Discovering molecular mechanisms of human disease through gene sets and networks

A dissertation

submitted by

Jisoo Park, BS, EWU; MS, WUSTL

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

May 2017

ADVISOR: Prof. Donna K. Slonim

Dedicated to the memories of my grandfather, Joon-han Park

Acknowledgments

This dissertation describes my work that would have never been possible without advice, contributions, collaborations, suggestions and support from many people in my life.

First, I owe my deepest gratitude to my thesis advisor Professor Donna Slonim for several years of guidance, which has been essential to my achievement. All the insightful advice she provided in the course of my doctoral work, combined with her encouragement and motivation, have greatly contributed to my development as a researcher. Her patience when faced with my countless raw ideas, and her intellectual ability to suggest ways to mature these ideas have always been inspirational. Her mentorship on teaching has helped realize my potential for being a good teacher. Her exemplary writing practice has contributed to my growth as a writer, and her encouragement has motivated me to improve my English as a second language. It is no exaggeration to say that none of my achievements over the past seven years would have been possible without her help and advice.

I thank my committee members, Professor Lenore Cowen, Professor Benjamin Hescott, Professor Kyongbum Lee and Dr. Teresa Przytycka, for their invaluable feedback and suggestions on this dissertation. A special thanks is extended to Professor Lenore Cowen, Professor Carla Brodley and Professor Robert Jacob for their kind and thoughtful advice throughout my Ph.D. years and their service as my qualifying exam committee members. I also thank Professor Norman Ramsey for his writing class, which initiated significant improvement in my writing.

My collaborators and colleagues have greatly contributed to my doctoral work. I thank all previous and current members of my research group including Noah Daniels, Mengfei Cao, Xian Feng, Andrew Gallant, Hao Zhang, Inbar Fried,

Chris Pietras, Jake Crawford, Rebecca Newman and Anselm Blumer, for their invaluable comments and suggestions throughout countless discussions and practice talks. I extend my appreciation to my co-authors, Heather Wick, Daniel Kee, Keith Noto and Jill Maron, for their contribution to the work published in the journal *PLOS Computation Biology*, and to fellow graduate students at Tufts University, Anzu Hakone, Xinmeng Li, Sara Amir, Diogenes Nunez, Mike Shah and Jason Wilson, for the constructive comments they provided throughout practice talks for the dissertation defense.

I cannot be more grateful for meeting great friends at Tufts University: Alvitta Ottley, Jordan Crouser, Nathan Ricci, Bradford Larsen, Evan Peck, Eli Brown and Jingjing Liu. I thank them for their friendship through all the moments we have shared, as it has been essential to my survival of the life as a graduate student. I also thank Professor Remco Chang for providing socializing opportunities through a series of inclusive events in his research lab.

Finally, I would have never been able to reach this point without the support of my husband, Collin Hong. I also would like to acknowledge the love and help of my family. I dedicate this dissertation to them.

Jisoo Park

TUFTS UNIVERSITY

May 2017

Discovering molecular mechanisms of human disease through gene sets and networks

Jisoo Park

ADVISOR: Prof. Donna K. Slonim

Molecular processes, such as genetic mutations and interactions between gene products, play a key role in the development of disease. Yet, there is still not enough known about molecular causes of disease. In this thesis, we introduce methods to discover hidden connections between biological functions and disease using gene sets and networks.

We first investigate a topological property of pathways in protein-protein interaction networks. We define a new measurement called *pathway centrality* which measures the amount of information flow between disease genes and differentially expressed genes handled by a pathway. We find mediating pathways for three pulmonary diseases (asthma; bronchopulmonary dysplasia (BPD); and chronic obstructive pulmonary disease (COPD)) using pathway centrality. Mediating pathways shared by all three pulmonary disorders are mostly related to inflammation or immune responses and include specific pathways such as cytokine production, NF Kappa B signaling, and JAK/STAT signaling. We confirm our findings, which suggest new treatment approaches, both with anecdotal evidence from the literature and via systematic evaluation using genetic interactions.

Second, we identify connections between developmental processes and disease by statistically testing overlaps between developmental gene sets and disease gene sets. To handle missing disease-gene association information, we pool disease

genes from specific disease terms to more general disease terms in a disease taxonomy. Our overlap analysis results for nine developmental gene sets confirms many expected connections, such as those between cardiovascular disorders and heart development genes. Closer investigation of our results highlights some unexpected connections, such as ones between bone development and dementia, heart development and polycystic ovary syndrome, and lung development and retinopathy of prematurity. These connections have been further supported by recent publications and again suggest novel therapeutic strategies.

While successful, this work highlights a need for more molecular disease taxonomies to improve the efficacy of gene pooling. We ran a pilot study to infer disease hierarchies *only* using disease-gene association information. We evaluate our inferred disease hierarchies by comparing to existing ones because there is no gold standard molecular disease taxonomy available. We find that our inference algorithm is able to recover much of the structure of existing disease taxonomies. While inference is easier for smaller sets of disease terms, we found some large disease categories where inference methods perform well, such as Endocrine System Diseases, Nutritional and Metabolic Diseases, and Respiratory Tract Diseases. We suspect that the existing hierarchies representing these disease categories incorporate larger amounts of molecular data, perhaps because they include several well-studied complex diseases.

Overall, we have introduced new computational methods that highlight novel connections between gene sets and diseases. We expect that our studies will lead to deeper understanding of underlying mechanisms of human disease, and ultimately to better support of molecular medicine.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Development of therapeutic solutions to human disease has been a primary aim of much interdisciplinary research. The crucial first step towards this goal is to identify underlying molecular mechanisms of human disease, as doing so helps identify potential new drug targets or therapeutic approaches. Our studies strive to achieve this same goal. This dissertation will introduce our computational approaches to discover new connections between gene functions and human disease and ultimately to better support molecular medicine.

The advent of high throughput molecular biology has brought an influx of data to the field of computational biology and bioinformatics. For example, new next-generation sequencing (NGS) techniques and decreasing costs have led to a vast amount of sequence data; a single sequencing run can produce several terabases of data. This growth in the amount of data has raised the need for innovation in at least two areas: algorithms for "big data" and data representations. Our focus is on utilizing data representations helpful for inferring underlying biological functions of human disease, and two data representations were extensively used in our studies: networks and sets. This chapter provides details of these two main components and introduces general background and terminology used in later chapters.

1.1 Network representation of biological data

Networks are the best suited data representation for complex structures in which the volume of interactions between participating entities is large. There are three major advantages of using networks to represent data. First, networks effectively represent a wide range of relationships between entities, from direct ones such as pairwise relationships (i.e., edges) to indirect ones such as group memberships (i.e., cliques or connected components). Second, a network representation of data enables analysis using various graph theoretic algorithms to learn new information. Third, networks allow the visualization of a large volume of data to facilitate exploration of the complex structure of the data. Until the early 1990s, network representations were more common in specific fields of study, such as the social sciences [JS11]. Many studies in social science have used networks to model complex interactions between people or groups of people. Network analysis has led to the inference of new knowledge to support various activities, including customer relationship management (CRM) or marketing. However, the scope of network analysis has been expanded to various fields of study including computational biology, prompted by the need for new analytical tools for a large volume of molecular data.

Mathematically, networks are represented as graphs. Here we provide basic mathematical notations related to graphs that underlie the notation used in later chapters. A simple graph, $G = (V, E)$, is a tuple containing a set of nodes (vertices), V, and a set of edges (links), E. A set is defined to be a collection of distinct elements. The vertices, $V = \{v_1, v_2, v_3, ...\}$, represent entities such as proteins and the edges, $E = \{(v_1, v_2), (v_1, v_3), ...\}$, represent relationships between pairs of entities. Graph properties can be expanded by adding more components to the tuple, such as with weighted directed graphs. We call a graph "weighted" when nodes or edges in the graph are labeled with associated values. A weighted graph will be defined as the triple, $G = (V, E, W)$, where W is a function that assigns some value to either nodes or edges. There can be multiple weight functions in case both nodes and edges are weighted. In directed graphs, the edges have an assigned direction going

from one node to another, maybe characterizing a directional relationship between nodes. The notation for a directed graph is identical to that for a simple graph, but the orders of nodes in an edge notation indicate the directionality of the edge. That is, (v_1, v_2) and (v_2, v_1) are two different edges with reverse directionality in a directed graph although they are same in a simple graph.

The strength of our methods for finding novel connections between human disease and biological functions comes from exploiting information about relationships between biological entities defined in different types of biological networks. Examples of biological networks include regulatory networks, where nodes represent potential transcription factors or their regulated genes and directed edges represent one gene affecting the regulation of another [JS11]. Metabolic networks are comprised of cell compounds connected by biochemical reactions, they are used to model the hierarchical organization of metabolic regulations [SKB+02]. Phylogenetic networks model the evolutionary relationships between different organisms [JS11]. Our studies extensively use both protein-protein interaction networks, and networks characterizing relationships between diseases.

1.1.1 Protein-protein interaction (PPI) networks

Proteins are basic molecular units which execute a wide range of cellular functions. Proteins are encoded by genes, and therefore we sometimes use the terms "protein" and "gene" interchangeably in this dissertation, although there is not necessarily a one-to-one relationship. Proteins interact with other proteins to execute cellular functions. The web of such known interactions is called the protein-protein interaction (PPI) network. In the PPI network, nodes are proteins and edges are known interactions between proteins. The importance of PPI networks has been demonstrated in recent studies aiming at finding the underlying biology of human disease. The complex nature of human disease cannot be explained just by individual genes, as the number is comparable to that of simpler model organisms such as fruit flies [GBK+02]. Therefore, researchers suspect that the complexity of human disease is derived from interactions between gene products. We therefore study the PPI

network to infer new knowledge about human disease.

We note that we distinguish protein-protein interactions from protein-protein associations, although these terms are occasionally used equivalently. Protein-protein associations include indirect interactions between two proteins [FSF+12] defined by shared functions or properties such as co-expression of coding genes and sequence homology. We exclude these associations from our experiments to avoid study bias and potential false positives, with the exception of alleviating genetic interactions used to evaluate our algorithm for identification of disease mediating pathways, as discussed in chapter 2.

Protein-protein interactions used in our studies are physical PPIs. To reduce noise in the PPI data, we only use experimentally verified interactions with high confidence scores in our experiments. Descriptions of both interactions and experimental processes for their collection appear below.

1.1.1.1 Physical interaction

Physical interactions of proteins characterize physical contacts between them. That is, physical interactions are defined between two proteins when they bind to each other biochemically within a cell. The majority of pairwise protein interactions in databases are mostly identified by two types of high-throughput experimental assays: the yeast two-hybrid system [FS89], and affinity capture followed by mass spectrometric analysis [PM00].

Yeast two-hybrid (Y2H) uses transcriptional activation to identify protein interactions. Two proteins of interest are fused to a DNA-binding domain and a transcriptional activation domain. If these two fusion proteins (bait and prey) interact with each other, the transcription of a reporter gene will be activated. More specifically, expression of the reporter gene confirms a bait-prey interaction between the pair of tested proteins. Figure 1.1 illustrates how Y2H identifies a pairwise protein interaction.

Affinity Purification (AP) followed by **Mass Spectrometry (MS)** identifies sets of proteins interacting with each other also using a fusion protein, again

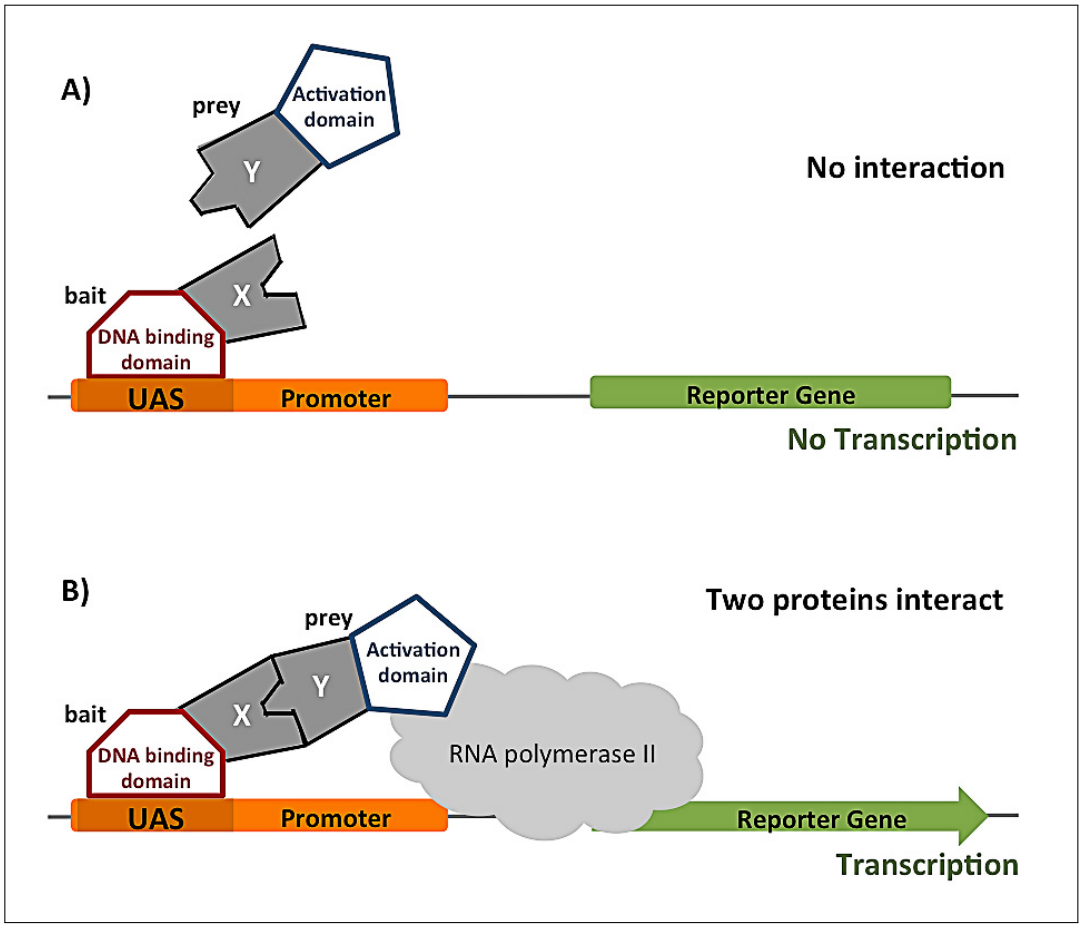

Figure 1.1: How yeast two-hybrid screening works to identify protein-protein interactions. A) If bait and prey do not interact, no transcription happens. B) If two fusion proteins (bait and prey) interact, the transcription for the reporter gene will be activated. Expression of the reporter gene will confirm the pairwise protein-protein interaction. The figure was adapted from [BPL+09].

called the "bait." The bait is inserted into a protein mixture, and later washed out, such that only proteins interacting with the bait will remain. Mass spectrometry on the remaining protein complex reveals identification of individual proteins. This approach specifically finds complexes, using one bait protein. Pairwise PPI's inferred from this may be all pairs (forming a complete graph or clique), or a hub-and spoke representation linking the bait protein with all others in the complex.

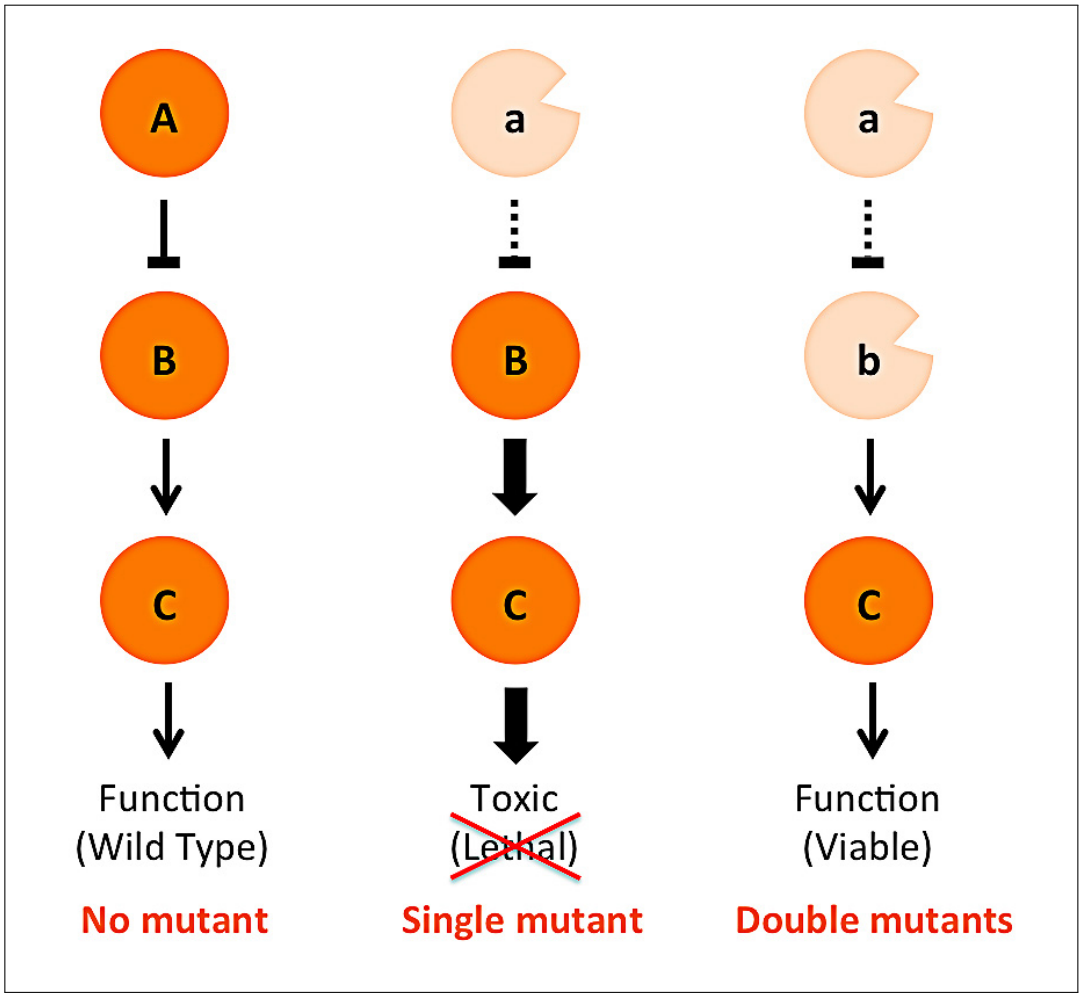

Figure 1.2: Positive (alleviating) genetic interactions. Gene A negatively regulates the pathway involving B and C, and C is a toxic gene product. Mutation in A results in hyperactivation of the pathway, and eventually causes excessive accumulation of C. However, if B is mutated subsequently, the flux through the pathway decreases and accumulation of toxic C is suppressed. That is, the outcome of single mutant (a) is worse than the outcome of double mutants (a,b). The figure was adapted from [DCB+09].

1.1.1.2 Genetic interaction

Genetic interactions are defined between two genes when variants of the genes lead to unexpected outcomes (phenotypes or diseases) when combined. Therefore, at a high level, we detect genetic interactions by comparing experimentally observed phenotypes in double mutants to the expected phenotypes from the combination of single-mutants, assuming independence of mutations [BCM+13]. Experimental

detection of genetic interactions typically requires a gene knockout or other modifications, therefore genetic interactions are identified most often in model organisms such as yeast and fruit fly.

Genetic interactions are distinct from physical interactions as they are indirect functional interactions between two proteins. Two proteins that have a genetic interaction may or may not physically interact. Genetic interactions most broadly can fall into two major classes: positive and negative genetic interactions [BCM+13]. Positive genetic interactions occur when the phenotype of double mutants is less severe than the combination of single mutants (illustrated in Figure 1.2). In the most extreme case of positive genetic interaction, a non-viable phenotype of single mutants is rescued by mutations of another gene, and we call this synthetic rescue. Negative genetic interactions are defined when the phenotype of double mutants is more severe than expected from the single mutants. The most extreme example of a negative genetic interaction is synthetic lethality, which describes double mutants resulting in cell death while the phenotypes of the single mutants cause little or no noticeable growth defects. In our studies, we use positive genetic interactions to evaluate our algorithm for identifying mediating pathways for a disease (described in Chapter 2), because positive genetic interactions describe relationships between two gene products where one may be a downstream gene of another. That is, if a pair of two proteins, (A, B), is known to have a positive genetic interaction, it is likely that B is a downstream gene of A along some functional pathway.

1.1.2 Networks characterizing relationships between diseases: disease ontologies and taxonomies

Disease *ontologies* both catalog disease terms and explicitly define relationships between the terms. The structure of a disease ontology is commonly a directed acyclic graph. Between two disease terms in a connected pair, the more specific term has an outgoing edge and the more general term has an incoming edge. The primary purpose of disease *taxonomies* is to catalog all the terms describing disease. However, in some cases the terms are organized hierarchically, in which case, specific-general

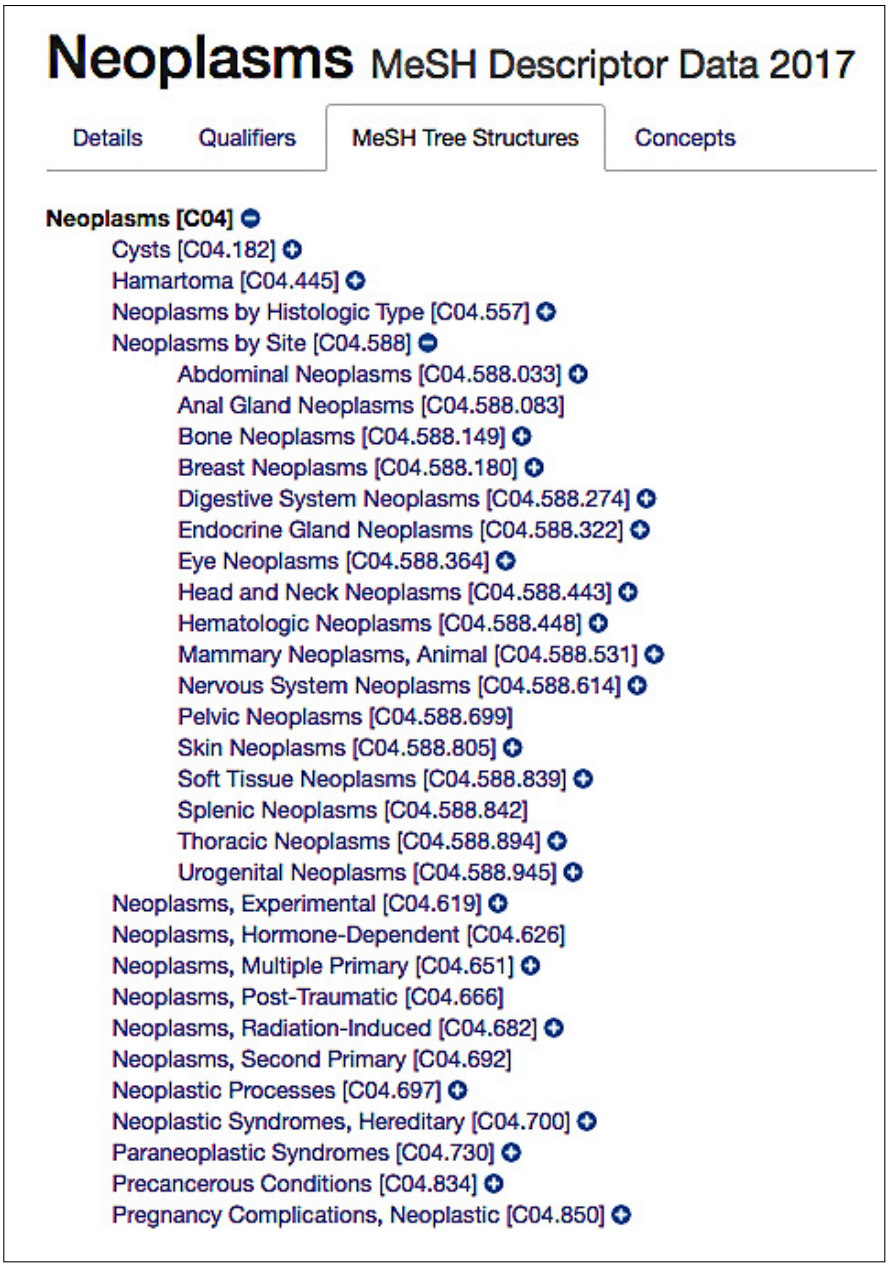

Figure 1.3: A MeSH tree organizing the terms related to "cancer." For the sake of space, only one branch (Neoplasms by Site) is expanded in the figure.

relationships between diseases can be inferred.

Regardless of structure, however, both disease ontologies and taxonomies describe disease terms (represented by nodes) and can define hierarchical relationships between diseases (represented by edges). For this reason, we introduce these hierarchical structures as network representations of diseases. Figure 1.3 shows an example

of a disease taxonomy, the Medical Subject Headings (MeSH) tree, containing the terms describing "neoplasms."

1.2 Gene sets represent biological functions

Genes work in a complex way in biological systems. In this thesis, "gene functions" refer to biological pathways interrupted by mutations of the corresponding genes, or to a biological functions known to be executed by their protein products. The process of mapping a gene to its function(s) is called "functional annotation." There is rarely a one-to-one correspondence between genes and gene functions. Multiple genes work in concert to perform a given biological function, and a gene may be involved in multiple functions. Conversely, biological processes work together, affecting multiple genes. Therefore, it is more meaningful to analyze a *set* of genes as a functional unit to find relationships between gene functions and human disease. Such an approach is broadly called gene set analysis.

The efficacy of gene set analysis is well justified. Gene Set Enrichment Analysis (GSEA) [STM+05a] was one of the first such approaches for expression analysis, a comparative approach to identifying genes whose expression levels are significantly different between control and phenotypic groups. GSEA tests enrichment of a gene function represented by a *set* of genes, by checking concentration of set members at both extremes of a list of genes sorted by t-score or other measures of differential expression.

Gene set analysis helps overcome shortcomings of performing enrichment on the results of single-gene analysis. Specifically, single-gene analysis may miss important effects of a gene, as many diseases are governed by modest regulation of a set of genes rather than a notable change of a single gene. In addition, single gene analysis has a caveat that the results are dependent on a threshold. Therefore, after correction for multiple testing, we may end up with no individual genes with statistical significance when each individual gene's effect is small compared to the noise induced by microarray experiments. Furthermore, gene set analysis has been

used in several genome-wide association studies (GWASs) analyzing multiple single nucleotide polymorphisms (SNPs) in genes grouped together based on shared function [FB11]. In this context, gene set analysis may help detect the combined small effect of multiple SNPs.

In dealing with functional annotation of genes, the importance of having a controlled vocabulary cannot be overlooked. Biological data is of large volume and often very noisy. One of the main contributors to such noisiness is the lack of unified, controlled vocabularies. That is, the same biological phenomena can be described in a hundred different ways, depending on the authors. Many controlled vocabulary systems for biological functions have been published; we utilize three of them in our studies. The Gene Ontology (GO) [Con14, ABB$^+$00] is the most commonly used vocabulary system for gene functions. It defines the functions and cellular locations of genes and gene products, and relationships between the terms. The Kyoto Encyclopedia for Genes and Genomes (KEGG) [KSK$^+$16] is a collection of pathways in which individual protein and interactions between them are identified. It also serves as a controlled vocabulary at a higher level because each pathway is associated with particular biological functions. Lastly, the Mammalian Phenotype (MP) Ontology [BEK$^+$16] defines terms describing phenotypic characteristics of mammals and relationships between the terms, as in GO.

1.2.1 Gene Ontology (GO)

The Gene Ontology (GO) database has been built through the collaborative efforts of multiple research communities for the purpose of unifying biology across different species. GO has two main components: detailed descriptions of biological terms and relationships between the terms. Figure 1.4 shows detailed a description of the term "lung development" in GO. Figure 1.5 graphically summarizes the relationship of the term "lung development" to other terms. The GO database is searchable through their website (http://geneontology.org), which expands its usability to broader users. GO's annotation and relationships are widely incorporated in other tools.

