: potential new therapeutical tools to improve chemotherapy, *Crit Rev Oncol Hematol*, **63**, 160-171.
Milenkovic, T., Memisevic, V., Ganesan, A.K. and Przulj, N. (2010) Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data, *J R Soc Interface*, **7**, 423-437.

Minn, A.J., Gupta, G.P., Siegel, P.M., et al. (2005) Genes that mediate breast cancer metastasis to lung, *Nature*, **436**, 518-524.
Mulder, N.J., Apweiler, R., Attwood, T.K., et al. (2003) The InterPro Database, 2003 brings increased coverage and new features, *Nucleic Acids Res*, **31**, 315-318.
Navlakha, S. and Kingsford, C. (2010) The power of protein interaction networks for associating genes with diseases, *Bioinformatics*, **26**, 1057-1063.
Nguyen, D.X. and Massague, J. (2007) Genetic determinants of cancer metastasis, *Nat Rev Genet*, **8**, 341-352.
Nibbe, R.K., Chowdhury, S.A., Koyuturk, M., et al. (2011) Protein-protein interaction networks and subnetworks in the biology of disease, *Wiley Interdiscip Rev Syst Biol Med*, **3**, 357-367.
Nibbe, R.K., Koyuturk, M. and Chance, M.R. (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer, *PLoS Comput Biol*, **6**, e1000639.
Nibbe, R.K., Markowitz, S., Myeroff, L., et al. (2009) Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer, *Mol Cell Proteomics*, **8**, 827-845.
Nitsch, D., Goncalves, J.P., Ojeda, F., et al. (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches, *BMC Bioinformatics*, **11**, 460.
Ogmen, U., Keskin, O., Aytuna, A.S., et al. (2005) PRISM: protein interactions by structural matching, *Nucleic Acids Res*, **33**, W331-336.
Ostlund, G., Lindskog, M. and Sonnhammer, E.L. (2010) Network-based Identification of novel cancer genes, *Mol Cell Proteomics*, **9**, 648-655.
Paget, S. (1889) The distribution of secondary growths in cancer of the breast., *Cancer Metastasis Rev*, **8**, 98-101.
Palotai, R., Szalay, M.S. and Csermely, P. (2008) Chaperones as integrators of cellular networks: changes of cellular integrity in stress and diseases, *IUBMB Life*, **60**, 10-18.
Parkinson, H., Sarkans, U., Kolesnikov, N., et al. (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments, *Nucleic Acids Res*, **39**, D1002-1004.
Peri, S., Navarro, J.D., Kristiansen, T.Z., et al. (2004) Human protein reference database as a discovery resource for proteomics, *Nucleic Acids Res*, **32**, D497-501.
Pujana, M.A., Han, J.D., Starita, L.M., et al. (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction, *Nat Genet*, **39**, 1338-1349.
Pujol, A., Mosca, R., Farres, J. and Aloy, P. (2010) Unveiling the role of network and systems biology in drug discovery, *Trends Pharmacol Sci*, **31**, 115-123.
Qiu, Y.Q., Zhang, S., Zhang, X.S. and Chen, L. (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data, *BMC Bioinformatics*, **11**, 26.
Quackenbush, J. (2006) Microarray analysis and tumor classification, *N Engl J Med*, **354**, 2463-2472.
Rhodes, D.R. and Chinnaiyan, A.M. (2005) Integrative analysis of the cancer transcriptome, *Nat Genet*, **37 Suppl**, S31-37.
Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., et al. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles, *Neoplasia*, **9**, 166-180.
Rhodes, D.R., Tomlins, S.A., Varambally, S., et al. (2005) Probabilistic model of the human protein-protein interaction network, *Nat Biotechnol*, **23**, 951-959.
Russell, R.B. and Aloy, P. (2008) Targeting and tinkering with interaction networks, *Nat Chem Biol*, **4**, 666-673.
Salwinski, L., Miller, C.S., Smith, A.J., et al. (2004) The Database of Interacting Proteins: 2004 update, *Nucleic Acids Res*, **32**, D449-451.
Sanz, R., Aragues, R., Stresing, V., et al. (2007) Functional pathways shared by liver and lung metastases: a mitochondrial chaperone machine is up-regulated in soft-tissue breast cancer metastasis, *Clin Exp Metastasis*, **24**, 673-683.
Sanz-Pamplona, R., Aragues, R., Driouch, K., et al. (2011) Expression of endoplasmic reticulum stress proteins is a candidate marker of brain metastasis in both ErbB-2+ and ErbB-2- primary breast tumors, *Am J Pathol*, **179**, 564-579.
Sawyers, C.L. (2008) The cancer biomarker problem, *Nature*, **452**, 548-552.
Schaefer, C.F. (2004) Pathway databases, *Ann N Y Acad Sci*, **1020**, 77-91.
Schlabach, M.R., Luo, J., Solimini, N.L., et al. (2008) Cancer proliferation gene discovery through functional genomics, *Science*, **319**, 620-624.
Scholl, C., Frohling, S., Dunn, I.F., et al. (2009) Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells, *Cell*, **137**, 821-834.
Schwartz, A.S., Yu, J., Gardenour, K.R., et al. (2009) Cost-effective strategies for completing the interactome, *Nat Methods*, **6**, 55-61.
Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol*, **18**, 1257-1261.

Segal, E., Friedman, N., Koller, D. and Regev, A. (2004) A module map showing conditional activity of expression modules in cancer, *Nat Genet*, **36**, 1090-1098.
Shedden, K., Taylor, J.M., Enkemann, S.A., et al. (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study, *Nat Med*, **14**, 822-827.
Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., et al. (2011) The BioGRID Interaction Database: 2011 update, *Nucleic Acids Res*, **39**, D698-704.
Steeg, P.S. (2006) Tumor metastasis: mechanistic insights and clinical challenges, *Nat Med*, **12**, 895-904.
Strausberg, R.L., Simpson, A.J. and Wooster, R. (2003) Sequence-based cancer genomics: progress, lessons and opportunities, *Nat Rev Genet*, **4**, 409-418.
Su, A.I., Wiltshire, T., Batalov, S., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*, **101**, 6062-6067.
Subramanian, A., Tamayo, P., Mootha, V.K., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci U S A*, **102**, 15545-15550.
Taylor, I.W., Linding, R., Warde-Farley, D., et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome, *Nat Biotechnol*, **27**, 199-204.
The-Uniprot-Consortium (2011) Ongoing and future developments at the Universal Protein Resource, *Nucleic Acids Res*, **39**, D214-219.
Tomlins, S.A., Mehra, R., Rhodes, D.R., et al. (2007) Integrative molecular concept modeling of prostate cancer progression, *Nat Genet*, **39**, 41-51.
Ulitsky, I. and Shamir, R. (2007) Identification of functional modules using network topology and high-throughput data, *BMC Syst Biol*, **1**, 8.
Ulitsky, I. and Shamir, R. (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions, *Bioinformatics*, **25**, 1158-1164.
Ulitsky, I., Shlomi, T., Kupiec, M. and Shamir, R. (2008) From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions, *Mol Syst Biol*, **4**, 209.
Valastyan, S. and Weinberg, R.A. (2011) Tumor metastasis: molecular insights and evolving paradigms, *Cell*, **147**, 275-292.
van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al. (2002) A gene-expression signature as a predictor of survival in breast cancer, *N Engl J Med*, **347**, 1999-2009.
Vanunu, O., Magger, O., Ruppin, E., et al. (2010) Associating genes and protein complexes with disease via network propagation, *PLoS Comput Biol*, **6**, e1000641.
Vazquez, A. (2009) Optimal drug combinations and minimal hitting sets, *BMC Syst Biol*, **3**, 81.
Venkatesan, K., Rual, J.F., Vazquez, A., et al. (2009) An empirical framework for binary interactome mapping, *Nat Methods*, **6**, 83-90.
Vidal, M., Cusick, M.E. and Barabasi, A.L. (2011) Interactome networks and human disease, *Cell*, **144**, 986-998.
Vogelstein, B. and Kinzler, K.W. (2004) Cancer genes and the pathways they control, *Nat Med*, **10**, 789-799.
von Mering, C., Huynen, M., Jaeggi, D., et al. (2003) STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res*, **31**, 258-261.
Wachi, S., Yoneda, K. and Wu, R. (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, *Bioinformatics*, **21**, 4205-4208.
Waltregny, D., Bellahcene, A., de Leval, X., et al. (2000) Increased expression of bone sialoprotein in bone metastases compared with visceral metastases in human breast and prostate cancers, *J Bone Miner Res*, **15**, 834-843.
Wang, Y., Klijn, J.G., Zhang, Y., et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet*, **365**, 671-679.
Weigelt, B., Mackay, A., A'Hern, R., et al. (2010) Breast cancer molecular profiling with single sample predictors: a retrospective analysis, *Lancet Oncol*, **11**, 339-349.
Weng, G., Bhalla, U.S. and Iyengar, R. (1999) Complexity in biological signaling systems, *Science*, **284**, 92-96.
Wood, L.D., Parsons, D.W., Jones, S., et al. (2007) The genomic landscapes of human breast and colorectal cancers, *Science*, **318**, 1108-1113.
Wu, X., Jiang, R., Zhang, M.Q. and Li, S. (2008) Network-based global inference of human disease genes, *Mol Syst Biol*, **4**, 189.
Yang, K., Bai, H., Ouyang, Q., et al. (2008) Finding multiple target optimal intervention in disease-related molecular network, *Mol Syst Biol*, **4**, 228.

Figure legends

Figure 1. Network-biology approach to cancer. First step in network-based cancer studies is data integration. (A) Known genetic descriptors (such as differentially expressed genes, mutations, etc.) for a given cancer type is either retrieved from the literature or experimentally identified. (B) An interactome is constructed using known PPIs. (C) Next, the gene-cancer associations obtained in (A) are overlaid on the products of these genes in the interaction network. (D) The phenotypic relevances (the likelihood of being involved in cancer) of the genes are ranked based on the topological characteristics of their products in the interaction network. (E) Promising candidates (top ranked genes) are computationally validated (e.g., by investigating the functional processes they are involved, co-expression with known genes associated with cancer and the tissues their products are localized). (F) The predictions are validated in wet-lab and depending on the results of the clinical trials (G) the predicted genes become a part of cancer therapies either as targets of drugs or biomarkers distinguishing the phenotype.

Figure 2. Combining topological properties of genes with structural and functional features to predict cancer gene candidates. (A) The method overview of (Aragues, et al., 2008). First, interactions of known cancer genes are fetched and cancer linkage degree is calculated. Second, gene expression from different cancer types is incorporated. Third, genes in the network are assigned probabilities based on their structural, functional and evolutionary properties. Finally, cancer genes are predicted for each cancer type. (B) Cancer linkage degree of a protein is the number of interaction partners that are known to be involved in cancer. (C) Positive predictive values (the percentages inside the diagrams) obtained when different strategies are applied separately or in combination with others to predict cancer gene candidates. Integration of multiple sources of data produced better predictions than the use of one single criterion. For example, combining the cancer linker degree with differential expression data increased the ratio of known cancer genes among all predictions from 17% to 73%. The figure is adopted from *BMC Bioinformatics* 2008, 9:172.

Figure 3. Identification of subnetworks discriminating cancer phenotype. (A) Gene expression and PPI data are integrated describing the dysregulation patterns as a result of certain cancer phenotype. (B) Possible subnetworks are searched typically starting from one node and then extending that node using several heuristics. (C) The subnetworks combinatorial expression of whose genes discriminate the cancer samples from normal samples are selected for further validation.

Figure 4. Deciphering distinctive organ-specific phenotype of breast cancer metastases. Based in comparative protein or gene expression analyses between primary tumor and each metastasis location, PPI networks can distinguish organ-specific preponderant functions. Coupling proteomic, transcriptomic and interactomic data, organ-specific PPI networks are reconstructed. The analysis of the networks reveals that although all metastases share common modules (represented in orange), each location trigger characteristic pathways and functions showed in red (brain), yellow (bone), green (liver) and blue (lung). Moreover, modules can be shared by several metastases according their cellular dissemination or colonization characteristics. Gray nodes represent proteins activated in both liver and lung metastases, indicating that important functions can be shared by soft-tissue metastases. Bone and brain have a tissular idiosyncrasy and metastasis in these organs derivate in a more specific and particular selection process.

