DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DISCOVERING DISEASE-CAUSING GENES BY NETWORK ANALYSIS

by
Samet TENEKECİ

August, 2019
İZMİR

DISCOVERING DISEASE-CAUSING GENES BY NETWORK ANALYSIS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of
Science in Computer Engineering

by
Samet TENEKECİ

August, 2019
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DISCOVERING DISEASE-CAUSING GENES BY NETWORK ANALYSIS"** completed by **SAMET TENEKECİ** under supervision of **ASST. PROF. DR. ZERRİN IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Zerrin IŞIK

Supervisor

Assoc. Prof. Dr. Derya BİRANT

(Jury Member)

Assist. Prof. Mustafa ÖZUYSAL

(Jury Member)

Prof. Dr. Kadriye ERTEKİN
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Asst. Prof. Dr. Zerrin IŞIK, for her guidance and useful advices that always show me the right direction in my research.

I would also like to thank my colleagues from Izmir Institute of Technology, who sincerely helped me whenever I need an advice throughout my M.Sc.

Of course, I would like to thank my beloved partner, *Seda*, who always listens to me and supports my decisions.

Last but not least, I would like to thank my parents, *Şenay* and *Erhan*, for being on my side in all important moments of my life.

A part of this study is presented as a poster in the 11th International Symposium on Health Informatics and Bioinformatics (HIBIT) and accepted as an oral presentation in ISCB RSG Turkey Student Symposium 2018.

Samet TENEKECİ

DISCOVERING DISEASE-CAUSING GENES BY NETWORK ANALYSIS

ABSTRACT

Identifying the common molecular mechanisms for metabolic disorders is crucial for early diagnosis and targeted drug therapies. However, the bioinformatics studies aiming to reveal shared disease genes remained limited because of the challenges arose from the complexity of the metabolic pathways. In this respect, we suggested an integrative bioinformatics model that combines multiple biological data sources and computational methods to identify shared disease genes in metabolic syndrome (MS), type 2 diabetes (T2D), and coronary artery disease (CAD).

We constructed weighted gene co-expression networks for each disease group by integrating peripheral blood gene expression data of 29 subjects, protein-protein interactions from STRING and INet, and Gene Ontologies. We clustered 90 disease networks, which are constructed using different parameters, by using MCL, SPICi, and Linkcomm algorithms and detected the disease modules. After comparatively evaluating the clustering results, we overlapped the networks providing the highest biological validity, and thus we obtained the common disease modules.

Our analyses revealed 22 shared genes in total for MS–CAD and T2D–CAD pairs. Moreover, 19 out of these 22 genes are directly or indirectly associated with relevant diseases in the previous medical studies. This integrative network based gene-disease association study on MS, T2D, and CAD offers potential insights into the common genetic mechanisms of the metabolic and cardiometabolic disorders.

Keywords: Gene expression, gene ontology, gene-disease association, protein-protein interaction, metabolic syndrome, type 2 diabetes, coronary artery disease

HASTALIĞA NEDEN OLAN GENLERİN AĞ ANALİZİ İLE BULUNMASI

ÖZ

Metabolik bozukluklardaki ortak moleküler mekanizmaların belirlenmesi, erken tanı ve hedefe yönelik ilaç tedavileri için çok önemlidir. Bununla birlikte, metabolik yolakların karmaşıklığı sebebiyle ortak hastalık genlerini ortaya çıkarmayı amaçlayan biyoinformatik çalışmaları sınırlı kalmıştır. Bu bağlamda, metabolik sendrom (MS), tip 2 diyabet (T2D) ve koroner arter hastalığı (KAH) tarafından paylaşılan ortak hastalık genlerini tanımlamak için çoklu biyolojik veri kaynaklarını ve hesaplama yöntemlerini birleştiren bütünleştirici bir biyoinformatik modeli önerdik.

29 hastanın periferik kan gen ifadelerini, STRING ve INet'ten protein-protein etkileşimlerini ve Gen Ontolojilerini birleştirerek her hastalık grubu için ağırlıklı gen ortak-ifade ağları oluşturduk. Farklı parametreler kullanarak oluşturduğumuz 90 hastalık ağını MCL, SPICi ve Linkcomm algoritmalarını kullanarak kümeledik ve hastalık modüllerini tespit ettik. Kümeleme sonuçlarını karşılaştırmalı olarak değerlendirdikten sonra, en yüksek biyolojik geçerliliği sağlayan ağları üst üste bindirdik ve böylece ortak hastalık modüllerini elde ettik.

