Towards constructing disease relationship networks using genome-wide association studies

Wenhui Huang

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in Partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science

Liqing Zhang, Chair
Liwu Li
Weiguo Fan

December 07, 2009
Blacksburg, Virginia

Keywords: Disease Relationship, Network, GWAS, SNP, Protein-Protein Interaction

Copyright 2009, Wenhui Huang
Towards constructing disease relationship networks using genome-wide association studies

Wenhui Huang

(ABSTRACT)

**Background:** Genome-wide association studies (GWAS) prove to be a powerful approach to identify the genetic basis of various human[1] diseases. Here we take advantage of existing GWAS data and attempt to build a framework to understand the complex relationships among diseases. Specifically, we examined 49 diseases from all available GWAS with a cascade approach by exploiting network analysis to study the single nucleotide polymorphisms (SNP) effect on the similarity between different diseases. Proteins within perturbation subnetwork are considered to be connection points between the disease similarity networks.

**Results:** shared disease subnetwork proteins are consistent, accurate and sensitive to measure genetic similarity between diseases. Clustering result shows the evidence of phenome similarity.

**Conclusion:** our results prove the usefulness of genetic profiles for evaluating disease similarity and constructing disease relationship networks.
Acknowledgements

I would like to thank my advisor Liqing Zhang for guidance.
## Contents

Chapter 1 ........................................................................................................................................................................1  

Introduction .....................................................................................................................................................................1  

1.1  The Human genome and full genome sequencing ............................................................................................1  

1.2  Categories of human genetic variation ..................................................................................................................1  

1.2.1  Single Nucleotide Polymorphisms......................................................................................................................2  

1.2.1.1  SNP selection in populations ......................................................................................................................2  

1.2.1.2  SNPs in coding and non-coding regions .......................................................................................................2  

1.2.2  Structural variants ..............................................................................................................................................3  

1.3  International HapMap project and the concept of LD .........................................................................................3  

1.4  Genome Wide Association (GWA) studies .............................................................................................................4  

1.5  Beyond statistical associations: understanding the functional implications of SNP distribution and a given complex trait .........................................................................................................................5  

1.5.1  Putative perturbed subnetworks based on GWA studies ..............................................................................5  

1.5.2  Phenome relationship based on GWA studies ...............................................................................................6  

1.6  Contributions of this thesis .................................................................................................................................7  

Chapter 2 ........................................................................................................................................................................8  

Methods ...........................................................................................................................................................................8  

2.1  Convert SNPs to genes ...........................................................................................................................................8  

2.2  Locate putative subnetworks in each disease .......................................................................................................10  

2.3  Measure disease similarity using Jaccard index and GO term IC scores .......................................................11  

2.4  Network clustering methods ...............................................................................................................................13  

Chapter 3 ........................................................................................................................................................................14  

Results and discussion ..................................................................................................................................................14  

3.1  Numbers of SNPs and genes associated with diseases ....................................................................................14  

3.2  Degree of similarity between diseases ..................................................................................................................16  

3.3  Network clustering methods ................................................................................................................................20  

3.4  Clustering disease into groups ............................................................................................................................23  

3.5  Conclusion and future work .................................................................................................................................28  

Bibliography ................................................................................................................................................................30
List of Figures

Figure 1: The disease relation network (DRN) based on the Jaccard index........21
Figure 2: The DRN based on GO-Term IC score .................................................22
List of Tables

Table 1: Disease abbreviated names ................................................................. 10
Table 2: Disease SNPs, genes, and PPI proteins .................................................... 15
Table 3: Shared SNP number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis) ........................................... 17
Table 4: Shared gene number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis) ........................................... 18
Table 5: Shared PPI number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis) ........................................... 18
Table 6: Ten disease pairs with the highest jaccard index .................................... 19
Table 7: Jaccard index between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis) .............................................. 20
Table 8: Network size and clusters ..................................................................... 25
Table 9: Parkinson's disease as a query disease .................................................. 27
Table 10: Systemic Lupus Erythematosus disease as a query disease ............... 28
Chapter 1

Introduction

With the advancement of genetics and molecular biology, the identification of genes related to disease pathogenesis becomes much easier than before. For example, large-scale human genome sequencing and genome-wide association studies (GWAS) help biologists to identify the variations of genomes as well as how these variations may affect genes and pathways[2]. Genetic information of diseases, for example, single nucleotide polymorphisms (SNPs) that are strongly associated with diseases, either located within genes or outside the genes, provides a blueprint for identifying genes and pathways that are underlying genetic mechanisms of diseases.

1.1 The Human genome and full genome sequencing

The human genome has twenty three chromosome pairs. The haploid human genome contains more than 3 billion DNA base pairs, including nearly 23,000 protein-coding genes. It is said that only 1.5% of the human genome codes for proteins, other regions are non-coding RNA, regulatory regions, introns and “junk” DNA[3]. Full Genome Sequencing (FGS) is a laboratory process that determines the complete DNA sequence of an organism’s genome at a single time [4-7]. It can use almost any biological samples and produce large sequence data.

1.2 Categories of human genetic variation

Human genetic variants usually are two types, called common variant and rare variants in terms of the frequency of the minor allele in the human population. Common variants are defined as genetic variants with a minor allele frequency (MAF) of more than one percent within population; and rare variants, however, with a MAF
less than one percent in population[8]. Generally speaking, variants in human genome can be divided into two different nucleotide composition classes: single nucleotide variants and structural variants[9]. Most of the variants in the genome are neutral, so they don’t have effect on the phenotypes of host.

1.2.1 Single Nucleotide Polymorphisms

The most common form of variation in human genome is Single Nucleotide Polymorphisms (SNPs). SNPs are single nucleotides which are substituted in fix positions. Human genome contains more than 11 million SNPs, with 7 million of these occurring with a Minor Allele Frequency (MAF) greater than 5%[8]. Among those SNPs, however, only a small number of which lead to phenotype differences within and between the populations, including the disease susceptibility and outcome[10].

1.2.1.1 SNP selection in populations

The occurrence of mutation events in genome is not even. It is more frequently of transition (a<->G or C<->T) than transversions(A<->C,A<->T,G<->C or G<->T) [11]. As the population size grows, the number of generations in which a new SNP will be observed in its heterozygous state will also increase[10]. Most SNPs are under “neutral selection” because they are making no effects to phenotypes. However, some SNPs are under “positive selection” thus being favored. In particular, enriched SNPs in coding regions given populations advantage in changing environment. Carriers of the variants have selective advantage over those are not. For example, immune system genes are under great environment pressure so remain relatively high frequency of SNPs.

1.2.1.2 SNPs in coding and non-coding regions

It has been estimated that 50,000-200,000 SNPs might have biological significance[12]. Some SNPs, called “Exonic SNPs”, existing in the coding region of
the genome, might lead to amino acid substitution (“nonsynonymous”). The consequence is obvious. They might alter biochemical processes of the organism. SNPs occurring in the exons of genes that do not alter protein primary sequences are called “synonymous” SNPs. Recent studies show that synonymous SNPs have effects on gene splicing, transcription factor binding, or the sequence of non-coding RNA [13-15].

