

# 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 $\leftrightarrow$ G or C $\leftrightarrow$ T) than transversions(A $\leftrightarrow$ C,A $\leftrightarrow$ T,G $\leftrightarrow$ C or G $\leftrightarrow$ 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  $D'$ (standard LD coefficient,  $D$ ) and  $r^2$ (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  $\beta$ -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/phenotypes 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.

Disease abbreviated name	Disease name
hbf	Adult fetal hemoglobin levels (HbF) by F cell levels
ad	Alzheimer's disease
als	Amyotrophic Lateral Sclerosis
af	Atrial Fibrillation/Atrial Flutter
bd	Bipolar disorder
bl	Blood Lipids
bpas	Blood Pressure and Arterial Stiffness
bmg	Bone mass and geometry
ba	Brain aging
bc	Breast cancer
qt	Cardiac repolarization (QT interval)
cdi	Celiac disease
ca	Childhood asthma
cc	Colorectal cancer
cad	Coronary Artery Disease
chd	Coronary Heart Disease
cs	Coronary spasm
cd	Crohn's disease
cvd	CVD outcomes
eo	Early onset extreme obesity

ecg	ECG and HR variability
ecgba	ECG dimensions
gd	Gallstone disease
gca	General cognitive ability
gla	Glaucoma
ht	Haematological (blood) traits
hesp	Hair
hdl	HDL cholesterol
hei	Height
hae	Hepatic adverse events with thrombin inhibitor ximelagatran
hiv1	HIV-1 disease progression
hem	Human episodic memory
hyp	Hypertension
iman	Immunoglobulin A nephropathy
ic	Iris color
is	Ischemic stroke
kfet	Kidney function and endocrine traits
load	Late-onset Alzheimer's disease
lm	Lipid measurements
long	Longevity and age-related phenotypes
lc	Lung cancer
mha	Minor histocompatibility antigenicity
ms	Multiple sclerosis
mi	Myocardial infarction
neu	Neuroticism
nd	Nicotine dependence
obe	Obesity-related traits
pd	Parkinson's disease
pa	Polysubstance addiction
psp	Progressive Supranuclear Palsy
pc	Prostate cancer
pr	Psoriasis
pf	Pulmonary function phenotypes
rls	Restless Leg Syndrome
ra	Rheumatoid Arthritis
sp	Schizophrenia
slcl	Serum LDL cholesterol levels
spm	Skin pigmentation
sle	Systemic Lupus Erythematosus
scp	Sleep and circadian phenotypes
sals	Sporadic Amyotrophic Lateral Sclerosis (ALS)
spbc	Sporadic post-menopausal breast cancer
str	Stroke

sa	Subclinical atherosclerosis
slew	Systemic Lupus Erythematosus (SLE)
tg	Triglycerides
t1d	Type I Diabetes
t2d	Type II Diabetes Mellitus
amd	Wet neovascular age-related macular degeneration (AMD)

**Table 1: Disease abbreviated names**

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,  $s_{\text{Resnik}}(T_i, T_j) = \text{IC}_{\text{corpus}}(T_{\text{lcta}})$ , 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.

<b>Disease</b>	<b>SNPs #</b>	<b>Gene #</b>	<b>Ppi protein #</b>
Ad	2325	782	3467
Ba	124	71	851
Bc	54	22	91
Bd	221	157	917
Bl	97	44	173
Bmg	99	60	603
Bpas	231	62	468
Ca	86	22	447
Cad	1059	348	1634
Cc	103	43	344
Cd	709	230	1624
Cdi	72	29	525
Cvd	94	44	328
Ecg	65	27	224
ecgba	115	55	400
Eo	43	30	663
Gd	251	148	1035
Hei	80	13	50
Hesp	65	23	151
hiv1	676	269	1523

hiv1-pro	676	268	1523
Ht	153	57	124
Hyp	16	8	48
Is	168	67	495
Kfet	84	45	285
Lm	138	79	595
Long	126	57	502
Mha	109	59	370
Ms	452	181	886
Nd	343	101	472
Neu	20	19	111
Obe	151	123	984
Pa	60	71	180
Pc	142	115	481
Pd	1604	995	2673
Pf	68	36	313
Qt	36	12	142
Ra	599	580	1680
Sa	98	83	688
Sals	137	67	192
Scp	30	38	177
Slcl	27	7	146
Sle	69	12	84
Slew	47	46	656
Sp	58	67	185
Spm	133	22	224
t1d	388	226	687
t2d	2036	1560	3837
Tg	15	11	96