BOX1: Data integration in cancer studies

Over the past years a vast amount of data from experimental cancer studies has been accumulated. Gene expression data is publicly available in two major repositories: Gene Expression Omnibus (GEO) (Barrett and Edgar, 2006) and ArrayExpress (Parkinson, et al., 2011). These two repositories contain high-throughput functional genomics data, including experiments related to different types of cancer (as of September 2011, a simple keyword search "cancer" restricting the organism to "Homo sapiens" results in 102,446 and 211,544 samples among 624,249 and 691,128 samples available in GEO and ArrayExpress respectively). In addition to these repositories, several databases such as Gene Expression Atlas (Kapushesky, et al., 2010) and Intogen (Gundem, et al., 2010) provide a platform for assessing importance of genes decided by statistical analysis on the integrated cancer expression data. Oncomine (Rhodes, et al., 2007) is another effort to compile gene expression data though access to advanced features requires a professional subscription. Moreover, initiatives such as International Cancer Genome Consortium (Hudson, et al., 2010), Cancer Genome Atlas (Collins and Barker, 2007) and Cancer Genome Project (Dickson, 1999) gather data to define the genetic landscape of various cancer phenotypes, improve the data quality and extend the catalog of genetic mutations in cancer.

Integrating molecular expression data is crucial to capture condition and time dependent behavior of genes *in vivo*. On the other hand, protein-protein interaction networks provide a snapshot of the relationships between the main actors in the cell. Unlike genomics data, the experimental data on relationships between biological macromolecules such as genes and proteins are spread across various data repositories. BIND(Alfarano, et al., 2005), BioGRID (Stark, et al., 2011), DIP (Salwinski, et al., 2004), HPID (Han, et al., 2004), HPRD (Keshava Prasad, et al., 2009), IntAct (Kerrien, et al., 2012), MINT (Licata, et al., 2012) , MIPS (Licata, et al., 2012), Mpact (Guldener, et al., 2006) are among widely used publicly available protein interaction databases. Furthermore, most of these databases lack a standard nomenclature and interface for the data they provide. To facilitate inter-operability among these databases, several software tools have been developed. These tools use equivalent identifiers or common features from different repositories (e.g. sequence identity or cross-reference identifiers) to merge data. A list of available biological data integration tools are given in **Table 1**. Some of these tools allow users to merge their own data with other available biological data to fetch species-wide genomic, proteomic and metabolomic annotation spread across various repositories such as UniProt (The-Uniprot-Consortium, 2011), KEGG (Kanehisa, et al., 2012), Reactome (Croft, et al., 2011) and major interaction data resources listed above.

BOX2: Basic concepts in graph theory

Relationships between discrete biological molecules (e.g. transcipts, proteins, metabolites) are typically represented as a network. Definitions and algorithms in graph theory provides a theoretical framework to characterize such networks. A **graph** (network) $\boldsymbol{G} = (V, E)$ is formally defined on a set of vertices (nodes) $\boldsymbol{V}$ and set of **edges** (links) $\boldsymbol{E}$ connecting a subset of *V*. G is called as a **directed graph** if its edges imply a directionality between the nodes they are connecting such that for two nodes u, v there are two possible edges: from u to v, (u,v); and from v to u, (v, u). In an **undirected graph** however, the edges between two nodes have no orientation. The **degree** of a node is the number of edges that connect to it and the vertices connected by these edges are called the **neighbors** of the node. Nodes with higher degrees are defined as **hubs**. The actual value of the degree threshold for defining a hub varies from study to study and depends on the biological context that the network represents. The **shortest path** between two nodes is the path in the network such that the sum of the weights of its constituent edges is minimum, thus there may be more than one shortest path connecting two nodes. The length of the shortest path between two nodes is called as the **distance** between these two nodes. G is a **connected graph** if there is a path between every pair of nodes in G. A subset S of the vertices V induces a **subgraph** (subnetwork), whose edges are a subset of E, that are the edges that connect two vertices inside S. All the connected subgraphs of G are called **connected components** of G. **Centrality** of a node in G determines the relative importance of that node within the graph and can be assessed using various metrics such as degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. A **clique** is a special type of subgraph where each node is connected with all other node and a clique with k nodes is referred as **k-clique**. The **clustering coefficient** of a node is the ratio of the number of the edges connecting any pair of its neighbors to the number of all possible edges that would exist between all possible pairs of neighbors. The **degeneracy** of G (k-core), the measure of how sparse G is, is given by the smallest value of k where in every possible subgraph of G, a node has a degree of at most k. The k-core of G is obtained by successively removing nodes with degrees less than k to the point that no further node removal is possible.

BOX3: Evaluating classifier performance

The prediction performance of a classifier is typically assessed with several metrics based on the number of **true positives** ("good" instances predicted as good), **true negatives** ("bad" instances predicted as bad), **false negatives** (good instances predicted as bad), **false positives** (bad instances predicted as good). **Accuracy** is the fraction of all correct predictions (true positives and true negatives) among all predictions. **Precision** is the ratio of true positives to true positive plus false positives (fraction of correct predictions for good instances among all predictions saying an instance is good). **Sensitivity** is

the ratio of true positives to true positives plus false negatives (fraction of correct predictions for good instances among all good instances). **Specificity** is the ratio of true negatives to true negatives plus false positives (fraction of correct predictions for bad instances among all bad instances). A **ROC** (receiver operating characteristic) curve compares sensitivity versus (1-specificity) while the threshold for being true positive is varied. The **AUC (**area under ROC curve) equals to the probability that a classifier will rank a randomly chosen true positive higher than a randomly chosen true negative. The practice of splitting the data into n groups and using (n-1) of these groups for training purposes while keeping the remaining group for testing is called **n-fold cross-validation**. During n-fold cross-validation, the process of choosing the groups are repeated n times such that each group is used once as the training set. This practice helps reducing the bias of the predictor towards the initial data used.

Table 1. Comparison of biological information integration software. *										
Feature		**BI**	**PI**	**PN**	**AP**	**BN**	**UH**	**MI**	**ON**	**iRI**
Data types	Supports multiple biomolecule types (protein, gene, compound...)	•				•		•	•	
	Supports multiple relation types (interaction, complex, pathway...)	•				•	•	•	•	
	Supports multiple data descriptor/identifiers types	•	•	•	•	•	•		•	•
	User extensible to new user defined data types and attributes	•								
Data Unification	*User specific data unification*	•							•	
	Standard user can extend to new data repositories	•							•	
User Interface	Standalone Graphical Interface					•			•	
	Scripting / Command line	•	•			•			•	
	Provides a webserver	•		•	•	•	•	•		•
	Provides a plugin for Cytoscape	•			•			•		•
Network analysis	Adds network analysis methods	•	•	•		•			•	
Availability	Open Source	•	•			•			•	
Installation	Does not require additional software			•	•		•	•	•	•
	Standalone application (runs locally)	•	•			•			•	

* Table adopted from Garcia-Garcia et al. (Garcia-Garcia, et al., 2010). Abbreviations: BIANA (BI); PIANA (PI); PINA(PN); APID / APID2NET (AP); BNDB (BN); UniHI (UH); MIMI (MI) ; ONDEX (ON) ; iRefIndex (iRI).

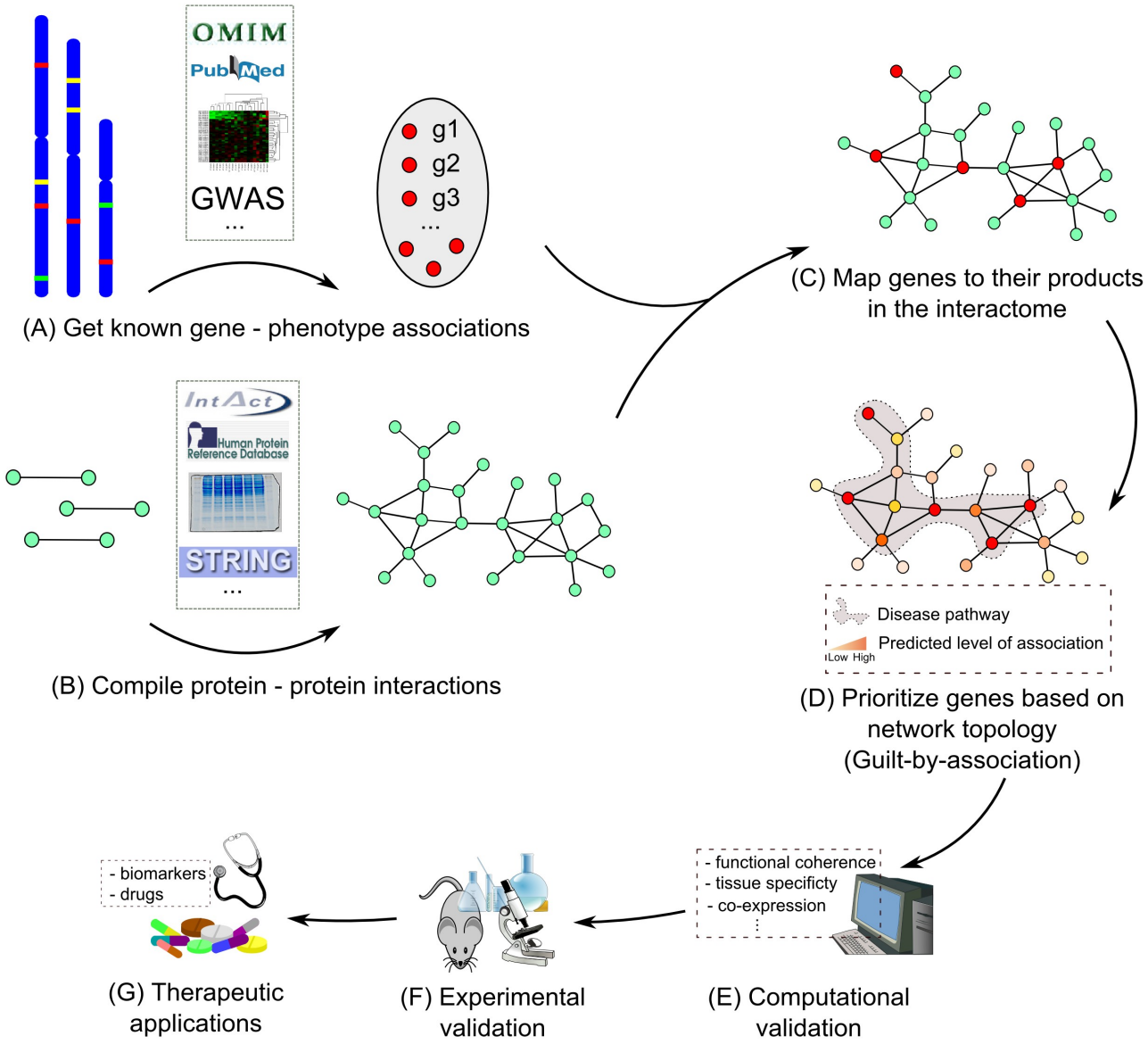

OMIM
PubMed
GWAS
...
g1
g2
g3
...
(A) Get known gene - phenotype associations
(C) Map genes to their products in the interactome
IntAct
Human Protein Reference Database
STRING
...
(B) Compile protein - protein interactions
Disease pathway
Predicted level of association
Low High
(D) Prioritize genes based on network topology (Guilt-by-association)
- functional coherence
- tissue specificty
- co-expression
(E) Computational validation
(F) Experimental validation
- biomarkers
- drugs
(G) Therapeutic applications