Analizlerimiz MS–CAD ve T2D–CAD çiftleri için toplamda 22 paylaşılan gen ortaya çıkardı. Dahası, bu 22 genin 19'u daha önceki tıbbi çalışmalarda alakalı hastalıklarla doğrudan veya dolaylı olarak ilişkilendirilmiştir. MS, T2D ve CAD üzerinde yapmış olduğumuz bu bütünleştirici ağ tabanlı gen-hastalık ilişkilendirme çalışması, metabolik ve kardiyometabolik bozuklukların ortak genetik mekanizmaları hakkında potansiyel bilgi arz etmektedir.

Anahtar kelimeler: Gen ifadesi, gen ontolojisi, gen-hastalık ilişkilendirme, protein-protein etkileşimi, metabolik sendrom, tip 2 diyabet, koroner arter hastalığı

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

1.1 Motivation

Metabolic disorders consist of a large set of genetic diseases and characterized by enzyme deficiencies that disrupt the normal metabolic process by causing abnormal chemical reactions in the body. Despite the fact that most of the metabolic disorders are *inherited*, they can be *acquired* by environmental conditions, particularly by diet-induced factors.

Most of the metabolic disorders derive from *metabolic syndrome* (MS) that is a set of pathological conditions including insulin resistance, hypertension, hyperlipidemia, and abdominal obesity. Patients with MS have two to three fold increased risk of developing *cardiovascular diseases* and five fold increased risk of developing *diabetes mellitus* (Eckel, Grundy, & Zimmet, 2005; Wilson, D'Agostino, Parise, Sullivan, & Meigs, 2005; Zimmet, Shaw, & Alberti, 2005). Besides, according to The International Diabetes Federation's report published in 2015, 25% of the global population has MS (O'Neill & O'Driscoll, 2015).

Diabetes mellitus is a chronic condition diagnosed by the increased level of blood glucose arising from the body's inability in producing enough hormone insulin or ineffective usage of insulin. Type 2 diabetes mellitus (T2D), which the most prevalent form of diabetes, accounts for approximately 95% of all cases. (Centers for Disease Control and Prevention, 2011). As of 2017, 8.8% (425 m) of the world population diagnosed with diabetes and the annual healthcare expenditures exceed USD 727 billion (International Diabetes Federation, 2017).

Cardiovascular diseases (CVDs) consist of a group of heart and vessel disorder that can be divided into five main groups: coronary artery disease (CAD), peripheral arterial disease (PAD), cerebrovascular disease, renal artery stenosis (RAS), and aortic aneurysm (AOA). Cardiovascular diseases causing one in every three deaths are the primary cause of death globally. Besides, coronary artery disease induced deaths account for 80% of all cardiovascular deaths (Mendis, Puska, Norrving, & World Health Organization, 2011).

Over the past three decades, many medical researches have been carried out to reveal the relationship between metabolic disorders (Carr & Brunzell, 2004; De Rosa et al., 2018; Eckel et al., 2005; Grundy, Hansen, Smith, Cleeman, & Kahn, 2004; Grundy, 2004; Hanson, Imperatore, Bennett, & Knowler, 2002; Hu et al., 2004; Isomaa et al., 2001; Laaksonen et al., 2002; Wilson et al., 2005). Their results clearly demonstrated the tie between the CVDs, T2D, and MS. However, the underlying mechanisms and the interactions in molecular level are not well understood yet due to the complexity of metabolic connections. The complexity in question arises from the fact that metabolic disorders act on many metabolic pathways that produce a large number of potential risk factors and therefore it is extremely difficult to distinguish between the more important and the less important. Nevertheless, the bioinformatics studies, which have remained limited to now, are rapidly increasing by means of the recent developments in computational biology (Chan et al., 2014; Ko, Cho, Lee, & Kim, 2016; Liu, Jing, & Tu, 2016; Shu et al., 2017; Wei Zhao et al., 2017).

Our motivation in this study is to propose an integrative bioinformatics approach that aims to combine multiple biological data sources and computational methods to identify common disease related protein complexes in MS, T2D, and CAD. While doing this, as well as discovering novel disease genes, we also aim to evaluate the performance of different protein-protein interaction networks, GO semantic similarity measures, orthogonal ontologies, and graph clustering algorithms in disease gene prediction.

1.2 Problem Definition

Understanding the underlying molecular mechanisms of metabolic disorders is crucial not only to reveal the course of the disease but also to design targeted drug therapies. In this respect, discovering the shared disease genes and protein complexes for multiple metabolic disorders provides an insight into the status of the disease and improves the accuracy of the early diagnoses.