**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.

	ba	bc	bd	bl	bmg	bpas	ca	cad	cc	cd	cdi	cvd	ecg	ecgba	eo	gca	gd	gla	hae	hbf	hdl	Hei
ad	0	0	0	0	0	0	0	5	0	1	0	0	0	0	4	0	3	0	1	0	0	0

ba	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bc		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bd			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bl				0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
bmg					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bpas						0	0	0	0	0	1	1	#	0	0	0	0	0	0	0	0
ca							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cad								0	0	0	0	0	0	0	0	0	1	0	0	0	0
cc									1	0	0	0	0	0	0	0	0	0	0	0	0
cd										0	0	0	0	0	0	0	0	0	0	0	0
cdi											0	0	0	0	0	0	0	1	0	0	0
cvd												1	1	0	0	0	0	0	0	0	0
ecg														0	0	0	0	0	0	0	0
ecgba															0	0	0	0	0	0	0
eo																0	1	0	0	0	0
gca																	0	0	0	0	0
gd																		0	0	0	0
gla																			0	0	0
hae																				0	0
hbf																					0

**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)**

	ba	bc	bd	bl	bmg	bpas	ca	cad	cc	cd	cdi	cvd	ecg	ecgba	eo	Gca	gd	gla	hae	hbf	hdl	hei
ad	8	5	24	8	8	4	1	48	5	26	7	4	5	8	10	0	17	0	2	1	0	0
ba		0	3	0	0	1	1	6	1	4	0	1	0	2	1	0	4	0	0	0	0	0
bc			2	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bd				2	1	2	0	7	1	3	0	3	0	4	0	0	5	0	0	0	0	0
bl					0	2	0	4	0	4	1	0	1	0	0	0	3	0	0	0	0	1
bmg						1	0	3	1	2	1	0	0	0	1	0	0	0	0	0	0	0
bpas							0	2	0	1	2	0	0	1	0	0	1	0	0	0	0	0
ca								0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cad									2	17	1	0	3	4	3	0	8	0	1	1	2	4
cc										43	0	0	0	0	0	0	1	0	0	0	0	0
cd											1	0	0	2	0	0	4	0	0	0	0	0
cdi												0	0	0	0	0	2	0	0	0	0	0
cvd													0	2	0	0	1	0	0	0	0	0
ecg														0	0	0	0	0	0	0	0	0
ecgba															0	0	2	0	0	0	0	0
eo																0	3	0	0	0	1	0
gca																	0	0	0	0	0	0
gd																		0	0	1	0	0

gla																			0	0	0	0
hae																				0	0	0
hbf																					0	0
hdl																						0

**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)**

	Ba	bc	bd	bl	bmj	bpas	ca	cc	cd	cdi	cvd	ecg	ecgba	eo	gd	hei	hesp	hiv1	ht
ad	632	60	591	107	445	291	319	1014	221	1016	343	199	161	248	561	636	19	59	870
ba		11	206	36	138	113	139	319	72	331	99	73	73	135	198	256	16	24	288
bc			22	8	15	9	10	26	5	23	4	7	6	5	9	14	0	1	44
bd				37	122	115	134	335	62	299	83	67	66	83	140	184	11	13	288
bl					30	29	15	57	11	75	22	22	11	17	22	42	1	5	52
bmj						75	110	253	42	204	107	64	37	60	311	146	3	16	155
bpas							38	178	42	191	36	27	18	41	67	98	3	4	149
ca								217	38	211	106	49	41	58	127	148	5	26	165
cad									132	604	152	106	69	133	331	350	38	39	472
cc										344	35	32	25	33	43	63	7	10	98
cd											216	118	79	145	218	376	24	50	522
cdi												41	55	55	107	169	4	21	252
cvd													19	38	55	71	8	10	64
ecg														13	39	64	4	6	83
ecgba															73	72	1	18	136
eo																274	7	18	175
gd																	16	29	430
hei																		0	12
hesp																			45