Figure 1

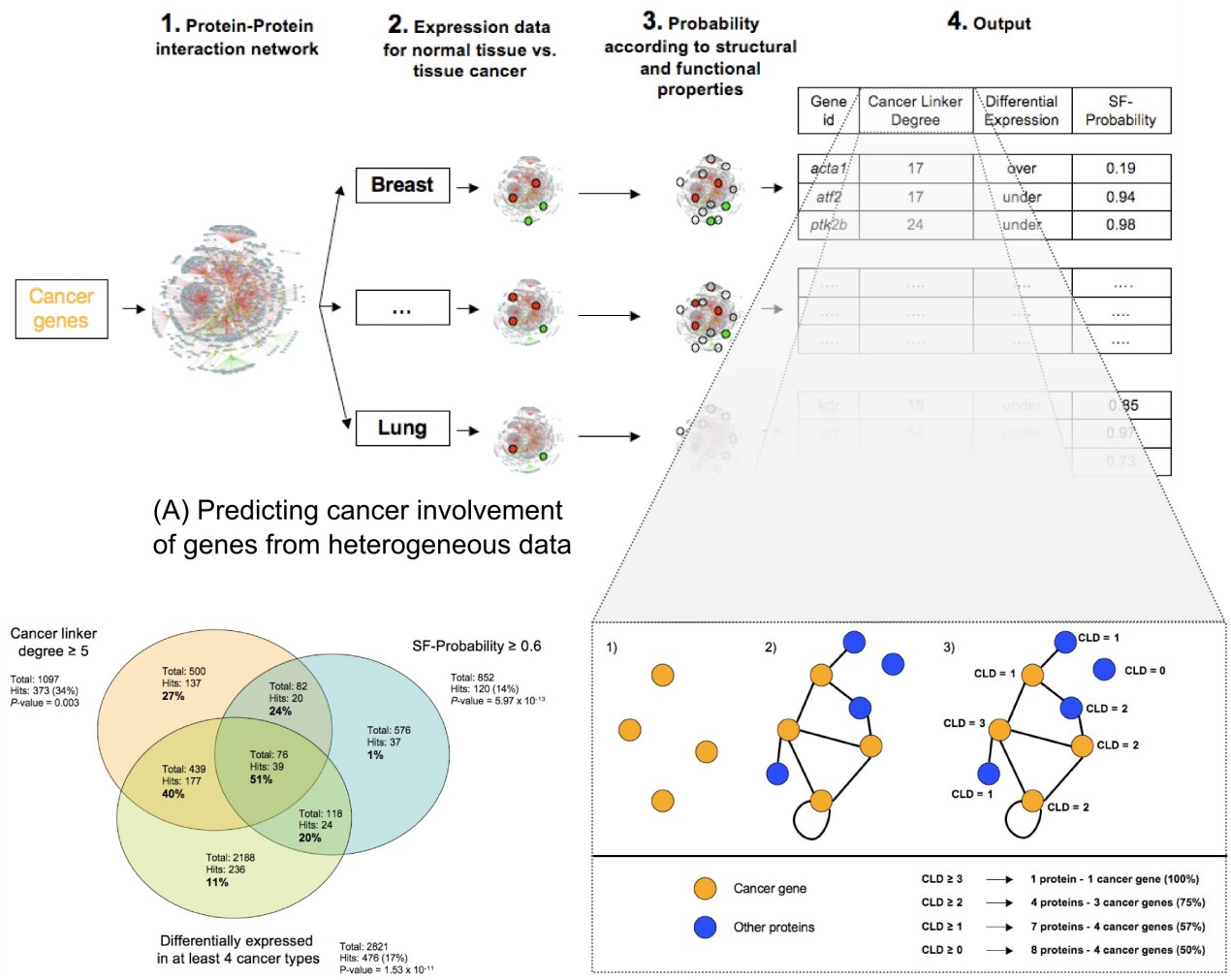

1. Protein-Protein interaction network
2. Expression data for normal tissue vs. tissue cancer
3. Probability according to structural and functional properties
4. Output
Cancer genes
Breast
...
Lung
Gene id
Cancer Linker Degree
Differential Expression
SF-Probability
acta1 17 over 0.19
atf2 17 under 0.94
ptk2b 24 under 0.98
kdr 15 under 0.85
(A) Predicting cancer involvement of genes from heterogeneous data
Cancer linker degree ≥ 5
Total: 1097
Hits: 373 (34%)
P-value = 0.003
SF-Probability ≥ 0.6
Total: 852
Hits: 120 (14%)
P-value = 5.97 x 10-13
Total: 500
Hits: 137
27%
Total: 82
Hits: 20
24%
Total: 576
Hits: 37
1%
Total: 439
Hits: 177
40%
Total: 76
Hits: 39
51%
Total: 118
Hits: 24
20%
Total: 2188
Hits: 236
11%
Differentially expressed in at least 4 cancer types
Total: 2821
Hits: 476 (17%)
P-value = 1.53 x 10-11
(C) Overlap among predictions using different strategies
1)
2)
3)
CLD = 1
CLD = 0
CLD = 2
CLD = 3
CLD = 2
CLD = 1
CLD = 2
Cancer gene
Other proteins
CLD ≥ 3 → 1 protein - 1 cancer gene (100%)
CLD ≥ 2 → 4 proteins - 3 cancer genes (75%)
CLD ≥ 1 → 7 proteins - 4 cancer genes (57%)
CLD ≥ 0 → 8 proteins - 4 cancer genes (50%)
(B) Cancer linker degree calculation

Figure 2

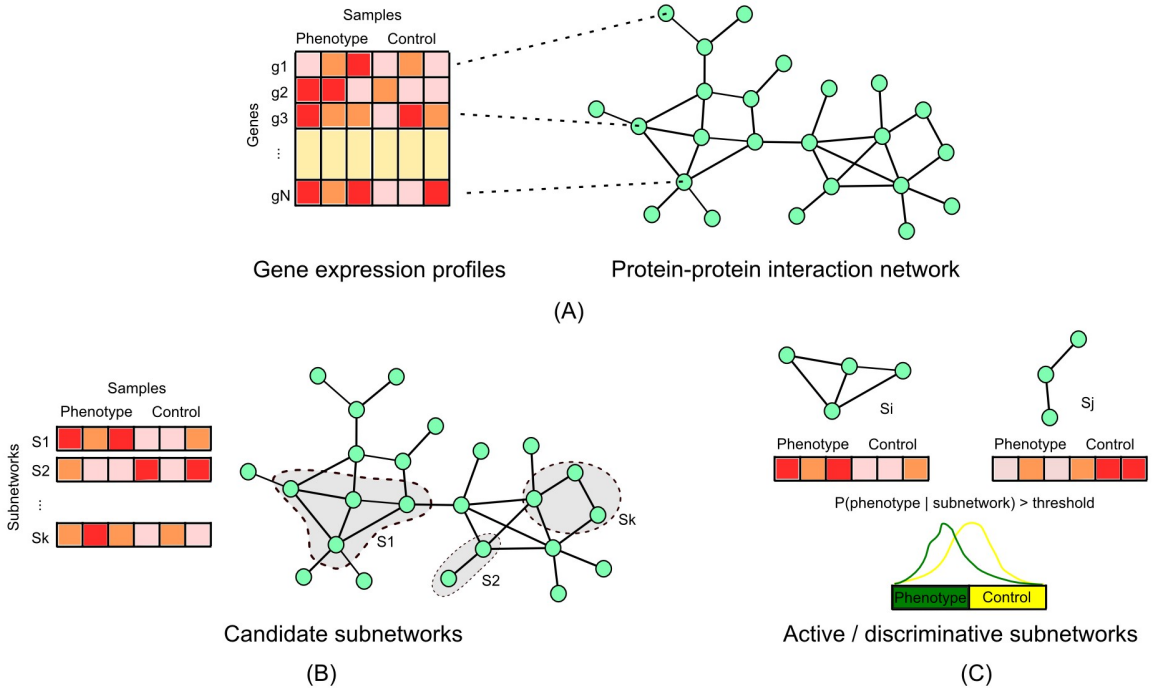

Samples
Phenotype
Control
g1
g2
g3
gN
Genes
Gene expression profiles
Protein-protein interaction network
(A)
Samples
Phenotype
Control
S1
S2
Sk
Subnetworks
S1
S2
Sk
Candidate subnetworks
(B)
Si
Sj
Phenotype
Control
Phenotype
Control
P(phenotype | subnetwork) > threshold
Phenotype
Control
Active / discriminative subnetworks
(C)

Figure 3

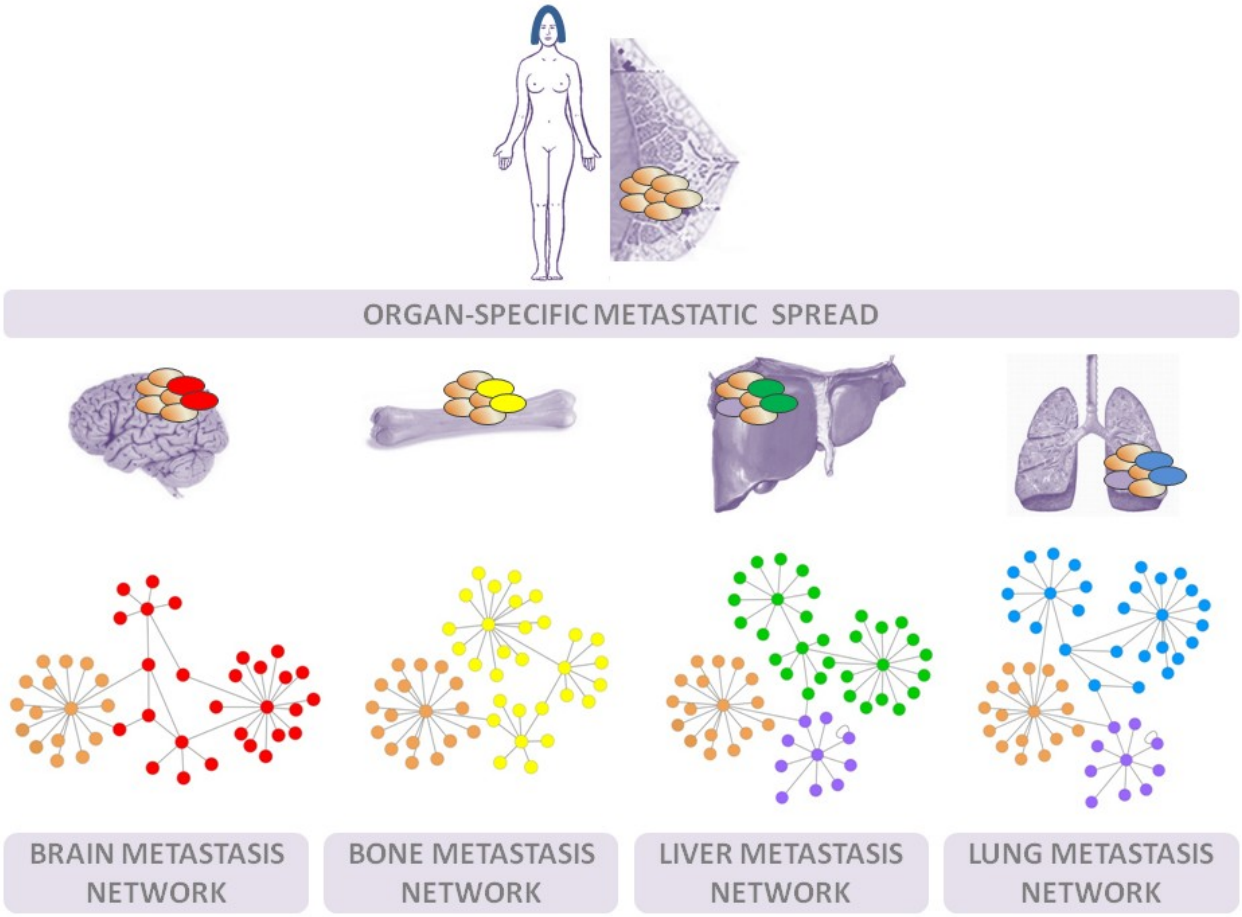

ORGAN-SPECIFIC METASTATIC SPREAD
BRAIN METASTASIS NETWORK
BONE METASTASIS NETWORK
LIVER METASTASIS NETWORK
LUNG METASTASIS NETWORK

Figure 4

Part IV

Discussion

A novel network-based disease-gene prioritization framework

In this thesis, I have presented a novel disease-gene prioritization framework, GUILD, which associated genes with diseases using the global topology of the network and an initial set of genes known to be implicated in the disease. Three algorithms have been designed and implemented in GUILD: *NetScore, NetZcore, NetShort*. These methods assigned a disease-association score to each node in the network considering the proximity to the known disease-gene associations (seeds).