1.2.2 Structural variants

Structural variants are defined as all based pairs that differ between individuals and that are not single nucleotide variants[9]. Most common structural variation including insertion-deletion (indels), block substitution, inversions of DNA sequences and copy number differences. Structural variation accounts for 20% of all genetic variants in humans and underlies greater than 70% of the variant bases. In this thesis, structural variants are not taken into consideration [14, 16, 17].

1.3 International HapMap project and the concept of LD

With the identification of millions of SNPs in the human genome, it remains a daunting task to genotype every single SNP, even with the latest genotyping technologies. To overcome this obstacle, the International HapMap Project was initiated in 2003 with the aim of characterizing LD patterns, and identifying haplotype-tagging SNPs in a total of 270 DNA samples that were collected from four major populations of European, African and Asian ancestry[18]. The Phase I and Phase II of the International HapMap Project were completed in 2005 and 2007 respectively. The application of the International HapMap Project is evident once we consider tagging SNPs that were identified in this global project were found to be ‘transferable’ in many populations around the world and in isolated populations. At the same time, Perlegens Sciences genotyped million SNPs on 71 individuals of European, African and Asian ancestry, and reported that these SNPs were able to
capture most of the common genetic variations based on LD. The major lesson that geneticists learn from these two studies is that it is not necessary to genotype every single SNP in the human genome because this would be redundant. SNPs that are close to each other within a genomic region tend to be inherited together more frequent than expected by chance in a block pattern (known as haplotype) due to the presence of LD. Several measures of pairwise LD are regularly used when describing marker-marker correlation and are central to SNP tagging[19]. The two most commonly used measures are $D'(\text{standard LD coefficient, } D)$ and $r^2(\text{correlation coefficient})$. Both $D'$ and $r^2$ have maximal values of one. When less than the total of four possible two-SNP haplotypes is observed in a population, A maximum $D'$ value will reach[14, 16, 19, 20]. When designing indirect testing studies, $r^2$ has useful property that the sample size adjustment required to achieve the equivalent power of a direct test is a function of the inverse of the correlation coefficient.

1.4 Genome Wide Association (GWA) studies

GWA studies published to date have used various commercial genotyping platforms containing about 300,000 to 500,000 common SNPs to detect differences in allele frequencies between cases and controls. Now that for over 80 phenotypes, including diseases and biological measurement, GWA studies provide significant statistical association for a total of 300 different loci in human genome [12]. So far, there are 280 reported studies in which almost all disease categories have been addressed. In a typical GWA study, a lot of markers should be tested to make sure the adequate coverage of the whole genome. In addition to the fixed content of genome-wide genotyping arrays, several custom made genotyping products are also introduced by illumina and affymetrix to accelerate the fine mapping of the genomic regions identified by GWA studies and linkage analysis [10]. The genome-wide genotyping products such as Illumina HumanHap550 and Affymetrix GeneChip 500K offer good coverage of the international HapMap Phase I and Phase II data in both Caucasians
and Asians. With the wealth of information on HGP and HapMap project; the data from a large number of case and controls studies; the fast developed genotyping technologies as well as the emergence of following efficient algorithms such PLINK, the GWA studies are quickly becoming available and economically feasible for everyone. The follow-up issues are huge capacity for data generation; high level of QA control and statistical methods development for data interpretation [19, 21]. One of the advantages of GWA studies is that such studies are hypothesis free, as there is no bias or presumptive list of candidate genes that are being tested [22]. In light of this, novel loci have been identified in a wide range of conditions, yielding many potential genes that are not identified previously involved in disease pathogenesis.

1.5 Beyond statistical associations: understanding the functional implications of SNP distribution and a given complex trait

Current data of GWAS provide us not only the statistic evidence of genetic risk within populations, but more importantly, whether those evidences could lead to the discovery of biologic pathways underlying polygenic diseases and traits. GWAS, in this way, can “recertify” many genes that have been experimentally identified to be important [23]. With the idea of biological network that are composed of genes and proteins, we can build a disease perturbation network in which all putative genes involved with that disease are displayed. It is of great interest to compare different diseases/phenotypes in terms of disease perturbation network. By doing this, we go beyond simple SNP associations and move forward to disease relationship by mining SNP knowledge.

1.5.1 Putative perturbed subnetworks based on GWA studies

It is now increasingly interested, for both biologists and medical practitioner, to further reveal the biological pathways or interaction networks underlying the surface of the statistical association studies of SNPs[5]. To state more concretely, how those
newly identified loci/relevant genes affected by SNPs are interwoven with an interaction pathway/network that can provide a biological reasons for common diseases? For example, the genetic variants that are associated with age-related macular degeneration strongly implicate components of the complement system, the loci associated with crohn’s disease point unambiguously to autophagy and interleukin-23-related pathways[23], and the height loci include genes encoding chromatin proteins and hedgehog signaling[24]. The notion of “subnetwork of perturbation based on genetics” is under the assumption that variation on the genome will eventually have accumulated alteration effects on biological network, causing those on disease state populations have distinctive subnetwork of perturbation comparing with non-disease state population. For example, the genetic propensity to develop Type 2 Diabetes (T2D) seems to involve genes in several different pathways that affect pancreatic β-cell formation and function, as well as pathways affecting fasting glucose levels and obesity[25]. Another example is that many of the loci associated with multiple sclerosis are identified to be involved with immune functions, including the interleukin receptor genes IL2RA and IL7RA, and the HLA-DRA locus[1]; those genes, when in “abnormal” state, are accumulated to affect the immune system subnetwork. Those studies raise a question whether those common diseases have shared genetic traits and relevant biological implications among them.

1.5.2 Phenome relationship based on GWA studies

The current challenge is whether we can exploit the GWA studies to diseases/pheontypes comparison. Of course, it is superficial if simply counting the shared SNPs or genes between interesting disease pairs. What we expect here is the biological network, or the “perturbation subnetwork based on genetics” which derived from SNPs that we could compare in order to find the commonalities. This concept was proposed by Atul et.al, for the creation of phenome-genome network. In the recent publication, it is supported that for a specific phenotype/disease, there exists a phenotype specific modules. For example, a module specific to “leukemia” datasets and a module specific to “skeletal muscle structure”, in which the former consists of 8
genes, both were strikingly homogeneous in gene functions of immune response. In this sense, we can say that phenotypically similar diseases are often caused by functionally related genes, being referred to as the modular nature of human genetic diseases. However, it is still unclear whether we can learn information from GWA studies and use network methods to compare phenotypes.

1.6 Contributions of this thesis

In this thesis, we conduct a large-scale disease comparison study by collecting all available GWAS data. Current SNP tagging and selection algorithm are effective in selecting the candidate representative SNPs for chromosomal regions that are in strong linkage disequilibrium (LD). However, the performance of tagged SNPs can be overestimated and as a result, current GWAS analysis might miss the important neighboring SNPs that are in fact contributing to disease pathogenesis [26]. For example, the two tagged SNPs around dynein 1 heavy chain 1 gene (Dync1h1) in a case-control study of a northern European derived population have no association with motor neuron degeneration (MND) whereas Dync1h1 has been experimentally proven to be associated with MND [27], suggesting that the actual causal SNPs may have been missed during the SNP tagging/selection process. In this paper, we address this problem based on the fact that the actual causal disease variants or SNPs might be in strong linkage disequilibrium with the tagged SNPs that have been identified to be associated with the diseases, and we can use linkage disequilibrium to fish out the possible missing genetic variants.