**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).

Interaction pairs	Protein 1	Protein 2	Share protein #	Jaccard Index
Eo-obe	663	984	462	0.389873
Ad-t2d	3467	3837	2028	0.384382
Pd-t2d	2673	3837	1690	0.350622
Ra-t1d	1680	687	607	0.344886
Ad-pd	3467	2673	1553	0.338565
bmj-sa	603	688	326	0.337824

bmg-eo	603	663	311	0.325654
Eo-sa	663	688	308	0.295302
Ra-t2d	1680	3837	1235	0.288417
hesp-spm	151	224	78	0.262626

**Table 6: Ten disease pairs with the highest jaccard index**

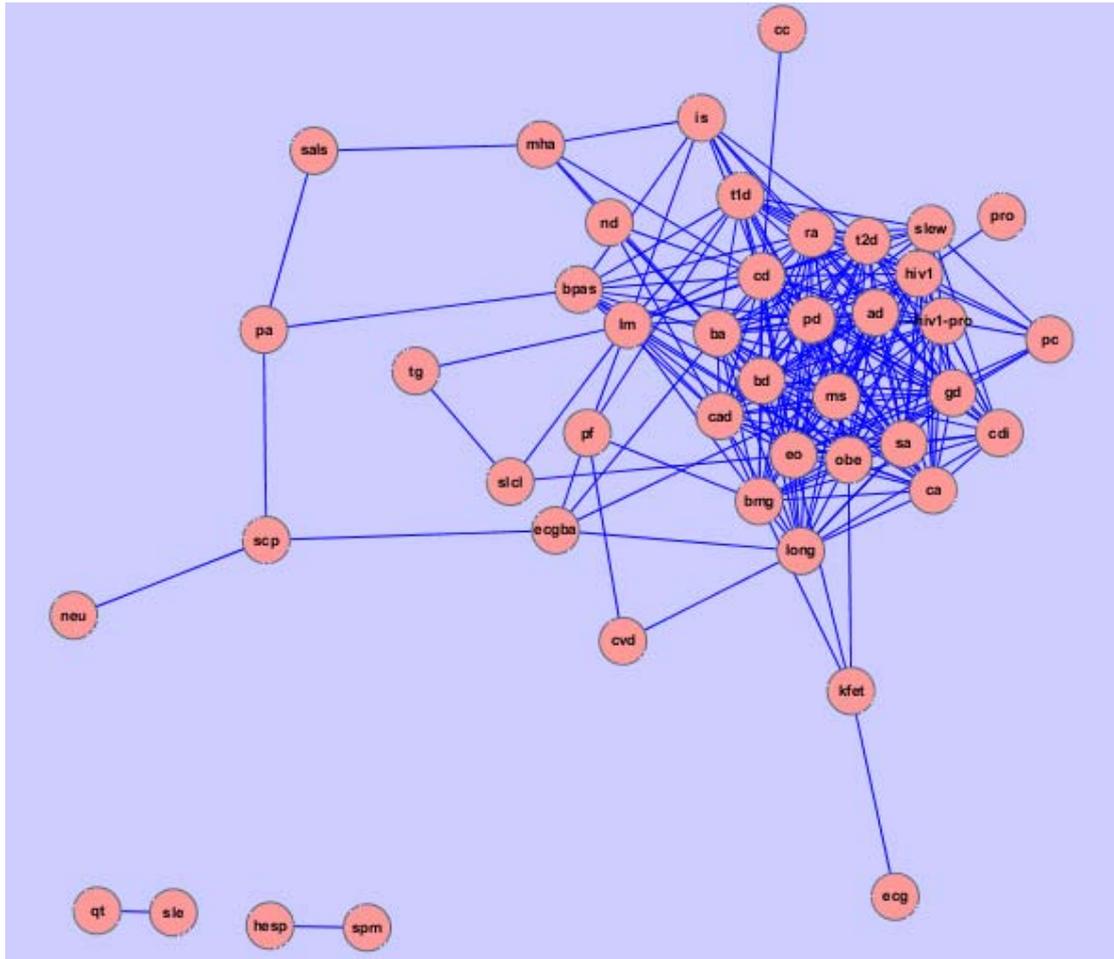
	ba	bc	bd	bl	bm g	Bpa s	ca	cc	cd	cdi	cvd	ecg	ecg ba	eo	gd	hei	hes p	hiv l	ht
ad	0.1 715	0.0 172	0.1 558	0.0 303	0.1 228	0.0 799	0.0 887	0.2 481	0.0 616	0.2 493	0.0 94	0.0 553	0.0 456	0.0 685	0.1 572	0.1 645	0.0 054	0.0 166	0.2 112
ba		0.0 118	0.1 319	0.0 364	0.1 049	0.0 937	0.1 199	0.1 473	0.0 641	0.1 544	0.0 775	0.0 66	0.0 729	0.1 21	0.1 505	0.1 571	0.0 181	0.0 245	0.1 381
bc			0.0 223	0.0 313	0.0 221	0.0 164	0.0 189	0.0 153	0.0 116	0.0 136	0.0 065	0.0 17	0.0 194	0.0 103	0.0 121	0.0 126	0 0	0.0 041	0.0 28
bd				0.0 351	0.0 873	0.0 906	0.1 089	0.1 512	0.0 517	0.1 334	0.0 611	0.0 569	0.0 614	0.0 673	0.0 972	0.1 041	0.0 115	0.0 123	0.1 338