In each case, the algorithms captured different aspects of the network topology by defining proximity in a different way. *NetScore* not only checked the shortest paths between nodes for calculating the distance in-between but also accounted for multiple shortest paths between nodes. *NetZcore* evaluated the proximity as the number of neighboring nodes but the adjacency is further refined to represent the biological significance of the neighborhood using an ensemble of networks in which nodes were shuffled randomly preserving the topology of the original network. The proximity in *NetShort* was assessed by the number of seeds contained in a path. These algorithms were combined in *NetCombo*, an algorithm which calculated average of the normalized scores coming from all three algorithms above.

In addition to these algorithms, four existing network-based prioritization algorithms, *Functional Flow* [Nabieva et al., 2005], *PageRank with priors* [Chen et al., 2009], *Random walk with restart* [Kohler et al., 2008] and *Network propagation* [Vanunu et al., 2010] have been implemented in GUILD (among the existing algorithms, only *PageRank with priors* algorithm had a publicly available implementation). GUILD was made publicly available as a standalone software package at `sbi.imim.es/GUILD.php`. A manual providing instructions on how to install and use the software, along with example input files was also provided.

Benchmarking network-centric guilt-by-association

The prediction capacity of the network-based prioritization methods in GUILD was tested throughly at genome-wide scale using disease-gene annotations from various datasets (such as OMIM, GAD, CTD and several available as expert curated datasets) covering around 100 different disease phenotypes and PPI networks (human interactomes at different levels of granularity integrated from several major PPI databases). On average *NetCombo*, the consensus method combining the results of *NetScore, NetZcore and NetShort*, outperformed state-of-the-art network-based prioritization methods included in GUILD.

The results suggested that *NetCombo* and *NetScore exploited the global topology of the network better than the existing methods.* Typically, available network-based prioritization methods consider all the paths between nodes equally relevant for a particular disease. In this regard, incorporating a refined definition of closeness while assessing the relevance of the nodes in the network improved the prediction accuracy.

Some of the prioritization algorithms, *NetScore, NetZcore* and *Functional Flow* required a user-defined parameter: the number of look-ahead links during the proximity calculation in the network. That is, how many links further from one node should the algorithm consider. A value of 6 in the case of *NetScore* and 5 in the case of *NetZcore* and *Functional Flow* have been found to be optimum during the parameter optimization process. Given that PPI networks have been found to be scale-free and bear "ultra-small world" effect (one can reach to any node with much less than 6 links), these values coincided with the expectation that the nodes further than $5 - 6$ links did not provide any information that would improve the annotation.

Several of the existing methods like *PageRank with priors* and *Random walk with restart* have addressed genome-wide disease-gene prioritization previously [Wu et al., 2008, Navlakha and Kingsford, 2010], however a comprehensive comparison of network-based prioritization methods using only network topology has not been available so far, probably due to two main reasons. First, most of these methods are either not available as a standalone tool or accessible through interfaces with limited functionality (i.e. as a web server [Kohler et al., 2008, Chen et al., 2009] or Cytoscape plugin [Gottlieb et al., 2011]). Second, these methods integrate a variety of additional data including phenotypic similarity between diseases. Furthermore, most of the candidate disease-gene prioritization studies focus on the prioritization of genes within the genetic linkage interval of a given disease-gene. The prediction performance of the method is then evaluated based on whether it could rank the disease-gene highest among the 100 closest genes (to the disease-gene) in the linkage interval. However, considering 100 closest genes as candidates could be too restrictive in some cases, since the linkage interval might contain as many as thousands of genes and multiple disease-genes can fall under the same genetic interval. Therefore, the prioritization methods were benchmarked assuming all the genes (whose products are in the interaction network) as candidates. To ensure a fair evaluation, the negative genes were selected in two different ways: *i)* assuming all non disease-genes as negatives (not associated with the disease) and *ii)* assuming a balanced distribution between disease-genes and non disease-genes (achieved by grouping non disease-genes such that there were equal number of disease-genes and non disease-genes).

Biological significance of the predictions

The network-based algorithms assigned a "proximity" score for each node in the network corresponding to the likelihood of that node being involved in the phenotype of interest. The nodes were then ranked/prioritized with respect to the calculated score. The validity of these scores was confirmed on disease-gene associations provided by CTD. The *NetCombo* scores calculated for Alzheimer's disease (AD), diabetes and AIDS, distinguished disease associated genes from genes that were not associated with the disease (no-association genes). The findings also demonstrated that *there were 4 to 8 times more disease-genes than no-association genes at a reasonable score cutoff* (e.g., 0.1 within the range of $[0, 1]$).

The analysis of top-ranking genes (i.e. genes encoding highest scoring 1% of the nodes) further proved the utility of the prioritization using *NetCombo*. For all of the

three disease phenotypes, top-ranking genes covered significantly more disease genes with respect to GAD. Among these three phenotypes, prioritization yielded slightly poorer results in diabetes which could be due to the definition of the diabetes phenotype merging diabetes type 1 and 2 in a single phenotype.

When the biological processes enriched among the top-ranking genes were taken into account, Notch and amyloid pathways stood out in AD and inflammatory response pathway was implicated in diabetes and AIDS. Moreover, in the case of AD, some of the predictions were consistent with the literature and network-based prioritization highlighted several genes that have recently been implicated with the disease pathology. It is also worth mentioning that some of these genes did not fall to any known linkage interval for AD pointing out the potential of genome-wide prioritization.

Dependence on current knowledge

Two main components of the network-based prioritization are the information on the seeds and the network through which the information of the seeds is transferred to the other nodes. Just as the way that there is no guilt if there is no proper evidence, the success of the prioritization would be dependent on the quality of seed and PPI information.

As the size of the PPI network went up, the prediction accuracy of the network-based prioritization methods increased with the exception of the interaction network which contained protein complexes in addition to binary (physical) interaction information. This finding was not surprising considering that the more was known on the relationships between genes (or their proteins) the more accurate would be the annotation. However, protein complexes might be introducing spurious interactions. Take a complex consisting of three proteins A, B and C for example. The interaction between A-B and B-C may exist while A and C need not necessarily be interacting with each other. Although complex involvement could be thought as functionally and thus phenotypically informative, it is not as informative as physical interaction and may hinder precise functional characterization. Another key finding arose from testing various types of PPI networks was that incorporating additional evidence for the functional importance of the interaction (e.g., genomics data relating two genes connected in the network) could further improve the prediction performance of the prioritization.

On the other side, the prediction capacity of the prioritization methods varied substantially over different disease phenotypes. The prioritization methods worked better on several diseases (i.e., OMIM disorders merged with respect to the first keyword) such as amyloidosis, myasthenic, myocardial and xeroderma whereas for mitochondrial, osteopetrosis and epilepsy they were less successful. The number of seeds in each disease was quite diverse, nevertheless, there was no correlation between the prediction accuracy and the number of seeds given for the disease. Rather surprisingly, several prioritization methods, including *NetCombo* achieved better performance for diseases with lower number of seeds. Nevertheless, *the topology of the network was observed to play a more important role than the number of seeds in defining the outcome of the prioritization.* In general, the prioritization performance was slightly higher for the diseases whose seeds were connected to each other with shorter paths.

Using a dataset in which the phenotype is defined in a broader sense (the dataset of [Goh et al., 2007]) such that the disease-gene associations of a number of phenotypes are grouped together allowed to evaluate the effect of the quality of the known disease-gene associations. The prediction performance were lower in this dataset compared to other disease-gene association datasets tested.

Robustness and fragility

Robustness of five of the network-based disease-gene prioritization algorithms mentioned above (*NetScore, NetZcore, NetShort, Functional Flow, PageRank with priors*) were further analyzed using randomly perturbed seed sets and PPI networks for a subset of diseases. As expected, for all the methods, the prediction capacity fell down linearly as the noise in the seed set increased. Consistent with the findings mentioned above, the quality of the seeds affected the outcome of the prioritization more than the number of seeds. A relevant implication of this finding for guilt-by-association approaches in general is that one should use a grain of salt while incorporating data that is not expert curated, as this may be worsening the power of the method rather than improving.

When the noise was introduced in the network via rewiring interactions, the predictions were also less accurate, though the pace of the drop in the accuracy with respect to the amount of perturbation introduced was rather slow compared to the drop in the accuracy due to the perturbation of the seeds. Moreover, the prediction performance of *PageRank with priors* method was less effected compared to the rest despite rewiring meaning that the probability of randomly reaching to a node over the links of the network remained similar. This might have emerged from the scale-free nature of the PPI network, where proteins are connected with each other with very short paths. Thus, rewiring interactions preserved considerably the probability of reaching from one node to another.

On the other hand, a rather unexpected behavior was observed with random removal of interactions in the network. The prediction performance of *NetScore, Functional Flow* and *PageRank with priors* went down as the interactions were deleted from the network, whereas, *NetZcore* and *NetShort* increased their prediciton performance with less number of interactions. Strikingly *NetZcore* and *NetShort* were able to predict disease-genes better than random even when half of the interactions in the network were removed.

This observation could be explained by the two aspects of network-based prioritization: first, the data underlying the prioritization and second, the way these two methods worked algorithmically. Firstly, the seeds were a tiny portion (typically much less than 1%) of all the genes in the network so were the interactions connecting a seed with another seed or any other node among all the interactions in the network. Therefore, the probability of removing an edge connecting two seed or a seed and non-seed was lower compared to the probability of removing an edge connecting two non-seeds. That would suggest that the perturbed networks would be more likely to preserve connections among seeds. Indeed, a case study on AD supported this explanation where the ratio of AD-related genes in the neighborhood increased as the interactions were removed.

Secondly, turning to the algorithmic properties of the prioritization methods, *NetShort* gave more weight to the paths containing more seeds. A performance improvement implied that random interaction removal eliminated paths with less number of seeds (since it was more likely to remove an edge connecting non-seeds). A similar rationale applied to *NetZcore*, which boosted the score of the node based on the number of neighboring seeds in the original network compared to the random ensemble of networks. Seed-seed connections were more likely to be preserved with random edge removal in the original network and thus would become more significant compared to the random ensemble of networks. With the same line of reasoning, the performance drop observed in *NetScore*, the method that considered all possible shortest paths between nodes, provided evidence for the existence of alternative routes between seeds. In other words, although removing interactions was more likely to eliminate non-seed connecting paths, the predicton accuracy of *NetScore* went down since alternative shortest paths connecting seeds –which possibly included non-seed connections in-between– were affected.

Taken together, these results suggested that disease phenotypes reflect the robustness of the underlying biological systems. The cell can maintain its functioning to a certain extent even under abnormal conditions and this is likely to be valid for disease states. Moreover, *redundancy, that is the existence of backup circuits within the interaction network played a key role in robustness.* Nevertheless, this does not mean that the biological system is immune to all kind of perturbations. Certain specialized perturbations can easily exploit the fragility of the system. In particular, under the context of diseases, the information of seeds and the connections between them could substantially affect the prioritization of disease-genes constituting the Achilles' heel of the network-based prioritization. Still, it should be taken into account that the analysis presented here is a rather indirect way of proving robustness involving both the disease state and the underlying biological system and further investigation (e.g., testing nodes that change seed connectivity most for the association to the disease) is required to biologically support these findings.

Identifying functional decoupling and its implications on the plasticity

The analysis of the prioritization methods with respect to their ability to identify groups of genes functionally related to the disease showed that *NetScore* was superior to the rest of the methods. The genes prioritized by *NetScore* tended to cluster in the network as connected set of genes and these clusters were more enriched in those GO functions that were enriched among the seeds. This method was then used to check whether there were certain features of diseases mediating the outcome of the prioritization with the hypothesis that it might be easier to recover the disease-genes for a particular group of diseases. It was assumed that investigating the capacity of the network-based prioritization to recover disease-genes under perturbed conditions would delineate the adaptation capacity of the disease to the changing environment (e.g., drugs attacking to certain PPIs).