There are three categories in GO: biological process, molecular function, and

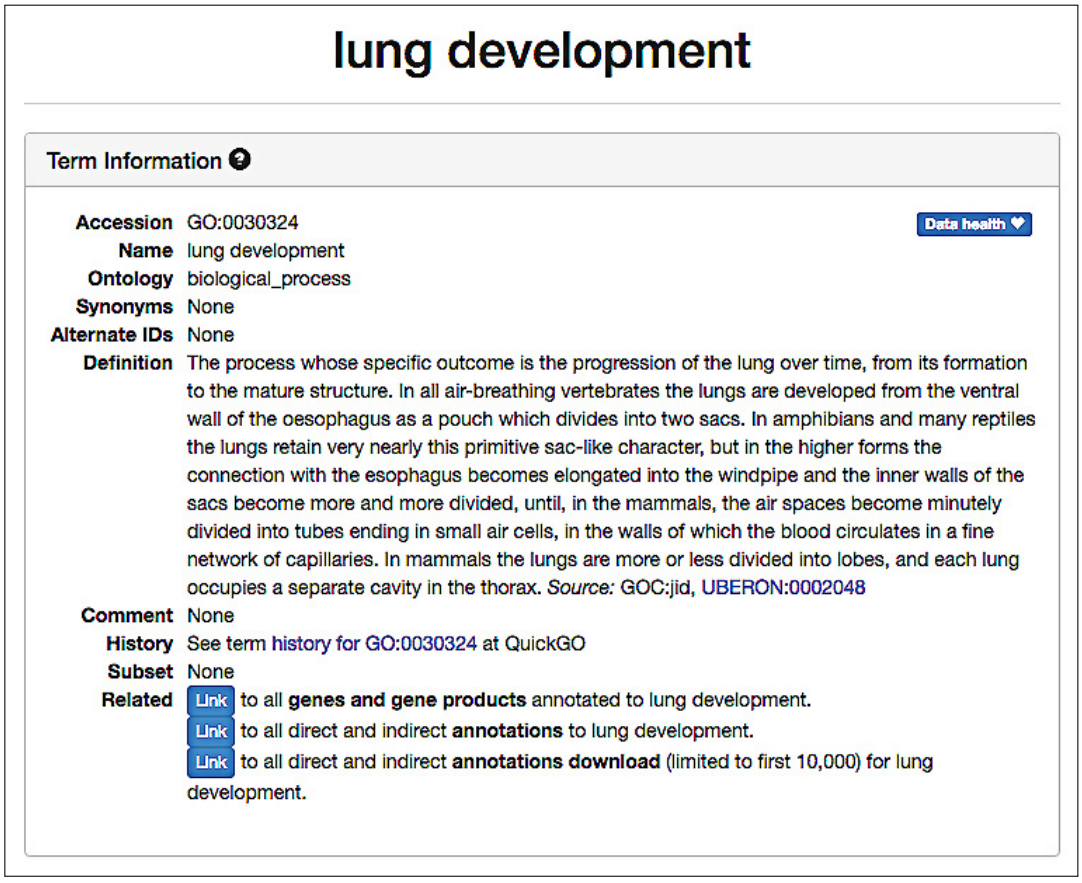

Figure 1.4: Query results for "lung development" in GO.

cellular component. Biological process (BP), the branch that we use for our study, represents gene functions. The GO also provides lists of associated genes (or gene products) for GO terms, if known. However, we instead used GO gene annotation information downloaded from the Molecular Signatures Database (MSigDB; http://software.broadinstitute.org/gsea/msigdb) [Lib14] because use of their data reduced the burden of unifying gene identifications in our work in chapter 2. The current version of MSigDB (version 5.2) contains 4,653 BP terms annotating 15,578 genes.

1.2.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [KSK+16] is a manually collected knowledge base helpful for systematic functional analysis of genes.

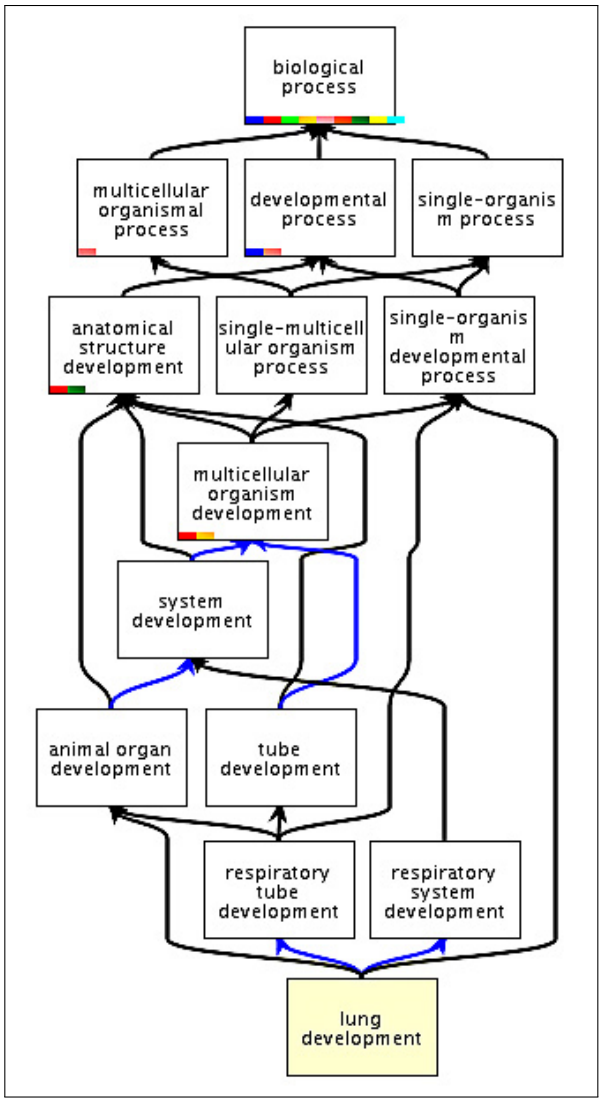

Figure 1.5: Relationships of "lung development" with other GO terms. Black lines are "is_a" relationships and blue lines represent "part_of" relationships.

KEGG provides multiple databases, including GENES for genomic information and DISEASE for human disease, but our studies are taking advantages of their "PATHWAY" database, where we can find a graphical representation of interactions between genes. Figure 1.6 shows an example of a KEGG pathway entry, the VEGF signaling pathway. The diagram provides detailed information about genes participating in the pathway and directional interactions between them.

Per statistics released by the KEGG developers, there are 504 pathways in their database (http://www.kegg.jp/kegg). The number of participating genes is

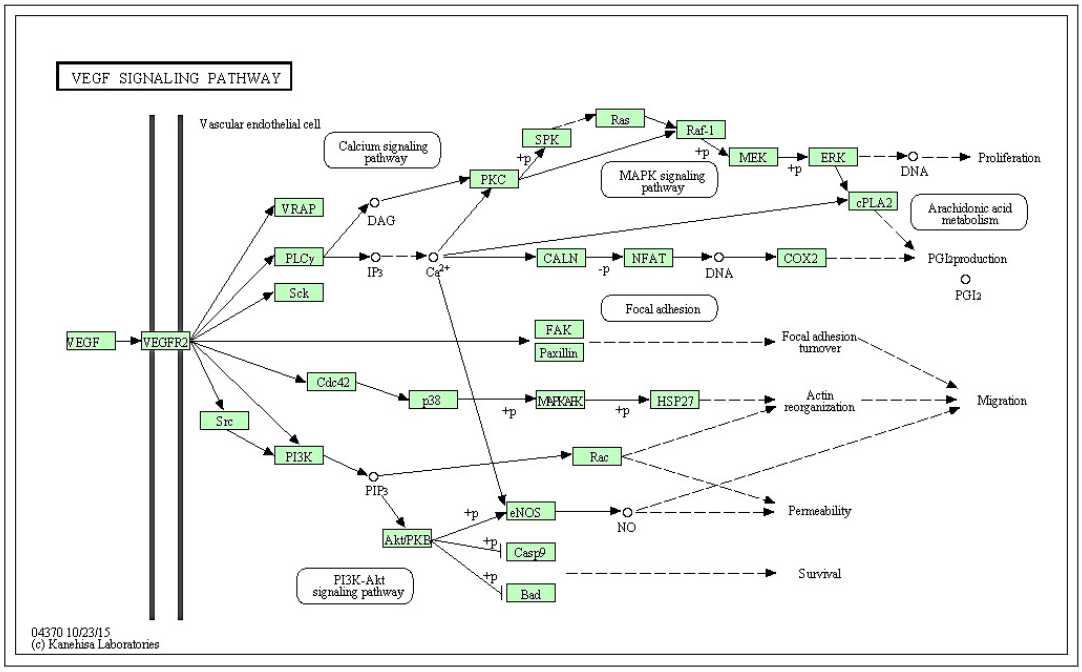

Figure 1.6: The KEGG VEGF signaling pathway.

not released. Some pathways including the MAPK signaling pathway have different diagrams for different organisms. Because our studies are more focused on the sets of participating genes than the interactions between them, we rely on the MSigDB [Lib14] as a data source for KEGG pathway gene sets. The current version of MSigDB (version 5.2) contains 186 KEGG pathway gene sets in which 12,875 genes are participating.

1.2.3 Mouse Genome Database

The Mouse Genome Database (MGD) [BEK+16] is an integrated data source for genomic and phenotypic information about the mouse (http://www.informatics.jax.org/). Among various types of data, the mammalian phenotype (MP) ontology [SE12] data can be used for functional annotation of genes. The MP ontology is structured in a similar way as the Gene Ontology (GO), and the MGD database [BEK+16] also provides a list of mouse genes and mutants annotated with MP terms. For our studies, all mouse genes were mapped to human genes using the homology information also provided by the MGD. The MGD provides a list of 18,464 human genes annotated

with 11,895 MP terms as of January 29, 2017.

1.3 Outline of This Work

In this dissertation, we present new methods we have developed to discover novel connections between human disease and biological functions using information embedded in biological networks including protein-protein interaction networks and disease hierarchies. We also include our efforts to enhance existing disease ontologies (or taxonomies) by adding additional molecular representations of disease.

Chapter 2 demonstrates our method to identify mediating pathways of a disease using topology in protein-protein interaction networks. Hypothesizing that mediating pathways are more likely to handle information flow between genetic mutations and gene expression, we define "pathway centrality (betweenness)," a variation of group betweenness, which only considers the shortest paths between disease genes and differentially expressed genes in calculation of betweenness. We tested our algorithm on three pulmonary diseases and identified mediating pathways of those pulmonary diseases supported by literature-based evidence. We further introduce a new method that systematically confirms our findings using alleviating genetic interactions. This work is to be submitted for publication in the near future.

In Chapter 3, we discuss a new method to enhance overlap analysis between developmental processes and diseases by enriching disease-gene association information. Disease-gene association is imperfect and incomplete. We show such limitations can be overcome by pooling disease genes using vertical relationships defined in disease taxonomies. Using this method, we demonstrated several interesting connections between developmental processes and human disease that invite further investigation. The efficacy of our method was systematically evaluated by comparing the ability of recovering withheld disease-gene association pairs to the same analysis without pooling. We also developed a web-based tool which runs the same analysis for user-defined gene sets representing any biological functions and visualizes the results. Most of this work, excluding the details about the extended

version of the web tool for general implications, was described in [PWK+14].

Chapter 4 introduces our work toward building more molecular disease taxonomies, motivated by our work in Chapter 3. Gene pooling from specific diseases to their parent diseases will work better if relationships between diseases reflect molecular contents. While our ultimate goal is to combine information in existing disease hierarchies with molecular genomics, we demonstrate what happens when we build disease hierarchies using only disease-gene association information. We introduce three metrics suitable for cross-ontology comparisons. Using these metrics, we evaluate our disease hierarchies by comparing to existing disease ontologies and taxonomies since there is no gold standard. While our method could recover existing disease hierarchies better than other ontology building algorithms, there is still a lot to improve. Our future work includes further improvement on our method and incorporation of other molecular information. A preliminary version of this work was presented at the ISMB 2016 SIG Meeting on Bio-ontologies [PHS16], and the extended version is being revised for Journal of Biomedical Semantics.

Lastly in Chapter 5, we discuss the results of our studies and summarize the key findings of this thesis. We conclude by discussing possible directions for future work.

Chapter 2

Pathway centrality in protein interaction networks identifies functional mediators of pulmonary disease

2.1 Introduction

It has long been noted that genes with variants implicated in disease are not necessarily differentially expressed, and that differential expression does not easily lead to the discovery of disease genes [HDR12]. In many cases, differential expression simply reflects the tissue-specific consequences or downstream result of a disease-causing process that integrates complex genetic and environmental responses. This makes differentially expressed genes useful as diagnostic markers, but often poor as therapeutic targets [FHBS11]. Conversely, functional analysis of causal disease genes, whether identified through highly specific cell or animal studies or systematically via GWAS, often won't fully explain how these downstream responses occur [Del15]. However, we know that other functional pathways may mediate the expression responses we see. We hypothesize that finding these can provide new insights

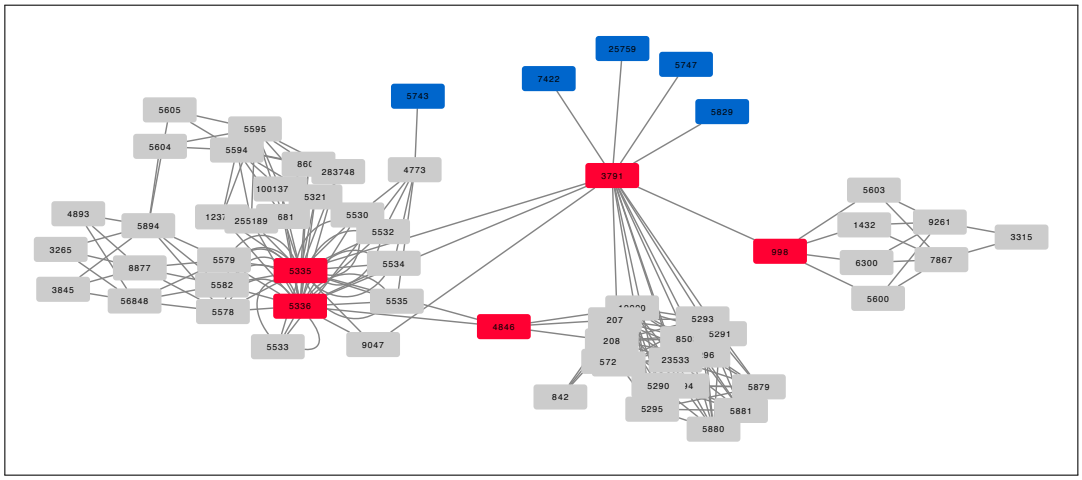

Figure 2.1: **Topological difference between high betweenness nodes and low betweenness nodes in a network.** Node betweenness counts the number of the shortest paths between all pairs of nodes pass through a given node. Blue nodes are low betweenness nodes. Red nodes are the top five nodes with high betweenness scores and they inter-connect different groups of nodes.

into disease processes and suggest novel therapeutic approaches. We test this hypothesis by examining three different pulmonary disorders from different life stages to identify common and unique mediating mechanisms in airway disease.

We start by looking at disease genes and differentially expressed genes in the context of protein-protein interaction networks, and consider the roles that the corresponding proteins play in these networks. We postulate that the functional pathways mediating disease response will disproportionately reflect communication between these two sets of proteins, and conversely, that if we look for such "central" pathways we will find mediators. To identify putative mediating pathways in disease response, we first introduce a generalized notion of network centrality called *pathway centrality*, which combines a variation of betweenness with a notion of *group centrality* [EB99]. Note that node "betweenness" characterizes how many of the shortest paths between all pairs of nodes pass through a given node [YKS+07]. (Figure 2.1 shows the topological differences between nodes with high betweenness and nodes with low betweenness.) Our variation of betweenness counts only the shortest paths between disease genes and differentially expressed genes passing through a given node. The notion of group betweenness or group centrality averages the between-

ness scores of a set of nodes to measure the centrality of the whole set. Therefore, our pathway centrality score for a pathway is the averaged value of the modified betweenness scores of a set of genes participating in the given pathway.

There are many other approaches to defining centrality in protein interaction networks [MV07]. Some have focused on modeling information flow between nodes by network or graph flows [FBW91]. Using network flows to calculate network distance has proven helpful for protein function prediction [NJA+05], and has been compared favorably to betweenness centrality for integrating function prediction with contextual data [MLZ+09]. However, this approach has not yet been applied to group centrality in protein interaction networks. Viewing group centrality as an optimization problem has also been used to discover new groups of important nodes in networks [Erd15], but not for identifying functional gene sets playing a pivotal role.

Most relevant to our efforts is a collection of prior results linking expression quantitative trait loci (eQTLs) to differentially expressed genes via protein-protein, protein-DNA, and phosphorylation networks. These studies were initially intended to find the exact causal genes in a linked locus using pathway information. One early approach [TWA+06] uses a node's accessibility in random walks to prioritize causal genes and identify regulatory paths through the network. Slightly later work [SBK+08] overcomes some limitations of the previous method by modeling the integrated network as an electrical circuit, and identifying putative causal genes by the flow of current through the network. The idea of using a model of current flow through a protein interaction network to infer protein properties is not new [MLZ+09], but using such models to link causal mutations to expression changes was valuable, in part because it allows identification of the proteins carrying high information flow connecting likely causal and differentially expressed genes.

Kim, et al. [KWP11] explicitly extended the eQTL-target work to a disease context. They again take a network-based approach linking copy number variations, through eQTL mapping, to differentially expressed genes in disease by modeling current flow through an integrated network. The method was applied to identify

likely causal mutations in glioblastoma multiforme. Such efforts as this are related to ours in the sense that they examine information flow between genes linked to disease and differentially expressed genes. Specifically, our disease genes come from a compilation of sources that includes but is not limited to GWAS data. However, here our focus is limited to the disease-related pathways, and our aim is to identify underlying biological functions that mediate cellular response in disease, rather than to identify causal mutations.

Finally, related work by Yeger-Lotem, et al. [YLRS+09] used integrated protein-protein and protein-DNA interaction networks to identify proteins and genes on high-probability paths between "genetic hits" and transcriptionally regulated genes. In this work, which focuses on a number of cellular perturbations in yeast, the emphasis is on finding individual proteins on these paths and exploring the functions they perform. Perturbation by exposure to alpha-synuclein, a protein implicated in Parkinson's disease, led to observations about functional pathways' roles in the disorder, including identifying mTOR as an enhancer of alpha-synuclein toxicity.

Our approach focuses on finding gene sets that have high centrality with respect to the communications between disease genes and differentially expressed genes in disease (i.e., we focus on identifying M, significantly participating in passing signals from D to E, in Figure 2.2). Given a disease, a set D of known disease genes, and a set E of genes differentially expressed in relevant tissues as a consequence of that disease, we calculate pathway centrality (PC) score for a pathway S considering only the paths from nodes in D to nodes in E. We use permutation tests to assess whether a gene set S has higher pathway-centrality than expected, creating a null distribution of centrality scores by repeatedly selecting sets of $|S|$ random pathway genes from a sufficiently-connected subset of the PPI network and computing their betweenness scores (see Methods). For each gene set S, the percentage of such random sets with higher pathway centrality scores than S is reported as $p_{cent}(S)$, a rough measure of significance.

In this study, we apply our pathway-centrality method to three pulmonary

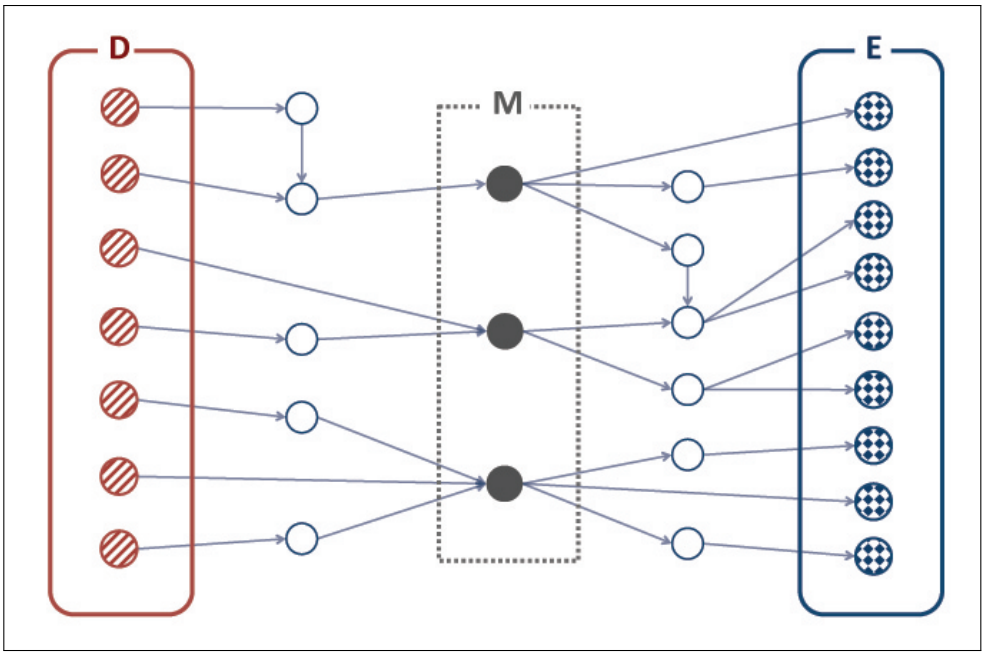

Figure 2.2: **Hypothesized topological property of mediating pathways.** Given that D is a set of disease genes and E is a set of differentially expressed genes, mediating pathway genes, M, will participate more significantly in passing signals from D to E.

diseases that primarily affect patients at different stages of life: bronchopulmonary dysplasia (BPD), a neonatal complication of preterm birth; asthma, which is relevant across the lifespan but is becoming increasingly common in children; and chronic obstructive pulmonary disease (COPD), a term that encompasses a number of progressive lung disorders that predominantly affect the elderly [SMK$^+$16]. We examine mediating pathways in each disease and look for common pathways relating all three.

One caveat is that pathways with low pathway centrality significance scores, p_{cent} (see Methods), for multiple diseases might be highly central in the network structure overall, and could appear to be "significant" for any disease considered. Therefore, as a further control, we applied the method to Down syndrome [DS] data, with the understanding that most functions relevant in pulmonary disorders are probably not that relevant to the etiology of DS, although we agree that there may be some exceptions. Our "disease genes" in this case are simply those genes

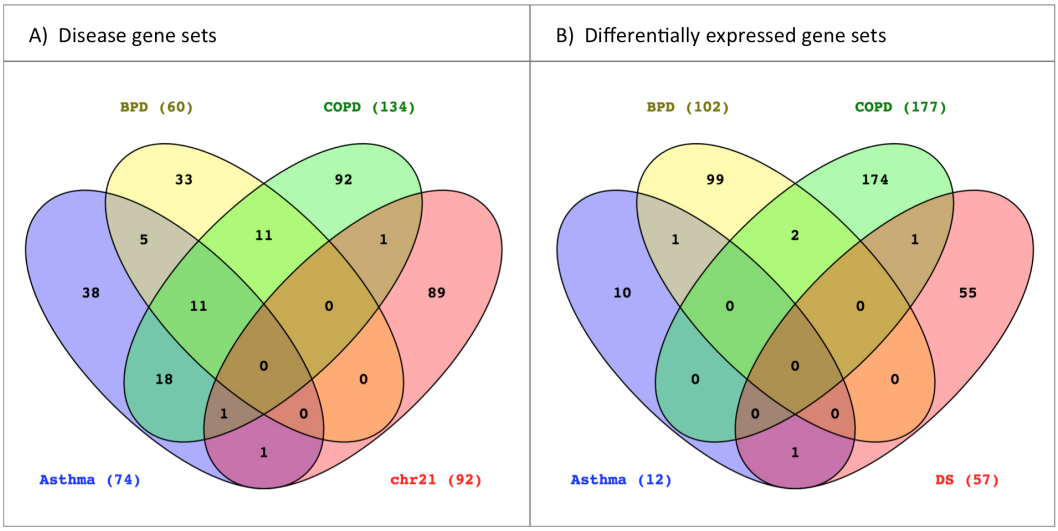

Figure 2.3: **Venn diagram showing overlaps between gene sets.** A) Overlap between three pulmonary disease gene sets (asthma [BMR11], BPD [ABSH09, YCKG10] , and COPD [Bos12]) and genes on chromosome 21 [LSP+11]. B) Overlap between differentially expressed gene sets in 3 pulmonary diseases [WBD+07, PKW+13, KLL+15] and Down syndrome [SKJ+09]. The figure is drawn using a web-based visualization tool (Venny) [Oli07].

located on chromosome 21, rather than genes that have been directly implicated in the etiology of DS symptoms.

Figure 2.3 shows the overlaps between the three pulmonary disease gene sets and the genes on chromosome 21, and the four differentially-expressed gene sets. Although there is some overlap between the sets of disease genes, particularly involving genes involved in inflammation and immune response, overall the disease gene sets are reasonably disjoint and the differentially-expressed gene sets even more so. Thus, common pathways across all three networks are unlikely to have arisen from shared shortest paths between identical sets of genes.

Pathway gene sets used in our experiments come from the Biological Processes [CDRV+15] terms in the Gene Ontology [ABB+00], pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [KSK+16], and mammalian phenotypes (MP) from the International Mouse Phenotyping Consortium in the Mouse Genome Database [BEK+16].

2.2 Methods and Resources

2.2.1 Disease-related genes

Disease genes for Asthma and COPD were collected from recent reviews, [BMR11] and [Bos12], respectively. A set of 361 Chromosome 21 genes was downloaded from the Molecular Signature DataBase (MSigDB) [LSP+11]. We collected genes associated with BPD from Online Mendelian Inheritance in Man (OMIM) [ABSH09] and Genopedia [YCKG10], as described in [PWK+14]. All datasets were downloaded on February 3, 2016.

Microarray gene expression profiles for Asthma and BPD were downloaded from the GEO database (accession numbers GSE4302 and GSE32472, respectively). The first measured differential expression in airway epithelial cells between healthy controls and asthma patients [WBD+07], while the second examined expression in peripheral blood cells from infants born preterm with or without BPD [PKW+13]. We used only the samples taken on postnatal day 5 for this study. We selected as differentially expressed genes between disease and control groups those with an adjusted Benjamini-Hochberg t-test p-value below 0.005, yielding 52 and 274 expression-related genes in asthma and BPD, respectively.

For COPD, we downloaded RNA-seq EdgeR results comparing expression in lung cells from COPD patients and controls (GSE57148) [KLL+15]. We identified 266 significantly expressed genes with an EdgeR q-value below 10^{-10} . For the control case, we chose a prior study from our group (GSE16176) [SKJ+09] that compared gene expression in amniotic fluid from second trimester fetuses with Down syndrome to those of age- and sex-matched controls. Using the same criteria as for the other Affymetrix microarray studies (all three of which used the U133 Plus 2.0 arrays) but paired t-tests to deal with the matched samples, we identified 129 differentially expressed genes in Down syndrome.

Note that there may be disease-associated genes that are also differentially expressed in that disease. To avoid confusion about how to use these in computing pathway centrality, we removed genes from the disease gene sets that also appeared

in the corresponding set of differentially expressed genes. The resulting disease gene sets contained 107 (asthma), 73 (BPD), 181 (COPD), and 360 (DS).

In our experiments, only genes participating in known protein-protein interactions are considered. Further filtering of gene sets in this way yields 12, 102, 177, and 57 differentially expressed genes in asthma, BPD, COPD and Down syndrome respectively. For disease gene sets, we end up with 74 asthma genes, 60 BPD genes, 134 COPD genes and 92 genes on chromosome 21.

2.2.2 Functional pathways and gene sets

Both the Gene Ontology and the KEGG gene set collections were downloaded from the Molecular Signature DataBase (MSigDB) [LSP+11] on August 18, 2015 (http://software.broadinstitute.org/gsea/msigdb). This GO gene set collection includes 825 biological process terms and 6,178 genes, and the KEGG collection includes 186 pathways and 5,267 genes. In addition, we created a collection of gene sets using data from the Mouse Genome Informatics site (http://www.informatics.jax.org), which includes a database characterizing mutant mouse strains with respect to the Mammalian Phenotype Ontology [BEK+16]. To create gene sets with mammalian phenotype (MP) data, we downloaded the files HMD_HumanPhenotype.rpt and MGI_GenePheno.rpt on February 21, 2016. The first file includes human / mouse orthology mapping and some MP data, while the second includes more MP data characterizing mouse gene mutant strains (other than conditional mutations), but no orthology information. We therefore used the human-mouse orthologs from the first file and then combined the phenotypes from both files, excluding the two normal phenotype labels MP:0002169 ("no abnormal phenotype detected") and MP:0002873 ("normal phenotype"). We then created for each phenotype a list of genes with at least one mutant allele with that phenotype. These lists were used as functional gene sets, which we describe as the "mammalian phenotype" gene set collection. There are 8,225 phenotypes in the mammalian phenotype data and 7,778 genes.

2.2.3 Protein-protein and genetic interaction networks

We use two biological networks in our experiments. To measure pathway centrality, physical protein-protein interactions were collected from the Human Integrated Protein-Protein Interaction rEference (HIPPIE) [ALANS17] database. HIPPIE contains experimentally verified protein interactions with confidence scores. We downloaded the interaction data (version 2) on July 18, 2016 and selected only those interactions reported as "high confidence" (≥ 0.73), as these interactions are supported by more reliable evidence. We worked with the largest connected component extracted from the network, which contains 43,475 interactions between 9,379 proteins.

To systematically evaluate our findings about disease-mediating pathways, we utilized genetic interaction data featuring alleviating (positive) genetic and phenotypic suppression interactions. Because relatively few of these types of genetic interactions are known for humans, we collected such genetic interactions from Saccharomyces cerevisiae, Drosophila melanogaster, and Caenorhabditis elegans. These interactions came from BioGRID [SBR+06] (version 3.4.141, downloaded October 14, 2016); the Saccharomyces Genome Database (SGD project, http://www.yeastgenome.org; downloaded October 10, 2016), and Flybase [AFG+16] (downloaded August 15, 2016). To find human homologous interaction pairs, we use a mapping downloaded from the HomoloGene database [Coo16] on July 19, 2016. This approach gave us 4,536 pairs of putative positive genetic interactions in humans.