Moreover, instead of doing SNP-wise comparison between each pair of disease, we do network-wise comparison between disease pairs, that is, we consider the putative perturbed subnetworks that lie within those SNPs and compare all the proteins in the subnetworks. Our method can discover potential relations between diseases that are often ignored by single disease SNPs data alone.
Chapter 2
Methods

2.1 Convert SNPs to genes

We obtained diseases and their associated SNPs from the open access database of genome-wide association results curated by Andrew et al. (http://www.ncbi.nlm.nih.gov/pubmed/19161620) [6]. The database contains SNP data for 118 diseases, but many diseases have only a handful of SNPs. Thus, we focused on the 49 diseases that have at least 15 SNPs that are associated with diseases to make sure that we were able to get sufficient number of corresponding proteins. For brevity, we assigned a short name for each GWAS disease, for example, “ad” for Alzheimer's disease. The full abbreviated names for all 49 GWAS phenotypes are in the Table 1.

<table>
<thead>
<tr>
<th>Disease abbreviated name</th>
<th>Disease name</th>
</tr>
</thead>
<tbody>
<tr>
<td>hbf</td>
<td>Adult fetal hemoglobin levels (HbF) by F cell levels</td>
</tr>
<tr>
<td>ad</td>
<td>Alzheimer's disease</td>
</tr>
<tr>
<td>als</td>
<td>Amyotrophic Lateral Sclerosis</td>
</tr>
<tr>
<td>af</td>
<td>Atrial Fibrillation/Atrial Flutter</td>
</tr>
<tr>
<td>bd</td>
<td>Bipolar disorder</td>
</tr>
<tr>
<td>bl</td>
<td>Blood Lipids</td>
</tr>
<tr>
<td>bpas</td>
<td>Blood Pressure and Arterial Stiffness</td>
</tr>
<tr>
<td>bmg</td>
<td>Bone mass and geometry</td>
</tr>
<tr>
<td>ba</td>
<td>Brain aging</td>
</tr>
<tr>
<td>bc</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>qt</td>
<td>Cardiac repolarization (QT interval)</td>
</tr>
<tr>
<td>cdi</td>
<td>Celiac disease</td>
</tr>
<tr>
<td>ca</td>
<td>Childhood asthma</td>
</tr>
<tr>
<td>cc</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>cad</td>
<td>Coronary Artery Disease</td>
</tr>
<tr>
<td>chd</td>
<td>Coronary Heart Disease</td>
</tr>
<tr>
<td>cs</td>
<td>Coronary spasm</td>
</tr>
<tr>
<td>cd</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>cvd</td>
<td>CVD outcomes</td>
</tr>
<tr>
<td>eo</td>
<td>Early onset extreme obesity</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ecg</td>
<td>ECG and HR variability</td>
</tr>
<tr>
<td>ecgba</td>
<td>ECG dimensions</td>
</tr>
<tr>
<td>gd</td>
<td>Gallstone disease</td>
</tr>
<tr>
<td>gca</td>
<td>General cognitive ability</td>
</tr>
<tr>
<td>gla</td>
<td>Glaucoma</td>
</tr>
<tr>
<td>ht</td>
<td>Haemotological (blood) traits</td>
</tr>
<tr>
<td>hesp</td>
<td>Hair</td>
</tr>
<tr>
<td>hdl</td>
<td>HDL cholesterol</td>
</tr>
<tr>
<td>hei</td>
<td>Height</td>
</tr>
<tr>
<td>hae</td>
<td>Hepatic adverse events with thrombin inhibitor ximelagatran</td>
</tr>
<tr>
<td>hiv1</td>
<td>HIV-1 disease progression</td>
</tr>
<tr>
<td>hem</td>
<td>Human episodic memory</td>
</tr>
<tr>
<td>hyp</td>
<td>Hypertension</td>
</tr>
<tr>
<td>iman</td>
<td>Immunoglobulin A nephropathy</td>
</tr>
<tr>
<td>ic</td>
<td>Iris color</td>
</tr>
<tr>
<td>is</td>
<td>Ischemic stroke</td>
</tr>
<tr>
<td>kfet</td>
<td>Kidney function and endocrine traits</td>
</tr>
<tr>
<td>load</td>
<td>Late-onset Alzheimer's disease</td>
</tr>
<tr>
<td>lm</td>
<td>Lipid measurements</td>
</tr>
<tr>
<td>long</td>
<td>Longevity and age-related phenotypes</td>
</tr>
<tr>
<td>lc</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>mha</td>
<td>Minor histocompatibility antigenicity</td>
</tr>
<tr>
<td>ms</td>
<td>Multiple sclerosis</td>
</tr>
<tr>
<td>mi</td>
<td>Myocardial infarction</td>
</tr>
<tr>
<td>neu</td>
<td>Neuroticism</td>
</tr>
<tr>
<td>nd</td>
<td>Nicotine dependence</td>
</tr>
<tr>
<td>obe</td>
<td>Obesity-related traits</td>
</tr>
<tr>
<td>pd</td>
<td>Parkinson's disease</td>
</tr>
<tr>
<td>pa</td>
<td>Polysubstance addiction</td>
</tr>
<tr>
<td>psp</td>
<td>Progressive Supranuclear Palsy</td>
</tr>
<tr>
<td>pc</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>pr</td>
<td>Psoriasis</td>
</tr>
<tr>
<td>pf</td>
<td>Pulmonary function phenotypes</td>
</tr>
<tr>
<td>rls</td>
<td>Restless Leg Syndrome</td>
</tr>
<tr>
<td>ra</td>
<td>Rheumatoid Arthritis</td>
</tr>
<tr>
<td>sp</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td>slcl</td>
<td>Serum LDL cholesterol levels</td>
</tr>
<tr>
<td>spm</td>
<td>Skin pigmentation</td>
</tr>
<tr>
<td>sle</td>
<td>Systemic Lupus Erythematosus</td>
</tr>
<tr>
<td>scp</td>
<td>Sleep and circadian phenotypes</td>
</tr>
<tr>
<td>sals</td>
<td>Sporadic Amyotrophic Lateral Sclerosis (ALS)</td>
</tr>
<tr>
<td>spbc</td>
<td>Sporadic post-menopausal breast cancer</td>
</tr>
<tr>
<td>str</td>
<td>Stroke</td>
</tr>
</tbody>
</table>
Because the SNPs identified by different studies could be only a subset of disease-causing SNPs or neighboring SNPs that are closely linked to the actual disease-causing SNPs, we fished out additional SNPs based on the fact that SNPs that are in strong linkage disequilibrium with the already identified SNPs are strong candidates for the potentially missing disease-causing SNPs. We used the SNP functional annotation portal, a web database for exploring SNP function [28], to search and identify new SNPs that are in strong linkage disequilibrium with the already identified disease-causing SNPs (hereafter called seed SNPs). We obtained all the SNPs that have LD scores of $0.9<r^2<1$ with the seed SNPs. The criterion has been suggested previously due to the observation that it is about 30kbp upstream region of target genes, which is enriched with regulatory elements [28]. We then converted all the SNPs of each disease to genes/proteins based on the simple requirement that the SNPs must fall within the genes, regardless of whether the SNPs are in coding or non-coding regions. This straightforward conversion of SNPs to genes might be somewhat conservative as the SNPs that are associated with the diseases may contribute to the diseases by influencing not the host genes in which they reside, but the genes that are either further downstream or upstream of the host genes.