bl					0.0 402	0.0 474	0.0 248	0.0 326	0.0 217	0.0 436	0.0 325	0.0 459	0.0 285	0.0 306	0.0 27	0.0 36	0.0 045	0.0 157	0.0 316
bm g						0.0 753	0.1 17	0.1 275	0.0 464	0.1 008	0.1 048	0.0 738	0.0 468	0.0 636	0.3 257	0.0 979	0.0 046	0.0 217	0.0 786
bpa s							0.0 433	0.0 925	0.0 545	0.1 005	0.0 376	0.0 351	0.0 267	0.0 496	0.0 63	0.0 698	0.0 058	0.0 065	0.0 809
ca								0.1 164	0.0 505	0.1 134	0.1 224	0.0 675	0.0 651	0.0 735	0.1 292	0.1 109	0.0 102	0.0 455	0.0 914
cad									0.0 715	0.2 276	0.0 757	0.0 571	0.0 386	0.0 7	0.1 684	0.1 509	0.0 231	0.0 223	0.1 758
cc										0.2 118	0.0 42	0.0 5	0.0 46	0.0 464	0.0 446	0.0 479	0.0 181	0.0 206	0.0 554
cd											0.1 117	0.0 643	0.0 447	0.0 772	0.1 054	0.1 647	0.0 145	0.0 29	0.1 989
cdi												0.0 505	0.0 793	0.0 632	0.0 99	0.1 215	0.0 07	0.0 321	0.1 403
cvd													0.0 356	0.0 551	0.0 588	0.0 55	0.0 216	0.0 213	0.0 358
ecg														0.0 213	0.0 46	0.0 536	0.0 148	0.0 163	0.0 499
ecg ba															0.0 737	0.0 528	0.0 022	0.0 338	0.0 761
eo																0.1 924	0.0 099	0.0 226	0.0 87
gd																	0.0 15	0.0 251	0.2 021
hei																		0 0	0.0 0.0