2.2.4 Pathway centrality measures how central a pathway is between disease genes and differentially expressed genes

Pathway centrality measures the amount of information a set of pathway genes handles by counting paths linking disease genes and differentially expressed genes. While the classical definition of fractional betweenness (FB) for a node v is the fraction of shortest paths between all pairs of nodes in a network passing through

v, our pathway centrality (PC) score will be based on a modified FB score (FB'), which for node v, only reflects the shortest paths between disease genes, $D(d)$, and differentially expressed genes, $E(d)$, that passes through v. Formally, this is defined as:

$$FB'(v) = \sum_{v \in V | v \neq s \in D(d), v \neq t \in E(d)} \frac{B_{s,t}(v)}{B_{s,t}} \tag{2.1}$$

where V is the set of all vertices in the protein-protein interaction network, $D(d)$ is the set of genes in V associated with disease d, $E(d)$ is the set of differentially expressed genes in V for disease d, $B_{s,t}$ is the total number of shortest paths between s and t, and $B_{s,t}(v)$ is the total number of shortest paths between s and t that pass through v.

We define pathway centrality as the average fractional betweenness score across all genes in a pathway, counting only the shortest paths from disease genes to differentially expressed genes. For a pathway S, pathway centrality(S) is defined as:

$$PC(S) = \frac{\sum_{v \in S} FB'(v)}{|S|} \tag{2.2}$$

2.2.5 Assessing significance of observed pathway centralities, p_{cent}

The significance of the observed pathway centrality score of a pathway or gene set S is assessed using a null distribution learned from a permutation with 10,000 random gene sets of size $|S|$. We calculate the probability that a gene set has the observed pathway centrality by chance, and call this p_{cent}. We used a threshold of 0.05 for the p_{cent} to identify significantly central pathways which we believe to be mediating signals between disease genes and differentially expressed genes. We initially observed, however, that almost a quarter of tested pathways were determined to be significant. Specifically, we identified 224, 201, and 227 out of 825 GO BP gene sets as significant mediating pathways for asthma, BPD, and COPD, respectively.

We reasoned that this was due to our random sampling process, where we

select random gene sets from all genes, including those without known functions. Pathway genes are known to be relatively central in protein interaction networks, and therefore, it is likely that pathway genes have higher fractional betweenness than those that are not involved in well-annotated functional processes. To decrease such biases that artificially reduce the p_{cent} values, we added a restriction to our random sampling process; that our random samples should be drawn from a collection of genes belonging to at least one pathway. Note that this restriction holds for all of the random sampling processes described below.

While this restriction did reduce the number of significant mediating pathways, there still seemed to be too many significant BP pathways for some data sets. One possible reason is that, in many protein interaction networks, the majority of nodes in the network have degree 1. To the extent that protein interaction networks have low-degree nodes, random samples will likely be enriched for samples of low degree unless we force the random sampling process to select higher degree genes. This is problematic, because the distribution of pathway centrality from permutation tests may then be right-skewed, given that the various centrality measurements tend to be correlated [GOKK03].

Therefore, we decided to force our random samples to have a degree distribution identical to that of the tested pathway. However, an issue arose immediately that protein interaction networks have many fewer high-degree nodes than low-degree nodes, so most high-degree nodes are only options for the sampling process to pick when choosing a node of a certain high degree. That is, if the mediating pathway has a high-degree node, the random samples are most likely to share that exact node with the mediating pathway because it may be the only one with the same degree. This can result in over-correcting the problem of the right-skewed distribution learned from low-degree random samples and determining most mediating pathways with high-degree nodes are *insignificant*, because the null distribution of pathway centralities is now likely to be left-skewed, making p_{cent} high for most mediating pathways with high-degree nodes.

A better way to simulate the degree distribution of the original pathway

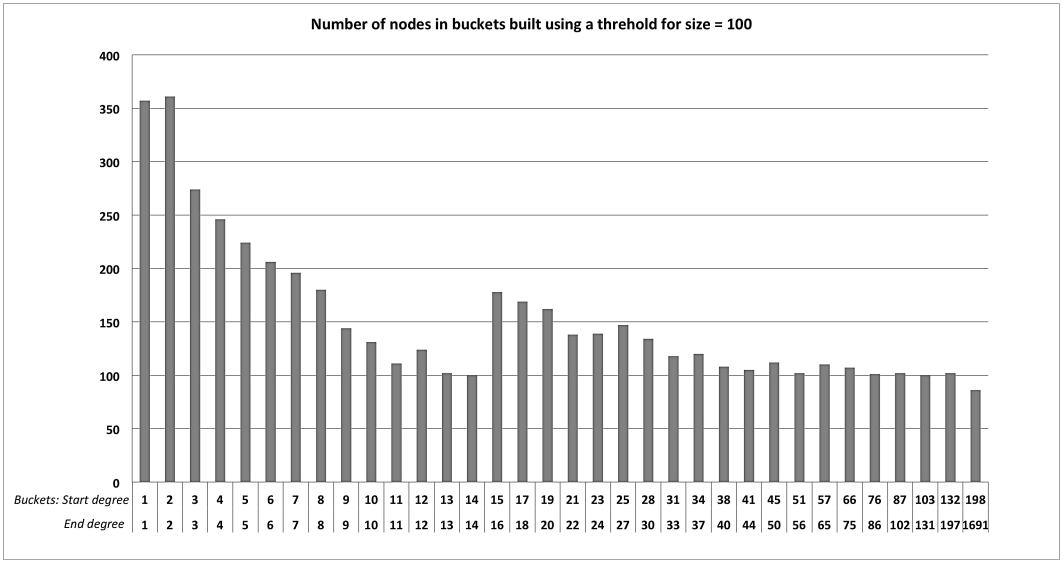

Figure 2.4: **Number of nodes in degree buckets built using a threshold of 100.** Some degree buckets may have more than 100 nodes because we keep same-degree nodes in a same bucket. The first multi-degree bucket is the one combining nodes of degree 15 and 16, and this merging happens because the number of nodes of degree 15 is less than 100.

with the random samples might be to "bucketize" node degrees such that high-degree nodes are placed into a single bucket for sampling. This way, the random sampling process will have more choices for approximately matching the degree of high-degree nodes in mediating pathways. To do this, we implemented an algorithm that sorted nodes by their degree and placed nodes of increasing degree into one bucket, until the size of the bucket is greater than some threshold. All nodes of the same degree are placed in the same bucket. The algorithm also checks the last bucket, and merges it with the previously built bucket if the size of the last bucket is smaller than half of the size threshold. Figure 2.4 shows an example of bucket sizes that we built using a size threshold of 100. Since we keep same-degree nodes together, low-degree buckets (i.e., degree 1 $\sim$ degree 14) have more than 100 nodes, all of the same degree. Multi-degree buckets start when we meet the node group with degree 15, whose size is less than 100 so we add all degree-16 nodes to the bucket. We repeat this process, increasing the degree and range for each bucket until we reach the highest degree node. The size of the last bucket is forced to be

at least half of the size threshold.

We tested this approach with five different thresholds for size: 20, 50, 100, 200, and 300. While this approach addresses the problem of the right-skewed distribution of pathway centralities from permutation test, we observed a pitfall which invites more through investigation. The statistical significance (p_{cent}) of a mediating pathway assessed from permutation tests using different bucket sizes can differ dramatically. Table 2.1 shows a few cases where p_{cent} for a mediating pathway varies depending on the thresholds for bucket size. We found this dependence of the significance on bucket sizes to be troubling. For this reason, we have abandoned the degree-match approach and moved on to an alternative sampling method which removes low-degree nodes from consideration in random sampling entirely. A through investigation of random sampling methods to simulate the degree distribution of tested pathways is left as one of my future research directions.

Table 2.1: **Difference in p_{cent} depending on thresholds for bucket size.**

Disease	Threshold for bucket size				
Pathway	20	50	100	200	300
Asthma					
Interleukin 2 production	**0.0326**	0.0337	0.1084	**0.1984**	0.0709
Adaptive immune response	**0.0475**	0.0506	0.1137	**0.1807**	0.0891
BPD					
Activation of immune response	0.0132	**0.0097**	0.0542	**0.1136**	0.0324
COPD					
Adaptive immune response	**0.0077**	0.0161	0.0912	**0.1716**	0.0293
Macromolecule biosynthetic process	**0.0057**	0.0147	0.0343	0.023	**0.1166**

p_{cent} for several pathways differs depending on the thresholds for bucket sizes. Lowest and highest p_{cent} values are highlighted. This is problematic especially when p_{cent} for a pathway can be either below or above threshold for significance (0.05) depending on the bucket size.

2.2.6 Identification of significant mediating pathways using p_{cent}

To correct for the overrepresentation of low-degree genes from permutation, we therefore decided to use only genes in the 2-core of the protein-protein interaction network in our permutation tests. A k-core of a network is a maximal group of

nodes, all of which are connected to at least k other nodes in the network [Sei83]. This sampling approach improves the observed distribution of p_{cent} values. It may also reduce the significance of pathways containing mostly low-degree genes, but this applies to relatively few of the gene sets in the considered collections. Note that the restriction that we only consider genes participating in at least one pathway still holds for our permutation tests. Finally, we use a threshold of 0.05 for p_{cent} to identify significant mediating pathways.

2.2.7 Evaluation of identified disease-mediating pathways

We evaluate our findings about disease-mediating pathways by measuring the excess of epistatic interactions between the pathways and the disease genes over the number that would be expected. To assess how surprising it is to see the observed number of epistatic interactions between the disease genes and a mediating pathway, we compute a probability, which we call p_{med}, from a null distribution of such counts between the same set of disease genes and 10,000 random gene sets (R) of equal size as the mediating pathway (see Figure 2.5). Again we impose restrictions on the source of random genes to avoid over-estimating significance. Here, random gene sets are drawn from a pool of genes that belong to at least one pathway in the collection and that are downstream genes of any alleviating genetic or phenotypic suppression interactions.

2.3 Results and Discussion

2.3.1 Pathway centrality finds mediating pathways and potential drug targets

Pathways showing significant pathway centrality in specific pulmonary data sets but not in the Down syndrome data include known disease mediators and potential targets. Table 2.2 shows the top few most significant pathways implicated in exactly one of the pulmonary disorders (full lists of results are available on http://bcb.cs.tufts.edu/jpark/pathway-centrality/). We highlight some of these top

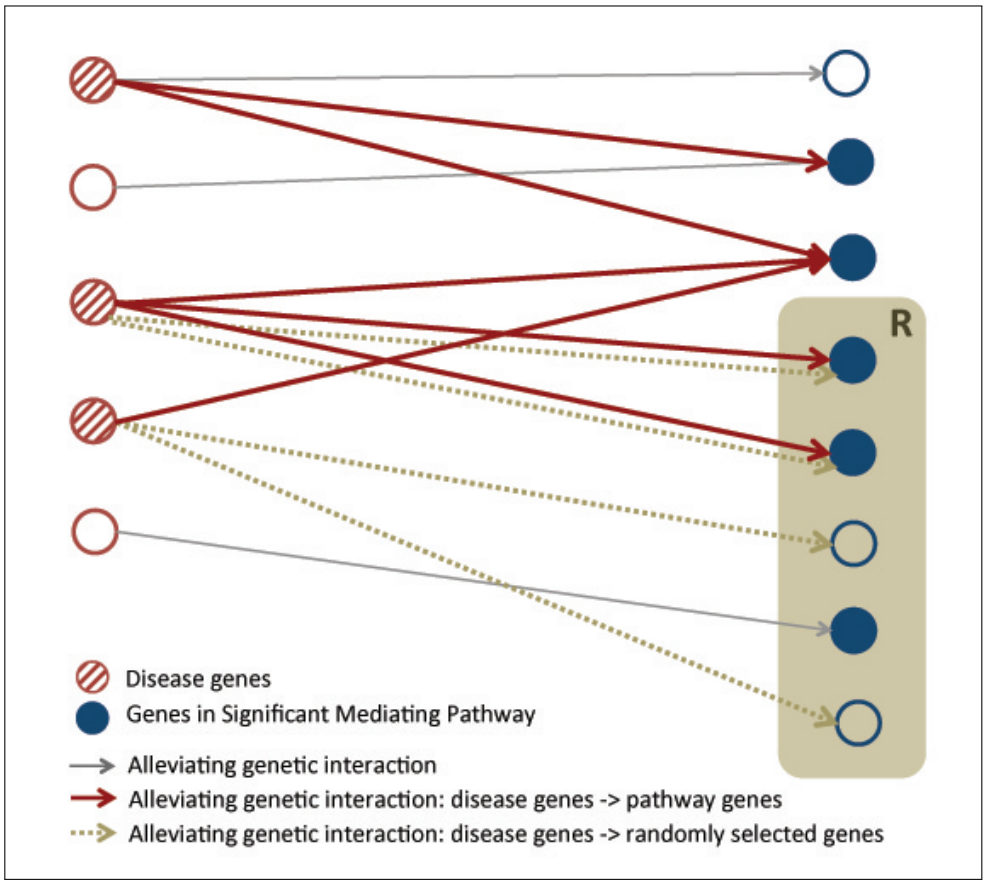

Figure 2.5: **Systematic confirmation of significant mediating pathways.** If identified pathways are truly mediating a disease, the pathways are likely to be downstream of the corresponding disease genes. This relationship can be captured by the excess of epistatic interactions between disease genes and pathway genes. We count alleviating genetic interactions between disease genes and our identified mediating pathways (the number of red solid arrows between the red circles filled with dashed lines and blue solid circles). We then assess the significance of the observed number of such interactions (i.e., counts of red solid arrows) by calculating the probability that a gene set of equal size has the same number of alleviating genetic interactions with disease genes. The null distribution is learned from 10,000 random samples (R) drawn from a pool of downstream genes of any known alleviating genetic interactions.

Table 2.2: **Significant pathways implicated in one of the pulmonary disorders.**

Pathway	Asthma p-value	BPD p-value	COPD p-value	DS p-value
GO Biological Process Terms:				
Regulation of DNA Metabolic Process	**0.0014**	0.1015	0.2034	0.3279
Leukocyte Chemotaxis	**0.0021**	0.3232	0.1634	0.2219
GI Phase	**0.0027**	0.0771	0.1375	0.087
Leukocyte Migration	**0.0032**	0.4966	0.2705	0.3176
Regulation of MAP Kinase Activity	0.1779	**0.0035**	0.0502	0.0874
Ras Protein Signal Transduction	0.0995	**0.0049**	0.0854	0.1305
Homeostatic Process	0.156	0.1313	**0.0004**	0.362
Nuclear Transport	0.2514	0.056	**0.0018**	0.089
Nucleocytoplasmic Transport	0.2455	0.0524	**0.002**	0.0872
Chemical Homeostasis	0.2365	0.1249	**0.002**	0.6049
KEGG pathways:				
FC Gamma R Mediated Phagocytosis	**0.0044**	0.18	0.111	0.3449
Tight Junction	**0.0128**	0.0564	0.135	0.0927
Regulation of Actin Cytoskeleton	0.2283	**0.0026**	0.103	0.2323
Antigen Processing and Presentation	0.2898	**0.0053**	0.0932	0.1658
Insulin Signaling Pathway	0.3315	**0.0195**	0.0966	0.2575
Cell Adhesion Molecules (CAMs)	0.7917	0.1658	**0.0198**	0.3768
RIG I Like Receptor Signaling Pathway	0.0726	0.1353	**0.0235**	0.1229
Viral Myocarditis	0.3229	0.0825	**0.0391**	0.3046

Most-significant GO Biological Processes and KEGG pathways in individual pulmonary disorders. These pathways are the top few with $p_{cent} < 0.05$ in exactly one of the pulmonary disorders and $p_{cent} > 0.05$ in the DS data. Cells containing p-values below 0.05 are in bold. Cutoffs shown here were determined by discussion in the text; full results are available as supplemental data on http://bcb.cs.tufts.edu/jpark/pathway-centrality/.

pathways here.

Highly significant in asthma is the BP gene set leukocyte chemotaxis. Neutrophil chemotaxis velocity has been suggested as a biomarker for asthma and incorporated into a diagnostic test to distinguish asthma from nonasthmatic allergic rhinitis [SBS+14]. The KEGG gene set cell adhesion molecules tops the list in COPD, pointing at similar processes, although the only genes shared between these two gene sets are ITGB2 and ITGA9. Prior work suggests that adhesion molecules also play a significant role in the pathogenesis of COPD [Ish99], and that regula-

tion of adhesion may be a new therapeutic approach for both COPD and asthma [CKRC08, WV08]. The KEGG pathway tight junction is also implicated in asthma. Although adhesion and leukocyte chemotaxis are important to all three disorders [FTWB$^+$16, ROSHW98, BDRT09], the pathway centrality approach highlights different sets of genes mediating these responses. Different pathways have been implicated in recruitment of neutrophils in different contexts in COPD [BDRT09], suggesting the plausibility of context-specific variation.

Topping the GO COPD list are two terms relating to homeostasis. Protein homeostatic imbalance has been implicated in the pathogenesis of COPD, although this approach to characterizing the disease has not yet been thoroughly explored [BTV12]. However, such imbalances have been successfully targeted through small molecules or gene therapies in other diseases such as cystic fibrosis. Addressing the imbalance of critical NFKB-regulated inflammatory proteins has thus been suggested as a possible therapeutic approach in COPD and other airways diseases [BB12].

Mitogen-activated protein kinases (MAPKs) regulate many developmental and cellular processes [KC10], including inflammation and apoptosis. While these pathways are likely involved to some degree in all three pulmonary disorders, the top GO BP gene sets implicated in BPD are regulation of map kinase activity and ras protein signal transduction. The extracellular signal-related kinases ERK1/2 are major components of the MAPK pathway, which can be activated by the Ras GTP-ase [KC10]. There is evidence that ERK/MAPK activation can protect against the negative effects of hyperoxia on alveolar development and lung epithelial cells [XGM$^+$06, BBWW05]. It has been suggested that drugs that promote this process are potential therapies for BPD [SVH$^+$13].

2.3.2 Commonalities across all three pulmonary disorders implicate immune processes and signaling pathways.

When we look for pathways that play a significant role across all three pulmonary disorders (Table 2.4), we are not surprised to find strong evidence pointing at in-

Table 2.3: **Significance of relationships between p_{cent} and p_{med}.**

	p-value of slope			Wilcoxon		
	BP	KEGG	MP	BP	KEGG	MP
asthma	**0.034**	**0.014**	**0.002**	**4.94E-04**	**4.56E-17**	**1.41E-08**
bpd	0.696	**0.032**	0.579	**0.047**	**2.21E-17**	**4.10E-20**
copd	0.607	**0.001**	**0.017**	**7.89E-07**	**1.47E-15**	**1.34E-12**
DS	0.86	0.893	0.158	**3.27E-07**	0.632	**1.73E-09**

The first set of numbers reflects the raw p-values reported by the glm() function in R showing that a regression line fit to the probabilities of having a $p_{med} < 0.05$ for each decile of p_{cent} values has a significantly negative slope. The second set shows the raw p-values for Wilcoxon tests comparing the p_{med} values in just the first decile to those in the remaining deciles combined. Raw p-values below 0.05 appear in bold.

flammation and immunity. GO terms topping the list implicate the adaptive immune response, cytokine production, and pro-inflammatory NF Kappa B signaling [KD00]. Recent research implicates IKK-driven NFKB activation of inflammation in both COPD and asthma, but suggests that the different components of the system involved in the two diseases could explain their differing pharmacological responses [GCP+11], and could suggest new avenues for therapy. NFKB-mediated inflammation has also been implicated in the pathogenesis of BPD, and the increased prevalence of early-onset emphysema in BPD survivors [WLL+08] supports the hypothesis that an NFKB-related inflammatory phenotype predisposes individuals to an increased reaction to environmental airway stress [PWK+14]. Links between BPD and asthma through NFKB1A promoter polymorphisms have also been identified [AHM+13].

The KEGG and MP gene set collections implicate several more specific pathways, including different types of lymphocyte responses, JAK/STAT signaling, toll-like receptor signaling, and FC Epsilon R1 signaling. The JAK/STAT pathway has been suggested as an asthma target through inhibitors of activing cytokines and receptors [Val16, WK98]. JAK pathway inhibitors are in development for a number of inflammatory disorders [OSV+15], and animal models have suggested that targeting this pathway can reduce airway hyperresponsiveness, reflecting the potential

Table 2.4: **Significant pathways implicated in all three pulmonary disorders, but not in the control data (DS).**

Pathway	Asthma p-value	BPD p-value	COPD p-value	DS p-value
GO Biological Process Terms:				
Cell Surface Receptor Linked Signal Transduction	0.0447	0	0.0091	0.2122
Protein Kinase Cascade	0.0058	0.0011	0.0075	0.2083
Positive Regulation of Response To Stimulus	0.0269	0.0017	0.0043	0.1066
Positive Regulation of Immune System Process	0.0343	0.0019	0.0059	0.1561
Regulation of Immune Response	0.02	0.0025	0.0023	0.0841
Positive Regulation of Immune Response	0.0174	0.0028	0.0024	0.0711
Regulation of Immune System Process	0.0443	0.0041	0.0173	0.2604
Regulation of Response To Stimulus	0.0383	0.0043	0.0198	0.199
I KappaB Kinase NF KappaB Cascade	0.0021	0.0046	0.0115	0.1327
Positive Regulation of Multicellular Organismal Process	0.0487	0.005	0.0153	0.2606
Activation of Immune Response	0.017	0.0074	0.0096	0.1492
Regulation of Cytokine Production	0.0144	0.019	0.0075	0.0546
Adaptive Immune Response GO 0002460	0.0221	0.0283	0.0384	0.0623
Adaptive Immune Response	0.025	0.0289	0.0476	0.0697
Protein Polyubiquitination	0.0133	0.0435	0.0167	0.0683
KEGG pathways:				
Leishmania Infection	0.0002	0.0032	0.0098	0.1635
JAK STAT Signaling Pathway	0.0003	0.0001	0.0097	0.1072
B Cell Receptor Signaling Pathway	0.0003	0.0004	0.0091	0.0662
FC Epsilon RI Signaling Pathway	0.0004	0.0006	0.0236	0.1027
Natural Killer Cell Mediated Cytotoxicity	0.0058	0.001	0.0473	0.5846
Chemokine Signaling Pathway	0.0065	0	0.0025	0.4633
Toll Like Receptor Signaling Pathway	0.0136	0.0094	0.0294	0.247
Acute Myeloid Leukemia	0.0175	0	0.0158	0.0637
T Cell Receptor Signaling Pathway	0.0275	0	0.0206	0.0574
Selected Mammalian Phenotype Terms:				
Immune System Phenotype	0	0	0.0001	0.0654
Decreased Bone Resorption	0.0003	0.011	0.0175	0.1103
Abnormal Circulating Complement Protein Level	0.0047	0.0245	0.0063	0.2421
Increased Circulating Angiotensinogen Level	0.0121	0.0117	0.0172	0.1478
Abnormal Circulating Chemokine Level	0.0176	0.015	0.018	0.1993
Abnormal Airway Responsiveness	0.0222	0.0067	0.0372	0.556
Abnormal Chemokine Level	0.0341	0.0131	0.017	0.2805
Abnormal Interleukin-4 Secretion	0.0427	0.0055	0.0424	0.1368

All GO Biological Process terms and KEGG pathways, and manually selected Mammalian Phenotype terms, significant ($p < 0.05$) in all three pulmonary disorders and with $p > 0.05$ in the DS data. There were 58 such MP terms in total.

for such compounds in both COPD and asthma [Bar16]. The role for JAK/STAT signaling in bronchopulmonary dysplasia is less clear, but it has been suggested that it plays a role in airway smooth muscle mitogenesis, implicated in both asthma and BPD [STS$^+$02], and postulated that it may be an alternative mediator of the oxidative stress response in both diseases [ZH03]. Thus, our work suggests that the effects of JAK pathway inhibitors may be worth exploring in pre-clinical models of bronchopulmonary dysplasia.

Toll-like receptor (TLR) signaling, which activates the innate immune response, is another familiar part of the story of airway hyperreactivity and fetal lung development [PGLT10]. TLR polymorphisms have been linked to an increased risk of developing BPD [CDRV$^+$15, MASS16], and TLR agonists are already being tested for therapeutic efficacy in asthma [BAB$^+$16]. However, the role of this system in COPD is not as clear. Aspects of the innate immune response are often demonstrably suppressed in COPD patients [SC13], and TLR polymorphisms play a role in disease susceptibility and severity [AKM$^+$16, Bar16]. Our work therefore also suggests a role for TLR pathways in the diagnosis, stratification, and treatment of COPD.

Finally, the FC epsilon RI signaling pathway raises interesting questions about common elements of these three diseases. Figure 2.6 shows a subset of this pathway and the network for BPD. The gene FCER1A is one of the primary receptors for immunoglobulin E (IgE) [SMG$^+$14], the key player in initiating allergic response. The role of allergy in these three disorders, however, is thought to be quite different. The significant impact of IgE response in allergic asthma is well-studied [GS08]. Its role in COPD is less clear. Many patients with COPD but no asthma diagnosis have one or more asthma-like symptoms, including atopy, and there are suggestions that allergic response plays a role in severity for a subset of those with COPD [SMK$^+$16]. In contrast, most evidence suggests that bronchopulmonary dysplasia is not linked to the development of allergies or to IgE response [KLK$^+$13]. Indeed preterm birth has been shown to correlate with a decreased risk of atopy [SKPS01], although with an increased risk of asthma and ultimately of obstructive

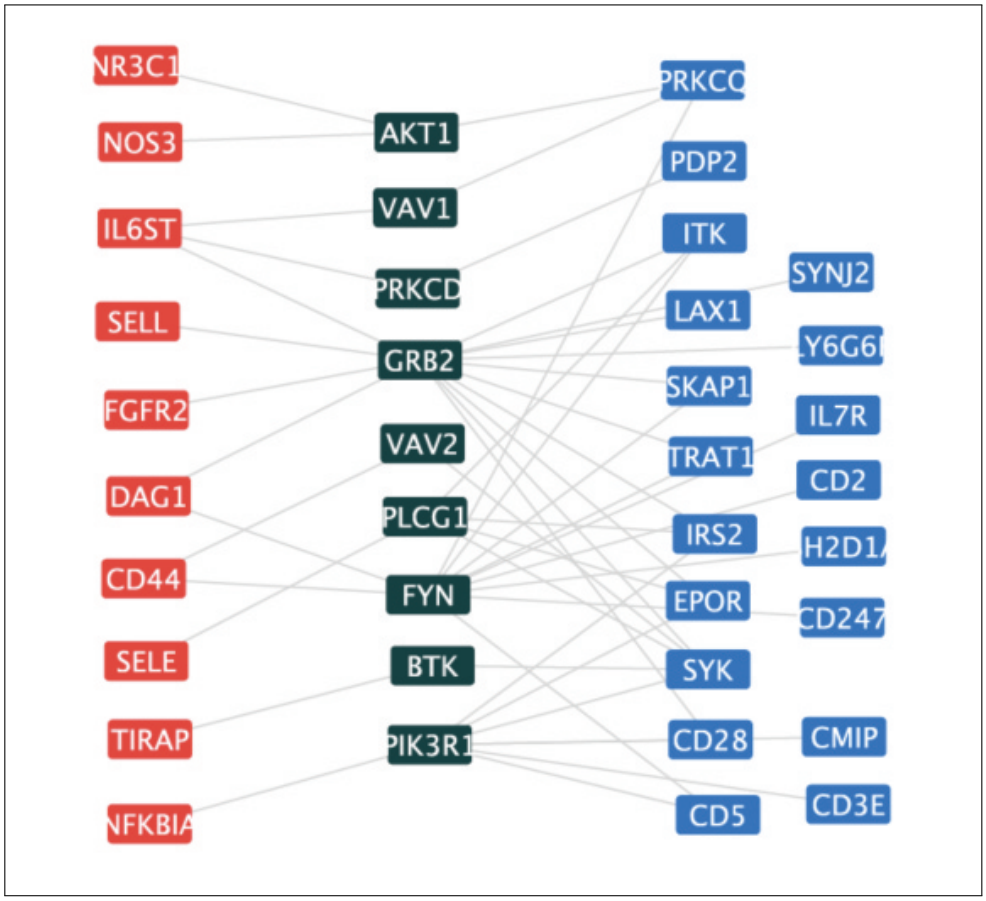

Figure 2.6: **Topology of BPD-related genes and FC epsilon RI signaling pathway (KEGG).** One of the significant mediating pathways for bronchopulmonary dysplasia [CDRV+15] is the FC epsilon RI signaling pathway (pathway genes are colored with dark green). The mediating pathway genes are in between BPD disease genes (in red) and differentially expressed genes (in blue).

pulmonary disease [SMW+07]. Further, the question of whether asthma in survivors of BPD is more likely to be non-atopic is still open, although there is some evidence supporting this hypothesis [VDWL+90, RKM+16]. In this context, the fact that this signaling pathway is nonetheless implicated as a mediator of gene expression changes in infants with BPD is intriguing and may shed light on the mechanisms involved in the disease. Further exploration of its role in this process is warranted.

2.3.3 Genetic interaction data confirms the identification of mediating pathways.

Finding specific examples consistent with existing knowledge provides anecdotal evidence that an approach is effective, but more systematic evaluation is needed. One possible way to show that the proposed mediating pathways are, at least in some respect, downstream of the disease genes, would be to identify an excess of epistatic relationships between them. For example, if a mediating pathway looks like that shown in Figure 2.2, one might expect a higher likelihood of certain kinds of genetic interactions between a disease gene d in set D and a mediating gene m from set M than between d and genes that are not in a mediating pathway for that disease. The genetic interactions of most interest would be "alleviating" or positive genetic interactions, where the deleterious effect of the double mutant of both d and m is less severe than would be predicted by combining the independent effects of individual mutations in d or m. Such relationships might arise when m is part of a pathway mediating the response of d.