2.2 Locate putative subnetworks in each disease

We are interested in knowing how genes that harbor candidate disease-causing SNPs are potentially involved in the molecular mechanism of the pathogenesis of a disease; specifically, what is the protein interaction subnetwork formed by the genes? In order to address this question, we downloaded the STRING database that contains all the
known and predicted protein-protein interaction data and also direct (i.e. physical) and indirect (i.e. functional) associations. The protein-protein interactions and associations in the database are evaluated by composite criteria of multiple sources including genomic context, high throughput experiments, conserved co-expression, and existing literature, and are thus quite robust. We put seed proteins (i.e. the proteins that are converted from the SNPs) into the PPI to identify additional new proteins (hereafter called prey proteins) that interact with the seed proteins. We require that the prey proteins must have direct interaction with seed proteins for the following two reasons: first, the interaction confidence score between indirect protein interaction pairs will become weaker when more hop proteins (i.e. proteins in the path of indirectly interacted protein pairs) are included, making the result hard to interpret [29, 30]; Second, perturbed subnetworks will grow too dense to allow for any meaningful interpretation of biological networks [30]. The subnetworks formed by the seed and prey proteins are thus the candidate of perturbed subnetworks in the diseases that may explain what part of the network is affected in the diseases.

2.3 Measure disease similarity using Jaccard index and GO term IC scores

We are interested in constructing a disease relationship network (DRN) where the nodes are diseases and the edge weights indicated the degree of similarity between diseases. DRN can therefore provide us information on how various diseases are related to one another and a global view on disease similarities. Depending on the specific measurements used for edge weights, we expect that the resulting DRN can provide insight into different perspectives of disease relationships. Here we constructed DRN using two weight schemes. One is the Jaccard Index, defined as the size of intersection divided by the size of union of two sets, which is commonly used to measure the degree of similarity between two sets. Specifically, in the disease case, the Jaccard index between two diseases is calculated by the number of shared genes
divided by the total number of unique genes involved in the diseases. Therefore, the higher the Jaccard index is, the higher genetic similarity two diseases show. We calculated the Jaccard index for all pairwise comparisons of the 49 diseases and constructed a DRN. The other is the GO term IC (information content) score, introduced in [31, 32] to measure the semantic similarity in taxonomy. The informativeness of the lowest common ancestor between GO terms can be used as a measurement of semantic similarity, $s_{\text{Resnik}}(T_i, T_j) = IC_{\text{corpus}}(T_{\text{lcta}})$, where $T_{\text{lcta}}$ denotes the lowest common taxonomic ancestor between ontological terms $T_i$ and $T_j$. Each disease can be expressed as a collection of GO terms, and the more similar between the sets of GO terms, the more functional similarity the two diseases share. For the 49 diseases, there are many GO terms derived from the proteins that are likely to be associated with the diseases. However, our observation suggests that some GO terms are more relevant to the diseases than others and thus might dictate more the functional implications of disease phenotypes. Therefore, to better quantify the functional similarity between diseases, we should choose and compare those GO terms that are more close to the diseases than other ones that are not. The goal is to select the top ten most frequently occurred GO terms as the representative GO terms for each disease and calculate IC scores using Resnik’s values [32]. To achieve this, we clustered all terms using heuristic fuzzy partition algorithm developed from DAVID package [33]; If cluster number is greater than ten, for the first ten clusters, we chose the ten highest kappa value GO terms, else we chose ($EASE$ score/total $EASE$ score of all clusters)*10 GO terms for each cluster. Once the ten GO terms is selected, we compared the IC scores using Resnik’s method [32] for all pairwise ten GO terms of each disease with another disease, and used the average IC scores as the final measurements of GO term-term similarities between each pair of diseases. In summary, the two measurements of edge weights in the disease relation network complement each other and provide different perspectives for disease relationships. The disease relation network using GO term similarity score as edge weights focuses more on the function perspective of diseases, whereas the one using the extent of shared genes/proteins between diseases focuses more on the genetic perspective.
2.4 Network clustering methods

We used the Restricted Neighborhood Search Clustering (RNSC) method developed by Andrew et al. [34] to perform the clustering. We also tried the Markov clustering (MCL) method and found that the MCL method produces only a single big cluster even with different parameter settings. It is not clear why the MCL method failed to cluster the diseases. The RNSC method requires a cutoff value for the number of edges in the resulting clustered networks. For example, 49 edges network means that we set the cutoff of 49, that is, we limited one disease per edge for the network. Likewise, 98 edges network means that one disease per two edges; 147 edges means that one disease per three edges. We tried several cutoffs and compared their results.
Chapter 3
Results and discussion

3.1 Numbers of SNPs and genes associated with diseases

We compiled SNPs of 49 diseases/phenotypes from the open access database of genome-wide association results curated by Andrew et al[6]. Table 2 shows the number of SNPs that have been identified to be associated with each of the 49 diseases. On average, each disease has about 297 SNPs associated with. Some diseases have more SNPs identified than others, for example, Alzheimer’s disease (ad) has the largest number of SNPs (2325) and is likely the most extensively studied disease among the 49 diseases. In contrast, the Triglycerides disease (Tg) has only 15, the least number of SNPs, associated with it, possibly due to the limited number of small-scale studies on it. It is also possible that some diseases are caused by fewer SNPs than others and are inherently simpler in their genetic causes.