																			077
He																			0.0
sp																			276

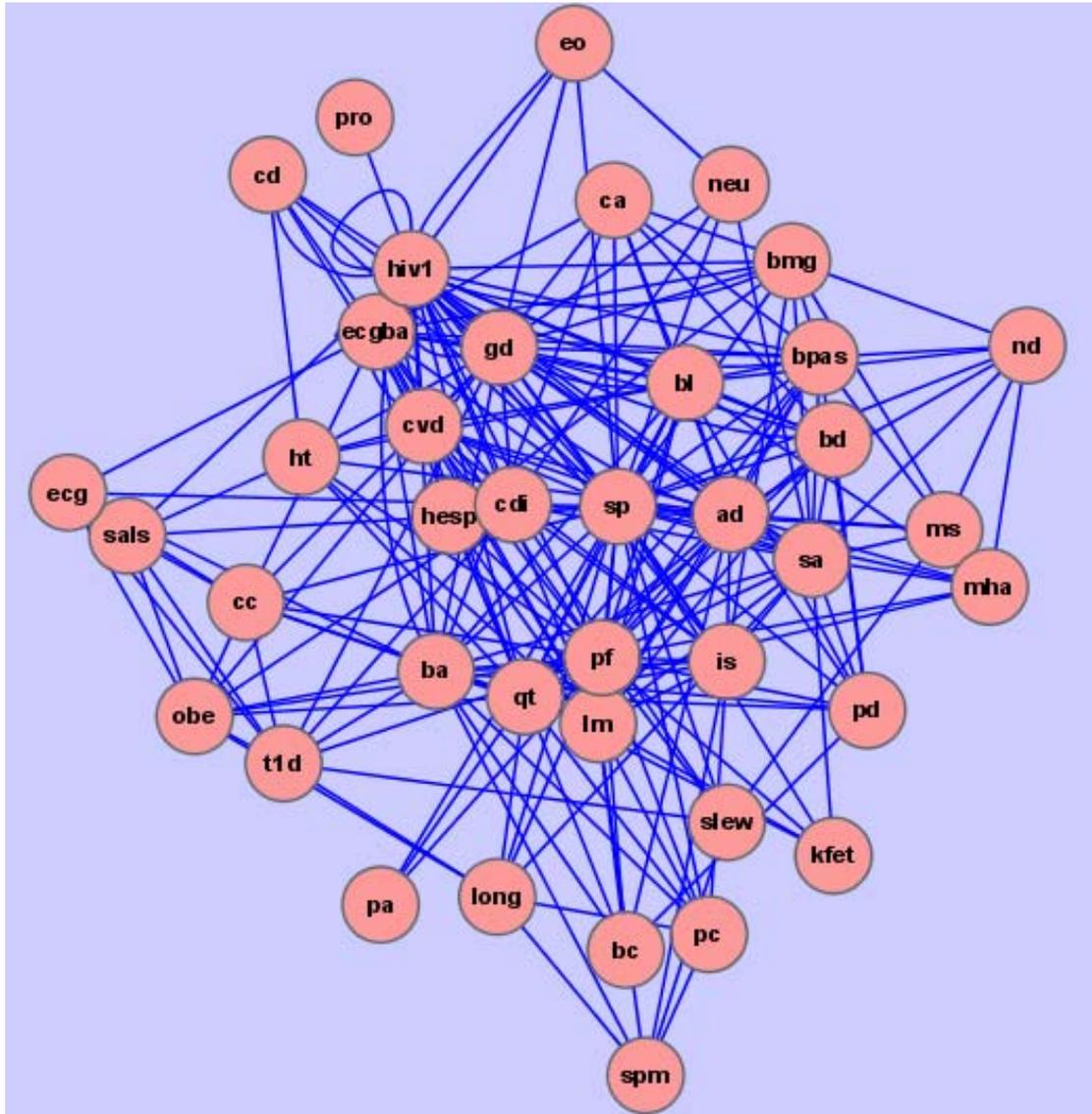
**Table 7: Jaccard index between disease pairs (Note: this is only part of the table, the actual table is too big to fit into this thesis)**

### 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**



**Figure 2: The DRN based on GO-Term IC score**

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 .

Network	Cluster number	cluster 1	cluster 2	cluster 3	cluster 4	Cluster 5	cluster 6
49 edge	5	hiv1	hesp	Lm	Neu	Kfet	
		eo	spm	Slcl	Scp	Ecg	
		obe	ecgba	Tg		t1d	
		ad	pf	Is		Bpas	
		t2d		Mha		Qt	
		pd				Sle	
		ra					
		bmg					
		sa					
		cd					
		cad					
		ms					
		slew					
		gd					
		ba					

		cdi					
		bd					
		long					
		ca					
98 edges	6	hiv1	bd	Hesp	Lm	Neu	kfet
		eo	long	Spm	Slcl	Scp	ecg
		obe	ca	Ecgba	Tg		t1d
		ad		Pf	Is		bpas
		t2d			Mha		qt
		pd					sle
		ra					
		bmj					
		sa					
		cd					
		cad					
		ms					
		slew					
		gd					
		ba					
		cdi					
147 edges	6	t1d	qt	Lm	hiv1	Kfet	hesp
		bpas	sle	Slcl	hiv1-pro	Ecg	spm
		neu		Tg	Eo		ecgba
		scp			Obe		pf
					Ad		is
					t2d		mha
					Pd		
					Ra		
					Bmj		
					Sa		
					Cd		
					Cad		
					Ms		
					Slew		
					Gd		
					Ba		
					Cdi		
					Bd		
					Long		