Recall that, for a pathway S, we defined the pathway centrality significance score, $p_{cent}(S)$ and the significance of the excess of epistatic relationships between the disease genes (D) and S, $p_{med}(S)$. The overall correlation between $p_{cent}(S)$ and $p_{med}(S)$ varies considerably across the data sets and annotation sets, with a range of 0.07 to 0.74. The lowest correlations come from the Down syndrome data, where this approach is more problematic due to the large number of "disease" (i.e., trisomic) genes that are not actually involved in the etiology of the DS phenotype. Instead of overall correlation, however, our goal is to assess the prob-

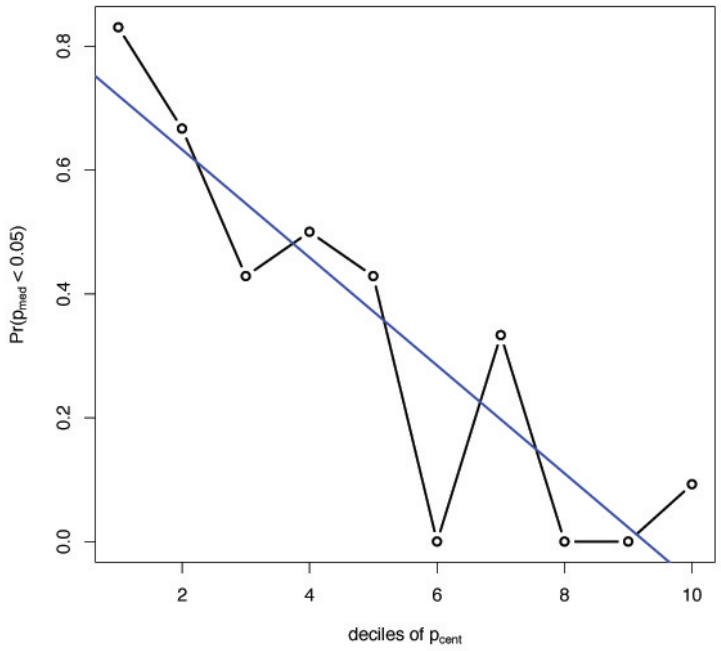

Figure 2.7: **Correlation between p_{cent} and p_{med} calculated for KEGG pathways and COPD.** We claim that pathways with low p_{cent} also have low p_{med} (for definitions, see Methods). That is, there is low probability (p_{med}) that a gene set of equal size has as many alleviating genetic interactions with COPD genes as a significant mediating pathway (with low p_{cent}). The plot supports our claim because the slope of the linear regression line (in blue) is significantly different from zero (with probability of 0.0006). However, a more interesting observation is that more than 80% of KEGG pathways in the first decile of p_{cent} have have $p_{med} < 0.05$.

ability that a pathway S having a low $p_{cent}(S)$ value is more likely to have a low $p_{med}(S)$ value as well. We therefore split the range of p_{cent} values up into deciles, and plotted the probability of finding a p_{med} score below 0.05 in each decile. A sample plot of these probabilities is shown in Figure 2.7; the rest are available on http://bcb.cs.tufts.edu/jpark/pathway-centrality/. While these plots are typically noisy, in most cases the slope of a linear regression line through the points (the blue line in Figure 2.7) is significantly greater than zero, suggesting that the probability of being a mediator is highest for the most significantly central pathways.

Table 2.5: **Significance of relationships between p_{cent} and p_{med}.**

	p-value of slope			**Wilcoxon**		
	BP	KEGG	MP	BP	KEGG	MP
asthma	**0.034**	**0.014**	**0.002**	**4.94E-04**	**4.56E-17**	**1.41E-08**
bpd	0.696	**0.032**	0.579	**0.047**	**2.21E-17**	**4.10E-20**
copd	0.607	**0.001**	**0.017**	**7.89E-07**	**1.47E-15**	**1.34E-12**
DS	0.86	0.893	0.158	**3.27E-07**	0.632	**1.73E-09**

The first set of numbers reflects the raw p-values reported by the glm() function in R showing that a regression line fit to the probabilities of having a $p_{med} < 0.05$ for each decile of p_{cent} values has a significantly negative slope. The second set shows the raw p-values for Wilcoxon tests comparing the p_{med} values in just the first decile to those in the remaining deciles combined. Raw p-values below 0.05 appear in bold.

Table 2.5 shows the p-values assessing whether the slope of the line differs from zero (as reported by the R generalized linear model function glm()). Both the tests and the plots support the conclusion that there is enrichment of alleviating genetic relationships between disease genes and pathway genes for the pathways whose p_{cent} values are in the most significant decile in most cases. The p-values are consistently high for the DS data, which makes sense for the reason noted above. The test is otherwise uniformly significant for the KEGG pathways, and in some of the cases where it is not, such as BPD with the MP gene sets, the first decile does in fact have the highest probability; the poor significance and relatively flat slope occur because there also happen to be many genetic relationships for p_{cent} values above 0.8.

To focus on just the most-significant pathways, we can simply compare the distribution of p_{med} values in the first bin to that in the remaining bins using a one-sided, non-parametric Wilcoxon test. These significance values are shown in Table 2.5 as well. Here the test identifies a mediating relationship for all the pulmonary disease cases and almost all the DS ones as well.

2.4 Conclusions

In this chapter, we introduced our algorithm for identifying disease mediating pathways based on the degree of their participation in signal mediating from disease genes to differentially expressed genes. The degree of such participation was measured by *pathway centrality*, a variation of *group centrality (betweenness)*, that only counts the shortest paths between disease genes and differentially expressed genes. Our experimental results for three pulmonary diseases include several interesting observations which are consistent with previous publications. Furthermore, many of our significant disease mediating pathways were confirmed by a method we developed to systematically evaluate such pathways using alleviating genetic or phenotypic suppression interactions.

Although there are many issues with the data used to produce these results (the conservation of genetic interactions across species being one of the most salient), the trends shown here add systematic evidence to the anecdotal evidence above that this approach can indeed find pathways playing a mediating role in disease processes, potentially leading to the discovery of novel therapeutic interventions. Such an approach may be applied in the context of any disease or phenotype of interest.

Chapter 3

Finding Novel Molecular Connections between Developmental Processes and Disease

3.1 Introduction

The study of the health implications of developmental processes has now entered the genomic era. The recent sequencing of an entire fetal genome [TOP+12] has demonstrated the possibility of applying molecular methods to design novel prenatal diagnostics. The development of therapeutic approaches for personalized fetal treatment of developmental disorders is now on the horizon [Bia12]. Genomic approaches are providing new insights into causes of and possible treatments for such widespread pediatric disorders as asthma [DAB+13] and autism [JYJ+13]. A growing awareness that development may influence lifelong health risk [Bar03, CD11] has led to closer examination of the molecular links between developmental processes and disease at multiple life stages.

Despite considerable progress, our understanding of the molecular etiology

of most complex diseases is still limited. Yet by combining weak signals from multiple genes, we may identify patterns that provide clinically significant insights into disease processes. We hypothesized that by examining the relationships between sets of genes related to specific developmental processes and reported disease genes, we could develop novel insights into developmental impacts on health. To test this hypothesis, we created a novel approach and tool to assess the overrepresentation of various developmental gene sets among groups of genes linked to specific diseases. Our approach derives its strength from combining signals of sets of genes and from pooling disease-gene links across disease subtypes using a hierarchical taxonomy of disease. We demonstrate that this pooling approach improves accuracy over a comparable enrichment-detection method without pooling. Our approach has the advantage of potentially generalizing incomplete disease gene data and overcoming variation in how genes are associated with specific disease terms, improving our ability to detect novel and interesting connections.

We note that a similar principle - that of pooling many weak signals to provide a stronger one - has led to the creation of many highly effective "gene-set analysis" methods for expression data [STM+05b, TGK+05] and genome wide association data [TTS08]. However, these approaches are inappropriate for assessing the overlap of disease-linked genes with genes involved in developmental pathways, because the members of our developmental gene sets cannot meaningfully be ranked by the strength of their participation in the set. Standard statistical enrichment methods such as the hypergeometric distribution might be more suitable, but their probabilities depend on inappropriate assumptions of gene independence [GB07]. Our approach avoids these problems.

The choice of a disease taxonomy for this analysis is vitally important, yet most existing hierarchies lack the molecular focus inherent in the proposed analysis [DHS+11]. We chose the MeSH hierarchy of diseases (category C) because it is widely used, it is relatively compatible with our disease-gene databases, and it represents diseases multiple times within different parts of the tree, thus potentially including somewhat molecularly homogeneous groupings [NTC+91]. For example,

type 1 diabetes mellitus appears multiple times in the taxonomy under categories corresponding to nutritional and metabolic diseases, endocrine disorders, and immune system diseases. The MeSH disease taxonomy can be represented as a "forest" of disease terms (a collection of "trees," in the computational sense [AHU83]), with 26 top-level categories Table 3.1 represented by "disease trees," and more specific disease terms located at increased tree depths.

Table 3.1: **List of 26 top-level categories in the MeSH disease forest**

MeSH Index	Disease Name
C01	Bacterial Infections and Mycoses
C02	Virus Diseases
C03	Parasitic Diseases
C04	Neoplasms
C05	Musculoskeletal Diseases
C06	Digestive System Diseases
C07	Stomatognathic Diseases
C08	Respiratory Tract Diseases
C09	Otorhinolaryngologic Diseases
C10	Nervous System Diseases
C11	Eye Diseases
C12	Male Urogenital Diseases
C13	Female Urogenital Diseases and Pregnancy Complications
C14	Cardiovascular Diseases
C15	Hemic and Lymphatic Diseases
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
C17	Skin and Connective Tissue Diseases
C18	Nutritional and Metabolic Diseases
C19	Endocrine System Diseases
C20	Immune System Diseases
C21	Disorders of Environmental Origin
C22	Animal Diseases
C23	Pathological Conditions, Signs and Symptoms
C24	Occupational Diseases
C25	Chemically-Induced Disorders
C26	Wounds and Injuries

We derive our disease-gene links from two sources: OMIM, a curated collection of genes linked to human disease [ABSH09], and the Genopedia data from the database of Human Genetic Epidemiology (HuGE), whose disease-gene information is obtained primarily by computational literature curation, but includes manual review of both abstracts and index terms [YCKG10]. We then pool genes linked to

descendants of a disease node in the MeSH trees, and we assess significance through permutation. Because of the current incomplete knowledge of gene-disease connections, enrichment of gene sets among genes linked to a specific disease node in the MeSH forest may not be detectable. By pooling gene links from related diseases, we are able to rescue some of these lost connections.

For this study, we focus on identifying connections to genes involved in developmental processes. The gene sets chosen were based on Biological Process terms from the Gene Ontology (GO), a hierarchically-organized collection of controlled-vocabulary functional annotation of genes and gene products [ABB$^+$00]. However, given our interest specifically in developmental gene sets, we chose to use the gene sets from DFLAT, a prior collaboration of ours that aimed to expand human developmental annotation in the Gene Ontology framework [Wic13]. Gene sets derived from the Gene Ontology that include the DFLAT annotation have been shown to improve the interpretability of gene expression data related to human development [WDN$^+$14], so they are a reasonable choice for the analysis described here. We refer to the developmental gene sets whose links to disease are being investigated as the *query* gene sets.

Additional related work assesses significant enrichment of GO functional annotation terms in query gene sets using the directed-acyclic graph structure of the Gene Ontology. Such approaches adjust enrichment calculations by accounting for relationships between the genes at a given annotation node and those at the parent or child [GBRV06, GM08]. But these methods are concerned with a different problem - that of spurious enrichment at higher levels of the GO hierarchy. Instead, the hazard in our case is false negatives that occur because of the incomplete knowledge of disease genes and the variable levels of precision used to map known disease genes to the MeSH forest. We therefore focus here on query sets representing top-level developmental processes (e.g., "heart development" rather than "atrial cardiac muscle cell development"), because highly specific terms typically include very few genes, rendering gene-set analyses powerless. Future efforts will include drilling down into specific developmental pathways. Yet even at this high level,

our analysis identifies both expected links and several unexpected ones, the latter leading to individual novel hypotheses about surprising molecular connections that may affect future disease research.

3.2 Methods

3.2.1 Gene-disease data

We assembled a combined set of disease-gene links for 11,831 genes using 116,117 human gene-disease associations from the Genopedia compendium in the HuGE database of Human Genetic Epidemiology [LCW+06] and 4,813 gene-disease associations from the OMIM database [McK12], both downloaded in November, 2013. Genes from the Genopedia database were mapped to their corresponding disease concepts in the MeSH hierarchy of medical subject headings (http://www.nlm.nih.gov/mesh/), using the Unified Medical Language System (UMLS) [NPH02] as a thesaurus to identify synonymous diseases. To find MeSH terms that best correspond to the OMIM phenotypes, we used the MEDIC merged disease vocabulary, an ongoing toxicogenomics effort to map OMIM disease terms into the MeSH disease hierarchy, downloaded from the Comparative Toxicogenomics Database [DGLH+14] in November, 2013. After removing one copy of the 1,530 duplicate associations found in both data sets, we were left with a total of 119,400 unique associations.

3.2.2 Estimating significance

We estimate the distribution of the expected number of shared genes between the query gene set and the genes associated with a disease under the null hypothesis that there is no meaningful relationship between the query gene set and the disease class. We do so by randomly choosing gene sets of the query-set size from among all the genes in our MeSH tree. This is equivalent to randomly permuting the labels of the genes in the data to determine whether or not they are in the query set. Such permutation leaves the gene-disease connections intact and maintains the complex

correlation structure of genes between related diseases. Assuming that S_N is the observed size of the real overlap at disease node N (i.e., the number of genes in the query gene set that are linked to node N), for each permuted query set we can then determine whether the number of genes at node N in that random query set is larger than S_N. We ran 10,000 permutations to compute a p-value at each node estimating the probability of seeing an overlap of the observed size at that node by chance.

3.2.3 Density of significant enrichment

Density of enrichment was computed between the 9 query gene sets and the 26 top-level MeSH disease categories, each represented by its own tree. Because many diseases are represented multiple times at different places in each tree, we first created a listing of all the unique MeSH disease terms in each tree. If different instances of the same disease in the same tree had different p-values, they were averaged. We then compared the p-values to the chosen significance cutoff of 0.005. The fraction of unique terms in the tree with lower significance was computed. This fraction represents the "density" of significant enrichment of the query gene set in the chosen MeSH category.

To create the heatmap, we z-score normalized the densities across each row (query gene set). To identify expected enrichment, we manually selected the 9 top-level MeSH disease categories thought to be most relevant to the 9 query gene sets (or to many/all developmental gene sets, as in the case of C4 - neoplasms and C16 - congenital, hereditary, and neonatal diseases and disorders).

3.2.4 Comparing the accuracy of the pooling and traditional approaches

We performed the following experiment to compare the accuracy of our proposed pooling approach to a comparable enrichment analysis using only the genes directly associated with a given disease term. To describe the experiment, we first introduce new terminology:

Assume that we are discussing only a single, fixed query gene set. Let G be the set of all gene-disease links in our combined database: $G = \{\langle g, d\rangle \mid$ gene g is associated with disease $d\}$. For any disease node i in the MeSH forest, let $p_{trad}(i, G)$ be the permutation-based significance score for enrichment of the query gene set among genes in G associated with that node using the traditional method (only those genes directly linked to node i). Similarly, let $p_{pool}(i, G)$ be the analogous score for node i under the pooling approach.

Then we will repeatedly randomly withhold some links from G. Specifically, for the jth random iteration, let R_j be a randomly chosen set of 100 $\langle g, d\rangle$ pairs from G, such that g is in the query gene set, and let $G'_j = G \backslash R_j$. We can then partition the disease nodes into those that are more significant under the pooling method (in the jth iteration) and those that are more significant under the traditional method. Formally, let $S_{pool}(j) = \{\text{nodes } i \mid p_{pool}(i, G'_j) < p_{trad}(i, G'_j)\}$, and let $S_{trad}(j) = \{\text{nodes } i \mid p_{pool}(i, G'_j) > p_{trad}(i, G'_j)\}$. (Note that in the many cases where $p_{pool}(i, G'_j) = p_{trad}(i, G'_j)$, the nodes contribute to neither set. Many of these are either leaves, or nodes with no associated genes under either method.)

We say a node i is *supported* by gene-disease link $\langle g, d\rangle$ from R_j if a node corresponding to d appears in the subtree rooted at i. We can then determine the probability that a node in the set $S_{pool}(j)$ or $S_{trad}(j)$ is supported by some link in R_j. Let indicator function $I(i, j) = 1$ if node i is supported by a link in R_j, and 0 otherwise. Then the probability that a node in $S_{pool}(j)$ is supported by R_j is defined as

$$P_{pool}(j) = \frac{\sum_{i \in S_{pool}(j)} I(i, j)}{|S_{pool}(j)|},$$

and $P_{trad}(j)$ is defined analogously, using $S_{trad}(j)$ Finally, we average over all random trials j to compute the averages P_{pool} and P_{trad} that are reported in Table 3.2. Figure 3.1 illustrates the process of calculating $P_{pool}(j)$ and $P_{trad}(j)$ with an example for the j^{th} random trial.

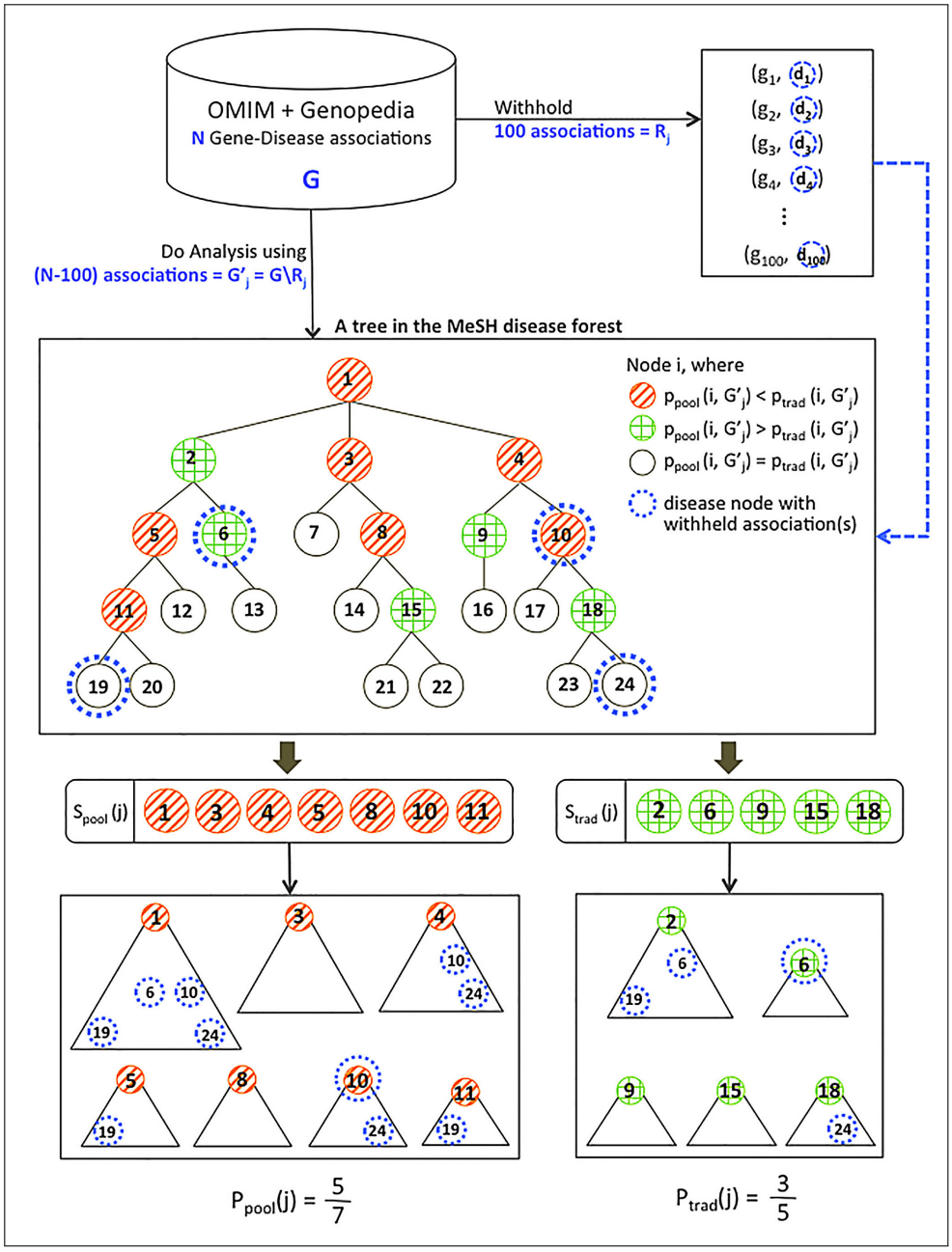

Figure 3.1: **Example of comparison between pooling approach and traditional approach.** Illustration of the process for calculating $P_{pool}(j)$ and $P_{trad}(j)$ for the j^{th} random trial. 100 gene-disease associations involving genes in the query gene set are withheld. Using the remaining associations, p-values for enrichment of the disease gene set at each node are computed using both the traditional and pooling approaches. Nodes are assigned to $S_{pool}(j)$ or $S_{trad}(j)$ based on which approach shows more significant enrichment, and the rate at which each set is supported by withheld links is computed. The idea is that if a disease class is correctly linked to the query gene set, it should be more likely to be supported by withheld gene-disease associations from that same query set.

3.3 Results/Discussion

3.3.1 A new approach linking gene sets and disease classes

To identify significant connections between gene sets and disease, we used a novel method of assessing overlaps between disease genes and the designated query gene sets. We first created a computational representation of the MeSH disease taxonomy in which each node represents a MeSH disease concept. We extracted and combined gene-disease links from the HuGE Genopedia database and from OMIM, and mapped the resulting 119,400 gene-disease links to the MeSH forest (see Methods). Taking advantage of the hierarchical representation of disease concepts in MeSH, we then created a version of the forest in which each disease node D contains any genes in the subtree rooted at D. For example, instead of identifying four lung development genes linked to neural tube defects, two to meningomyelocele, and three to spinal dysraphism, pooling them together identifies seven distinct lung development genes implicated in neural tube defects (Figure 3.2).

For this study we considered nine DFLAT gene sets, broadly representing development in brain, bone, heart, kidney, liver, lung, nerve, blood vessels, and skin. We identified the overlaps between each of these gene sets and the disease genes at each node of our MeSH tree by counting the number of genes in both. Assessing the significance of these overlaps must account for gene set sizes and multiple testing. However, such adjustment is non-trivial because of the complex dependencies between the tests. (For example, any method that assumes the probability of enrichment at node D is independent of the probability of enrichment at D's parent or child is going to be wildly inaccurate.) We therefore use a permutation test (described in the Methods section) to assess the significance of each observed overlap, given the number of genes in the query set and the disease-gene mappings in the MeSH forest. This test produces a p-value at each node estimating the probability of seeing an overlap of the observed size at that node by chance.

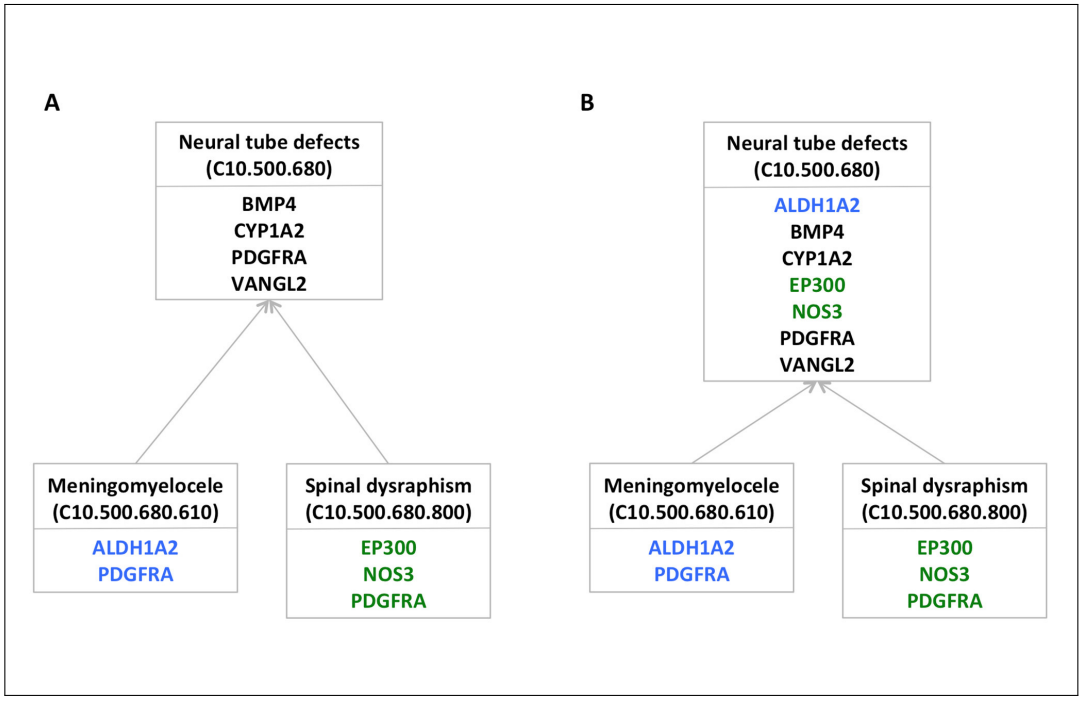

Figure 3.2: **Pooling genes across related diseases to assess enrichment.** a) Lung development genes linked directly to three related MeSH terms. The genes associated with each term are shown in a different color. b) By pooling the lung development genes from the subtree rooted at the *Neural tube defects* node, we obtain enough genes to identify significant enrichment at that node. Colors, the same as those in part a, indicate the disease terms with which the genes were associated before pooling.

3.3.2 Pooling genes from disease subtrees improves accuracy

Our hypothesis was that mapping disease genes to broader disease terms in the MeSH tree as described above would improve our power to detect actual enrichment by mitigating the effects of varying precision in gene annotation. However, it is also possible that pooling might lead to less-accurate results by incorrectly mapping genes to unrelated disease classes. Assessing which happens more frequently is challenging because the right answers are rarely known. Thus, to compare our pooling approach to a more traditional enrichment analysis, we performed the following experiment.

The intuition behind this experiment is that disease classes that are *correctly* linked to the query gene set should be more likely to be supported by withheld data from the same query set. So we use support by withheld data as a rough way to approximate correctness. Our "pooling" approach computes the significance of the query gene set's enrichment at disease node D by pooling data from the genes in the subtree rooted at D. For fairness, we chose (as the "traditional" method) to assess significance of linkage using exactly the same random permutations of gene labels, but counting only the genes *directly* linked to disease node D (rather than those linked to the node or any of its descendants).

We note that the traditional method used here is really just a randomized approximation to the classical hypergeometric calculation, but one that maintains the correlation structure of genes between different diseases. We have separately computed the hypergeometric probabilities (data not shown), and found them to give very similar overall results to those derived using permutation. Accordingly, we present just the permutation-based method, which is the most direct control for our pooling approach, in the comparison below.

We withheld 100 randomly chosen links, each connecting a gene in the query gene set to a specific associated disease. We recomputed enrichment at each disease node without the withheld links, using both the pooling method and the traditional one. Counting then allows us to estimate the probability P_{pool} that a randomly-chosen node found to be more significant under the pooling approach than the

traditional approach would be supported by a randomly withheld link, and P_{trad}, the probability that a node more significant by the traditional method would be. (See Methods for further details.)

Table 3.2: **Advantage of the pooling approach**

Query Gene Set	P_{trad}	P_{pool}
Blood Vessel Development Gene Set	0.0698	0.2428
Bone Development Gene Set	0.1930	0.4574
Brain Development Gene Set	0.1252	0.2887
Heart Development Gene Set	0.0990	0.2781
Kidney Development Gene Set	0.1532	0.3507
Liver Development Gene Set	0.2350	0.5632
Lung Development Gene Set	0.1438	0.3460
Nerve Development Gene Set	0.3296	0.6140
Skin Development Gene Set	0.3007	0.5176

Average probabilities (over 100 trials) that random, withheld gene-disease links support nodes more significant by the traditional method (P_{trad}) or the proposed pooling method (P_{pool}) for the 9 query gene sets. Significance in each trial was computed without the withheld links. When P_{pool} is larger than P_{trad}, the nodes that are more significant under the pooling approach tend to be more consistently supported by the withheld data, our proxy for correctness.

We repeated this experiment with a different set of 100 withheld links 100 times for each of the 9 developmental gene sets. Table 3.2 shows the average values of P_{pool} and P_{trad} for each of the development gene sets, and Figure 3.3 shows histograms of the distribution of $P_{pool} - P_{trad}$ for all of the development gene sets. If P_{pool} is larger than P_{trad} then the nodes that are more significant under the pooling approach tend to be more consistently supported by the withheld data, which is our proxy for correctness. In other words, when P_{pool} is larger, it suggests that the pooling method tends to make correct links appear more significant. For all nine query sets, we found that the averaged P_{pool} is greater than the averaged P_{trad}, suggesting that the pooling method is better able to identify true links between developmental gene sets and disease.