<table>
<thead>
<tr>
<th>Disease</th>
<th>SNPs #</th>
<th>Gene #</th>
<th>Ppi protein #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad</td>
<td>2325</td>
<td>782</td>
<td>3467</td>
</tr>
<tr>
<td>Ba</td>
<td>124</td>
<td>71</td>
<td>851</td>
</tr>
<tr>
<td>Bc</td>
<td>54</td>
<td>22</td>
<td>91</td>
</tr>
<tr>
<td>Bd</td>
<td>221</td>
<td>157</td>
<td>917</td>
</tr>
<tr>
<td>Bt</td>
<td>97</td>
<td>44</td>
<td>173</td>
</tr>
<tr>
<td>Bmg</td>
<td>99</td>
<td>60</td>
<td>603</td>
</tr>
<tr>
<td>Bpas</td>
<td>231</td>
<td>62</td>
<td>468</td>
</tr>
<tr>
<td>Ca</td>
<td>86</td>
<td>22</td>
<td>447</td>
</tr>
<tr>
<td>Cad</td>
<td>1059</td>
<td>348</td>
<td>1634</td>
</tr>
<tr>
<td>Cc</td>
<td>103</td>
<td>43</td>
<td>344</td>
</tr>
<tr>
<td>Cd</td>
<td>709</td>
<td>230</td>
<td>1624</td>
</tr>
<tr>
<td>Cdi</td>
<td>72</td>
<td>29</td>
<td>525</td>
</tr>
<tr>
<td>Cvd</td>
<td>94</td>
<td>44</td>
<td>328</td>
</tr>
<tr>
<td>Ecg</td>
<td>65</td>
<td>27</td>
<td>224</td>
</tr>
<tr>
<td>ecgba</td>
<td>115</td>
<td>55</td>
<td>400</td>
</tr>
<tr>
<td>Eo</td>
<td>43</td>
<td>30</td>
<td>663</td>
</tr>
<tr>
<td>Gd</td>
<td>251</td>
<td>148</td>
<td>1035</td>
</tr>
<tr>
<td>Hei</td>
<td>80</td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td>Hesp</td>
<td>65</td>
<td>23</td>
<td>151</td>
</tr>
<tr>
<td>hiv1</td>
<td>676</td>
<td>269</td>
<td>1523</td>
</tr>
<tr>
<td>Disease</td>
<td>Seeds</td>
<td>Preys</td>
<td>Total</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>hiv1-pro</td>
<td>676</td>
<td>268</td>
<td>1523</td>
</tr>
<tr>
<td>Ht</td>
<td>153</td>
<td>57</td>
<td>124</td>
</tr>
<tr>
<td>Hyp</td>
<td>16</td>
<td>8</td>
<td>48</td>
</tr>
<tr>
<td>Is</td>
<td>168</td>
<td>67</td>
<td>495</td>
</tr>
<tr>
<td>Kfet</td>
<td>84</td>
<td>45</td>
<td>285</td>
</tr>
<tr>
<td>Lm</td>
<td>138</td>
<td>79</td>
<td>595</td>
</tr>
<tr>
<td>Long</td>
<td>126</td>
<td>57</td>
<td>502</td>
</tr>
<tr>
<td>Mha</td>
<td>109</td>
<td>59</td>
<td>370</td>
</tr>
<tr>
<td>Ms</td>
<td>452</td>
<td>181</td>
<td>886</td>
</tr>
<tr>
<td>Nd</td>
<td>343</td>
<td>101</td>
<td>472</td>
</tr>
<tr>
<td>Neu</td>
<td>20</td>
<td>19</td>
<td>111</td>
</tr>
<tr>
<td>Obe</td>
<td>151</td>
<td>123</td>
<td>984</td>
</tr>
<tr>
<td>Pa</td>
<td>60</td>
<td>71</td>
<td>180</td>
</tr>
<tr>
<td>Pc</td>
<td>142</td>
<td>115</td>
<td>481</td>
</tr>
<tr>
<td>Pd</td>
<td>1604</td>
<td>995</td>
<td>2673</td>
</tr>
<tr>
<td>Pf</td>
<td>68</td>
<td>36</td>
<td>313</td>
</tr>
<tr>
<td>Qt</td>
<td>36</td>
<td>12</td>
<td>142</td>
</tr>
<tr>
<td>Ra</td>
<td>599</td>
<td>580</td>
<td>1680</td>
</tr>
<tr>
<td>Sa</td>
<td>98</td>
<td>83</td>
<td>688</td>
</tr>
<tr>
<td>Sals</td>
<td>137</td>
<td>67</td>
<td>192</td>
</tr>
<tr>
<td>Scp</td>
<td>30</td>
<td>38</td>
<td>177</td>
</tr>
<tr>
<td>Slcl</td>
<td>27</td>
<td>7</td>
<td>146</td>
</tr>
<tr>
<td>Sle</td>
<td>69</td>
<td>12</td>
<td>84</td>
</tr>
<tr>
<td>Slew</td>
<td>47</td>
<td>46</td>
<td>656</td>
</tr>
<tr>
<td>Sp</td>
<td>58</td>
<td>67</td>
<td>185</td>
</tr>
<tr>
<td>Spm</td>
<td>133</td>
<td>22</td>
<td>224</td>
</tr>
<tr>
<td>t1d</td>
<td>388</td>
<td>226</td>
<td>687</td>
</tr>
<tr>
<td>t2d</td>
<td>2036</td>
<td>1560</td>
<td>3837</td>
</tr>
<tr>
<td>Tg</td>
<td>15</td>
<td>11</td>
<td>96</td>
</tr>
</tbody>
</table>

**Table 2: Disease SNPs, genes, and PPI proteins**

We used the SNP functional annotation portal to get additional SNPs (prey SNPs) that are strongly linked to the set of SNPs (seed SNPs) that we compiled for the 49 diseases. Using the LD criterion, all the RefSNPs that are in strong linkage disequilibrium with the seed SNPs (i.e., 0.9<r^2<1.0) were obtained. Altogether, we were able to obtain additional number of SNPs for all the diseases, and the total number of SNPs for each disease is shown in Table 2.

All the SNPs were then converted to genes based on the simple criterion that they must fall within genes, regardless of whether the SNPs are in coding or noncoding regions. Table 2 shows the number of genes that are likely to be associated with each
of the 49 diseases.

3.2 Degree of similarity between diseases

We observed that for the number of shared SNPs (i.e. the SNPs that have been identified to be associated with both diseases of interest) between 49 diseases, many values are zero (Table 3) indicating that many disease pairs have no SNPs in common. When including prey SNPs, there are less zero values for the number of shared proteins between diseases (Table 4). The number of shared proteins is positively correlated with the number of shared SNPs (Pearson correlation coefficient: 0.353827, p-value = 0). Despite the apparent and significant consistency between the number of shared SNPs and shared genes, there is also notable disagreement. For example, for some disease pairs, the number of shared SNPs might be zero but the number of shared proteins may not be. HIV-1 and Crohn's disease shared no SNPs while they do share nine proteins. Table 5 shows the number of shared PPI proteins (i.e. proteins within perturbation subnetworks shared by two diseases) that there are rarely zero shared proteins between each disease pair, so we can more easily compare the difference of disease pair without null value. Moreover, the number of shared PPI proteins is highly correlated to the previous two measurements. For example, “Alzheimer's disease”(ad) and “Triglycerides”(tg) shared 186 proteins in Table 4, which is the highest among all diseases pairs with ad and they also shared 2028 proteins in supplement Table 5, which is also the highest. Some great differences in the number of shared proteins in Table 4 between disease pairs are not so obvious in Table 5. For example, in Table 4, “Alzheimer's disease”(ad) and “Brain aging”(ba) share eight proteins and “Alzheimer's disease” and “breast cancer”(bc) share five proteins, so from the perspective of shared proteins, “ad” and “ba” have similar degree of similarity to “ad and “bc”. In Table 5, “ad” and “ba” share 632 proteins, whereas “ad” and “bc” only 60 proteins.
Table 3: Shared SNP number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis)
Table 4: Shared gene number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis)