**Table 8: Network size and clusters**

<a href="#">1</a>	<a href="#">168600</a>	1	PARKINSON DISEASE	PD	SNCAIP TBP UCHL1 MAPT NDUFV2 NR4A2 SNCA LRRK2 PARK7 PARK2 PINK1
<a href="#">2</a>	<a href="#">168601</a>	0.5556	PARKINSON DISEASE, FAMILIAL, TYPE 1	PARK1	<a href="#">SNCA</a>
<a href="#">3</a>	<a href="#">127750</a>	0.5537	DEMENTIA, LEWY BODY	DLB	PRNP APOE CYP2D6
<a href="#">4</a>	<a href="#">260540</a>	0.5292	PARKINSON-DEMENTIA SYNDROME		<a href="#">MAPT</a>
<a href="#">5</a>	<a href="#">600116</a>	0.5219	PARKINSON DISEASE, JUVENILE, AUTOSOMAL RECESSIVE	PDJ	<a href="#">PARK2</a>
<a href="#">6</a>	<a href="#">183090</a>	0.5048	SPINOCEREBELLAR ATAXIA 2	SCA2	<a href="#">ATXN2</a>
<a href="#">7</a>	<a href="#">600274</a>	0.4841	FRONTOTEMPORAL DEMENTIA		MAPT PSEN1
<a href="#">8</a>	<a href="#">128230</a>	0.4761	DYSTONIA, PROGRESSIVE, WITH DIURNAL VARIATION		<a href="#">GCH1</a>
<a href="#">9</a>	<a href="#">168100</a>	0.464	PARALYSIS AGITANS, JUVENILE, OF HUNT		
<a href="#">10</a>	<a href="#">601104</a>	0.4622	SUPRANUCLEAR PALSY, PROGRESSIVE	PSP	<a href="#">MAPT</a>
<a href="#">11</a>	<a href="#">213600</a>	0.4513	FAHR DISEASE		
<a href="#">12</a>	<a href="#">109150</a>	0.4477	MACHADO-JOSEPH DISEASE	MJD	<a href="#">ATXN3</a>
<a href="#">13</a>	<a href="#">606324</a>	0.445	PARKINSON DISEASE, TYPE 7, AUTOSOMAL RECESSIVE EARLY-ONSET	PARK7	<a href="#">PARK7</a>
<a href="#">14</a>	<a href="#">168605</a>	0.4333	PARKINSONISM WITH ALVEOLAR HYPOVENTILATION AND MENTAL DEPRESSION		
<a href="#">15</a>	<a href="#">190300</a>	0.431	TREMOR, HEREDITARY ESSENTIAL, 1	ETM1	
<a href="#">16</a>	<a href="#">607485</a>	0.4231	DEMENTIA, HEREDITARY DYSPHASIC DISINHIBITION	HDDD	<a href="#">MAPT</a>
<a href="#">17</a>	<a href="#">172700</a>	0.4176	PICK DISEASE OF BRAIN		MAPT PSEN1

<a href="#">18</a>	<a href="#">168610</a>	0.4161	PALLIDOPONTONIGRAL DEGENERATION	PPND	<a href="#">MAPT</a>
<a href="#">19</a>	<a href="#">607822</a>	0.4128	ALZHEIMER DISEASE, FAMILIAL, TYPE 3	AD3	<a href="#">PSEN1</a>
<a href="#">20</a>	<a href="#">164500</a>	0.411	SPINOCEREBELLAR ATAXIA 7	SCA7	<a href="#">SCA7</a>
<a href="#">21</a>	<a href="#">105500</a>	0.4076	AMYOTROPHIC LATERAL SCLEROSIS-PARKINSONISM/DEMENTIA COMPLEX OF GUAM		
<a href="#">22</a>	<a href="#">604326</a>	0.4047	SPINOCEREBELLAR ATAXIA 12	SCA12	<a href="#">PPP2R2B</a>
<a href="#">23</a>	<a href="#">311510</a>	0.4026	PARKINSONISM, EARLY-ONSET, WITH MENTAL RETARDATION		
<a href="#">24</a>	<a href="#">535000</a>	0.3995	LEBER OPTIC ATROPHY		MIHSA
<a href="#">25</a>	<a href="#">164400</a>	0.3987	SPINOCEREBELLAR ATAXIA 1	SCA1	<a href="#">ATX1</a>