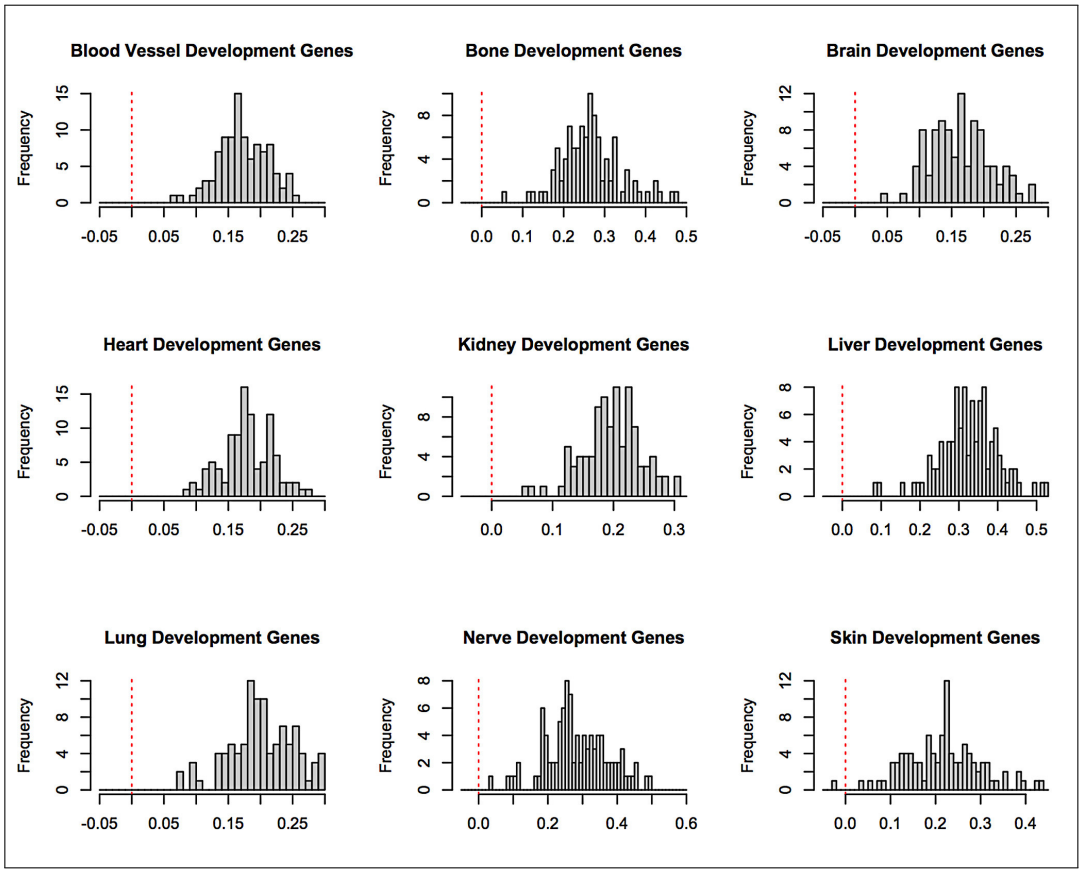

Figure 3.3: **Histogram showing P_{pool}-P_{trad} for each query gene set.** The red lines show a difference of zero; values to the left of these lines represent individual random trials in which the traditional method outperformed the pooling method. This occurred only once, in one trial for the skin development gene set.

3.3.3 A visualization tool for connecting gene sets and disease

While it is relatively easy to provide a list, for each developmental gene set, of MeSH terms whose gene set enrichment p-value is below some cutoff, interpreting those lists is complex. Because enrichment calculations are based on subtrees, there is important information available at different scales, ranging from high-level overviews of the MeSH disease forest to specific enriched gene-disease links, their significance scores, and the genes involved. For these results to lead to new discoveries, we must select from this large collection of significant links a few that are surprising yet plausible. Doing this requires a considerable amount of domain knowledge in molecular medicine.

To facilitate data exploration by collaborators with such expertise, we developed a web-based tool that provides both an abstract and a detailed view of the associations (available at http://gda.cs.tufts.edu/development). For a high-level overview, we visualize each disjoint hierarchy of disease terms (i.e., each tree of the MeSH disease forest) in a simplified triangular form (Figure 3.4). Each significant disease association with the given gene set is represented as a dot in this triangle, whose color represents the degree of significance. This abstract view helps highlight the broad overall patterns of association between development gene sets and disease classes.

Clicking on a particular disease subtree leads to a detailed tree view (Figure 3.5). The tree visualization is implemented using Cytoscape Web [LFK+10]. Color again corresponds to significance, with darker nodes indicating more significant enrichment of the developmental gene set in the disease genes associated with the subtree rooted at that node. For clarity, this view by default only displays disease nodes significantly associated with the query gene set (and their ancestors in the chosen tree). However, users can adjust parameters to view the full tree if desired. Specific genes and p-values for individual links can be identified by selecting nodes in this view. The associated gene lists are easily selected and pasted into functional analysis tools for pathway identification.

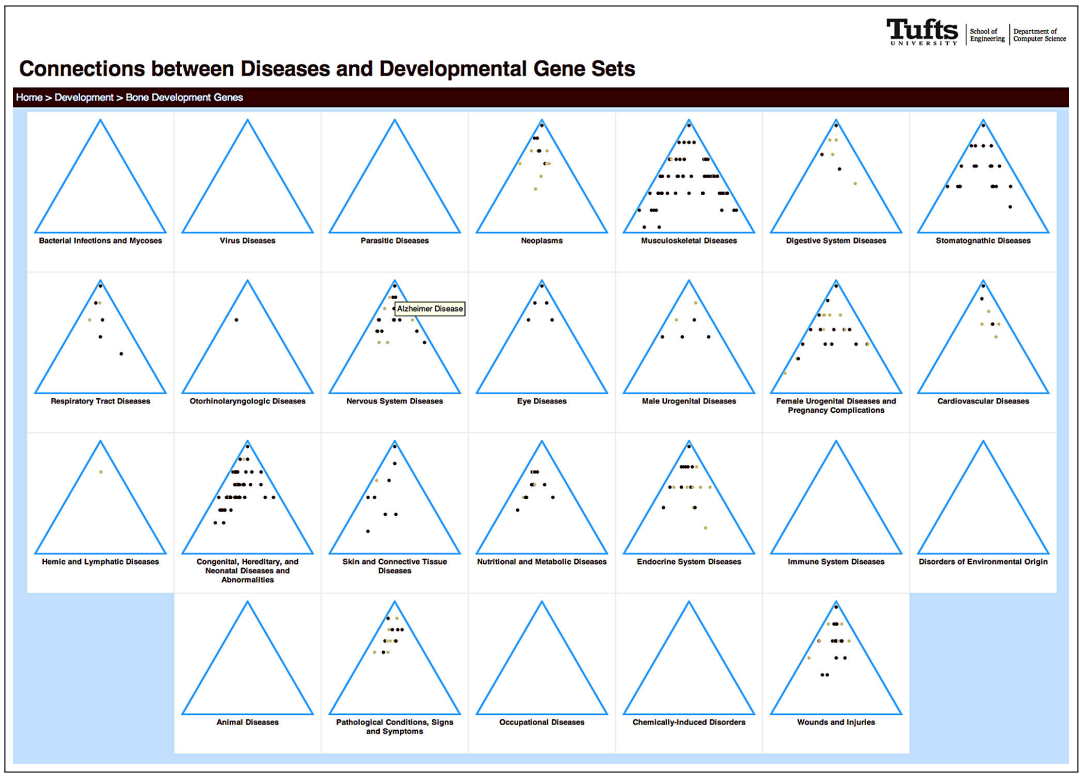

Figure 3.4: **Triangle view of disease enrichment for the bone development gene set.** Each triangle represents one of the 26 top-level categories in the MeSH disease forest. Each dot represents a disease node with significant enrichment of brain development genes. To clearly indicate the significance of relationships between diseases and the query gene set in these small images, we used two colors: light brown dots indicate $p < 0.005$, and darker brown dots, $p < 0.001$. Mousing over the dots reveals a pop-up of the disease term associated with that node (Alzheimer's Disease is shown). Clicking on the category name leads to a detailed view of that tree.

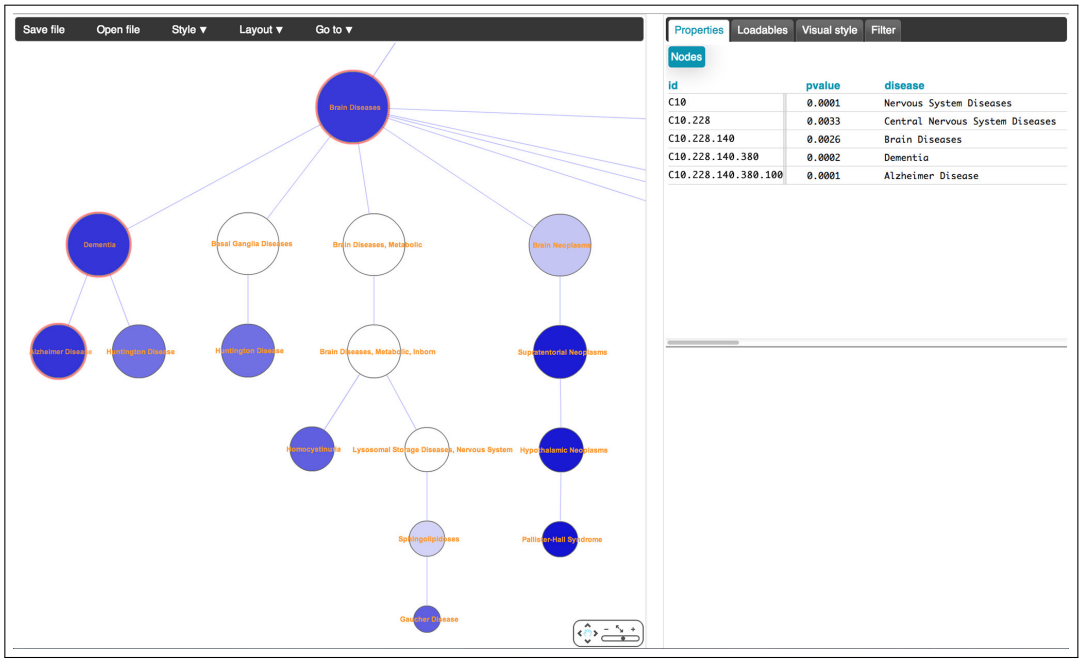

Figure 3.5: **Detailed view of part of the Nervous System Disease subtree, showing enrichment of bone development genes.** Links to dementia and Alzheimer's disease are shown. Significance of each node in the tree is represented by color; a gradient of shades of blue indicates p-values ranging from 0 (darkest blue) to 1.0 (white). Clicking on a node or selecting a set of nodes allows users to see, in the box in the upper right corner, the selected disease terms, p-values, and genes shared between those diseases and the developmental gene set.

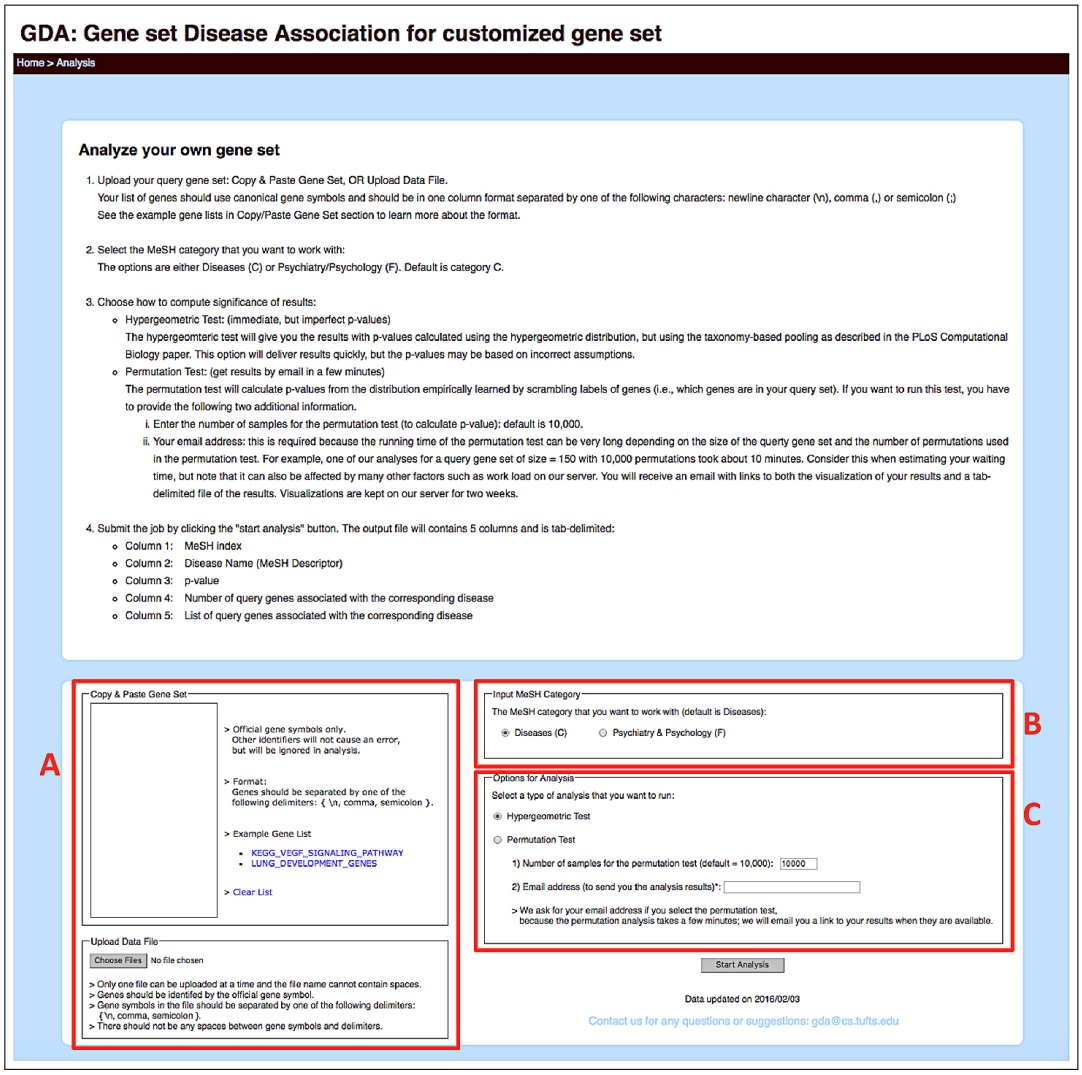

Figure 3.6: **Visualization tool extended for general implications.** The visualization tool allows repeating our analysis for user-defined gene sets. A) Query gene sets can be uploaded as a list or a file. B) The analysis can be done using two different MeSH trees: one rooted at "Diseases" and another rooted at "Psychiatry & Psychology". C) Our analysis uses permutation to assess significance of enriched diseases within user-defined gene sets, but the test might take up to several minutes. An option for approximation using the hypergeometric test is available for faster analysis.

We further extended the web-based visualization tool such that users can repeat the same analysis for their own gene sets of interest (available at http://gda.cs.tufts.edu/analysis). Figure 3.6 shows the front page of the web application. The query gene sets can be uploaded as a list or a file. Users can run the analysis to find enriched diseases, both psychological and non-psychological, within their query gene set. Significance of enriched diseases can be assessed through permutation as our original analysis does, or can be approximated using the hypergeometric test to shorten the processing time. Permutation tests can take up to several minutes, and the analysis results and a link to the corresponding triangular visualization will be sent to users via e-mail. The hypergeometric test is a real-time analysis and web links to the results including visualization will be provided without delay.

In the next two sections, we describe some results from our initial explorations using this tool. The first section provides a sanity-check by demonstrating that we find the broad patterns of connections that one would expect, while the next shows that we can use this approach and the tool described here to make novel but plausible discoveries with potential clinical impact.

3.3.4 Developmental gene sets implicated in expected disease trees

We first take a high-level view of all the results together. Generally speaking, one would expect to see connections between tissue-specific developmental gene sets and broad categories of diseases known to involve those particular tissues. For example, it seems likely that many cardiovascular disorders would be linked to a significant number of heart development genes. Figure 3.7 shows a heatmap of the relative "density" of disease terms significantly linked to each of the gene sets (see Methods) for several MeSH disease trees. We see high enrichment that essentially mirrors our expectations: bone development genes are over-represented in musculoskeletal disorders, brain development genes in nervous system disorders, heart development genes in cardiovascular disorders, etc.

There are a few interesting exceptions. For example, the percentage of ner-

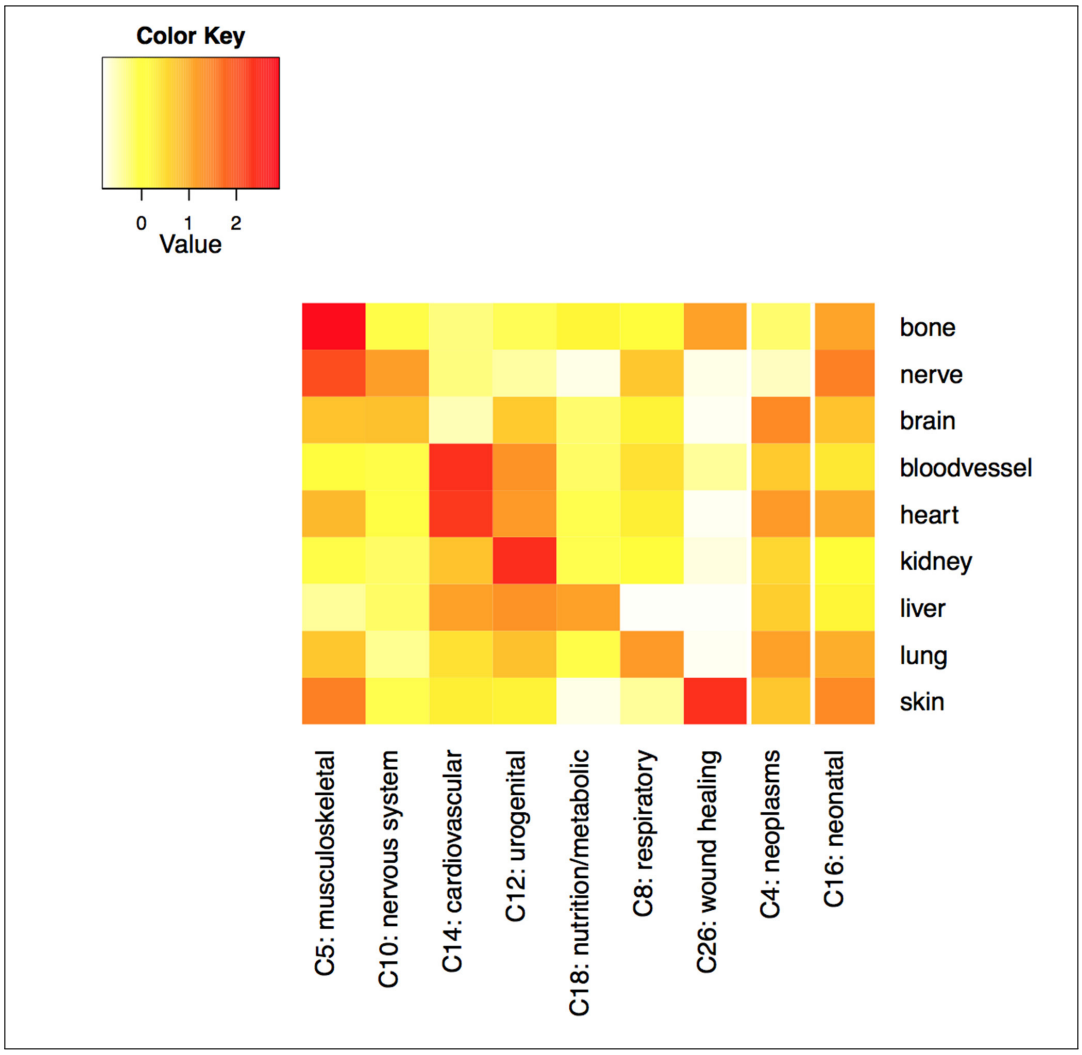

Figure 3.7: **Expected results by tissue.** Density of enrichment of developmental gene sets (labels on the right) in major disease subtrees. Darker squares indicate that a larger fraction of the disease terms in the MeSH category have significant enrichment ($p < 0.005$) of genes in the indicated gene set. Expected connections appear approximately along the diagonal in the first 7 columns, and throughout the rightmost two columns.

vous system disorders significantly enriched for nerve development genes is relatively high, but not quite high as the percentage of musculoskeletal diseases enriched for nerve development genes. This seems to be in part an artifact of the large number of distinct nervous system disorders listed in MeSH category C despite having little or no molecular information, artificially decreasing the normalized density values (the maximum density score in the C10 category is lower than the maximum score in any of the other MeSH disease trees shown in the figure).

The root node of MeSH category C4, "Neoplasms", is significantly associated ($p \leq 0.0001$) with all of the developmental gene sets except for nerve and skin (the two smallest of the gene sets and therefore the least likely to have significant overlaps). This observation reflects the fact that the regulation of cell growth and differentiation that comprise normal developmental processes are typically disrupted and dysregulated during the onset of malignancy [BC07, Moo09]. A range of signaling proteins that play roles in directing both developmental processes and tumorigenesis are likely to blame for these interactions [DPW06, AIS09, SD99]. However, the specific signaling processes implicated in the different tumor types, as well as those known to be involved in developmental processes but not yet implicated in specific tumor types, may be of interest.

Similarly, given that the query gene sets are all involved in developmental processes, it is not surprising that the C16 MeSH subtree, described as "Congenital, Hereditary, and Neonatal Diseases and Abnormalities," shows significant enrichment at the root node ($p \leq 0.0001$) for all of the tested developmental gene sets. A wide range of molecular developmental processes are implicated in this MeSH category. The density measurement shown in Figure 3.7 provides a broader way of assessing a similar property. The density measure for the C16 tree is above average (i.e., the z-score normalized density metric is positive) for each of the nine gene sets considered here.

By confirming that we find expected and reasonable high-level results, the observations in this section provide evidence of the efficacy of our approach.

3.3.5 Unexpected connections and implications

Delving more closely into specific results, we identified several findings that seemed, at first glance, less predictable than those described above. Here we describe three such links. All of them identified surprising connections that, since our initial discovery of them using this approach, have been further supported by new publications.

3.3.5.1 Bone development and dementia

One surprising link is a significant overlap ($p = 0.0002$) between bone development genes and genes involved in dementia (MeSH term C10.228.140.380). There are 24 genes involved in this overlap. One might suspect that the connection would be through BMP signaling proteins, which play developmental roles in a variety of processes including bone formation and neurogenesis. Yet although *BMP4* is among the 24 genes, it is the only BMP family member on the list. Functional analysis (in DAVID [DSH$^+$03], v6.7) of the linking gene set indicates enrichment of a broader set of proteins involved in bone morphogenesis (*COL1A1, COL13A1, HSPG2, PEX7*, and *RUNX2*). This is to be expected in a subset of genes involved in bone development. Yet we also saw enrichment of retinoic acid receptor proteins (*RARA, RARB, RARG*) and heparin-binding proteins (*BMP4, COMP, COL13A1, FGFR2*). These links are of interest because both heparin derivatives and retinoic acid are candidates for new Alzheimer's therapies [SFNY09, PEYT08], yet both are also known to contribute to osteoporosis [SL73, LLW$^+$03].

Evidence supporting this connection has recently been proposed in empirical observations of an association between lower bone mineral density and dementia in postmenopausal women [LNS$^+$12]. Although molecular pathways supporting this link were not identified, a role for estrogen deficiency was suggested. Our observations are consistent with this hypothesis – five of the 24 shared genes (*ALPL, BMP4, COL1A1, GH1,*and *RARA*) are among those with the GO Biological Process annotation "response to steroid hormone stimulus," a finding whose adjusted false discovery rate (as computed in DAVID via the Benjamini-Hochberg method)

is below 0.015.

This analysis also suggests a possible connection between dementia and bone density through additional signaling pathways. For example, the growth factor *MDK*, a still relatively unstudied, retinoic acid-responsive, heparin-binding protein appears to be involved in both neuron and bone growth [MMMT95]. Elevated levels have been observed in serum from Alzheimer's patients [SMS+05]. Our observations suggest that molecular connections through this and related signaling pathways may be worth exploring in the quest for novel therapeutic approaches to dementia.

3.3.5.2 Heart development and polycystic ovary syndrome

The link between heart development and polycystic ovary syndrome (PCOS; MESH term C19.391.630.580.765) has a p-value below 0.0001. PCOS is an endocrine disorder that causes hormonal changes, ovarian "cysts" (that are actually immature follicles), and decreased female fertility. It has been associated with an increased risk of diabetes, dyslipidemia, and cardiovascular disease [Dok13]. There are 31 genes responsible for the connection we observed between PCOS and heart development. Functional analysis of this gene list shows enrichment of genes annotated with the GO Molecular Function term "SMAD binding" and those in the KEGG "TGF-beta signaling" pathway. TGF-beta ($TGFB$) is the canonical member of a family of cytokines that play regulatory roles in many developmental, homeostatic, and immune processes. It regulates apoptotic pathways, in part through SMAD binding [tDH04].

It has long been known that cardiovascular symptoms are associated with PCOS, but the molecular etiology of this connection is not clear. One study proposed that oxidative stress caused by insulin resistance may lead to cardiovascular injury in nonobese PCOS patients, but did not implicate specific molecular pathways [MSL+11]. Given that *TGFB/SMAD* complexes are known to mediate the DNA damage response [WSH+13, HKBH12], dysregulation of $TGFB$ is a possible mechanism to be considered.

A role for $TGFB$ in PCOS through mutations in fibrillin 3, a gene linked

to PCOS, has also recently been suggested. Fibrillin 3 expression changes in fetal ovaries of PCOS patients have been shown to affect *TGFB* binding, perhaps leading to changes in follicle formation [HBIR+11]. The same paper suggested that the PCOS phenotype was consistent with increased *TGFB* activity. A more direct role for *TGFB* itself in PCOS was recently proposed, despite largely circumstantial evidence [RKURL13]. Our analysis also seems to support this hypothesis, which potentially explains both the observed cardiovascular outcomes and the early developmental origins of ovarian cysts in PCOS.

We looked for further corroborating evidence in mouse models, but could not identify an existing, well-characterized mutant that is a good model of *TGFB upregulation*. However, there are four genes in the KEGG TGFb pathway that are known to inhibit *TGFB* activity: *LTBP1, DCN, Lefty,* and *Activin*. For all of these genes there are mouse mutant strains (in a variety of backgrounds) that disrupt the homologous proteins' expression, thus potentially upregulating *TGFB*. These mutant strains have differing degrees of phenotypic characterization, but two of them (*DCN* and *Activin*) have knockout mutations that cause reduced female fertility [PJT+07, MLCTIL11], and the activin knockouts are even characterized as having ovarian cysts. These findings are unexpected by chance: the hypergeometric probability of seeing at least two of four randomly-selected proteins whose knockout strains are characterized as having reduced female fertility in the Mouse Genome Database is below 0.0001, as is the probability of seeing at least one of four with an ovarian cyst phenotype. We therefore suggest that further work on the role of the *TGFB* pathway in the development of PCOS may prove fruitful.

3.3.5.3 Lung development and retinopathy of prematurity

The lung development gene set was linked, with a p-value below 0.0001, to retinopathy of prematurity (MeSH term C16.614.521.731). Retinopathy of prematurity (ROP) is a complication that occurs primarily in infants delivered before approximately 28 weeks' gestational age, before the infants' visual system has been fully formed [SD08]. While early detection and treatment often lead to a full recovery,

severe cases may lead to permanent nearsightedness or vision loss [ABB+12]. Yet we still know too little about why some infants develop this complication of prematurity, while others born at the same age and with similar clinical characteristics do not. Although most neonates with retinopathy of prematurity also have immature lungs, a molecular connection between ROP and lung development is not readily apparent.

The significant connection we observed was based on five genes linked to both ROP and lung development: *IGF1, NOS3, EPAS1, KDR,* and *VEGFA*. These genes are all related to blood vessel or tube development. Excessive but disordered VEGF-mediated vascularization of the retina is known as the cause of ROP [SC04], and indeed ROP has been successfully treated in recent pilot studies by intravitreal administration of the VEGF inhibitor bevacizumab [MHB09]. It appears likely that these genes may be playing a specific role in alveolar development and lung function.

We therefore hypothesized that there might be similar molecular enrichment of the ROP genes in bronchopulmonary dysplasia (BPD), another complication of prematurity characterized by extended need for supplemental oxygen and, in extreme cases, long-term respiratory insufficiency. Like ROP, BPD also affects some, but not all, extremely premature infants. Its exact cause is unknown. Current hypotheses include one in which inflammation plays a major role [WDSM96], as well as the so-called vascular hypothesis of BPD in which decreased vascularization impairs alveolar formation at a critical time [SA05].