<table>
<thead>
<tr>
<th></th>
<th>Ba</th>
<th>bc</th>
<th>bd</th>
<th>bl</th>
<th>bmg</th>
<th>bpas</th>
<th>ca</th>
<th>cd</th>
<th>cdi</th>
<th>cvd</th>
<th>ecg</th>
<th>ecgba</th>
<th>eo</th>
<th>gd</th>
<th>hei</th>
<th>hesp</th>
<th>hiv1</th>
<th>lt</th>
</tr>
</thead>
<tbody>
<tr>
<td>ad</td>
<td>632</td>
<td>60</td>
<td>591</td>
<td>107</td>
<td>445</td>
<td>291</td>
<td>319</td>
<td>1014</td>
<td>221</td>
<td>1016</td>
<td>343</td>
<td>199</td>
<td>161</td>
<td>248</td>
<td>561</td>
<td>636</td>
<td>19</td>
<td>59</td>
</tr>
<tr>
<td>ba</td>
<td>11</td>
<td>206</td>
<td>36</td>
<td>138</td>
<td>113</td>
<td>139</td>
<td>319</td>
<td>72</td>
<td>331</td>
<td>99</td>
<td>73</td>
<td>73</td>
<td>135</td>
<td>198</td>
<td>256</td>
<td>16</td>
<td>24</td>
<td>288</td>
</tr>
<tr>
<td>bc</td>
<td>22</td>
<td>8</td>
<td>15</td>
<td>9</td>
<td>10</td>
<td>26</td>
<td>5</td>
<td>23</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bd</td>
<td>37</td>
<td>122</td>
<td>115</td>
<td>134</td>
<td>335</td>
<td>62</td>
<td>299</td>
<td>83</td>
<td>67</td>
<td>66</td>
<td>83</td>
<td>140</td>
<td>184</td>
<td>11</td>
<td>13</td>
<td>288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bl</td>
<td>30</td>
<td>29</td>
<td>15</td>
<td>57</td>
<td>11</td>
<td>75</td>
<td>22</td>
<td>22</td>
<td>11</td>
<td>17</td>
<td>22</td>
<td>42</td>
<td>1</td>
<td>5</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bmg</td>
<td>75</td>
<td>110</td>
<td>253</td>
<td>42</td>
<td>204</td>
<td>107</td>
<td>64</td>
<td>37</td>
<td>60</td>
<td>311</td>
<td>146</td>
<td>3</td>
<td>16</td>
<td>155</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bpas</td>
<td>38</td>
<td>178</td>
<td>42</td>
<td>191</td>
<td>36</td>
<td>27</td>
<td>18</td>
<td>41</td>
<td>67</td>
<td>98</td>
<td>3</td>
<td>4</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ca</td>
<td>217</td>
<td>38</td>
<td>211</td>
<td>106</td>
<td>49</td>
<td>41</td>
<td>58</td>
<td>127</td>
<td>148</td>
<td>5</td>
<td>26</td>
<td>165</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cad</td>
<td>132</td>
<td>604</td>
<td>152</td>
<td>106</td>
<td>69</td>
<td>133</td>
<td>331</td>
<td>350</td>
<td>38</td>
<td>39</td>
<td>472</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cc</td>
<td>344</td>
<td>35</td>
<td>32</td>
<td>25</td>
<td>33</td>
<td>43</td>
<td>63</td>
<td>7</td>
<td>10</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd</td>
<td>216</td>
<td>118</td>
<td>79</td>
<td>145</td>
<td>218</td>
<td>376</td>
<td>24</td>
<td>50</td>
<td>522</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cdi</td>
<td>41</td>
<td>55</td>
<td>55</td>
<td>107</td>
<td>169</td>
<td>4</td>
<td>21</td>
<td>252</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cvd</td>
<td>19</td>
<td>38</td>
<td>55</td>
<td>71</td>
<td>8</td>
<td>10</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ecg</td>
<td>13</td>
<td>39</td>
<td>64</td>
<td>4</td>
<td>6</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ecgba</td>
<td>73</td>
<td>72</td>
<td>1</td>
<td>18</td>
<td>136</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eo</td>
<td>274</td>
<td>7</td>
<td>18</td>
<td>175</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gd</td>
<td>16</td>
<td>29</td>
<td>450</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hei</td>
<td>0</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hesp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Shared PPI number between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis)

Table 6 shows disease pairs with top ten ranked Jaccard indexes and the number of shared proteins (see Table 7 for the complete results).

<table>
<thead>
<tr>
<th>Interaction pairs</th>
<th>Protein 1</th>
<th>Protein 2</th>
<th>Share protein #</th>
<th>Jaccard Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eo-tobe</td>
<td>663</td>
<td>984</td>
<td>462</td>
<td>0.389873</td>
</tr>
<tr>
<td>Ad-t2d</td>
<td>3467</td>
<td>3837</td>
<td>2028</td>
<td>0.384382</td>
</tr>
<tr>
<td>Pd-t2d</td>
<td>2673</td>
<td>3837</td>
<td>1690</td>
<td>0.350622</td>
</tr>
<tr>
<td>Ra-t1d</td>
<td>1680</td>
<td>687</td>
<td>607</td>
<td>0.344886</td>
</tr>
<tr>
<td>Ad-pd</td>
<td>3467</td>
<td>2673</td>
<td>1553</td>
<td>0.338565</td>
</tr>
<tr>
<td>bmg-sa</td>
<td>603</td>
<td>688</td>
<td>326</td>
<td>0.337824</td>
</tr>
</tbody>
</table>

18
### Table 6: Ten disease pairs with the highest Jaccard index

<table>
<thead>
<tr>
<th>ba</th>
<th>bc</th>
<th>bd</th>
<th>bl</th>
<th>bm</th>
<th>Bpa</th>
<th>ca</th>
<th>ce</th>
<th>cd</th>
<th>cd2</th>
<th>cvd</th>
<th>ecg</th>
<th>eeg</th>
<th>eg</th>
<th>eo</th>
<th>gd</th>
<th>hei</th>
<th>hes</th>
<th>hiv</th>
<th>htc</th>
</tr>
</thead>
<tbody>
<tr>
<td>715</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>223</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>351</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>402</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>755</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>433</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>715</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>505</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>356</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from the image provided.
Table 7: Jaccard index between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis)

3.3 Network clustering methods

We built DRNs using two measurements for edge weights, one is the Jaccard index (Figure 1) and the other is GO term IC scores (Figure 2). In these networks, each node represents a disease and each edge the degree of similarity measured by either the Jaccard index or GO term IC scores for the relationship between disease pairs. Disease relationship measured by the Jaccard index is only weakly correlated with that based on functional similarity Pearson correlation coefficient: 0.2157, p-value >0.05).
Figure 1: The disease relation network (DRN) based on the Jaccard index
Intuitively, if two diseases are similar enough in terms of shared genes, it is expected that they also have high degree of functional similarity regarding to their corresponding protein functions within “perturbation subnetwork”. Our result shows that this is not necessarily the case. For those disease pairs that have high Jaccard index, their “between GO scores” is not necessarily high. This seems to be contradictory to the recent study of Mehan et.al. [35], in which nearly all genes within “phenotype-specific modules” are homogeneous in functions. However, there are two reasons that can explain the difference. First, they used microarray data rather than genetic information of diseases. The modules within microarray expression data are
more consistent in functional implication during particular cell states. Second, since the data of our study comes from GWA studies on complex diseases, which are usually caused by accumulate and coordinate effects of multiple genes. These genes could be diversified regarding to their functions albeit participated in coordinated pathways that contributed together to the pathogenesis of a disease. It is valuable that we could identify disease similarity solely from their genetic information, thus rule out other complicated factors as environment and cell development. By doing that, we can trace back to whether phenotypically different diseases might have the same genetic root.