**Table 9: Parkinson's disease as a query disease**

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

<a href="#">11</a>	<a href="#">266600</a>	0.3988	INFLAMMATORY BOWEL DISEASE 1	IBD1	<a href="#">CARD15</a>
<a href="#">12</a>	<a href="#">607624</a>	0.3945	GRISCELLI SYNDROME, TYPE 2	GS2	<a href="#">RAB27A</a>
<a href="#">13</a>	<a href="#">249100</a>	0.393	FAMILIAL MEDITERRANEAN FEVER	FMF	<a href="#">MEFV</a>
<a href="#">14</a>	<a href="#">208900</a>	0.3911	ATAXIA-TELANGIECTASIA	AT	<a href="#">ATM</a>
<a href="#">15</a>	<a href="#">234700</a>	0.3852	HEART BLOCK, CONGENITAL		<a href="#">SSA2</a>
<a href="#">16</a>	<a href="#">209920</a>	0.3821	BARE LYMPHOCYTE SYNDROME, TYPE II		MHC2TA RFX5 RFXAP RFXANK
<a href="#">17</a>	<a href="#">214500</a>	0.3812	CHEDIAK-HIGASHI SYNDROME	CHS	<a href="#">LYST</a>
<a href="#">18</a>	<a href="#">102700</a>	0.38	ADENOSINE DEAMINASE	ADA	<a href="#">ADA</a>
<a href="#">19</a>	<a href="#">260400</a>	0.3779	SHWACHMAN-DIAMOND SYNDROME	SDS	<a href="#">SBDS</a>
<a href="#">20</a>	<a href="#">192310</a>	0.3747	VASCULITIS, HEREDITARY INFLAMMATORY, WITH PERSISTENT NODULES		
<a href="#">21</a>	<a href="#">242900</a>	0.3724	IMMUNOSKELETAL DYSPLASIA, SCHIMKE TYPE		<a href="#">SMARCAL1</a>
<a href="#">22</a>	<a href="#">187360</a>	0.3719	TEMPORAL ARTERITIS		
<a href="#">23</a>	<a href="#">225750</a>	0.371	AICARDI-GOUTIERES SYNDROME 1	AGS1	
<a href="#">24</a>	<a href="#">182410</a>	0.3663	SNEDDON SYNDROME		
<a href="#">25</a>	<a href="#">230800</a>	0.366	GAUCHER DISEASE, TYPE I		<a href="#">GBA</a>