Overall prediction accuracy and being able to distinguish between groups of connected genes with functions identical to those of the seeds were two key criteria to

choose the prioritization method for investigating disease-mediated features. *NetZcore* also satisfied these criteria however, as mentioned above, it proved intrinsically more robust against the perturbations on the seeds and the network. Thus, the observed performance would be more likely to account for the robustness of the method in addition to any disease-mediated feature. To minimize the bias that might be introduced by the method *NetScore* was chosen instead.

Based on the performance of *NetScore* prioritization over perturbed networks, the diseases were grouped into two main categories: diseases whose associations were easier to recover using network-based prioritization (robust) and those that were harder (non-robust). Interestingly, robust diseases were more likely to include diseases with high prevalence in society including breast cancer, diabetes and obesity. Robust diseases were distinguished from non-robust diseases by the connectedness of their seeds in the network. A disease was more likely to be robust if its seeds were connected with shorter paths to each other. The top-ranking genes of the robust diseases showed slightly increased modularity compared to the top-ranking genes of the non-robust diseases. Moreover, both the seeds and the top-ranking clusters of the robust diseases tended to cover a larger number of biological functions than non-robust diseases.

These findings pointed out two other aspects of robustness: functional decoupling and functional diversity. *The biological processes relevant for the diseases are handled through designated units (modules) diverse in function and the modularization is probably achieved by having shorter connections between disease-genes at the network level.* Consequently, the network-based prioritization method recovered the mediators of the phenotype (disease-genes) easier in spite of perturbations in the underlying network for the diseases whose seeds were more connected in the network. Therefore, it is claimed that these diseases would be more likely to show plasticity and adopt to changing conditions than the rest of the diseases. The reason why polypharmacological approaches targeting multiple gene products simultaneously work better on the diseases such as AIDS or cancer might be due to such plasticity these diseases may have.

Extending the focus to non-disease phenotypes

Though mainly tested on disease phenotypes, the network-based prioritization methods presented in this thesis can be applied to any phenotype given a set of phenotypic annotations (seeds, e.g., disease-genes in the case of diseases) to be transferred over the network. Nonetheless, a common challenge in phenotypic characterization using guilt-by-association principle is often the lack of such phenotypic annotations.

This challenge has been addressed by integrating phenotype-gene information from various available biological data repositories such as UniProt, GO and OMIM and by applying network-based prioritization method in order to characterize genes for their relevance to any given phenotype. GUILDify, a free and easy to use web server has been made publicly available at `sbi.imim.es/GUILDify.php`.

GUILDify locates phenotype-gene annotations by searching for the user-provided keywords in its knowledge base and provides a genome-wide ranking for all genes in the PPI network. The ranking can be used to shortlist the set of candidate genes that needs to be further validated. In addition to the data extracted from biological databases, GUILDify lets users to input their own set of annotations to be used in the prioritization.

Systems-level characterization of dysregulation patterns in cancer

Cancer is a complex, pathway-centric phenotype. Even slightest changes in interconnected pathways can cause abnormal regulation events affecting whole system. Characterization of dysregulation patterns in cancer requires the integration of multiple types of data including information on gene regulation, signaling and cell metabolism as well as the cross-talk between them. Systems biology approaches aim to foresee the outcome of perturbations in regulation patterns by incorporating protein-protein interactions in addition to proteomic, genomic, metabolic, physiological and environmental information defining a dynamic context and possibly suggesting points of action [Pujol et al., 2010, Fliri et al., 2010]. These approaches can have multiple biological and clinical applications: first, they may identify disease genes and disease pathways; second, they can be used to discover new targets and to develop new drugs; and third, some of these new targets can be used as more accurate cancer biomarkers leading to improved personalized therapies and treatments [Barabasi et al., 2011].

Future directions

Network-based prioritization offers a great opportunity to rank genes with respect to their relevance to the phenotype of interest. However, the accuracy of the methods depends considerably on the seeds (known associations) and the underlying network. Using only PPI information to describe the network may fail to identify associations for the genes whose proteins do not interact with other proteins. Therefore, incorporating gene expression, functional annotations or phenotypic similarity profiles to employ a network where the links are ”functional associations” rather than solely physical interactions could significantly improve the prioritization accuracy. Several works have already reported favorable results using functional association networks [Franke et al., 2006]. It would be very interesting to see how the methods presented here work using such kind of networks.

Towards a comprehensive understanding of biological pathways underlying diseases, it would be also engaging to use the methods presented here to discover so called ”active subnetworks”, clusters of genes in the PPI network that account for the observed phenotypic difference. A recent study has employed *PageRank with priors* to identify such subnetworks [Nibbe et al., 2010b]. Considering that the methods presented here outperformed this algorithm, biologically more sound results can be achieved.

Another application area of the methods presented here would be prioritization of SNP data coming from GWAS. Using network-based prioritization is tempting considering the collaborative role of genetic variants in diseases. Rare variants that are frequently skipped during the analysis of GWAS data can be pinpointed using presented methods albeit some inherent limitations such as most of the SNPs falling under a non-coding region. In fact, *PageRank with priors* method has already been applied to this problem as well [Lee et al., 2011, Akula et al., 2011].

A topic gaining attraction in the research community lately is investigation of pathways shared among disease. Identifying the genetic elements common to diseases and

reusing existing drugs accordingly (drug repurposing) has been central to research efforts recently. These methods can characterize the interconnected pathways implicated in diseases and possibly suggest points of action to compensate the changes induced by the disease [Zanzoni et al., 2009].

At last but not least, given that the gene-gene interactions are the key descriptors of biological function, the network-based prioritization methods can be adopted to predict genetic interactions. Efforts towards this direction has also been exerted recently, in particular using *Random walk with restart* method [Chipman and Singh, 2009] to calculate the proximity in the network and score interactions in non-human species. However, in order to predict genetic interactions in human and to take into account the biological context that makes genetic interactions plausible, annotations of relevant biological functions identified in model organisms can be transferred to human and gene pairs prioritized with respect to these functions can be checked for their enrichment in genetic interactions.

Part V

Conclusions

The main contributions of this thesis can be summarized as follows:

- A novel network-based disease-gene prioritization framework, GUILD, has been implemented and made publicly available. GUILD associates genes with diseases using the global topology of the network and an initial set of genes known to be implicated in the disease (seeds).
- The prediction of accuracy of the algorithms in GUILD is evaluated on human genetic diseases using PPI networks. The results show that proposed algorithms exploit the global topology of the network better than existing algorithms.
- The genes prioritized by GUILD for a particular disease are more likely to be implicated in the disease than the rest of the genes encoding the proteins in the interaction network.
- The outcome of the network-based prioritization depends on the network and seeds. A combination of these two features, the connectedness of the seeds in the network, defines the success of the prioritization.
- The perturbations on the input data affect all the prioritization algorithms however, some of the prioritization algorithms in GUILD are more robust against perturbations than the others.
- For several diseases, disease-gene annotations are easier to be recovered using network-based prioritization methods suggesting robustness at pathophenotypic level. The robustness of disease phenotypes emerging from the underlying biological system is probably due to *i)* alternative routes in the interaction network connecting the seeds and *ii)* functionally diverse modules in which the seeds are connected with shorter paths.
- GUILD has been extended as an online and user-friendly tool (GUILDify) where the genes can be prioritized for their association to any user-provided phenotype. GUILDify ranks genes in human interactome fetching initial gene-phenotype associations for a given phenotype from BIANA, a biological data integration platform.
- The current state-of-the-art systems-biology approaches towards delineating disease mechanisms –particularly in various types of cancer– have been reviewed and further uses of the developed framework have been discussed.

Part VI

Appendix

Appendix A

BIANA: A software framework for compiling biological interactions and analyzing networks

Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010 Jan 27;11:56-2105-11-56.

Appendix B

Extending signalling pathways with protein-interaction networks. Application to apoptosis

Planas-Iglesias J, Guney E, Garcia-Garcia J, Robertson KA, Raza S, Freeman TC, et al. Extending signaling pathways with protein-interaction networks. Application to apoptosis. OMICS. 2012 May;16(5):245-256.

Appendix C

Networks of Protein-Protein Interactions: from uncertainty to molecular details

Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas J, Oliva B. Networks of Protein- Protein Interactions: From Uncertainty to Molecular Details. Mol Infor. 2010 May; 31(5): 342- 362.

Appendix D

Identifying genes involved in human cell fate determination

Introduction

Therapeutic benefits of stem cells given their unrestricted differentiation potential are out of question. Though, the mechanism beneath the developmental process these stem cells undertake is far less understood. Embryonic stem cell (ESC) fate is shown to be dynamically regulated by transcription factors, their interactions and epigenetic modifiers (mainly miRNAs and methylation events). Widely accepted view on cell fate determination suggests that state of a cell is determined by the interplay of these transcriptional regulation elements yielding in transitions between several coexisting stable states [1].

Core regulatory circuitry mediating ESC self renewal and differentiation has been fairly explained [2] and interest is recently shifting towards identifying transcription factors involved in the later stages of the embryonic development [3]. Studies demonstrate that cells at similar differentiation state cluster together with respect to their gene expression profiles [4,5]. Furthermore, it has been shown that group of "actively communicating" genes (bearing differential expression patterns and whose products interact) has more discriminative power in terms distinguishing a certain phenotype or cell type than individual genes alone [6,7]. Thus, combining gene expression data with transcription factor interaction data and annotation information at the developmental level can provide insights to the key transcription factors mediating cell fate.

Associating transcription factors with their role in cell lineage requires a well structured relationships hierarchy between the steps of the lineage. Though, the Gene Ontology (GO) [8] is among one of the most comprehensive functional annotation source, it does not contain developmental lineage information [9] but rather focus on defining biological processes behind the development. On the other hand, the Cell Ontology [10] and more recently an anatomy oriented ontology [11] also aim to provide a basic framework for cell types but fail to capture the human developmental lineage categorize genes falling under each specific category.

Integration of gene expression data over multiple experiments can benefit addressing the cell types or developmental stages a gene is expressed in [12,13]. One prominent approach is Bgee database which collects information in which organs and at which developmental stages a gene is expressed [12] through ontology mapping. Recently, Novershtern et al. [14] identified modules of highly coexpressed genes in investigated roles of these modules in hematopoietic differentiation. They showed that some of the identified modules are specific to certain lineages and controls changes in differentiation. On the other hand, Ravasi et al. [7] identified modules involved in tissue development and classified the cell lines according to their embryonic origin (endoderm, mesoderm, ectoderm).

Here, we use a large compendium of microarray experiments and develop a bioinformatics method to identify gene sets as the mediators of differentiation over human developmental cell lineage. We then investigate to which extent these gene sets can be used to distinguish the cell line of given sample.

Results and Discussion

Human developmental lineage tree

Throughout human development stem cells go through various steps of differentiation yielding in different tissues and cell types. We built an ontology containing all the cell types we had experimental gene expression data to capture these possible paths of differentiation (Figure 1). This ontology was referred as the *human lineage tree*, since the path from the root to each leaf constituted a possible path that a cell can undertake during differentiation. Therefore, the tree described all possible lineages that a cell could commit (e.g., from embryonic stem cell to liver cell). The lineage tree covered a diverse set of cell types including both fully differentiated cells such as tissue specific smooth muscle cells and rather less differentiated germ layer cells like ectoderm, mesoderm and endoderm. The expression data containing 697 samples was laid over the lineage tree (see Methods). Fully undifferentiated cells, that is ESCs, were a lot more represented in the gene expression dataset with 425 samples, whereas there was no sample from germ layers and there were only a handful of samples for fully differentiated cells. This is because controlling the state of the germ layer cells is often not possible due to unavailability of biomarkers distinguishing these kind of cell types. Consequently, we did not have gene expression data for some of these cell states and we imputed information for these states using their children cell states in the lineage tree as a proxy.