3.4 Clustering disease into groups

From the disease relationship networks, it is clear that some diseases are more related than others. It is therefore interesting to see whether we can cluster them into groups based on the degree of similarity among diseases. We used RNSC clustering methods as basis for clustering disease network. We limited the node per edge to one, two, and three cutoff respectively, and produce three networks. The first one is made up with 49 highest ranked Jaccard index edges; the second one 98, and the third one 147. Then the RNSC was employed to identify potential clusters with disease similarity. The detailed clustering result is shown in Table 8. Briefly, in the DRN with 49 edges, we identified five groups of diseases, with {HIV-1 disease progression; Alzheimer's disease; Type II Diabetes Mellitus; Parkinson's disease; Crohn's disease; Multiple sclerosis; Type I diabetes; Gallstone disease}, {hair, eye and skin pigmentation; skin pigmentation; brachial artery endothelial function; pulmonary function phenotypes}, {lipid measurements; serum LDL cholesterol levels; triglycerides; Ischemic stroke; minor histocompatibility antigenicity}, {neuroticism; sleep and circadian phenotypes}, and {kidney function and endocrine traits; ECG and HR variability; blood pressure and arterial stiffness; QT interval; systemic lupus erythematosus} Here the biggest cluster contains eight disease nodes including hiv1, ad (Alzheimer's disease), t2d (Type II
Diabetes Mellitus), pd (Parkinson's disease), cd (Crohn's disease), ms (Multiple sclerosis), t1d (Type I Diabetes) and gd (Gallstone disease), which suggests that they are closer to each other. It is difficult to evaluate the quality of the disease groups biologically as there seems to be no easy way that we could validate them computationally. Nevertheless, we decided to use the mimminer program, a program that analyzes the disease relationships based on literature mining[21], as an independent source to evaluate our results. Because there are different nomenclatures for the same disease, and there seems no easy way to cross reference them, we limited this analysis to a few diseases. Using Pd as the query disease, we found that ad and ms also appear in the list of 25 most similar diseases list, which indicates that Pd, ad, and ms are phenotypically connected (Table 9). Also, using Sle as the query disease, we found that cad is appear in the list of 25 most similar disease list, which indicates that Sle and cad are phenotypically connected (Table 10). Thus, at least for the two small clusters, there is independent literature support for their relationships. We noted that t1d and t2d belong to the same cluster. It is unclear why they are clustered together. We found that these two diseases have a large number of proteins, which may cause bias and increase the likelihood of them sharing genes and proteins, regardless of whether they are indeed related or not.

<table>
<thead>
<tr>
<th>Network</th>
<th>Cluster number</th>
<th>cluster 1</th>
<th>cluster 2</th>
<th>cluster 3</th>
<th>cluster 4</th>
<th>Cluster 5</th>
<th>cluster 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>49 edge</td>
<td>5</td>
<td>hiv1</td>
<td>hesp</td>
<td>Lm</td>
<td>Neu</td>
<td>Kfet</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>eo</td>
<td>spm</td>
<td>Sle</td>
<td>Sep</td>
<td>Ecg</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>obe</td>
<td>eegba</td>
<td>Tg</td>
<td></td>
<td>t1d</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ad</td>
<td>pf</td>
<td>Is</td>
<td></td>
<td>Bpas</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>t2d</td>
<td></td>
<td>Mha</td>
<td></td>
<td>Qt</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>pd</td>
<td></td>
<td></td>
<td></td>
<td>Sle</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ra</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>bmg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>cad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slew</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98 edges</td>
<td>6</td>
<td>hiv1</td>
<td>bd</td>
<td>Hesp</td>
<td>Lm</td>
<td>Neu</td>
<td>kfet</td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td>------</td>
<td>----</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>eo</td>
<td>long</td>
<td>Spm</td>
<td>Slel</td>
<td>Scp</td>
<td>ecg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>obe</td>
<td>ca</td>
<td>Ecgba</td>
<td>Tg</td>
<td>t1d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ad</td>
<td>Pf</td>
<td>Is</td>
<td>bpas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t2d</td>
<td>Mha</td>
<td>qt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pd</td>
<td>sle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ra</td>
<td>bmg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sa</td>
<td>cd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd</td>
<td>cad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ms</td>
<td>slew</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gd</td>
<td>ba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cdi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>147 edges</th>
<th>6</th>
<th>t1d</th>
<th>qt</th>
<th>Lm</th>
<th>hiv1</th>
<th>Kfet</th>
<th>hesp</th>
</tr>
</thead>
<tbody>
<tr>
<td>bpas</td>
<td>sle</td>
<td>Slel</td>
<td>hiv1-pro</td>
<td>Ecg</td>
<td>spm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neu</td>
<td>Tg</td>
<td>Eo</td>
<td>ecgba</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scp</td>
<td>Obe</td>
<td>pf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ad</td>
<td>is</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t2d</td>
<td>mha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pd</td>
<td>Ra</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmg</td>
<td>Sa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cd</td>
<td>Cad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ms</td>
<td>Slew</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gd</td>
<td>Ba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bd</td>
<td>Cdi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 8: Network size and clusters**
<table>
<thead>
<tr>
<th>#</th>
<th>Code</th>
<th>Score</th>
<th>Disorder Description</th>
<th>Gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>168600</td>
<td>1</td>
<td>PARKINSON DISEASE</td>
<td>PD SNCAIP TBP UCHL1 MAPT NDUFV2 NR4A2 SNCA LRRK2 PARK7 PARK2 PINK1</td>
</tr>
<tr>
<td>2</td>
<td>168601</td>
<td>0.5556</td>
<td>PARKINSON DISEASE, FAMILIAL, TYPE 1</td>
<td>PARK1 SNCA</td>
</tr>
<tr>
<td>3</td>
<td>127750</td>
<td>0.5537</td>
<td>DEMENTIA, LEWY BODY</td>
<td>DLB PRNP APOE CYP2D6</td>
</tr>
<tr>
<td>4</td>
<td>260540</td>
<td>0.5292</td>
<td>PARKINSON-DEMENTIA SYNDROME</td>
<td>MAPT</td>
</tr>
<tr>
<td>5</td>
<td>600116</td>
<td>0.5219</td>
<td>PARKINSON DISEASE, JUVENILE, AUTOSOMAL RECESSIVE</td>
<td>PDJ PARK2</td>
</tr>
<tr>
<td>6</td>
<td>183090</td>
<td>0.5048</td>
<td>SPINOCEREBELLAR ATAXIA 2</td>
<td>SCA2 ATXN2</td>
</tr>
<tr>
<td>7</td>
<td>600274</td>
<td>0.4841</td>
<td>FRONTOTEMPORAL DEMENTIA</td>
<td>MAPT PSEN1</td>
</tr>
<tr>
<td>8</td>
<td>128230</td>
<td>0.4761</td>
<td>DYSTONIA, PROGRESSIVE, WITH DIURNAL VARIATION</td>
<td>GCH1</td>
</tr>
<tr>
<td>9</td>
<td>168100</td>
<td>0.464</td>
<td>PARALYSIS AGITANS, JUVENILE, OF HUNT</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>601104</td>
<td>0.4622</td>
<td>SUPRANUCLEAR PALSY, PROGRESSIVE</td>
<td>PSP MAPT</td>
</tr>
<tr>
<td>11</td>
<td>213600</td>
<td>0.4513</td>
<td>FAHR DISEASE</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>109150</td>
<td>0.4477</td>
<td>MACHADO-JOSEPH DISEASE</td>
<td>MJD ATXN3</td>
</tr>
<tr>
<td>13</td>
<td>606324</td>
<td>0.445</td>
<td>PARKINSON DISEASE, TYPE 7, AUTOSOMAL RECESSIVE EARLY-ONSET</td>
<td>PARK7 PARK7</td>
</tr>
<tr>
<td>14</td>
<td>168605</td>
<td>0.4333</td>
<td>PARKINSONISM WITH ALVEOLAR HYPOVENTILATION AND MENTAL DEPRESSION</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>190300</td>
<td>0.431</td>
<td>TREMOR, HEREDITARY ESSENTIAL, 1</td>
<td>ETM1</td>
</tr>
<tr>
<td>16</td>
<td>607485</td>
<td>0.4231</td>
<td>DEMENTIA, HEREDITARY DYSPHASIC DISINHIBITION</td>
<td>HDDD MAPT</td>
</tr>
<tr>
<td>17</td>
<td>172700</td>
<td>0.4176</td>
<td>PICK DISEASE OF BRAIN</td>
<td>MAPT PSEN1</td>
</tr>
<tr>
<td>18</td>
<td>168610</td>
<td>0.4161</td>
<td>PALLIDOPONTONIGRAL DEGENERATION</td>
<td>PPND</td>
</tr>
<tr>
<td>19</td>
<td>607822</td>
<td>0.4128</td>
<td>ALZHEIMER DISEASE, FAMILIAL, TYPE 3</td>
<td>AD3</td>
</tr>
<tr>
<td>20</td>
<td>164500</td>
<td>0.411</td>
<td>SPINOCEREBELLAR ATAXIA 7</td>
<td>SCA7</td>
</tr>
<tr>
<td>21</td>
<td>105500</td>
<td>0.4076</td>
<td>AMYOTROPHIC LATERAL SCLEROSIS-PARKINSONISM/DEMENTIA COMPLEX OF GUAM</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>604326</td>
<td>0.4047</td>
<td>SPINOCEREBELLAR ATAXIA 12</td>
<td>SCA12</td>
</tr>
<tr>
<td>23</td>
<td>311510</td>
<td>0.4026</td>
<td>PARKINSONISM, EARLY-ONSET, WITH MENTAL RETARDATION</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>535000</td>
<td>0.3995</td>
<td>LEBER OPTIC ATROPHY</td>
<td>MIHSA</td>
</tr>
<tr>
<td>25</td>
<td>164400</td>
<td>0.3987</td>
<td>SPINOCEREBELLAR ATAXIA 1</td>
<td>SCA1</td>
</tr>
</tbody>
</table>