Our observations are consistent with the hypothesis that both complications may be caused in part by perturbations of the *VEGF* pathway (Figure 3.8), which provides a molecular link between the vascular and inflammatory BPD hypotheses. In support of this theory, we observe that 7 of the 26 ROP genes are among the 49 BPD genes listed in our disease-gene collection, an overlap that would occur by chance with a hypergeometric probability of $\approx 8 \times 10^{-12}$. We also wondered whether the two disorders tend to occur in the same infants more often than expected by chance. This question was recently answered in the affirmative, in a paper that did not identify the cause of such clustering but hypothesized that both ROP and BPD

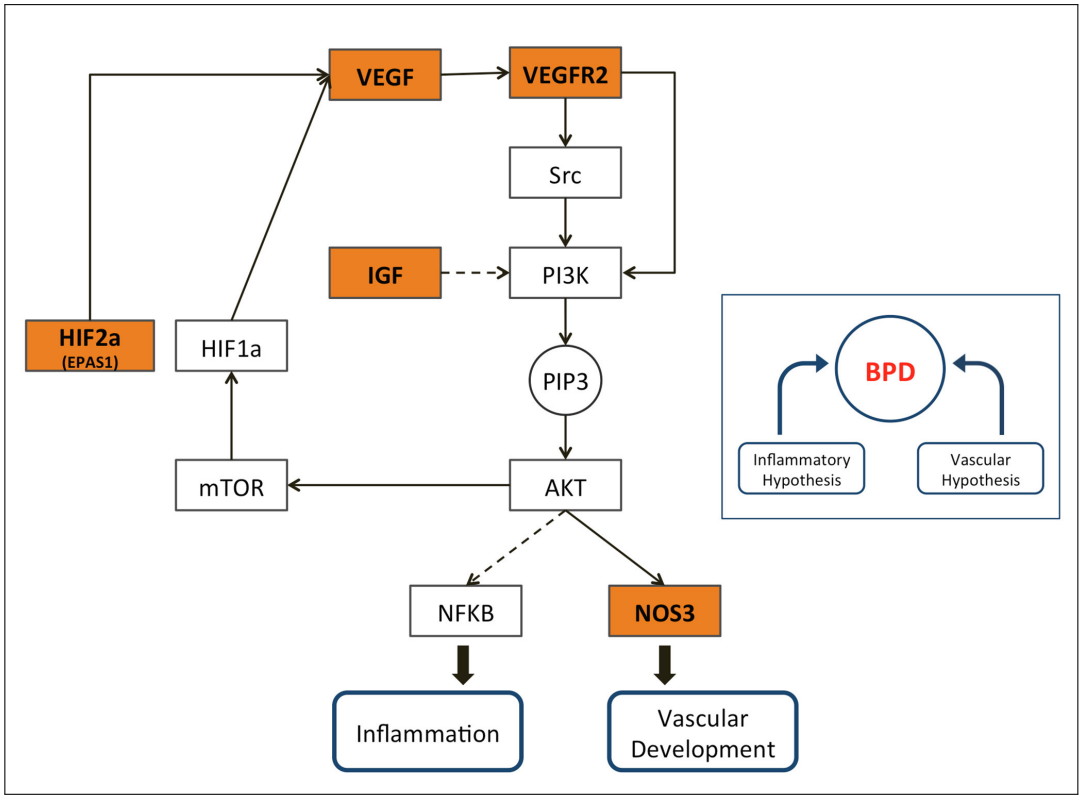

Figure 3.8: **The VEGF pathway and its relevance to both BPD hypotheses.** The relationships shown here are derived from the VEGF, PI3K-AKT, mTOR, and HIF-1 signaling pathways and the "Pathways in Cancer" map in the KEGG Pathway database. Dashed lines represent indirect regulation. Genes highlighted in orange are the five lung development genes implicated in ROP.

might be consequences of NICU-acquired infection [LDE+10]. More recent work shows an association between low VEGF protein levels in urine and the eventual development of both ROP and BPD [LKW+13], suggesting a non-invasive early approach to predict both outcomes.

However, the picture appears to be more complex than this. If both ROP and BPD were associated with uniformly low VEGF activity, then the current practice of treating ROP with VEGF *inhibitors* presumably would not have arisen. BPD, on the other hand, does appear to be associated with reduced *VEGF* expression levels in the lungs of human neonates [BPH+01], and administration of VEGF ameliorates symptoms of respiratory distress in a mouse model of BPD caused by inhibition of the *VEGF* pathway [CBA+02]. It is possible that an interaction between VEGF levels and NFKB-mediated response to neonatal infection accounts for the observed co-occurrence of these disorders.

There is also a possible connection to adult pulmonary disease [VVT06]. Drug-induced suppression of VEGF has been used to create a model of emphysema in adult rats [CTSS+03]. A prior analysis of gene expression in BPD implicated chromatin remodeling and histone acetylation pathways [CMS+07]. We therefore investigated possible molecular links between BPD and chronic obstructive pulmonary disease (COPD), in which histone acetylation also plays a role [MTM+11, MTT+11]. We observed that 26 of the 49 BPD genes in our data set are among the genes linked to COPD (an event that would occur under the null hypothesis with hypergeometric probability $< 10^{-15}$). It is difficult to directly assess co-occurrence of BPD and COPD in the same individuals, because COPD typically affects older patients, most of whom were born before the neonatal diagnosis of BPD was formally defined in 1967 [NRP67]. However, the observation of early-onset emphysema (a form of COPD) in BPD survivors [WLL+08] supports our theory that this molecular overlap does not occur by chance. Our evidence leads to the hypothesis that there might be a shared molecular mechanism predisposing individuals to an excessive response to alveolar damage, whether the damage is caused by oxidative stress from neonatal ventilation, or cigarette smoke in adults.

3.4 Conclusions and future work

We have introduced a new approach that identifies significant overlap of gene sets with groups of related diseases in a hierarchical disease taxonomy. To evaluate this approach, we implemented a tool that allows users to explore connections between disease subtrees in MeSH and several developmental gene sets. Our observations in this analysis have helped identify surprising molecular connections between disparate processes. They have also more generally served to validate the approach of pooling incomplete information about disease genes across related disorders to strengthen our ability to identify such connections. With a growing interest in research into the developmental origins of adult disease, this resource should prove a valuable source of information for generating hypotheses about such connections at the molecular level.

Our work has assumed only that query gene sets are lists of genes that share some common property [TVM+11]. However, for this study we have chosen query sets whose genes share common annotations in the Gene Ontology. An interesting future direction would be to consider the possibility of creating hierarchically-structured queries representing related query terms in the Gene Ontology's directed acyclic graph structure, while still looking for significant links to disease classes or subtrees in the MeSH forest.

While our implementation relies on a particular set of disease-gene information and a small group of developmental gene sets, the power of the approach will be best exploited by the inclusion of a more comprehensive set of disease-associated genes. One key limitation of the current approach is due to the nature of the available data linking genes to diseases. OMIM is an excellent resource created largely by computer-assisted manual review of the literature [ABSH09]. However, it is limited in scope and is curated by locus rather than by disease, so that even identifying all genes related to, for example, type 2 diabetes, can be complicated. Conversely, the HuGE database, which provides the majority of the disease-gene data used in this project, derives most of its information from computational screening of PubMed

(along with some manual review) [LCW+06, YCD+08]. This raises the possibility that, in addition to being incomplete, our gene-disease database may include a substantial number of false positives due not only to false-positive experimental results but also to inappropriate interpretation of the text. There is prior work on reducing the rate of false positives when mining such information from the literature [GUT+07], and the HuGE database creators worked to assess and improve accuracy [YCD+08], but any data set derived from computational literature analysis will always have this concern. On the other hand, the success of our initial analysis in identifying expected connections suggests that false positives are so far not interfering significantly with the use of this tool for discovery. Further improving the quality of the data and characterizing the impact of different types of noise on the results will be an important area to investigate in the future.

Finally, we note that while there are many disease taxonomies that are widely used for different purposes, there is growing dissatisfaction with most of them, in part because of the lack of a molecular representation of disease relationships [DHS+11]. A need for development of molecular disease taxonomies is consequently recognized because such disease taxonomies would improve analyses like ours and eventually lead to better support for molecular medicine. Specifically, our disease gene pooling process will gain more strength if it is based on molecularly defined relationships between diseases. Therefore, we initiated a study where we attempt to build a hierarchical structure of disease only using disease-gene association information. The following chapter will discuss this study in greater detail.

Chapter 4

Towards a More Molecular Taxonomy of Disease

4.1 Introduction

The recent growth in availability of genomic and clinical data allows for the discovery of new molecular-level mechanistic models of disease. However, existing disease taxonomies and ontologies are often focused on either physiological characterizations of disease, sometimes using decades-old criteria, or on the organizational and billing needs of hospital. Automatically inferring common molecular links between related diseases is made more difficult by the limited molecular representation in current taxonomies [PWK+14], leading some researchers to manually group related disorders for individual projects (for example, PheWAS analysis [DRB+10] or network-based disease gene prioritization [KBHR08]). Yet such manual efforts limit consistency and reproducibility. To further advance such research and biomedical knowledge in the genomic era, a recent National Academy of Sciences working group has called for the development of new disease taxonomies better suited to incorporate molecular information [DHS+11].

A truly modern taxonomy would presumably combine clinical, physiological, and molecular data. The question we address here is the degree to which we can

infer a meaningful disease taxonomy simply using disease gene information. In this, we were inspired by efforts by Trey Ideker's group to infer a version of the Gene Ontology using pairwise similarity scores between genes [KDY+14, DKS+13]. Their CliXO algorithm, for example, sorts gene pairs by a pairwise similarity score and incrementally uses these scores to group together cliques of similar genes. The resulting ontology forms a Directed Acyclic Graph (DAG) of sets of genes. As in that work, here we are *not* arguing that we should ultimately construct a disease hierarchy automatically in this way. However, learning how we can discover the relationships in existing disease taxonomies from disease gene data is a first step towards developing new hierarchies of disease that integrate the clinical information used in today's taxonomies with genomic data. Such integrated taxonomies are needed to better support research in molecular medicine [KB11].

To infer a disease taxonomy, we would like to simply cluster diseases hierarchically based on associated genes from a large gene-disease database. However, if the items we are clustering are diseases, the internal nodes of any hierarchical clustering method will correspond to unnamed sets of diseases. While some of these may be informative, identifying them is a challenge. We therefore introduce here an algorithm called Parent Promotion, based on hierarchical clustering, that addresses this problem.

We acknowledge that we are deliberately blurring the distinction here between an ontology of disease [SAR+07] and a disease taxonomy [Lam07]. In this manuscript, we focus on learning a hierarchical characterization of disease using *existing* disease terminology yet incorporating molecular relationships. Such a description may be able to better identify novel relationships between disorders that do not appear clinically similar but that arise from similar underlying genotypes. Yet we are not expecting here to comprehensively infer disease relationships as in most ontologies, in part because the current project ignores the clinical and anatomical characteristics built into many existing taxonomies. Accordingly, we frequently use the term "disease hierarchy" to encompass our inferred hierarchies as well as those to which we compare.

One important question is how to evaluate our inferred hierarchies of disease when there is no existing gold standard. However, there are a handful of existing taxonomies and disease ontologies that are somewhat suitable for molecular analyses and comparisons [DHS+11]. MeSH is a hierarchical structure of controlled biological vocabularies used to index articles in MEDLINE [LB94]. MeSH includes many medical concepts beyond diseases, but here we refer to MeSH category C, a comprehensive set of 26 trees that represent relationships between diseases. SNOMED-CT provides an organized terminology for clinical terms [WSS02]; this is one of the most detailed terminologies available, but there are restrictions on its distribution. The Unified Medical Language System (UMLS) metathesaurus includes disease terms from multiple taxonomies; while it is not intended to be an ontology, its semantic network can identify some relationships between terms [Bod04]. The Disease Ontology (DO) also integrates the knowledge and relationships from several taxonomies, including MeSH, SNOMED-CT, and ICD [SAN+11a].

Initially, because of the high coverage and availability of MeSH and its simple structure, we chose to compare our inferred hierarchies to the MeSH forest of disease terms. Although it is not necessarily a gold standard for the problem we are trying to solve, we can use such a comparison to identify the strengths and limitations of different inference methods. In addition, identifying individual MeSH disease trees that are more consistent with the hierarchies inferred from disease-gene data helps in assessing the molecular content of existing domains in MeSH. We have also extended our assessments by comparison to the Disease Ontology, which is a more complex process for reasons detailed below.

Even after fixing a "reference" hierarchy for comparison, the question of how to assess correctness remains. Many of the standard network and graph comparison metrics are inappropriate for our problem. Of these, we use a strict variant of edge correctness [SXB08] that asks how many parent-child relationships we get right.

One limitation of edge correctness, however, is that the distances between pairs of terms are not uniform [SBIV12]. That is, two diseases that are separated by more than one taxonomic link may be more closely related to each other than two

other diseases in a direct parent-child relationship. We therefore also introduce the notion of ancestor correctness, a feature-based similarity measurement [PVHR06] that assesses our ability to properly identify ancestry without concern about distances.

Finally, neither edge correctness nor ancestor correctness penalizes an algorithm for false positives (inferred edges not in the reference hierarchy). This is fine for inference methods like Parent Promotion that build trees, which all have the same number of edges for a fixed set of disease nodes, but not for comparison to ontology-learning approaches that can add arbitrary numbers of edges. Accordingly, we also compute a variation of hierarchical precision and recall [VCSM06], analagous to ancestor correctness, that accounts for both false positives and false negatives.

4.2 Methods and Materials

4.2.1 Reference taxonomies

Table 4.1: **Subproblems of the Disease Ontology**

Root Disease	**#Diseases (Nodes)**	**#Edges**	**#Nodes with 1 parent**	**#Nodes with 2 parents**	**#Nodes with 3 parents**
Disease	2,039	2,095	1,982	55	1
Cardiovascular Disease	141	141	139	1	0
Gastrointestinal Disease	115	118	110	4	0
Musculoskeletal Disease	133	135	129	3	0
Nervous System Disease	308	324	291	15	1

The entire Disease Ontology (root = "Disease") and four subproblems of various sizes extracted from it. The original DO and its subproblems are tree-like: 1) the numbers of edges are close to $n - 1$ while n is the number of nodes and 2) small fraction of nodes have 2 or more parents.

To quantify performance of various disease hierarchy inference methods, we compare our inferred taxonomies to the 2016 Medical Subject Headings (MeSH)

disease trees [LB94] and the Disease Ontology (DO) [SAN+11b], downloaded on August 5, 2016. From both datasets, we exclude diseases for which we cannot find any associated genes, because our methods would then have no way to learn about how they relate to other diseases. However, excluding diseases can disconnect our reference hierarchies. To reconnect them, we therefore add edges from a deleted node's parents to all of its closest descendants that do have associated genes.

We note that the MeSH trees allow repeated disease names, resulting in multiple nodes with the same name in different parts of the tree. We treat these terms as if they were the same node, effectively matching against the corresponding DAG. However, given that the original structure is a tree, most of these DAGs end up being fairly tree-like.

Because the Disease Ontology is substantially larger than any of the individual MeSH trees, we extracted smaller DAGs from the full DO to facilitate algorithm comparison. To find these smaller DAGs, we searched through the DO starting at the most general term. A term became a root of a DO subset if its name approximately corresponded to the name of the root of one of the 26 MeSH trees and if it had at least 100 DO terms as descendants. This approach identified four new DAGs that can be described as covering mostly "Cardiovasular Disease," "Gastrointestinal Disease," "Musculoskeletal Disease," and "Nervous System Disease."

Table 4.1 reports the sizes and topology of these four subsets of the DO. All are fairly tree-like; only small numbers of nodes have more than one parent, and the total number of edges is not that much larger than the number of nodes. We note that it is not necessarily the case that all disease nodes in the DAG labeled Musculoskeletal Disease, for example, actually correspond to musculoskeletal disorders, because the Disease Ontology and MeSH are organized according to different principles. We therefore acknowledge that each subset of the DO may contain terms that map to several different MeSH disease trees. Nonetheless, we use these labels as shorthand ways to refer to the chosen DO subgraphs.

Table 4.2: **Four MeSH subtrees of various sizes used for method development.**

Root Disease	**#Diseases (Nodes)**	**#Edges**
Infant, Premature, Diseases	6	5
Dementia	13	12
Respiration Disorders	23	22
Eye Diseases	149	178

4.2.2 Withheld MeSH subtrees for method development

We selected four small subtrees from MeSH that we used for refining our computational methods. These are the MeSH subtrees rooted at the terms "Infant Premature Diseases," "Dementia," "Respiration Disorders," and "Eye Diseases," giving us a range of subtrees of different sizes and complexity (Table 4.2). Note that the MeSH tree rooted at "Eye Diseases" includes 149 disease terms and 178 edges, indicating that several terms appear multiple times, although we allow a node with a given name to appear only once in each inferred hierarchy.

Although we show the performance of the inference methods on these subtrees in a separate table in the Supplemental Material, we did not think it fair to include them in our overall MeSH results because we used them to tune our methods. Accordingly, we removed the subtrees rooted at these nodes from the relevant disease trees in MeSH before evaluating the different methods' performance. Only one whole disease tree, C11 ("Eye Diseases"), was removed, because the entire C11 tree was used for method development.

There are two other MeSH disease trees that were also removed before evaluation: C21, "Diseases of Environmental Origin," which included only 3 diseases with associated genes, and C22, "Animal Diseases," which contained no diseases with associated genes. We therefore report averaged MeSH results over the remaining 23 MeSH disease categories.

4.2.3 Disease genes

We use disease genes to calculate pairwise similarity of diseases. For our comparison to MeSH, we gathered disease-gene associations from the Online Mendelian Inheritance in Man (OMIM) database [McK12] and the Genopedia compendium in the HuGE database of Human Genetic Epidemiology [LCW+06], both downloaded on February 3rd, 2016. We combined disease-gene associations from the two databases as in our previous work [PWK+14], using the MEDIC merged disease vocabulary (downloaded from the Comparative Toxicogenomics Database [DGLH+14] on February 3rd, 2016). This combined data set contains 2,755 diseases and 12,873 genes.

To infer hierarchies based on DO terms with this disease-gene data, however, required converting the MeSH disease terms to DO terms. The DO obo file provides synonym information for this conversion. However, because not every MeSH term has a DO equivalent, nor vice-versa, the mapped disease gene data set included 1,790 DO terms with 12,230 associated genes. The Disease Ontology actually includes 6,932 disease nodes, so the resulting DAG of diseases with associated genes was largely disconnected.

We therefore augmented the disease gene data with disease-gene associations from the DISEASES database [PFSP+15] (downloaded on August 5th, 2016) which directly uses DO terms. We used the filtered version of the DISEASES database which provides non-redundant disease-gene association pairs, and selected only associations derived from experiments or database curation ("knowledge"), which we expect to be of relatively high confidence. The DISEASES data included 772 disease terms and 13,059 genes. When combined with the mapped data from the MeSH comparison, the total yielded 2,039 DO terms with 16,404 associated genes, producing a sufficiently connected ontology for our purposes.

4.2.4 Measuring pairwise similarity

For our inference algorithms we need methods to measure similarities both between pairs of diseases and between pairs of genes. To calculate pairwise similarity between diseases A and B, we use the Jaccard Index [Jac01] as follows:

$$Jaccard(G_A, G_B) = \frac{|G_A \cap G_B|}{|G_A \cup G_B|} \tag{4.1}$$

where G_A is the set of associated genes for disease A and G_B is the set of associated genes for disease B.

To calculate pairwise similarity between genes g_1 and g_2, we do the opposite, as we are interested in measuring the similarity of diseases with respect to their associated genes:

$$Jaccard(D_{g_1}, D_{g_2}) = \frac{|D_{g_1} \cap D_{g_2}|}{|D_{g_1} \cup D_{g_2}|} \tag{4.2}$$

where D_{g_1} is the set of diseases associated with gene g_1 and D_{g_2} is the set of diseases associated with gene g_2.

Note that no information about the relationships between diseases other than this measure of overlapping disease genes is incorporated into this similarity matrix or used by our inference algorithms.

4.2.5 Inference strategies

4.2.5.1 Clique Extracted Ontology (CliXO)

To use CliXO to generate disease ontologies, we begin by creating a matrix containing the Jaccard similarity score between genes as defined above. CliXO uses this similarity matrix as input. It also relies on two parameters: α, which represents the amount of noise allowed in forming cliques, and β, which represents missing data. The algorithm is demonstrated to be relatively robust to variation in β, so we set $\beta = 0.5$ as done by the CliXO team [KDY+14]. Variation in α has higher impact on the results, so tuning it to the data set is suggested. We chose $\alpha = 0.05$ because

it produced reasonable-sized output graphs in our initial experiments on the four MeSH subtrees in Table 4.2.

Initially, CliXO returns a DAG whose internal nodes correspond to sets of genes, not to specific disease terms in the reference ontology. We then used the ontology alignment technique of [DKS+13] to align the resulting ontology to the MeSH reference or to the Disease Ontology, in order to identify disease terms in the output DAG. Accordingly, some of the disease terms may not be represented in the CliXO output, because they fail to map to any node. (Figure 4.2 demonstrates the topological difference for a small example; note that the CliXO output on the right maps only 5 of the 6 disease nodes.)

4.2.5.2 Parent Promotion

We introduce a new technique we call Parent Promotion that focuses on similarities in disease genes. The idea is to group diseases by their similarity scores and use hierarchical clustering to form subgroups. Parent-child relations are then created from these subgroups by counting citation frequency in PubMed.

Specifically, we transform the pairwise similarity score into a distance by subtracting it from 1. We then perform complete-linkage hierarchical clustering on the disease terms using the `hclust` function in R with these distances. Internal nodes in this dendrogram correspond to sets of diseases. To convert the resulting dendrogram to a hierarchy with a single disease at each node, we identify the number of disease-related articles in PubMed for each disease in a cluster using the NCBI's E-utilities (http://www.ncbi.nlm.nih.gov/books/NBK25501/).

Working up from the bottom of the dendrogram, the disease term with the most citations is promoted to become the parent, with all other diseases in the cluster left as its children. Once defined as a child, a disease does not have another chance to be promoted. That is, we only consider the most recently promoted disease and its siblings in a cluster when deciding the next parent. Figure 4.1 shows an example of how the dendrogram guides the Parent Promotion process.

Notice that the inferred tree created by the Parent Promotion technique

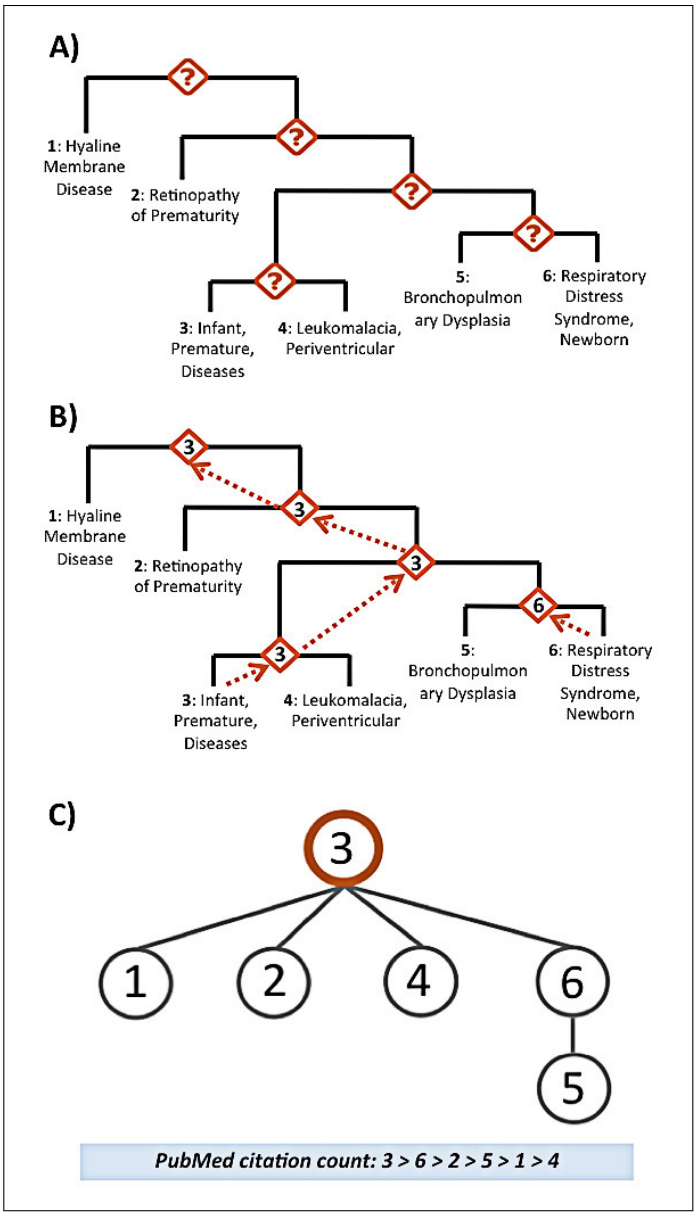

Figure 4.1: **How the Parent Promotion method transforms a dendrogram created by hierarchical clustering.** A) Dendrogram for premature birth complications. Hierarchical clustering builds a tree whose internal nodes are hard to interpret. B) Parent Promotion finds the most general disease term from each cluster and promotes it as an internal node. An internal node becomes the parent of all other nodes in the same cluster. Disease term 3 has the most citations and keeps being selected for promotion until it becomes the root. Disease term 6 has more citations than 5 and is promoted as the parent of 5. However, it later becomes a child of 3 because it has fewer citations than 3. C) Final tree built by Parent Promotion.

always has the same number of diseases (nodes) as the reference. However, the number of edges may differ from that of the reference, which may be either implicitly or explicitly a DAG. In either case, Parent Promotion may therefore produce a result with fewer edges.

4.2.5.3 Minimum Weight Spanning Tree

We also compared our new Parent Promotion method to the standard technique of finding a Minimum Weight Spanning Tree (MWST) [CLRS09] over the complete network of disease terms, with pairwise similarity scores between diseases as edge weights. The idea behind this is that a representation of the relationships between diseases that connects all the disease terms by their highest disease gene similarity represents a minimum-length description of the data that seems likely to capture real disease relationships. The MWST is unrooted, so we choose the disease with the most related PubMED articles as the root.

4.2.6 Evalution metrics

Comparing the inference methods remains challenging due to the topological differences of the output. In particular, both Parent Promotion and MWST produce trees whose n nodes are exactly those of the reference hierarchy. In contrast, the DAG output by the CliXO method may be much larger (as in Figure 4.2). We use multiple methods to quantify and compare performance despite these differences.

4.2.6.1 Edge Correctness (EC)

Inspired by the notion of edge correctness (EC) used in network alignment [SXB08] we measure the number of edges that are identical to those in the reference hierarchy. Unlike in the network alignment problem, which uses edge correctness as a proxy for node correctness, for this problem we know the node correctness and wish to measure correctly inferred edges. We count edges as correctly matched if and only if the parent child relations (both the edges and the directions of the edges) are

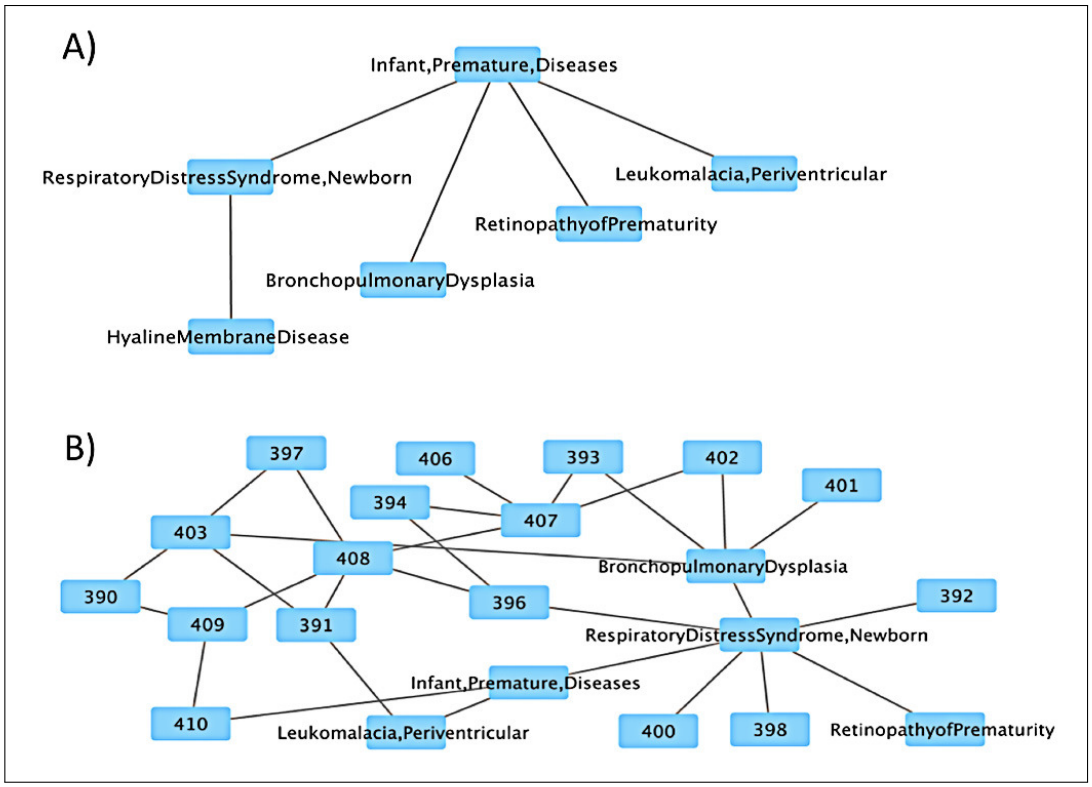

Figure 4.2: **Topological difference between MeSH and the corresponding inferred ontology using CliXO.** A) A MeSH subtree containing prematurity complications. B) Corresponding disease ontology inferred using CliXO and ontology alignment. Drawn in Cytoscape v. 3.3.0 [SMO+03].

preserved. To create an overall score we calculate the percentage of edges in the reference that also appear in the inferred ontology.