**Table 10: Systemic Lupus Erythematosus disease as a query disease**

### 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

1. Maier LM, Lowe CE, Cooper J, Downes K, Anderson DE, Severson C, Clark PM, Healy B, Walker N, Aubin C *et al*: **IL2RA Genetic Heterogeneity in Multiple Sclerosis and Type 1 Diabetes Susceptibility and Soluble Interleukin-2 Receptor Production.** *Plos Genet* 2009, **5**(1):-.
2. Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JPT *et al*: **Genome-Wide Association Studies, Field Synopses, and the Development of the Knowledge Base on Genetic Variation and Human Diseases.** *Am J Epidemiol* 2009, **170**(3):269-279.
3. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations.** *Science* 2005, **307**(5712):1072-1079.
4. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**(7145):661-678.
5. Li Y, Agarwal P: **A pathway-based view of human diseases and disease relationships.** *PLoS One* 2009, **4**(2):e4346.
6. Johnson AD, O'Donnell CJ: **An Open Access Database of Genome-wide Association Results.** *Bmc Med Genet* 2009, **10**:-.
7. Huang W, Wang P, Liu Z, Zhang L: **Identifying disease associations via genome-wide association studies.** *BMC Bioinformatics* 2009, **10 Suppl 1**:S68.
8. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease.** *Science* 2008, **322**(5903):881-888.
9. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits.** *Nat Rev Genet* 2009, **10**(4):241-251.
10. Seng KC, Seng CK: **The success of the genome-wide association approach: a brief story of a long struggle.** *Eur J Hum Genet* 2008, **16**(5):554-564.
11. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M *et al*: **STRING 8-a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**:D412-D416.
12. Hardy J, Singleton A: **Genomewide association studies and human disease.** *N Engl J Med* 2009, **360**(17):1759-1768.
13. Teng S, Michonova-Alexova E, Alexov E: **Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions.** *Curr Pharm Biotechnol* 2008, **9**(2):123-133.
14. Capon F, Allen MH, Ameen M, Burden AD, Tillman D, Barker JN, Trembath RC: **A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups.** *Hum Mol Genet* 2004, **13**(20):2361-2368.
15. Kim LH, Lee HS, Kim YJ, Jung JH, Kim JY, Park BL, Shin HD: **Identification of novel SNPs in the interleukin 6 receptor gene (IL6R).** *Hum Mutat* 2003, **21**(4):450-451.
16. Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI: **An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets.** *Nat Genet* 2005, **37**(12):1320-1322.

17. Goldstein DB: **Common genetic variation and human traits.** *N Engl J Med* 2009, **360**(17):1696-1698.
18. **Integrating ethics and science in the International HapMap Project.** *Nat Rev Genet* 2004, **5**(6):467-475.
19. Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A: **An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population.** *Plos Genet* 2006, **2**(3):e27.
20. Stram DO: **Tag SNP selection for association studies.** *Genet Epidemiol* 2004, **27**(4):365-374.
21. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F: **Prediction of human disease genes by human-mouse conserved coexpression analysis.** *Plos Comput Biol* 2008, **4**(3):e1000043.
22. Gail MH, Pfeiffer RM, Wheeler W, Pee D: **Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies.** *Biostatistics* 2008, **9**(2):201-215.
23. Abecasis GR, Yashar BM, Zhao Y, Ghiasvand NM, Zareparsari S, Branham KE, Reddick AC, Trager EH, Yoshida S, Bahling J *et al*: **Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease.** *Am J Hum Genet* 2004, **74**(3):482-494.
24. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C *et al*: **Identification of ten loci associated with height highlights new biological pathways in human growth.** *Nat Genet* 2008, **40**(5):584-591.
25. Perry JRB, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, Frayling TM, Consor WTCC: **Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach.** *Diabetes* 2009, **58**(6):1463-1467.
26. Visscher PM, Montgomery GW: **Genome-wide association studies and human disease: from trickle to flood.** *JAMA* 2009, **302**(18):2028-2029.
27. Shah PR, Ahmad-Annur A, Ahmadi KR, Russ C, Sapp PC, Horvitz HR, Brown RH, Goldstein DB, Fisher EMC: **No association of DYNC1H1 with sporadic ALS in a case-control study of a northern European derived population: A tagging SNP approach.** *Amyotroph Lateral Sc* 2006, **7**(1):46-56.
28. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F: **SNP Function Portal: a web database for exploring the function implication of SNP alleles.** *Bioinformatics* 2006, **22**(14):e523-529.
29. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci U S A* 2005, **102**(6):1974-1979.
30. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks.** *Nat Biotechnol* 2005, **23**(5):561-566.
31. Resnik P: **Using information content to evaluate semantic similarity in a taxonomy.** *Int Joint Conf Artif* 1995:448-453
- 2077.
32. Resnik P: **Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.** *J Artif Intell Res* 1999, **11**:95-130.

33. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biol* 2007, **8**(9):R183.
34. King AD, Przulj N, Jurisica I: **Protein complex prediction via cost-based clustering.** *Bioinformatics* 2004, **20**(17):3013-3020.
35. Mehan MR, Nunez-Iglesias J, Kalakrishnan M, Waterman MS, Zhou XJ: **An integrative network approach to map the transcriptome to the phenome.** *J Comput Biol* 2009, **16**(8):1023-1034.