Table 9: Parkinson's disease as a query disease

| 1  | 152700 | 1 | LUPUS ERYTHEMATOSUS, SYSTEMIC | SLE | FCGR3A PDCD1 PTPN22 TNFSF6 CTLA4 |
| 2  | 217000 | 0.5621 | COMPLEMENT COMPONENT 2 DEFICIENCY | C2 |  |
| 3  | 601744 | 0.5127 | SYSTEMIC LUPUS ERYTHEMATOSUS, SUSCEPTIBILITY TO, 1 | SLEB1 | C1QA C2 C4A FCGR2A |
| 4  | 301000 | 0.4335 | WISKOTT-ALDRICH SYNDROME | WAS | WAS |
| 5  | 601859 | 0.4325 | AUTOIMMUNE LYMPHOPROLIFERATIVE SYNDROME | ALPS | TNFRSF6 |
| 6  | 306400 | 0.415 | GRANULOMATOUS DISEASE, CHRONIC | CGD | CYBB |
| 7  | 216950 | 0.413 | COMPLEMENT COMPONENT C1r DEFICIENCY | C1R |  |
| 8  | 306700 | 0.4085 | HEMOPHILIA A | F8 |  |
| 9  | 107320 | 0.4024 | ANTIPHOSPHOLIPID SYNDROME |  |  |
| 10 | 308300 | 0.3991 | INCONTINENTIA PIGMENTI | IP | IKBKG |
### Table 10: Systemic Lupus Erythematosus disease as a query disease

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>266600</td>
<td>0.3988</td>
<td>INFLAMMATORY BOWEL DISEASE 1</td>
<td>IBD1</td>
<td>CARD15</td>
</tr>
<tr>
<td>12</td>
<td>607624</td>
<td>0.3945</td>
<td>GRISCHELLI SYNDROME, TYPE 2</td>
<td>GS2</td>
<td>RAB27A</td>
</tr>
<tr>
<td>13</td>
<td>249100</td>
<td>0.393</td>
<td>FAMILIAL MEDITERRANEAN FEVER</td>
<td>FMF</td>
<td>MEFV</td>
</tr>
<tr>
<td>14</td>
<td>208900</td>
<td>0.3911</td>
<td>ATAXIA-TELANGIECTASIA</td>
<td>AT</td>
<td>ATM</td>
</tr>
<tr>
<td>15</td>
<td>234700</td>
<td>0.3852</td>
<td>HEART BLOCK, CONGENITAL</td>
<td>SSA2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>209920</td>
<td>0.3821</td>
<td>BARE LYMPHOCYTE SYNDROME, TYPE II</td>
<td>MHC2TA RFX5 RFXAP RFXANK</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>214500</td>
<td>0.3812</td>
<td>CHEDIAK-HIGASHI SYNDROME</td>
<td>CHS</td>
<td>LYST</td>
</tr>
<tr>
<td>18</td>
<td>102700</td>
<td>0.38</td>
<td>ADENOSINE DEAMINASE</td>
<td>ADA</td>
<td>ADA</td>
</tr>
<tr>
<td>19</td>
<td>260400</td>
<td>0.3779</td>
<td>SWACHMAN-DIAMOND SYNDROME</td>
<td>SDS</td>
<td>SBDS</td>
</tr>
<tr>
<td>20</td>
<td>192310</td>
<td>0.3747</td>
<td>VASCULITIS, HEREDITARY INFLAMMATORY, WITH PERSISTENT NODULES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>242900</td>
<td>0.3724</td>
<td>IMMUNOOSSEOUS DYSPLASIA, SCHIMKE TYPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>187360</td>
<td>0.3719</td>
<td>TEMPORAL ARTERITIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>225750</td>
<td>0.371</td>
<td>AICARDI-GOUTIERES SYNDROME 1</td>
<td>AGS1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>182410</td>
<td>0.3663</td>
<td>SNEDDON SYNDROME</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>230800</td>
<td>0.366</td>
<td>GAUCHER DISEASE, TYPE I</td>
<td>GBA</td>
<td></td>
</tr>
</tbody>
</table>

3.5 Conclusion and future work

In a recent study, Mehan et. al.[35] presented an integrative network approach for the study of similarity of phenotypes. Our method is comparable with their studies. However, instead of exploring the microarray data for each phenotype, we used genetic information gathered from GWA studies, or SNP set for each interesting disease or phenotype. We intend to identify the genetic basis of disease relationships
rather than based on expression state and environment fluctuation similarity between diseases that is essentially another dimension. It is in this aspect, we conclude, that the degree of similarity between disease pairs in our studies is based uniformly on genetic information. Future work will focus on validating our results using possibly microarray gene expression data and see how the disease relationship networks compare to one another.
Bibliography


2077.