4.2.6.2 Ancestor Correctness (AC)

While Edge Correctness (EC) can measure how well two networks are aligned, it may not be the best method for evaluating disease taxonomies. In particular, diseases separated by multiple taxonomic links may still be closely related to each other, so EC can underestimate performance by ignoring the ancestor-descendant relationship. EC also rewards successfully matched edges with no penalty for incorrect ones. This property may favor CliXO, which tends to produce DAGs with many edges.

To address the first shortcoming, we introduce the notion of ancestor cor-

rectness (AC). For a disease x, let x_{ref} be a node representing x in the reference ontology and x_{inf} be a node representing x in our inferred hierarchy. Also let $A(x)$ be the set of all ancestors of x in the appropriate hierarchy. Then for a specific disease x_{inf} in the inferred taxonomy we can measure how well it matches the reference by calculating $AncestorJaccard = Jaccard(A(x_{ref}), A(x_{inf}))$. We can then apply $AncestorJaccard$ globally by averaging across all diseases in the inferred network. We report this average as our AC score for the inferred network. Note that we only consider diseases existing in both hierarchies. However, we exclude diseases that are roots in both because they do not have any ancestors.

4.2.6.3 Ancestor Precision and Recall (AP and AR)

Ancestor Correctness (AC) provides a good estimate of topological similarity in terms of the number of preserved ancestors of mapped nodes. However, it still does not penalize false positives.

To address this problem, we adapt the Hierarchical Precision (HP) and Hierarchical Recall (HR) measurements from Verspoor et al. [VCSM06]. These measurements compare the sets of all ancestors of a disease in the inferred hierarchy to the ancestors of the same term in the reference. Informally, HP is the fraction of x's ancestors in the inferred hierarchy that are correct, while HR is the fraction of true ancestors of x that are also predicted by an inference method to be ancestors of x.

More specifically, for a disease x, let x_{ref} be the node in the reference and x_{inf} be the node in the inferred ontology. Then our HP and HR are calculated as follows:

$$HP(x_{ref}, x_{inf}) = \frac{|A(x_{ref}) \cap A(x_{inf})|}{|A(x_{inf})|} \tag{4.3}$$

$$HR(x_{ref}, x_{inf}) = \frac{|A(x_{ref}) \cap A(x_{inf})|}{|A(x_{ref})|} \tag{4.4}$$

We also calculate an F score using HP and HR as:

$$F(x) = 2 \times \frac{HP(x) \times HR(x)}{HP(x) + HR(x)} \tag{4.5}$$

Finally, we define Ancestor Precision (AP) and Ancestor Recall (AR) to be the average of HP and HR across all diseases in our reference hierarchy.

4.3 Results

Table 4.3: **Average performance of inference methods across the MeSH trees**

Method	**EC (± stdev)**	**AC (± stdev)**	**AP (± stdev)**	**AR (± stdev)**	**F (± stdev)**
Parent Promotion	**0.13 (± 0.06)**	**0.30 (± 0.10)**	**0.46 (± 0.16)**	0.47 (± 0.14)	**0.46 (± 0.14)**
CliXO	0.12 (± 0.10)	0.22 (± 0.12)	0.30 (± 0.14)	0.38 (± 0.17)	0.33 (± 0.15)
MWST	0.07 (± 0.04)	0.11 (± 0.07)	0.13 (± 0.08)	**0.48 (± 0.18)**	0.21 (± 0.11)

Average Edge Correctness (EC), Ancestor Correctness (AC), Ancestor Precision (AP), Ancestor Recall (AR) and F-score across the different trees in the MeSH forest. Standard deviation is shown in parentheses. Best performance across different inference techniques is highlighted.

4.3.1 Comparison to MeSH

We ran all three algorithms on the disease gene data and disease terms from each of the 23 MeSH trees. Table 4.3 reports the averaged performance across all 23 trees for each method and the different evaluation criteria. Across this data set, we see that Parent Promotion on average outperforms CliXO and MWST for almost all evaluation measures. The only exception is Ancestor Recall, for which MWST slightly edges out Parent Promotion. Detailed performance on each MeSH disease tree is shown in Additional File 1; in most cases the methods' relative performance is similar to that in Table 4.3. The detailed table also shows that, for each evaluation criterion, performance of the different methods is highly correlated across the 23

disease trees, suggesting that some trees are more consistent with the disease gene data than others.

4.3.2 Comparison to the Disease Ontology

We first attempted to reconstruct all of the Disease Ontology reflected in our disease-gene data set (2,095 edges connecting 2,039 DO terms). However, we could not compare the performance of all three inference methods on this full data set because running CliXO, which has at its core the computationally hard problem of finding cliques, was infeasible on a data set this large and complex. Nonetheless, we found that Parent Promotion consistently outperformed MWST on this data set. Specifically, Parent Promotion had an EC of 0.07 compared to MWST's EC of 0.05, an AC of 0.23 compared to MWST's AC of 0.04, and an F score of 0.40 compared to MWST's 0.08.

We used the subsets of DO listed in Table 4.1 to compare all three methods. Table 4.4, 4.5 and 4.6 show the results of all three methods on these subsets of DO. We again see that in most cases Parent Promotion outperforms CliXO and MWST for each evaluation measure, with the exception of "Musculosketal Disease," where CliXO outperforms Parent Promotion and MWST. Again, MWST often has good Ancestor Recall despite unimpressive performance on most other metrics.

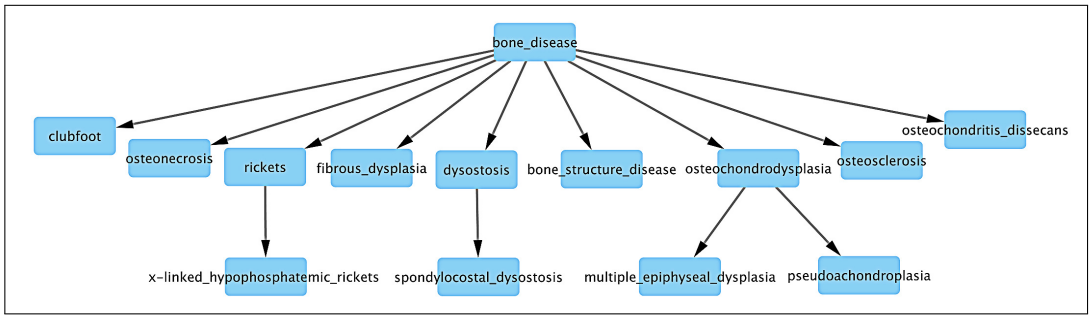

Figure 4.3: **Parent Promotion tree using DO data.** Subtree of the disease tree built by Parent Promotion on DO "musculoskeletal system disease" data that is an exact match to nodes and edges in the DO.

Figure 4.3 shows an example of one of the larger connected components inferred by Parent Promotion using the DO data. All edges in the figure occur

in both the Disease Ontology and the inferred tree. Although the inferred tree is relatively flat, the figure demonstrates that inference method is capturing some logical relationships between diseases.

Table 4.4: **Edge Correctness (EC) for four sub-DOs**

Root Disease	**Parent Promotion**	**CliXO**	**MWST**
Cardiovascular Disease	0.06	**0.09**	0.07
Gastrointestinal Disease	**0.17**	0.13	0.03
Musculoskeletal Disease	**0.16**	0.08	0.10
Nervous System Disease	**0.13**	0.07	0.09

Table 4.5: **Ancestor Correctness (AC) for four sub-DOs**

Root Disease	**Parent Promotion**	**CliXO**	**MWST**
Cardiovascular Disease	**0.32**	0.18	0.11
Gastrointestinal Disease	**0.37**	0.26	0.14
Musculoskeletal Disease	0.15	**0.26**	0.09
Nervous System Disease	**0.29**	0.17	0.10

Table 4.6: **Ancestor Precision / Recall (APR) with F score for four sub-DOs**

	F (AP, AR)		
Root Disease	**Parent Promotion**	**CliXO**	**MWST**
Cardiovascular Disease	**0.50** (**0.57**, 0.44)	0.27 (0.24, 0.30)	0.21 (0.13, **0.48**)
Gastrointestinal Disease	**0.55** (**0.56**, **0.53**)	0.39 (0.36, 0.42)	0.26 (0.18, 0.48)
Musculoskeletal Disease	0.26 (**0.44**, 0.18)	**0.41** (0.42, **0.40**)	0.17 (0.16, 0.19)
Nervous System Disease	**0.46** (**0.70**, 0.34)	0.30 (0.26, 0.34)	0.19 (0.13, 0.34)

4.4 Discussion

Overall, these experiments have provided some important insights into what can and cannot be learned about disease relationships from disease genes alone.

The correlations observed across the MeSH trees suggest that disease relationships some MeSH categories are easier to learn than others. Correctness appears

to be higher for smaller trees, perhaps simply because there are fewer possibilities. However, there are some large disease subtrees with higher AC and EC scores, especially Endocrine System Diseases (C19), Nutritional and Metabolic Diseases (C18), and Respiratory Tract Diseases (C08). It is possible that the MeSH hierarchy in these areas is better defined by molecular data, or that there are simply more disease genes known in these areas than in some others. One observation is that these categories include several well-studied complex diseases with high public health impact. For example, C19 includes diabetes and ovarian and pancreatic cancer; C18 also includes diabetes, plus obesity and related conditions; and C08 features asthma, COPD, and several types of lung cancer. Which exact properties of a set of diseases contribute most to the success of inference algorithms is an important question for future work.

On the "Musculoskeletal Disease" subset of DO, CliXO outperforms Parent Promotion by several criteria. Parent Promotion struggles with this region of the Disease Ontology, in part because the term "Musculosketal Disease" has fewer PubMed citations than the less general term "Bone Disease." The latter is therefore promoted incorrectly to become the root, while the former remains low in the inferred tree.

We also notice that despite its relatively poor performance overall, MWST seems to have good Ancestor Recall in many cases, sometimes even beating other methods. This may be because MWST tends to infer tall, thin trees rather than short and broad ones. Figure 4.4 illustrates this tendency. A node has more ancestors in tall, thin trees than in broad trees, and as a result, is more likely to share ancestors with the reference.

4.5 Conclusions

We have demonstrated that it is possible to recover much of the structure of both MeSH disease trees and the DO from molecular data alone.

One ultimate goal for a 21st-century disease taxonomy is the inference of

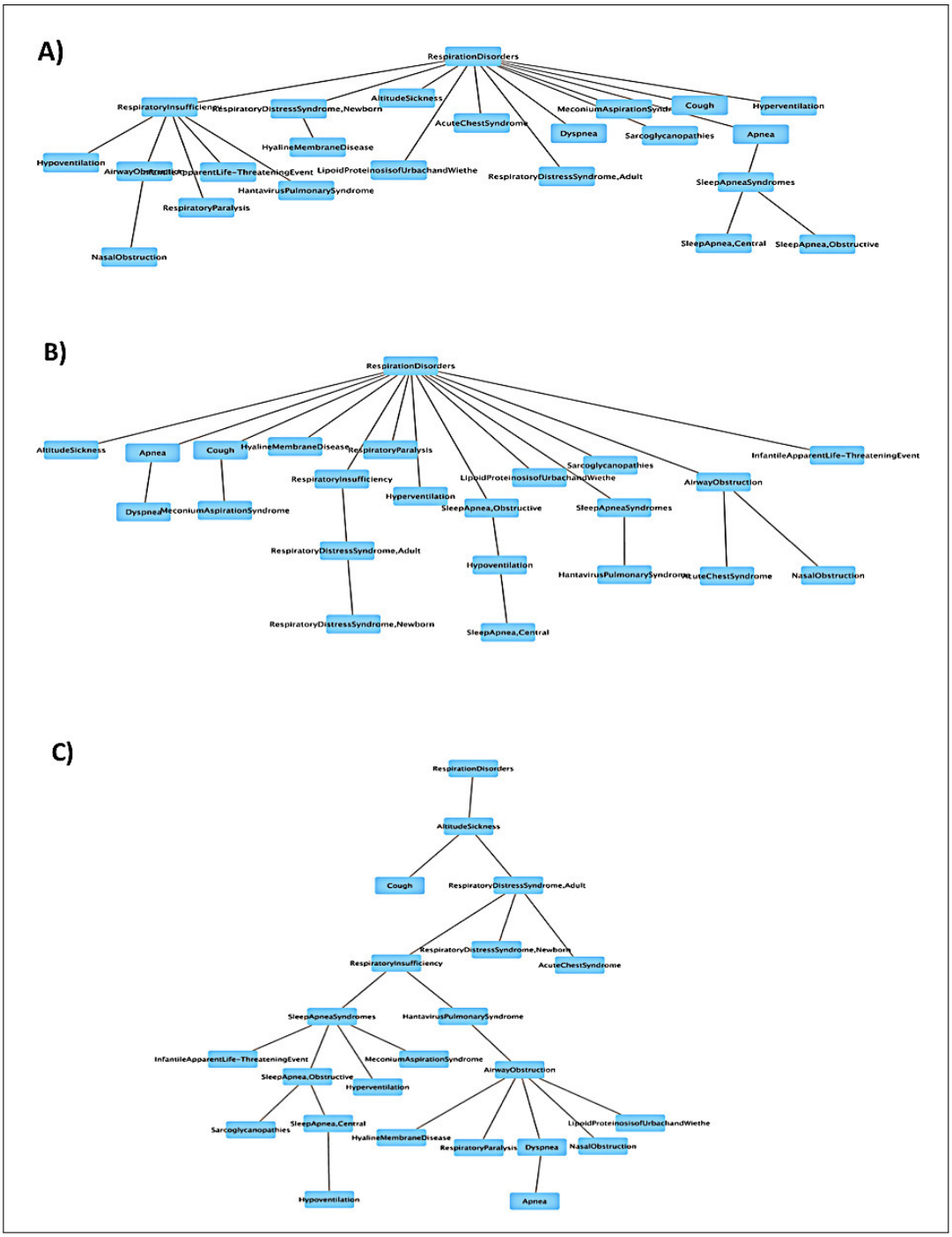

Figure 4.4: **A MeSH tree rooted at "Respiration Disorder" and corresponding inferred disease trees.** A) The MeSH tree containing "Respiration Disorder" and its descendants. B) The disease tree inferred by Parent Promotion on data from the tree in A). C) The disease tree inferred by MWST from the same data. MWST builds a taller and slimmer tree. As a result, each disease has more ancestors in C) than in A) or B). This leads MWST to have good performance with respect to Ancestor Recall (AR).

new disease terms based on molecular information [DHS+11, KB11]. Classification of cancer or autism subtypes based on underlying genetic contributions, for example, might be possible in such a system. We note that CliXO is the only method of those described here that could be used directly to address this problem, by creating internal nodes corresponding to sets of genes and then by finding new methods to map these gene sets into either known or previously unknown disease types. Further exploration of its abilities to do so, or extension of clustering-based methods analogous to Parent Promotion to incorporate comparable possibilities, is warranted.

Future work may also include exploring the incorporation of tissue specific gene expression to integrate relevant tissues and organs with the molecular level data, and to look more broadly at ways to combine clinical and molecular data. We also have not yet fully explored the range of relevant tree- and DAG-inference methods from the machine-learning community. However, the current results show that that it will be possible to construct integrated disease taxonomies that better support medical research in the genomic era.

Chapter 5

Conclusions and Future Work

5.1 Identification of disease mediating pathways in a weighted protein-protein interaction network

In chapter 2, we demonstrated a computational method to identify disease mediating pathways using *pathway centrality*. Our pathway centrality for each pathway was calculated in an unweighted protein-protein interaction (PPI) network. However, our pathway centrality analysis could be extended by incorporating additional features of PPI networks such as edge weights.

Pairwise protein interactions, especially ones identified by high throughput techniques such as yeast two-hybrid assay or mass spectrometry, are noisy, and the importance of quality control of PPIs has been addressed through many comparative studies of PPIs [BTD+09, VMKS+02, YBY+08, EKJ+02, HJB07, BCRC04]. While these comparative studies are focused on assessing the quality of each interacting pair of proteins by comparing to known protein complexes, there have been many attempts to control the quality of PPIs in big combined data sets. PPI networks are usually compiled by combining multiple data sources, and the curators of the networks have attempted to assign "confidence scores" on edges, each representing a pair of proteins, to inform users the degree of confidence they have for each pair [FSF+13, SFV+12, KAB+12]. The "confidence score" is often defined to reflect the

amount of supporting evidence of the pair.

The efficacy of incorporating confidence scores of a protein pair as edge weights in PPI networks has been well justified in many applications. The most common way to utilize the confidence scores is already adopted in our study described in chapter 2, which is to filter out pairs of proteins with low confidence to reduce the number of false positive edges in PPI networks. Another widely used method is to use the confidence scores as edge weights. This method has been heavily applied in studies where diffusion-based methods are used to measure proximity between proteins [CPF+14, VMR+10, WCZ+15] for inferring new knowledge, including protein functions or disease genes, based on the "guilt by association" assumption.

We may extend our study by calculating pathway centrality in PPI networks where edges are weighted by reciprocal values of confidence scores, as the core elements of our pathway centrality are the shortest paths between disease genes and differentially expressed genes.

5.2 Pathway centrality in a directed molecular interaction network

Protein-protein interaction network used in chapter 2 can be extended to a larger scale network incorporating other molecular interactions such as transcription factor(TF)-target interactions [TWA+06, KWP11, SBK+08]. In preliminary work not included here, we found that using a regulatory network, as compiled from TF-target interactions identified by a probabilistic algorithm quantitatively measuring the regulatory potential between TFs and targets from ChIP-seq or ChIP-chip data [CMG11], did not work as well for our analysis, because the resulting network was comprised of multiple large disconnected components. However, our analysis for finding disease mediating pathways may yield more precise predictions, if repeated in an extended network combining protein-protein interaction networks with regulatory networks. We could also explore using other TF-target network data as it

becomes available.

While the direction of protein-protein interactions are often not specified, transcription factor-target interactions have explicit directions. Therefore, the combined molecular network will be partially directed. We can further add precision to the network by adding signal flow directions to some pairs of proteins, which can be predicted using naïve Bayesian learning [VSF$^+$11] or constituent domains [LLW$^+$09]. The added precision to the molecular network might improve the accuracy of our algorithm to discover disease mediating pathways.

5.3 Identification of disease mediating pathways through functional enrichment analysis of a central module

In chapter 2, we calculated pathway centrality for all pre-defined gene sets collected from the Gene Ontology, KEGG, and Mouse Genome Database, and defined ones with significantly high pathway centrality as mediating pathways for the tested disease. We have an approach with a different view for identifying disease mediating pathways that we are interested in implementing and testing. The new approach identifies a central module using pathway centrality first and learns the functions of the module through functional enrichment analysis.

A central module can be learned by solving the k-Group Centrality Maximization (k-GCM) problem, which is to find a set of k nodes in a network for which the group centrality is highest of all groups of k nodes. The k-GCM problem is proven to be NP complete, but Ishakian et al. have published a greedy heuristic approximation algorithm [IETB12]. A variation of the k-GCM problem where traditional group centrality is replaced by pathway centrality will be able to identify a module of size k handling most information flows between disease genes and differentially expressed genes. Comparative studies of its results to those of our previous study are of interest because they may either confirm our findings for disease medi-

ating pathways or provide more interesting insights about topological properties of disease mediating pathways.

5.4 Enhanced gene pooling using more molecular disease taxonomies

Our analysis results demonstrated in chapter 3 show that enrichment of disease-gene association information through pooling disease genes from specific disease terms to general ones improves prediction of novel connections between biological processes and diseases. Such gene pooling will gain more strength when the disease hierarchies defining general-specific relationships between diseases contain more molecular contexts. This intuition led us to run the pilot study where we built disease hierarchies only using molecular information (i.e., disease-gene association information) introduced in chapter 4.

While there is lot more to do toward building more molecular disease hierarchies, we intend to repeat our analysis for predicting new relationships between biological processes and classes of diseases with empowered gene pooling by more molecular disease hierarchies, once such disease hierarchies are available. Preliminary work could use our inferred hierarchies even from the pilot work described in chapter 4, and look at new discoveries that might be validated experimentally.

5.5 Towards better supported molecular medicine

In this dissertation, we discussed several computational approaches to discover novel underlying biological processes of human disease, utilizing two different types of biological networks: protein-protein interaction networks and disease hierarchies (which we defined as networks in a broad sense). Yet, there exist other biological networks which may broaden the range of our discovery when similar approaches are applied. Furthermore, we may still be left with a large volume of high-throughput data that can be effectively modeled by networks and lead us to better inference of

underlying disease biology.

Our methods also utilized sets of genes representing biological functions and will gain more strength as the gene sets grow owing to continuing community efforts to predict functions of gene products. While there is much to be done to learn more about human disease, we hope that our studies in this dissertation contribute to discovery of interesting findings worthy of further investigation, and eventually lead to better support of molecular medicine.

Bibliography

[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, and Janan T Eppig. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[ABB+12] F. Althabe, Z. Bhutta, H. Blencowe, V. Chandra-Mouli, D. Chou, A. Costello, S. Cousens, et al. Born too soon: The global action report on preterm birth. Technical report, The World Health Organization, 2012.

[ABSH09] J. Amberger, C.A. Bocchini, A.F. Scott, and A. Hamosh. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, (Database issue):D793–6, 2009.

[AFG+16] H. Attrill, K. Falls, J. L. Goodman, G. H. Millburn, G. Antonazzo, A. J. Rey, S. J. Marygold, and Consortium FlyBase. Flybase: establishing a gene group resource for drosophila melanogaster. *Nucleic Acids Res*, 44(D1):D786–92, 2016.

[AHM+13] S. Ali, A. F. Hirschfeld, M. L. Mayer, 3rd Fortuno, E. S., N. Corbett, M. Kaplan, S. Wang, J. Schneiderman, C. D. Fjell, J. Yan, L. Akhabir, F. Aminuddin, N. Marr, T. Lacaze-Masmonteil, R. G. Hegele, A. Becker, M. Chan-Yeung, R. E. Hancock, T. R. Kollmann, D. Daley, A. J. Sandford, P. M. Lavoie, and S. E. Turvey. Functional genetic variation in NFKBIA and susceptibility to childhood

asthma, bronchiolitis, and bronchopulmonary dysplasia. *J Immunol*, 190(8):3949–58, 2013.

[AHU83] A.V. Aho, J.E. Hopcroft, and J.D. Ullman. *Data Structures and Algorithms*. Addison-Wesley, Reading, MA, 1983.

[AIS09] V.D. Acevedo, M. Ittmann, and D.M. Spencer. Paths of FGFR-driven tumorigenesis. *Cell Cycle*, 8(4):580–8, 2009.

[AKM+16] A. Apostolou, T. Kerenidi, A. Michopoulos, K. I. Gourgoulianis, M. Noutsias, A. E. Germenis, and M. Speletas. Association between tlr2/tlr4 gene polymorphisms and copd phenotype in a greek cohort. *Herz*, 2016.

[ALANS17] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer. Hippie v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res*, 45(D1):D408–D414, 2017.

[BAB+16] K. Biggadike, M. Ahmed, D. I. Ball, D. M. Coe, D. A. Dalmas Wilk, C. D. Edwards, B. H. Gibbon, C. J. Hardy, S. A. Hermitage, J. O. Hessey, A. E. Hillegas, S. C. Hughes, L. Lazarides, X. Q. Lewell, A. Lucas, D. N. Mallett, M. A. Price, F. M. Priest, D. J. Quint, P. Shah, A. Sitaram, S. A. Smith, R. Stocker, N. A. Trivedi, D. C. Tsitoura, and V. Weller. Discovery of 6-amino-2-[(1S)-1-methylbutyl]oxy-9-[5-(1-piperidinyl)pentyl]-7,9-dihydro-8h-pu rin-8-one (GSK2245035), a highly potent and selective intranasal toll-like receptor 7 agonist for the treatment of asthma. *J Med Chem*, 59(5):1711–26, 2016.

[Bar03] D.J. Barker. The developmental origins of adult disease. *Eur J Epidemiol*, 18(8):733–6, 2003.

[Bar16] P. J. Barnes. Kinases as novel therapeutic targets in asthma and chronic obstructive pulmonary disease. *Pharmacol Rev*, 68(3):788–815, 2016.

[BB12] M. Bouchecareilh and W. E. Balch. Proteostasis, an emerging therapeutic paradigm for managing inflammatory airway stress disease. *Curr Mol Med*, 12(7):815–26, 2012.

[BBWW05] S. Buckley, L. Barsky, K. Weinberg, and D. Warburton. In vivo inosine protects alveolar epithelial type 2 cells against hyperoxia-induced dna damage through map kinase signaling. *Am J Physiol Lung Cell Mol Physiol*, 288(3):L569–75, 2005.

[BC07] AI Baba and C. Câtoi. *Comparative Oncology*. The Publishing House of the Romanian Academy, Bucharest, 2007. Ch. 2.3.

[BCM+13] Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda Andrews, and Charles Boone. Genetic interaction networks: toward an understanding of heritability. *Annu Rev Genomics Hum Genet*, 14:111–133, 2013.

[BCRC04] Joel S Bader, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotech*, 22(1):78–85, 01 2004.

[BDRT09] P. J. Barnes, J. M. Drazen, S. I. Rennard, and N. C. Thomson. Asthma and copd basic mechanisms and clinical management second edition preface to the 2nd edition. *Asthma and Copd: Basic Mechanisms and Clinical Management, 2nd Edition*, page pp. 178ff, 2009.

[BEK+16] Judith A. Blake, Janan T. Eppig, James A. Kadin, Joel E. Richardson, Cynthia L. Smith, and Carol J. Bult. Mouse genome database (mgd)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, 45(D1):D723–D729, 11 2016.

[Bia12] D.W. Bianchi. From prenatal genomic diagnosis to fetal personalized medicine: progress and challenges. *Nat Med*, 18(7):1041–51, 2012.

[BMR11] M. Bijanzadeh, P. A. Mahesh, and N. B. Ramachandra. An understanding of the genetic basis of asthma. *Indian J Med Res*, 134:149–61, 2011.

[Bod04] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–270, Jan 2004.

[Bos12] Y. Bosse. Updates on the copd gene list. *Int J Chron Obstruct Pulmon Dis*, 7:607–31, 2012.

[BPH+01] A.J. Bhatt, G.S. Pryhuber, H. Huyck, R.H. Watkins, L.A. Metlay, and W.M. Maniscalco. Disrupted pulmonary vasculature and decreased vascular endothelial growth factor, Flt-1, and TIE-2 in human infants dying with bronchopulmonary dysplasia. *Am.J.Respir.Crit.Care Med.*, pages 1971–1980, 2001.

[BPL+09] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6):2763–2788, 2009.

[BTD+09] Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, and Anne-Sophie De Smet. An experimentally derived confidence score for binary protein-protein interactions. *Nature methods*, 6(1):91–97, 2009.

[BTV12] M. Bodas, I. Tran, and N. Vij. Therapeutic strategies to correct proteostasis-imbalance in chronic obstructive lung diseases. *Curr Mol Med*, 12(7):807–14, 2012.

[CBA+02] V. Compernolle, K. Brusselmans, T. Acker, P. Hoet, M. Tjwa, H. Beck, S. Plaisance, Y. Dor, E. Keshet, et al. Loss of HIF-2alpha and inhibition of VEGF impair fetal lung maturation, whereas treatment

with VEGF prevents fatal respiratory distress in premature mice. *Nat Med*, 8:702–710, 2002.

[CD11] K. Calkins and S.U. Devaskar. Fetal origins of adult disease. *Curr Probl Pediatr Adolesc Health Care*, 41(6):158–76, 2011.

[CDRV+15] P. Carrera, C. Di Resta, C. Volonteri, E. Castiglioni, S. Bonfiglio, D. Lazarevic, D. Cittaro, E. Stupka, M. Ferrari, M. Somaschini, Bpd, and Group Genetics Study. Exome sequencing and pathway analysis for identification of genetic variability relevant for bronchopulmonary dysplasia (bpd) in preterm newborns: A pilot study. *Clin Chim Acta*, 451(Pt A):39–45, 2015.

[CKRC08] Z. H. Chen, H. P. Kim, S. W. Ryter, and A. M. Choi. Identifying targets for copd treatment through gene expression analyses. *Int J Chron Obstruct Pulmon Dis*, 3(3):359–70, 2008.

[CLRS09] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition.* The MIT Press, Cambridge, MA, 3rd edition, 2009.

[CMG11] Chao Cheng, Renqiang Min, and Mark Gerstein. Tip: A probabilistic method for identifying transcription factor target genes from chip-seq binding profiles. *Bioinformatics*, 27(23):3221, 2011.

[CMS+07] J. Cohen, L.J. Van Marter, Y. Sun, E. Allred, A. Leviton, and I.S. Kohane. Perturbation of gene expression of the chromatin remodeling pathway in premature newborns at risk for bronchopulmonary dysplasia. *Genome Biol*, 8(10):R210, 2007.

[Con14] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 11 2014.

[Coo16] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 44(D1):D7–19, 2016.

[CPF+14] Mengfei Cao, Christopher M Pietras, Xian Feng, Kathryn J Doroschak, Thomas Schaffner, Jisoo Park, Hao Zhang, Lenore J Cowen, and Benjamin J Hescott. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, 30(12):i219–i227, 2014.

[CTSS+03] K.H. Choe, L. Taraseviciene-Stewart, R. Scerbavicius, L. Gera, R.M. Tuder, and N.F. Voelkel. Methylprednisolone causes matrix metalloproteinase-dependent emphysema in adult rats. *Am J Respir Crit Care Med*, 167:1516–1521, 2003.