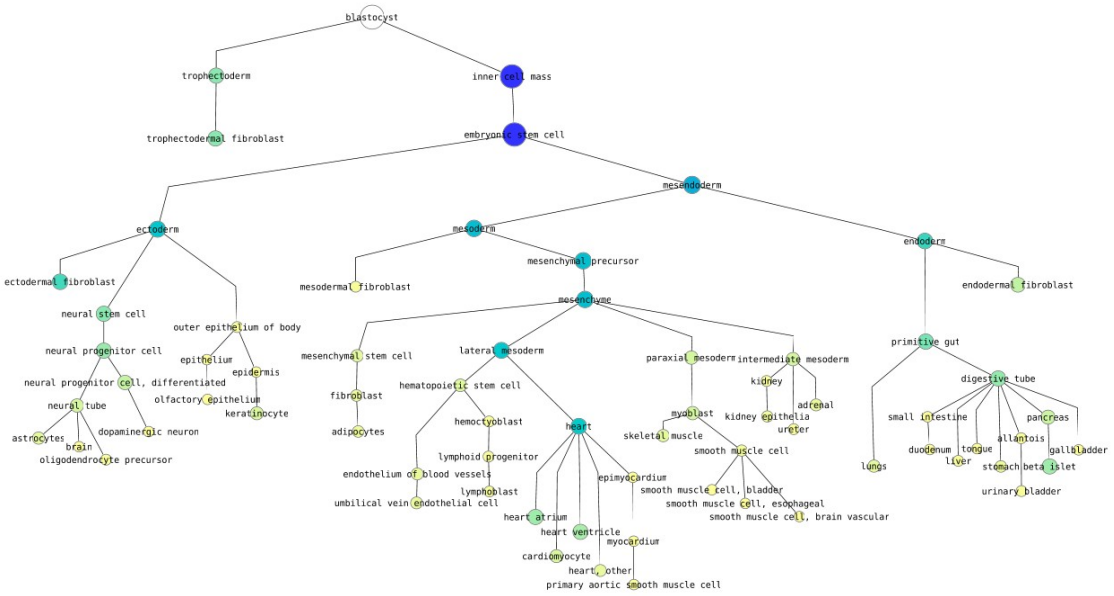


Figure 1. Lineage tree for the cell states in the human cell differentiation. Node size is scaled in accordance to the number of total samples. The nodes are colored with respect to the number of microarray samples that belong to those nodes (cell states).

Expression coherence among arrays

We confirmed the homogeneity of the samples for the ESCs and the three germ layer cells, namely ectoderm, mesoderm, endoderm cells by calculating the sample-wide correlation of the expression of all the genes in these samples with each other (Figure 2). The gene expression among the samples of the three germ layers was less coherent, due to using the samples of more differentiated tissues to describe

these cell states (e.g., using neural stem and brain cells to proxy ectoderm). Furthermore, there were groups of ESC samples that were rather correlated with ectoderm, mesoderm and endoderm samples. Although these annotations were spurious, they were not removed from the dataset since all of these samples were experimentally verified as stem cells (i.e. via checking the existence of various stem cell markers).

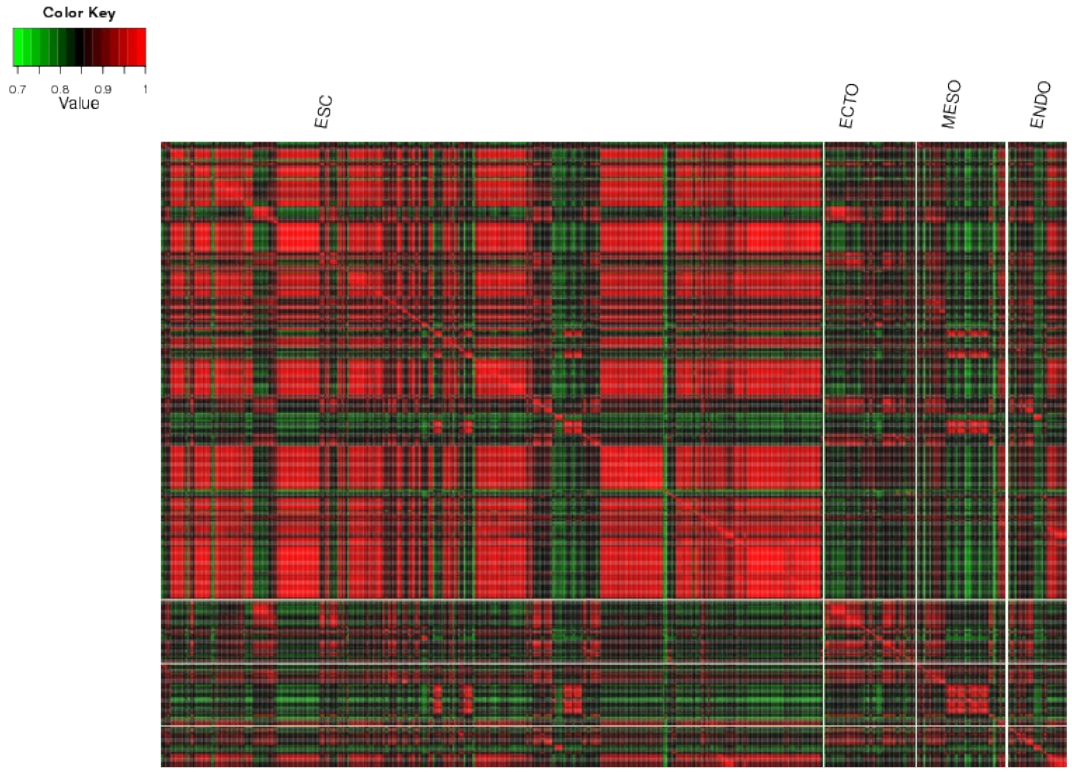


Figure 2. Correlation matrix for the gene expression samples of ESC and the three germ layer cells.

Gene expression patterns at the top of the lineage tree

In order to identify the genes mediating the cell fate decisions, we applied a statistical difference-of-mean test comparing the samples of one cell state and all the remaining samples (see Methods). Figure 3 shows top differentially expressed five genes among the ESC and three germ layer samples. Consistent with known stem cell maintenance and commitment mechanisms, POUF51 (a.k.a. OCT4) stood out as an important player showing dominantly up-regulation pattern during maintenance (among ESC samples) and down-regulation during commitment (among the samples of germ layers). Other known mediators of stem cell differentiation such as SOX2 and NANOG were also highly differentially expressed in ESCs (see Supplementary Material).

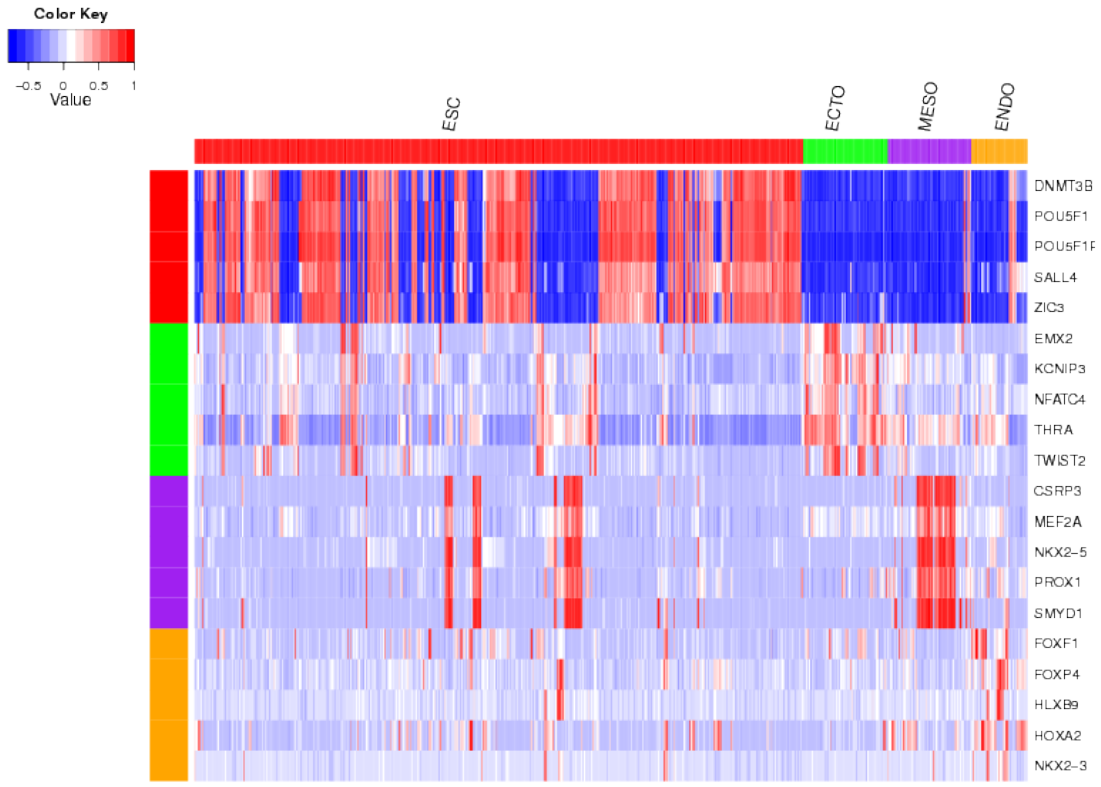


Figure 3. Top differentially up-regulated genes in ESC and the germ layer cells.

Next, we checked whether there was an enrichment in terms of up-regulated transcription factors in these cell states, i.e. if several cell states were more “active” compared to others. Figure 4 shows the distribution of the number of up-regulated transcription factors as a function of the standard deviation of the gene expression distribution within the samples of each cell state. The number of active transcription factors were similar in all cell types, though endoderm showed a slightly increased tendency to have more up-regulated transcription factors. Nevertheless, none of the differences in number of up-regulated transcription factors was significant (assessed by a two-sided t-test, alpha value 0.05) suggesting that similar number of transcription factors were involved in the cell fate determination at the top of the lineage tree.

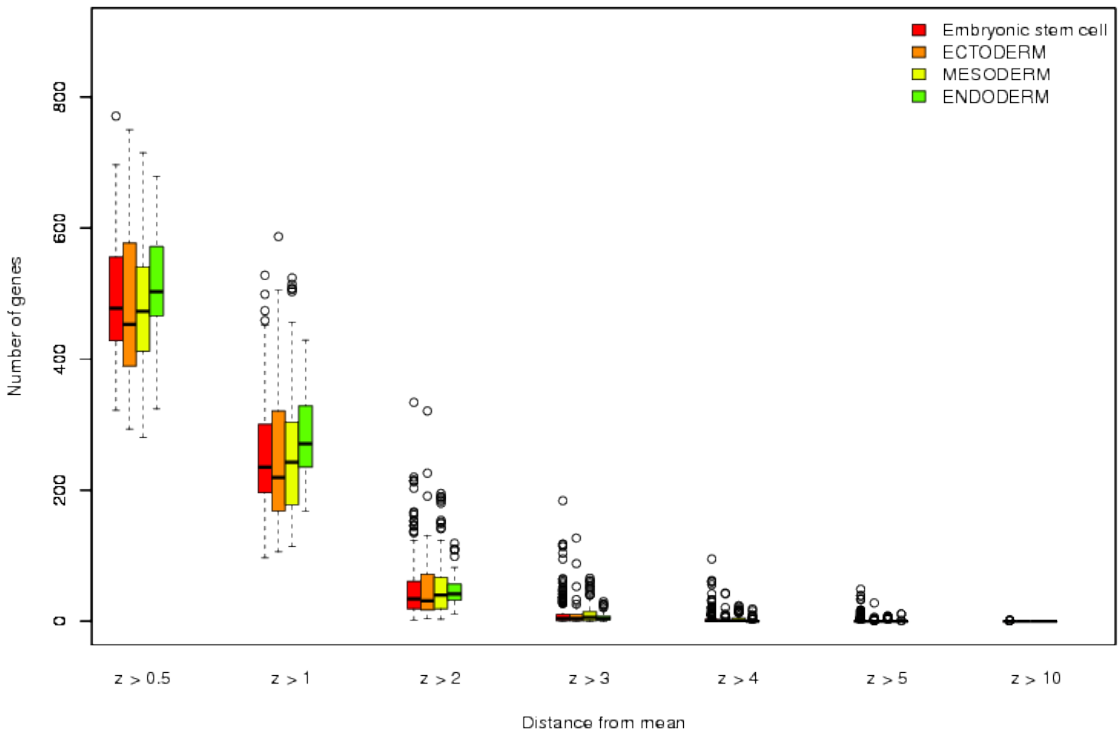


Figure 4. Number of up-regulated transcription factors in ESC and the germ layer cells at different standard deviation cutoffs of normalized gene expression distribution (z scores). For each sample, the gene expression values were standardized using the expression values of each gene in order to have a normal-like distribution with mean 0 and standard deviation 1.