[DAB+13] L. Ding, T. Abebe, J. Beyene, R.A. Wilke, A. Goldberg, J.G. Woo, L.J. Martin, M.E. Rothenberg, M. Rao, G.K. Hershey, R. Chakraborty, and T.B. Mersha. Rank-based genome-wide analysis reveals the association of Ryanodine receptor-2 gene variants with childhood asthma among human populations. *Hum Genomics*, 7:16, 2013.

[DCB+09] Scott J Dixon, Michael Costanzo, Anastasia Baryshnikova, Brenda Andrews, and Charles Boone. Systematic mapping of genetic interaction networks. *Annual review of genetics*, 43:601–625, 2009.

[Del15] C. M. Delude. Deep phenotyping: The details of disease. *Nature*, 527(7576):S14–5, 2015.

[DGLH+14] Allan Peter Davis, Cynthia J. Grondin, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L. King, Thomas C. Wiegers, and Carolyn J. Mattingly. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, pii(1):gku935, October 2014.

[DHS+11] Susan Desmond-Hellmann, Charles L. Sawyers, et al. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. Technical report, National Research Council, 2011.

[DKS+13] Janusz Dutkowski, Michael Kramer, Michal A Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan, and Trey Ideker. A gene ontology inferred from molecular networks. *Nature Biotechnology*, 31:34–35, 2013.

[Dok13] A. Dokras. Cardiovascular disease risk in women with PCOS. *Steroids*, 78(8):773–6, April 2013.

[DPW06] V.C. Daniel, C.D. Peacock, and D.N. Watkins. Developmental signalling pathways in lung cancer. *Respirology*, 11(3):234–40, 2006.

[DRB+10] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9):1205–1210, May 2010.

[DSH+03] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H.C. Lane, and R. A. Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[EB99] M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181–201, 1999.

[EKJ+02] Aled M Edwards, Bart Kus, Ronald Jansen, Dov Greenbaum, Jack Greenblatt, and Mark Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *TRENDS in Genetics*, 18(10):529–536, 2002.

[Erd15] Dora Erdos. Centrality measures and analyzing dot-product graphs. 2015.

[FB11] Brooke L Fridley and Joanna M Biernacka. Gene set analysis of snp data: benefits, challenges, and future directions. *Eur J Hum Genet*, 19(8):837–843, 08 2011.

[FBW91] L. C. Freeman, S. P. Borgatti, and D. R. White. Centrality in valued graphs - a measure of betweenness based on network flow. *Social Networks*, 13(2):141–154, 1991.

[FHBS11] Andrew D. Fox, Benjamin J. Hescott, Anselm C. Blumer, and Donna K. Slonim. Connectedness of ppi network neighborhoods identifies regulatory hub proteins. *Bioinformatics*, 27(8):1135–1142, 2011.

[FS89] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 07 1989.

[FSF+12] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 11 2012.

[FSF+13] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, and Christian Von Mering. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

[FTWB+16] G. Faura Tellez, B. W. Willemse, U. Brouwer, S. Nijboer-Brinksma, K. Vandepoele, J. A. Noordhoek, I. Heijink, M. de Vries, N. P. Smithers, D. S. Postma, W. Timens, L. Wiffen, F. van Roy, J. W. Holloway, P. M. Lackie, M. C. Nawijn, and G. H. Koppelman.

Protocadherin-1 localization and cell-adhesion function in airway epithelial cells in asthma. *PLoS One*, 11(10):e0163967, 2016.

[GB07] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[GBK+02] Anne-Claude Gavin, Markus Bosche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jorg Schultz, Jens M. Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Hofert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R. Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 01 2002.

[GBRV06] Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. An improved statistic for detecting over-represented gene ontology annotations in gene sets. *Lecture Notes in Bioinformatics*, 3909:85–98, 2006.

[GCP+11] R. Gagliardo, P. Chanez, M. Profita, A. Bonanno, G. D. Albano, A. M. Montalbano, F. Pompeo, C. Gagliardo, A. M. Merendino, and M. Gjomarkaj. Ikappab kinase-driven nuclear factor-kappab activation in patients with asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol*, 128(3):635–45 e1–2, 2011.

[GM08] J. J. Goeman and U. Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4):537–544, Feb 2008.

[GOKK03] K. I. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(1 Pt 2):017101, 2003.

[GS08] H. J. Gould and B. J. Sutton. Ige in allergy and asthma today. *Nat Rev Immunol*, 8(3):205–17, 2008.

[GUT+07] Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, and Chitta Baral. Mining gene-disease relationships from biomedical literature: Weighting protein-protein interactions and connectivity measures. *Pacific Symposium on Biocomputing*, 12:28–39, 2007.

[HBIR+11] Nicholas Hatzirodos, Rosemary A. Bayne, Helen F. Irving-Rodgers, Katja Hummitzsch, Laetitia Sabatier, Sam Lee, Wendy Bonner, Mark A. Gibson, William E. Rainey, Bruce R. Carr, Helen D. Mason, Dieter P. Reinhardt, Richard A. Anderson, and Raymond J. Rodgers. Linkage of regulators of TGF-β activity in the fetal ovary to polycystic ovary syndrome. *FASEB J.*, 25(7):2256–2265, 2011.

[HDR12] N. J. Hudson, B. P. Dalrymple, and A. Reverter. Beyond differential expression: the quest for causal mutations and effector molecules. *BMC Genomics*, 13:356, 2012.

[HJB07] Hailiang Huang, Bruno M Jedynak, and Joel S Bader. Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol*, 3(11):e214, 2007.

[HKBH12] S. Hubackova, K. Krejcikova, J. Bartek, and Z. Hodny. IL1- and TGFbeta-Nox4 signaling, oxidative stress and DNA damage response

are shared features of replicative, oncogene-induced, and drug-induced paracrine 'Bystander senescence'. *Aging*, 4(12):932–51, 2012.

[IETB12] Vatche Ishakian, Dóra Erdös, Evimaria Terzi, and Azer Bestavros. A framework for the evaluation and management of network centrality. pages 427–438. SIAM, 2012.

[Ish99] Y. Ishii. [role of adhesion molecules in the pathogenesis of copd]. *Nihon Rinsho*, 57(9):1965–71, 1999.

[Jac01] Paul Jaccard. Distribution de la flore alpine dans le bassin des drouces et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(140):241—272, 1901.

[JS11] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011.

[JYJ+13] Y. H. Jiang, R. K. Yuen, X. Jin, M. Wang, N. Chen, X. Wu, J. Ju, J. Mei, Y. Shi, M. He, G. Wang, J. Liang, Z. Wang, D. Cao, M. T. Carter, C. Chrysler, I. E. Drmic, J. L. Howe, L. Lau, C. R. Marshall, D. Merico, T. Nalpathamkalam, B. Thiruvahindrapuram, A. Thompson, M. Uddin, S. Walker, J. Luo, E. Anagnostou, L. Zwaigenbaum, R. H. Ring, J. Wang, C. Lajonchere, J. Wang, A. Shih, P. Szatmari, H. Yang, G. Dawson, Y. Li, and S. W. Scherer. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet*, 93(2):249–63, 2013.

[KAB+12] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The in-

tact molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846, 01 2012.

[KB11] Ismail Kola and John Bell. A call to reform the taxonomy of human disease. *Nat Rev Drug Discov*, 10(9):641–642, 09 2011.

[KBHR08] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, 82(4):949–958, Apr 2008.

[KC10] E. K. Kim and E. J. Choi. Pathological roles of mapk signaling pathways in human diseases. *Biochim Biophys Acta*, 1802(4):396–405, 2010.

[KD00] M. Karin and M. Delhase. The i kappa b kinase (ikk) and nf-kappa b: key elements of proinflammatory signalling. *Semin Immunol*, 12(1):85–98, 2000.

[KDY+14] Michael Kramer, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. Inferring gene ontologies from pairwise similarity data. *Bioinformatics*, 30:i34–i42, 2014.

[KLK+13] P. Kwinta, G. Lis, M. Klimek, A. Grudzien, T. Tomasik, K. Poplawska, and J. J. Pietrzyk. The prevalence and risk factors of allergic and respiratory symptoms in a regional cohort of extremely low birth weight children (<1000 g). *Ital J Pediatr*, 39:4, 2013.

[KLL+15] W. J. Kim, J. H. Lim, J. S. Lee, S. D. Lee, J. H. Kim, and Y. M. Oh. Comprehensive analysis of transcriptome sequencing data in the lung tissues of copd subjects. *Int J Genomics*, 2015:206937, 2015.

[KSK+16] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1):D457–62, Jan 2016.

[KWP11] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLOS Computational Biology*, 7(3):e1001095, 2011.

[Lam07] P. Lambe. *Organising knowledge: Taxonomies, knowledge and organisational effectiveness*. Chandos Publishing, Oxford, UK, 1st edition, 2007.

[LB94] Henry J. Lowe and G. Octo Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *JAMA*, 271(14):1103–1108, April 1994.

[LCW+06] B.K. Lin, M. Clyne, M. Walsh, O. Gomez, W. Yu, M. Gwinn, and M.J. Khoury. Tracking the epidemiology of human genes in the literature: the HuGE published literature database. *Am J Epidemiol*, 164(1):1–4, 2006.

[LDE+10] A. Leviton, O. Dammann, S. Engelke, E. Allred, K.C. Kuban, T.M. O'Shea, and N. Paneth. The clustering of disorders in infants born before the 28th week of gestation. *Acta Paediatr*, 99(12):1795–1800, 2010.

[LFK+10] C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader. Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2348, Sep 2010.

[Lib14] Arthur Liberzon. A description of the molecular signatures database (msigdb) web site. *Stem Cell Transcriptional Networks: Methods and Protocols*, pages 153–160, 2014.

[LKW+13] B. M. Levesque, L. A. Kalish, A. B. Winston, R. B. Parad, S. Hernandez-Diaz, M. Phillips, A. Zolit, J. Morey, M. Gupta, A. Mammoto, D. E. Ingber DE, and L. J. Van Marter. Low urine vascular endothelial growth factor levels are associated with mechanical

ventilation, bronchopulmonary dysplasia and retinopathy of prematurity. *Neonatology*, 104(1):56–64, 2013.

[LLW+03] E.Y. Liao, X.H. Luo, W.B. Wang, X.P. Wu, H.D. Zhou, R.C. Dai, H.J. Liao, et al. Effects of different nylestriol/levonorgestrel dosages on bone metabolism in female Sprague-Dawley rats with retinoic acid-induced osteoporosis. *Endocr Res*, 29(1):23–42, 2003.

[LLW+09] Wei Liu, Dong Li, Jian Wang, Hongwei Xie, Yunping Zhu, and Fuchu He. Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains. *Molecular & cellular proteomics*, 8(9):2063–2070, 2009.

[LNS+12] D.Y. Lee, D. L. Na, S. W. Seo, J. Chin, S.J. Lim, D. Choi, Y.K. Min, and B. K. Yoon. Association between cognitive impairment and bone mineral density in postmenopausal women. *Menopause*, 19(6):636–41, 2012.

[LSP+11] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–40, 2011.

[MASS16] A. H. Malash, A. A. Ali, R. M. Samy, and R. A. Shamma. Association of tlr polymorphisms with bronchopulmonary dysplasia. *Gene*, 592(1):23–8, 2016.

[McK12] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM. `http://www.omim.org/`, Dec. 2012.

[MHB09] H.A. Mintz-Hittner and L.M. Best. Antivascular endothelial growth factor for retinopathy of prematurity. *Curr Opin Pediatr*, 21(2):182–7, 2009.

[MLCTIL11] E. E. Macksoud M. L. Calmus, R. Tucker, R. V. Iozzo, and B. E. Lechner. A mouse model of spontaneous preterm birth based on the genetic ablation of biglycan and decorin. *Reproduction*, 142(1):183–94, 2011.

[MLZ+09] P. V. Missiuro, K. Liu, L. Zou, B. C. Ross, G. Zhao, J. S. Liu, and H. Ge. Information flow analysis of interactome networks. *PLoS Comput Biol*, 5(4):e1000350, 2009.

[MMMT95] T. A. Mitsiadis, T. Muramatsu, H. Muramatsu, and I. Thesleff. Midkine (MK), a heparin-binding growth/differentiation factor, is regulated by retinoic acid and epithelial-mesenchymal interactions in the developing mouse tooth, and affects cell proliferation and morphogenesis. *J Cell Biol*, 129(1):267–81, 1995.

[Moo09] S.W. Moore. Developmental genes and cancer in children. *Pediatr Blood Cancer*, 52(7):755–60, 2009.

[MSL+11] D. Macut, T. Simic, A. Lissounov, M. Pljesa-Ercegovac, I. Bozic, T. Djukic, J. Bjekic-Macut, M. Matic, M. Petakov, S. Suvakov, S. Damjanovic, and A. Savic-Radojevic. Insulin resistance in nonobese women with polycystic ovary syndrome: relation to byproducts of oxidative stress. *Exp Clin Endocrinol Diabetes*, 119(7):451–5, 2011.

[MTM+11] D. Malhotra, R.K. Thimmulappa, N. Mercado, K. Ito, P. Kombairaju, S. Kumar, J. Ma, et al. Denitrosylation of HDAC2 by targeting Nrf2 restores glucocorticosteroid sensitivity in macrophages from COPD patients. *J Clin Invest*, 121(11):4289–302, 2011.

[MTT+11] N. Mercado, R. Thimmulappa, C.M. Thomas, P.S. Fenwick, K.K. Chana, L.E. Donnelly, S. Biswal, et al. Decreased histone deacetylase 2 impairs Nrf2 activation by oxidative stress. *Biochem Biophys Res Commun*, 406(2):292–8, 2011.

[MV07] O. Mason and M. Verwoerd. Graph theory and networks in biology. *IET Syst Biol*, 1(2):89–119, 2007.

[NJA+05] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–10, 2005.

[NPH02] S.J. Nelson, T. Powell, and B.L. Humphreys. The Unified Medical Language System (UMLS) project. In A. Kent and C.M. Hall, editors, *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, Inc., New York, 2002.

[NRP67] W. H. Northway, R. C. Rosan, and D. Y. Porter. Pulmonary disease following respirator therapy of hyaline-membrane disease: Bronchopulmonary dysplasia. *N Engl J Med*, 276(7):357–68, 1967.

[NTC+91] S. J. Nelson, M. S. Tuttle, W. G. Cole, D. D. Sherertz, W. D. Sperzel, M. S. Erlbaum, L. L. Fuller, and N. E. Olson. From meaning to term: semantic locality in the UMLS metathesaurus. *Proc Annu Symp Comput Appl Med Care*, pages 209–213, 1991.

[Oli07] Juan C Oliveros. Venny. an interactive tool for comparing lists with venn diagrams. 2007.

[OSV+15] J. J. O'Shea, D. M. Schwartz, A. V. Villarino, M. Gadina, I. B. McInnes, and A. Laurence. The jak-stat pathway: impact on human disease and therapeutic intervention. *Annu Rev Med*, 66:311–28, 2015.

[PEYT08] S.J. Patey, E.A. Edwards, E.A. Yates, and J.E. Turnbull. Engineered heparins: novel beta-secretase inhibitors as potential Alzheimer's disease therapeutics. *Neurodegener Dis*, 5(3-4):197–9, 2008.

[PFSP+15] Pletscher-Frankild, Sune, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, and Lars Juhl Jensen. Diseases: text mining and data integration of disease-gene associations. *Methods*, 74:83–89, 2015.

[PGLT10] J. E. Petrikin, R. Gaedigk, J. S. Leeder, and W. E. Truog. Selective toll–like receptor expression in human fetal lung. *Pediatr Res*, 68(4):335–8, 2010.

[PHS16] Jisoo Park, Benjamin Hescott, and Donna Slonim. Towards a more molecular taxonomy of disease. Bio-ontologies at ISMB 2016, 2016.

[PJT+07] S. A. Pangas, C. J. Jorgez, M. Tran, J. Agno, X. Li, C. W. Brown, T. R. Kumar, and M. M. Matzuk. Intraovarian activins are required for female fertility. *Mol Endocrinol*, 21(10), 2007.

[PKW+13] J. J. Pietrzyk, P. Kwinta, E. J. Wollen, M. Bik-Multanowski, A. Madetko-Talowska, C. C. Gunther, M. Jagla, T. Tomasik, and O. D. Saugstad. Gene expression profiling in preterm infants: new aspects of bronchopulmonary dysplasia development. *PLoS One*, 8(10):e78585, 2013.

[PM00] Akhilesh Pandey and Matthias Mann. Proteomics to study genes and genomes. *Nature*, 405(6788):837–846, 06 2000.

[PVHR06] Euripides G.M. Petrakis, Giannis Varelas, Angelos Hliaoutakis, and Paraskevi Raftopoulou. X-similarity: Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management*, 4:233—237, 2006.

[PWK+14] Jisoo Park, Heather C. Wick, Daniel E. Kee, Keith Noto, Jill L. Maron, and Donna K. Slonim. Finding novel molecular connections between developmental processes and disease. *PLoS Computational Biology*, 10(5):e1003578, May 2014.

[RKM+16] E. Ronkainen, T. Kaukola, R. Marttila, M. Hallman, and T. Dunder. School-age children enjoyed good respiratory health and fewer allergies despite having lung disease after preterm birth. *Acta Paediatr*, 105(11):1298–1304, 2016.

[RKURL13] N. Raja-Khan, M. Urbanek, R.J. Rodgers, and R.S. Legro. The role of TGF-beta in Polycystic Ovary Syndrome. *Reprod Sci*, April 2013.

[ROSHW98] P. L. Ramsay, E. O'Brian Smith, S. Hegemier, and S. E. Welty. Early clinical markers for the development of bronchopulmonary dysplasia: soluble e-selectin and icam-1. *Pediatrics*, 102(4 Pt 1):927–32, 1998.

[SA05] K.R. Stenmark and S.H. Abman. Lung vascular development: implications for the pathogenesis of bronchopulmonary dysplasia. *Annu Rev Physiol*, 67:623–61, 2005.

[SAN+11a] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40:D940–D946, 2011.

[SAN+11b] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(Database issue):D940–D946, 2011.

[SAR+07] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, 25(11):1251–1255, Nov 2007.

[SBIV12] David Sanchez, Montserrat Batet, David Isern, and Aida Valls. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9):7718—7728, 2012.

[SBK+08] Silpa Suthram, Andreas Beyer, Richard M Karp, Yonina Eldar, and Trey Ideker. eqed: an efficient method for interpreting eqtl associations using protein networks. *Molecular Systems Biology*, 4(1), 2008.

[SBR+06] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–9, 2006.

[SBS+14] E. K. Sackmann, E. Berthier, E. A. Schwantes, P. S. Fichtinger, M. D. Evans, L. L. Dziadzio, A. Huttenlocher, S. K. Mathur, and D. J. Beebe. Characterizing asthma from a drop of blood using neutrophil chemotaxis. *Proc Natl Acad Sci U S A*, 111(16):5813–8, 2014.

[SC04] F. Sennlaub and S. Chemtob. VEGFR-1: a safe target for prophylaxis of retinopathy of prematurity? *Pediatr Res*, 55(1):1–2, 2004.

[SC13] R. Shaykhiev and R. G. Crystal. Innate immunity and chronic obstructive pulmonary disease: a mini-review. *Gerontology*, 59(6):481–9, 2013.

[SD99] M.J. Smalley and T.C. Dale. Wnt signalling in mammalian development and cancer. *Cancer Metastasis Rev.*, 18(2):215–30, 1999.

[SD08] S. Saigal and L.W. Doyle. An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet*, 371(9608):261–9, 2008.

[SE12] Cynthia L. Smith and Janan T. Eppig. The mammalian phenotype ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome*, 23(9):653–668, 2012.

[Sei83] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[SFNY09] K. Shudo, H. Fukasawa, M. Nakagomi, and N. Yamagata. Towards retinoid therapy for Alzheimer's disease. *Curr Alzheimer Res*, 6(3):302–11, 2009.

[SFV+12] Martin H Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E Wanker, and Miguel A Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS one*, 7(2):e31826, 2012.

[SKB+02] Jorg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 11 2002.

[SKJ+09] D. K. Slonim, K. Koide, K. L. Johnson, U. Tantravahi, J. M. Cowan, Z. Jarrah, and D. W. Bianchi. Functional genomic analysis of amniotic fluid cell-free mrna suggests that oxidative stress is significant in down syndrome fetuses. *Proc Natl Acad Sci U S A*, 106(23):9425–9, 2009.

[SKPS01] M. Siltanen, M. Kajosaari, M. Pohjavuori, and E. Savilahti. Prematurity at birth reduces the long-term risk of atopy. *J Allergy Clin Immunol*, 107(2):229–34, 2001.

[SL73] J. P. Sackler and L. Liu. Heparin-induced osteoporosis. *Br J Radiol*, 46(547):548–50, 1973.

[SMG+14] V. Sharma, S. Michel, V. Gaertner, A. Franke, C. Vogelberg, A. von Berg, A. Bufe, A. Heinzmann, O. Laub, E. Rietschel, B. Simma, T. Frischer, J. Genuneit, D. P. Potaczek, and M. Kabesch. A role of fcer1a and fcer2 polymorphisms in ige regulation. *Allergy*, 69(2):231–6, 2014.

[SMK+16] M. Suzuki, H. Makita, S. Konno, K. Shimizu, H. Kimura, H. Kimura, M. Nishimura, and Copd Cohort Study Investigators Hokkaido. Asthma-like features and clinical course of chronic obstructive pulmonary disease. an analysis from the hokkaido copd cohort study. *Am J Respir Crit Care Med*, 194(11):1358–1365, 2016.

[SMO+03] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498—504, November 2003.

[SMS+05] R.H. Salama, H. Muramatsu, E. Shimizu, K. Hashimoto, S. Ohgake, H. Watanabe, N. Komatsu, N. Okamura, and K. Koike. Increased midkine levels in sera from patients with Alzheimer's disease. *Prog Neuropsychopharmacol Biol Psychiatry*, 29(4):611–6, 2005.

[SMW+07] D. A. Stern, W. J. Morgan, A. L. Wright, S. Guerra, and F. D. Martinez. Poor airway function in early infancy and lung function by age 22 years: a non-selective longitudinal cohort study. *Lancet*, 370(9589):758–64, 2007.

[STM+05a] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 10 2005.

[STM+05b] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based

approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, October 2005.

[STS+02] A. R. Simon, S. Takahashi, M. Severgnini, B. L. Fanburg, and B. H. Cochran. Role of the jak-stat pathway in pdgf-stimulated proliferation of human airway smooth muscle cells. *Am J Physiol Lung Cell Mol Physiol*, 282(6):L1296–304, 2002.

[SVH+13] R. Sakurai, P. Villarreal, S. Husain, J. Liu, T. Sakurai, E. Tou, J. S. Torday, and V. K. Rehan. Curcumin protects the developing lung against long-term hyperoxic injury. *Am J Physiol Lung Cell Mol Physiol*, 305(4):L301–11, 2013.

[SXB08] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *PNAS*, 105(35):12763–12768, September 2008.

[tDH04] P. ten Dijke and C. S. Hill. New insights into TGF-beta-Smad signalling. *Trends Biochem Sci*, 29(5):265–73, 2004.

[TGK+05] Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38):13544–13549, September 2005.

[TOP+12] M. E. Talkowski, Z. Ordulu, V. Pillalamarri, C. B. Benson, I. Blumenthal, S. Connolly, C. Hanscom, N. Hussain, S. Pereira, J. Picker, J. A. Rosenfeld, L. G. Shaffer, L. E. Wilkins-Haug, J. F. Gusella, and C. C. Morton. Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N Engl J Med*, 367(23):2226–32, 2012.

[TTS08] A. Torkamani, E. J. Topol, and N. J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272, Nov 2008.

[TVM+11] S. Turcan, D. E. Vetter, J. L. Maron, X. Wei, and D. K. Slonim. Mining functionally relevant gene sets for analyzing physiologically novel clinical expression data. *Pac Symp Biocomput*, 16:50–61, 2011.

[TWA+06] Zhidong Tu, Li Wang, Michelle N Arbeitman, Ting Chen, and Fengzhu Sun. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, 22(14):e489–e496, 2006.

[Val16] K. Vale. Targeting the JAK-STAT pathway in the treatment of 'Th2-high' severe asthma. *Future Med Chem*, 8(4):405–19, 2016.

[VCSM06] Karin Verspoor, Judith Cohn, and Cliff Joslyn Susan Mniszewski. A categorization approach to automated ontological function annotation. *Protein Science*, 15:1544–1549, 2006.

[VDWL+90] M. Verhaeghe, M. De Wolf, A. Lagrou, G. Van Dessel, H. Hilderson, and W. Dierick. Identification of essential amino acids in the active center of thyroidal nad+ glycohydrolase. *Int J Biochem*, 22(2):197–202, 1990.

[VMKS+02] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[VMR+10] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, 6(1):e1000641, 2010.

[VSF+11] Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E Assmus, Miguel A Andrade-Navarro, and Erich E Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.*, 4(189):rs8–rs8, 2011.

[VVT06] Norbert F. Voelkel, R. William Vandivier, and Rubin M. Tuder. Vascular endothelial growth factor in the lung. *Am J Physiol Lung Cell Mol Physiol*, 290:L209–L221, 2006.

[WBD+07] P. G. Woodruff, H. A. Boushey, G. M. Dolganov, C. S. Barker, Y. H. Yang, S. Donnelly, A. Ellwanger, S. S. Sidhu, T. P. Dao-Pick, C. Pantoja, D. J. Erle, K. R. Yamamoto, and J. V. Fahy. Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. *Proc Natl Acad Sci U S A*, 104(40):15858–63, 2007.

[WCZ+15] Sheng Wang, Hyunghoon Cho, ChengXiang Zhai, Bonnie Berger, and Jian Peng. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics*, 31(12):i357–i364, 2015.

[WDN+14] H.C. Wick, H. Drabkin, H. Ngu, M. Sackman, C. Fournier, J. Haggett, J.A. Blake, D.W. Bianchi, and D.K. Slonim. DFLAT: functional annotation for human development. *BMC Bioinformatics*, 15(45), February 2014.

[WDSM96] K.L. Watterberg, L.M. Demers, S.M. Scott, and S. Murphy. Chorioamnionitis and early lung inflammation in infants in whom bronchopulmonary dysplasia develops. *Pediatrics*, 97(2):210–215, 1996.

[Wic13] H.C. Wick. Dflat: Developmental functional annotation at tufts. `http://bcb.cs.tufts.edu/dflat`, 2013.

[WK98] J. K. Weltman and A. S. Karim. Interleukin-5: a proeosinophil cytokine mediator of inflammation in asthma and a target for antisense therapy. *Allergy Asthma Proc*, 19(5):257–61, 1998.

[WLL+08] P.M. Wong, A.N. Lees, J. Louw, F.Y. Lee, N. French, K. Gain, C.P. Murray, et al. Emphysema in young adult survivors of moderate-to-severe bronchopulmonary dysplasia. *Eur Respir J*, 32(2):321–8, 2008.

[WSH+13] M. Wang, J. Saha, M. Hada, J. A. Anderson, J. M. Pluth, P. O'Neill, and F. A. Cucinotta. Novel Smad proteins localize to IR-induced double-strand breaks: interplay between TGFβ and ATM pathways. *Nucleic Acids Res*, 41(2):933–42, 2013.

[WSS02] Amy Y. Wang, Jeremiah H. Sable, and Kent A. Spackman. The snomed clinical terms development process: Refinement and analysis of content. In *Proc AMIA Symp*, pages 845–9, 2002.

[WV08] D. G. Woodside and P. Vanderslice. Cell adhesion antagonists: therapeutic potential in asthma and chronic obstructive pulmonary disease. *BioDrugs*, 22(2):85–100, 2008.

[XGM+06] D. Xu, J. R. Guthrie, S. Mabry, T. M. Sack, and W. E. Truog. Mitochondrial aldehyde dehydrogenase attenuates hyperoxia-induced cell death through activation of ERK/MAPK and PI3K-Akt pathways in lung epithelial cells. *Am J Physiol Lung Cell Mol Physiol*, 291(5):L966–75, 2006.

[YBY+08] Haiyuan Yu, Pascal Braun, Muhammed A. Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R. Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E. Hudson, Juyong Park, Xiaofeng Xin, Michael E. Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P. Roth, Albert-László Barabási, Jan Tavernier, David E. Hill, and Marc Vidal. High-quality binary protein interaction

map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[YCD+08] W. Yu, M. Clyne, S.M. Dolan, A. Yesupriya, A. Wulf, T. Liu, M.J. Khoury, and M. Gwinn. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9(205), 2008.

[YCKG10] W. Yu, M. Clyne, M. J. Khoury, and M. Gwinn. Phenopedia and genopedia: disease-centered and gene-cenered views of the evolving knowledge of human genetic associations. *Bioinformatics*, 26(1):145–148, 2010.

[YKS+07] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, 2007.

[YLRS+09] Esti Yeger-Lotem, Laura Riva, Linhui Julie Su, Aaron D Gitler, Anil G Cashikar, Oliver D King, Pavan K Auluck, Melissa L Geddie, Julie S Valastyan, and David R Karger. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–323, 2009.

[ZH03] L. Zhou and M. B. Hershenson. Mitogenic signaling pathways in airway smooth muscle. *Respir Physiol Neurobiol*, 137(2-3):295–308, 2003.