Predicting mediators of cell fate

Differential expression plays an important role in defining the cell state during differentiation. Using gold standard dataset for ESC, and the germ layers, we investigated the capability of test of differential expression to distinguish the cell state. Based on the hypothesis that the genes involved in one cell state should be differentially expressed compared to other states in general, we compared the means of distribution of gene expression in the samples of one cell state compared to the rest of the samples. The prediction performance was assessed as the area under ROC curve (Table 1). Among all the approaches evaluated (see Methods), Welch's test with versus rest grouping strategy and considering only up-regulated transcription factors achieved the best prediction performance. That is, a gene was more likely to be involved in the differentiation of a cell state if it was a transcription factor and its expression in that cell state was significantly higher than its expression in the rest of the cell types using Welch's test. Moreover, we are currently working on *i)* improving predictions using regulatory relationships between genes (e.g., gene co-expression) and *ii)* incorporating topology-based scores coming from the neighborhood of the genes in the protein-protein interaction network in addition to differential gene expression.

Table 1. Prediction accuracy of test of differential expression for embryonic stem, ectoderm, mesoderm and endoderm cell states.

Cell state	Area under ROC (%)
Embryonic stem cell	78.7
Ectoderm	68.5
Mesoderm	64.6
Endoderm	72.4

Predicting cell line of origin for the validation samples

The genes involved in the differentiation of cell types at the top of the lineage tree were predicted with a reasonable accuracy using differential expression. We extended this rationale to annotate the cell type of a given sample. Accordingly, the cell state of a given sample was predicted by comparing the expression of genes in the sample with the expression of genes in each cell state in the lineage tree (see Methods). First, we applied a ten-fold cross-validation procedure to find the optimum number of differentially up-regulated genes that discriminate the cell states. The average number of correctly classified arrays using one array from each cell state (24 in total) at three different runs of cross-validation is given in Table 2.

Next, we predicted the cell types of 16 additional arrays provided in the validation set (two samples in the validation set were not considered since it was not possible to map these samples to any of the cell states in the lineage tree) using most up-regulated 31 genes. This yielded in a relatively low precision of 3 out of 16 samples. However, considering that there were 70 possible cell states in the data set, this prediction performance is considerably better than random. We have also tested using gene-sets such as genes belonging to GO terms or complexes in CORUM database for their capacity to distinguish the cell state. Using the genes associated with certain GO terms improved the number of correctly predicted samples substantially for various GO terms (see Supplementary Material).

Table 2. 10-fold cross validation using up-regulated k transcription factors.

Cross validation run id	Average # of correctly predicted arrays averaged among folds	k
1	8.0	59
2	8.5	31
3	8.0	31

Methods

Human developmental lineage tree

Possible paths of human stem cell differentiation have been captured as an ontology we curated combining cell fate maps from Gilbert (2000) and Galvao et al. (2010). In case of inconsistency or lack of information in these two sources, we looked for further evidence from literature. We provided the ontology in Open Biological and Biomedical Ontology (OBO) format (see Supplementary Material). The ontology contained 70 terms, each of which corresponded to a distinct cell state. These terms were connected by “lineage of” relationships.

Gene expression and transcription factor data

The gene expression data was kindly provided by Jeanne Loring's Regenerative Medicine group in Scripps Institute through a collaboration. They have screened expression of genes on 709 samples from

human corresponding to 66 different cell types at various levels of human development using Illimuna BeadArray platform (31007 probes covering 23659 genes). The samples and the cell types they were assigned is given in the Supplementary Material. They provided bead-level expression information for the probes that passed the detection p-value cutoff (less than or equal to 0.05). This data was then quantile normalized and converted to log-ratios. Among 709 samples, 12 samples were tumor-like samples and thus discarded during the analyzes. In addition to these samples, a validation data set consisting of 18 samples created with using another Illimuna-based platform were provided independently (see Supplementary Material). These latter samples contained 28113 probes covering 17858 genes.

For the cell states for which no experimental data available, we used the samples falling under these states to describe the gene expression patterns in these cell states. For each of such states, we considered the expression data assigned to their children states in the lineage tree as a proxy.

During analyzes, we used the list of human transcription factors provided by Ravasi et al. [Ravasi10]. We also added DNMT3B and FGF5 to this dataset for which the transcriptional activity was already demonstrated. The complete list of transcription factors consisted of 1723 genes (see Supplementary Material).

Determining differential gene expression patterns

We used several statistical tests to compare gene expression levels of each gene under various parts of the created lineage tree based on the hypothesis that genes deriving a cell type to another should bear expression patterns characteristic to these cell types. These statistical tests include Kolmogorov-Smirnov divergence, Welch's test (a modification of t-test accounting for distributions with unequal variances), Jensen-Shannon divergence as well as simply amount of absolute expression and amount of change in the mean expression level (referred as abundance test and abundance difference test respectively). We have also checked 3 different strategies to group the samples over the tree in order to identify the most plausible way of capturing the differentiation process through the statistical test where the first group being the samples corresponding to the cell type for which a gene in concern was mediating differentiation and the second being other cell types assumed to be involved in the determination of the cell fate for the cell type in concern. The first strategy, *versus parent and siblings* compared the expression of a gene *g* given in a cell type *c*, with respect to expression of *g* in the cell types corresponding to *c*'s parent and sibling nodes in the lineage tree. Similarly, the second strategy, *versus siblings*, compared the expression of g in c with respect to siblings of c. The final strategy, *versus rest*, compared the expression of g in c with the expression of g in the rest of the samples.

We assessed the prediction performance of these methods on a small set of genes known to be involved in the differentiation of embryonic stem and the germ layer cells. This gold standard dataset was mainly curated from the literature (Takahashi et al., 2006; Jaenisch and Young, 2008; Lemischka et al., 2009). The genes included from literature were further curated by experts by the group that provided the expression data. All the genes in the gold standard that were annotated with a cell type other than the cell type in concern were considered as negative cases (genes that were not involved in that cell type). The prediction performance was evaluated using area under ROC curve (AUC).

Predicting mediators of cell fate

A centroid-based clustering approach was taken to classify a given sample based on existing samples. Instead of using all possible genes, we compared only the expression of the most differentially up-regulated k genes in each cell state with the expression of these genes in the sample of interest. A centroid vector for each state was generated using average expression of top differentially up-regulated k genes in all samples for that state. For each state, the similarity between the centroid vector and the expression of the same genes included in the centroid vector for the given sample was calculated using a correlation-based distance metric. The array was then predicted as the cell state with the closest distance. To find the optimum number of most differentially up-regulated genes, k, we employed a ten-fold

cross-validation procedure and split the samples in the gene expression dataset for which we had more than 10 samples into ten groups. There were 24 cell states in the dataset for which there were more than 10 samples. A cross validation group was created using randomly selecting one sample among all of the samples associated with each of these states and this procedure was repeated 10 times (without replacing the selected samples). We used 9 of these groups to identify differentially expressed genes for each cell state and then predicted the cell state of the left-out samples (10th group) for varying values of k in the range of 1 to 80. In addition to using k up-regulated genes, we used the genes involved in known complexes in CORUM database as well as genes associated with GO terms. In this gene-set oriented approach, the average expression of the genes of the complex/GO term in the samples of each cell state was compared with the expression of the same genes in the sample.

Acknowledgments

This work was supported by Generalitat de Catalunya – AGAUR through a fellowship (2009 BE2 00115) granted to E.G. for a research stay in Trey Ideker's group in University California, San Diego.

References

1. MacArthur BD, Ma'ayan A, Lemischka IR (2009) Systems biology of stem cell fate and cellular reprogramming. Nature Reviews Molecular Cell Biology 10: 672–681. doi:10.1038/nrm2766.

2. Jaenisch R, Young R (2008) Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. Cell 132: 567–582. doi:10.1016/j.cell.2008.01.015.

3. Sareen D, Svendsen CN (2010) Stem cell biologists sure play a mean pinball. Nature biotechnology 28: 333.

4. Muller F-J, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455: 401–405. doi:10.1038/nature07213.

5. Aiba K, Nedorezov T, Piao Y, Nishiyama A, Matoba R, et al. (2009) Defining Developmental Potency and Cell Lineage Trajectories by Expression Profiling of Differentiating Mouse Embryonic Stem Cells. DNA Res 16: 73–80. doi:10.1093/dnares/dsn035.

6. Chuang HYL (2007) Network-based classification of breast cancer metastasis. Molecular Systems Biology 3.

7. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, et al. (2010) An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell 140: 744–752. doi:10.1016/j.cell.2010.01.044.

8. Ashburner MB (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

9. Hill DP, Berardini TZ, Howe DG, Van Auken KM (2010) Representing ontogeny through ontology: A developmental biologist's guide to the gene ontology. Molecular Reproduction and Development 77: 314–329. doi:10.1002/mrd.21130.

10. Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome biology 6: R21.

11. Bard J (2011) A systems biology representation of developmental anatomy. Journal of Anatomy 218: 591–599. doi:10.1111/j.1469-7580.2011.01371.x.

12. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. Data Integration in the Life Sciences. pp. 124–131.

13. Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, et al. (2010) A global map of human gene expression. Nature Biotechnology 28: 322–324. doi:10.1038/nbt0410-322.

14. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, et al. (2011) Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell 144: 296–309. doi:10.1016/j.cell.2011.01.004.

Bibliography

[Aerts et al., 2006] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat Biotech*, 24(5):537–544.

[Agrawal, 2001] Agrawal, A. (2001). Phenotypic plasticity in the interactions and evolution of species. *Science*, 294(5541):321–6.

[Akula et al., 2011] Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M., Singleton, A., Ferrucci, L., Tanaka, T., Bandinelli, S., Cho, Y., et al. (2011). A Network-Based approach to prioritize results from Genome-Wide association studies. *PloS one*, 6(9):e24220.

[Ala et al., 2008] Ala, U., Piro, R., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., and Di Cunto, F. (2008). Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol*, 4(3):e1000043.

[Albert et al., 2000] Albert, Jeong, and Barabasi (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382. PMID: 10935628.

[Altshuler et al., 2008] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

[Amberger et al., 2009] Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online mendelian inheritance in man (OMIM). *Nucleic Acids Research*, 37(Database issue):D793–D796. PMID: 18842627 PMCID: PMC2686440.

[Aragues et al., 2006] Aragues, R., Jaeggi, D., and Oliva, B. (2006). PIANA: protein interactions and network analysis. *Bioinformatics*, 22(8):1015—1017.

[Aragues et al., 2008] Aragues, R., Sander, C., and Oliva, B. (2008). Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, 9:172.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[Barabasi and Albert, 1999] Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Barabasi et al., 2011] Barabasi, A., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68.

[Barabasi and Oltvai, 2004] Barabasi, A. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

[Bauer-Mehren et al., 2010] Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze Gene-Disease networks. *Bioinformatics*, 26(22):2924–2926.

[Becker et al., 2004] Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nature Genetics*, 36(5):431–432.

[Ben-Porath et al., 2008] Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., and Weinberg, R. A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nature Genetics*, 40(5):499–507.

[Broeckel and Schork, 2004] Broeckel, U. and Schork, N. J. (2004). Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping. *The Journal of Physiology*, 554(1):40–45.

[Capriotti et al., 2012] Capriotti, E., Nehrt, N. L., Kann, M. G., and Bromberg, Y. (2012). Bioinformatics for personal genome interpretation. *Briefings in Bioinformatics*.

[Carlson et al., 2004] Carlson, C., Eberle, M., Kruglyak, L., and Nickerson, D. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452.

[Carlson and Doyle, 2002] Carlson, J. and Doyle, J. (2002). Complexity and robustness. *Proc Natl Acad Sci U S A*, 99 Suppl 1:2538–45.

[Caspi et al., 2011] Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P., and Karp, P. D. (2011). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753.

[Chen et al., 2009] Chen, J., Aronow, B., and Jegga, A. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, 10(1):73.

[Chipman and Singh, 2009] Chipman, K. C. and Singh, A. K. (2009). Predicting genetic interactions with random walks on biological networks. *BMC bioinformatics*, 10(1):17.

[Chuang et al., 2007] Chuang, H., Lee, E., Liu, Y., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1).

[Clarke et al., 2011] Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133. PMID: 21293453 PMCID: PMC3154648.

[Croft et al., 2011] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl 1):D691.

[Davis et al., 2010] Davis, A. P., King, B. L., Mockus, S., Murphy, C. G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., and Mattingly, C. J. (2010). The comparative toxicogenomics database: update 2011. *Nucleic Acids Research*, 39(Database):D1067–D1072.

[Dezso et al., 2009] Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C., and Bugrim, A. (2009). Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol*, 3:36.

[Dowell et al., 2010] Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D. A., Rolfe, P. A., Heisler, L. E., Chin, B., Nislow, C., Giaever, G., Phillips, P. C., Fink, G. R., Gifford, D. K., and Boone, C. (2010). Genotype to phenotype: A complex problem. *Science*, 328(5977):469–469.

[Eichler et al., 2007] Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., Church, D. M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J. R., Mullikin, J. C., Pritchard, J. K., Sebat, J., Sherry, S. T., Smith, D., Valle, D., and Waterston, R. H. (2007). Completing the map of human genetic variation. *Nature*, 447(7141):161–165.

[Ergun et al., 2007] Ergun, A., Lawrence, C. A., Kohanski, M. A., Brennan, T. A., and Collins, J. J. (2007). A network biology approach to prostate cancer. *Molecular Systems Biology*, 3(1).

[Erten et al., 2011] Erten, S., Bebek, G., Ewing, R., Koyuturk, M., et al. (2011). DADA: Degree-Aware algorithms for Network-Based disease gene prioritization. *BioData mining*, 4(1):19.

[Feuk et al., 2006] Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97.

[Fliri et al., 2010] Fliri, A. F., Loging, W. T., and Volkmann, R. A. (2010). Cause-effect relationships in medicine: a protein network perspective. *Trends in Pharmacological Sciences*, 31(11):547–555.

[Franke et al., 2006] Franke, L., van Bakel, H., Fokkens, L., de Jong, E., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene

network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011—1025.

[Frazer et al., 2009] Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251.

[Freedman et al., 2011] Freedman, M. L., Monteiro, A. N. A., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., Casey, G., Biasi, M. D., Carlson, C., Duggan, D., James, M., Liu, P., Tichelaar, J. W., Vikis, H. G., You, M., and Mills, I. G. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, 43(6):513–518.

[Gandhi et al., 2006] Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293.

[Garcia-Garcia et al., 2012] Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J., and Oliva, B. (2012). Networks of Protein-Protein interactions: From uncertainty to molecular details. *Molecular Informatics*, 31(5):342—362.

[Garcia-Garcia et al., 2010] Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J., and Oliva, B. (2010). Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, 11(1):56.

[Goh et al., 2007] Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabasi, A. (2007). The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685.

[Gottlieb et al., 2011] Gottlieb, A., Magger, O., Berman, I., Ruppin, E., and Sharan, R. (2011). PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics*, 27(23):3325–3326.

[Groth et al., 2007] Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H., and Weiss, B. (2007). PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic acids research*, 35(Database issue):D696–699. PMID: 16982638.

[Guldener et al., 2006] Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H., and Stumpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436–41.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

[Hardy and Singleton, 2009] Hardy, J. and Singleton, A. (2009). Genomewide association studies and human disease. *New England Journal of Medicine*, 360(17):1759–1768.

[Hartwell et al., 1999] Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52.

[Hernandez-Boussard et al., 2007] Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J. M., Gong, L., Owen, R., Gong, M., Gor, W., Liu, F., Truong, C., Whaley, R., Woon, M., Zhou, T., Altman, R. B., and Klein, T. E. (2007). The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Research*, 36(Database):D913–D918.

[Hirschhorn and Daly, 2005] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108.

[Hirschhorn et al., 2002] Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, 4(2):45–61.

[Ideker and Sharan, 2008] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res*, 18(4):644—652.

[Isserlin et al., 2011] Isserlin, R., El-Badrawi, R. A., and Bader, G. D. (2011). The biomolecular interaction network database in PSI-MI 2.5. *Database: The Journal of Biological Databases and Curation*, 2011. PMID: 21233089 PMCID: 3021793.

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

[Jr and Scherer, 2008] Jr, E. H. C. and Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(7215):919–923.

[Kanehisa et al., 2012] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109—D114.

[Kann, 2010] Kann, M. G. (2010). Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, 11(1):96–110.

[Karchin, 2009] Karchin, R. (2009). Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics*, 10(1):35–52.

[Kerrien et al., 2011] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2011). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846.

[Keshava Prasad et al., 2009] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman,

B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772.

[Kingsmore et al., 2008] Kingsmore, S. F., Lindquist, I. E., Mudge, J., Gessler, D. D., and Beavis, W. D. (2008). Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews Drug Discovery*, 7(3):221–230.

[Kitano, 2004] Kitano, H. (2004). Biological robustness. *Nat Rev Genet*, 5(11):826–37.

[Kohler et al., 2006] Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383—1390.

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

[Lage et al., 2007] Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tumer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotech*, 25(3):309–316.

[Laubenbacher et al., 2009] Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. V., Shulaev, V., Mendes, P., Torti, F. M., and Akman, S. (2009). A systems biology view of cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1796(2):129–139.

[Lee et al., 2011] Lee, I., Blom, U., Wang, P., Shim, J., and Marcotte, E. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*, page advance online article.

[Liberzon et al., 2011] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740.

[Licata et al., 2011] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. (2011). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861.

[Lindvall et al., 2012] Lindvall, O., Barker, R. A., Brustle, O., Isacson, O., and Svendsen, C. N. (2012). Clinical translation of stem cells in neurodegenerative disorders. *Cell Stem Cell*, 10(2):151–155.

[Linghu et al., 2009] Linghu, B., Snitkin, E., Hu, Z., Xia, Y., and Delisi, C. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol*, 10(9):R91.

[Ma et al., 2007] Ma, X., Lee, H., Wang, L., and Sun, F. (2007). CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, 23(2):215–221.

[MacArthur et al., 2009] MacArthur, B. D., Ma'ayan, A., and Lemischka, I. R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10):672–681.

[Mani et al., 2008] Mani, K. M., Lefebvre, C., Wang, K., Lim, W. K., Basso, K., Dalla-Favera, R., and Califano, A. (2008). A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Molecular Systems Biology*, 4(1).

[McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

[Mewes et al., 2010] Mewes, H. W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K. F. X., Stumpflen, V., and Antonov, A. (2010). MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Research*, 39(Database):D220–D224.

[Milenkovic et al., 2010] Milenkovic, T., Memisevic, V., Ganesan, A. K., and Przulj, N. (2010). Systems-Level cancer gene identification from protein interaction network topology applied to Melanogenesis-Related functional genomics data. *Journal of The Royal Society Interface*, 7(44):423–437.

[Morohashi et al., 2002] Morohashi, M., Winn, A., Borisuk, M., Bolouri, H., Doyle, J., and Kitano, H. (2002). Robustness as a measure of plausibility in models of biochemical networks. *Journal of theoretical biology*, 216(1):19—30.

[Mottaz et al., 2010] Mottaz, A., David, F. P. A., Veuthey, A., and Yip, Y. L. (2010). Easy retrieval of single Amino-Acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, 26(6):851–852.

[Nabieva et al., 2005] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl 1):i302–i310.

[Navlakha and Kingsford, 2010] Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063.

[Nibbe et al., 2010a] Nibbe, R. K., Chowdhury, S. A., Koyuturk, M., Ewing, R., and Chance, M. R. (2010a). Protein-protein interaction networks and subnetworks in the biology of disease. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(3):357–367.

[Nibbe et al., 2010b] Nibbe, R. K., Koyuturk, M., and Chance, M. R. (2010b). An integrative -omics approach to identify functional Sub-Networks in human colorectal cancer. *PLoS Comput Biol*, 6(1):e1000639.

[Nitsch et al., 2010] Nitsch, D., Goncalves, J. P., Ojeda, F., de Moor, B., and Moreau, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460.

[Oti et al., 2006] Oti, M., Snel, B., Huynen, M., and Brunner, H. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691—698.

[Park et al., 2008] Park, I., Arora, N., Huo, H., Maherali, N., Ahfeldt, T., Shimamura, A., Lensch, M. W., Cowan, C., Hochedlinger, K., and Daley, G. Q. (2008). Disease-Specific induced pluripotent stem cells. *Cell*, 134(5):877–886.

[Pujana et al., 2007] Pujana, M., Han, J., Starita, L., Stevens, K., Tewari, M., Ahn, J., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W., Rual, J., Levine, D., Rozek, L., Gelman, R., Gunsalus, K., Greenberg, R., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernandez, P., Lazaro, C., Nathanson, K., Weber, B., Cusick, M., Hill, D., Offit, K., Livingston, D., Gruber, S., Parvin, J., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*, 39(11):1338—1349.

[Pujol et al., 2010] Pujol, A., Mosca, R., Farres, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences*, 31(3):115—123.

[Qiu et al., 2010] Qiu, Y. Q., Zhang, S., Zhang, X. S., and Chen, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC bioinformatics*, 11(1):26.

[Reya et al., 2001] Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature*, 414(6859):105–111.

[Safran et al., 2010] Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards version 3: the human gene integrator. *Database: The Journal of Biological Databases and Curation*, 2010. PMID: 20689021 PMCID: PMC2938269.

[Salomonis et al., 2007] Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., Doniger, S. W., Stuart, J., Conklin, B. R., and Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8(1):217.

[Salwinski et al., 2004] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451.

[Sam et al., 2009] Sam, L. T., Mendonca, E. A., Li, J., Blake, J., Friedman, C., and Lussier, Y. A. (2009). PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. *BMC bioinformatics*, 10 Suppl 2:S8. PMID: 19208196.

[Shastry, 2002] Shastry, B. S. (2002). SNP alleles in human disease and evolution. *Journal of Human Genetics*, 47(11):0561–0566.

[Stankiewicz and Lupski, 2010] Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455.

[Stark et al., 2010] Stark, C., Breitkreutz, B., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. (2010). The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(Database):D698–D704.

[Stenson et al., 2009] Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., and Cooper, D. N. (2009). The human gene mutation database: 2008 update. *Genome medicine*, 1(1):13. PMID: 19348700.

[Tranchevent et al., 2011] Tranchevent, L., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32.

[Vanunu et al., 2010] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641.

[Vidal et al., 2011] Vidal, M., Cusick, M., and Barabasi, A. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.

[Vogelstein and Kinzler, 2004] Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799.

[Wang et al., 2005] Wang, W., Barratt, B., Clayton, D., and Todd, J. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6(2):109–18.

[Wu et al., 2008] Wu, X., Jiang, R., Zhang, M., and Li, S. (2008). Network-based global inference of human disease genes. *Mol Syst Biol*, 4:189.

[Xu and Li, 2006] Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800—2805.

[Young, 2011] Young, R. (2011). Control of the embryonic stem cell state. *Cell*, 144(6):940–954.

[Zanzoni et al., 2009] Zanzoni, A., Soler-Lopez, M., and Aloy, P. (2009). A network medicine approach to human disease. *FEBS Letters*, 583(11):1759–1765.