Detection of common disease related modules for multiple metabolic disorders essentially requires identification of disease genes in separate disease networks, which is already a challenging task for several reasons. These challenges arise from computational difficulty as well as the biological complexity. The problem is computationally complex owing to the fact that biological data sets are usually very large and noisy, that it makes them difficult to analyze. On the other hand, it is biologically complex, since the metabolic reactions act in so many pathways and identification of the disease related ones is quite challenging.

To overcome such problems, many statistical and computational approaches have been proposed. As well as these approaches get use of biological interaction data, gene expression data, sequence data, or Gene Ontology information, they may also integrate multiple data sources to improve the integrity and reliability of the results. Due to the nature of the problem, any disease gene discovery study requires some essential data mining steps such as data preprocessing, data mapping / integration, feature selection, clustering / classification, and validation of the results.

On the other hand, selecting the best method for disease gene discovery is another challenge since the performance of a method is highly dependent to the data set used. Thus, it is usually needed to perform a comparative evaluation on as many methods and configurations as possible.

In conclusion, pathway analysis and disease gene discovery in metabolic disorders are biologically and computationally challenging problems that require us to develop some integrative models combining multiple computational methods and biological data sources. In this research, we will answer whether such an integrative model could successfully identify common pathways in metabolic disorders, and then detect protein complexes that can be used in development of targeted drug therapies.

1.3 Contribution

In this study we worked on three different metabolic disorders and integrated multiple biological data sources including protein-protein interaction networks, gene expressions, and Gene Ontologies. After constructing functional interaction networks using 30 different configurations for each disease subject, we detected the disease related protein complexes using three different graph clustering approaches. Through the utilization of the best configuration, we analyzed the overlapping disease modules and revealed the common disease modules for the metabolic disorders in question.

As well as presenting a comparative performance evaluation of different biological databases, network construction methods, and graph clustering algorithms in disease gene prediction, this study is novel in virtue of being the initial effort to collectively analyze MS, T2D, and CAD in bioinformatics perspective.

A part of this study (Tenekeci & Isik, 2018) is presented as a poster in the 11th International Symposium on Health Informatics and Bioinformatics (HIBIT) and accepted as an oral presentation in ISCB RSG Turkey Student Symposium 2018.

1.4 Organization of the Thesis

This thesis consists of five chapters organized as follows:

In Chapter 2, we provide a detailed background information and a literature review in order to describe some essential concepts such as metabolic disorders, microarray data analysis, gene ontologies, protein-protein interaction networks, graph clustering algorithms, and integrated methods for disease gene prediction; and to present the related work on common disease-gene discovery for metabolic disorders.

In Chapter 3, we introduce our methodology in seven main sections including a general system overview, the gene expression data and protein-protein interaction data we have used, utilization of the GO semantic similarities, integration of various biological data sources, clustering of the disease networks, and overlapping of the clusters for common module identification.

In Chapter 4, we present the results of our analyses that cover a comparative evaluation of the network construction and clustering methods, clusters discovered for each disease network, common modules identified for each disease pair, and the biological evaluation of the findings.

In Chapter 5, we conclude our study and offer the future work.

CHAPTER TWO
LITERATURE REVIEW

In this chapter, we will first introduce the metabolic disorders and the relationship between them. Following the presentation of medical and genomic studies on metabolic disorders, we will explain the process of microarray data analysis and differential gene expression analysis in depth. Then, we will describe the gene ontology (GO), GO enrichment, and GO similarity analyses respectively. In addition to listing and comparing the protein-protein interaction networks (PPINs) and some of the well known PPI databases, we will introduce and evaluate different clustering algorithms available for PPIN clustering. After referring the integrated methods and databases for gene-disease association, we will present the novelty of our study.

2.1 Metabolic Disorders

Metabolic disorders form a large class of genetic diseases and characterized by enzyme deficiencies that alter the normal metabolic process by causing abnormal chemical reactions in the body. Most of the metabolic disorders are inherited and they also known as *congenital metabolic disorders* or *inborn errors of metabolism* (Garrod, 1908). However, environmental conditions can induce non-hereditary, or in other words *acquired* metabolic disorders. Some recent studies have established that environmental factor-related changes in the germ line of parents can be transmitted to future generations through epigenetic mechanisms. (Carone et al., 2010; Daxinger & Whitelaw, 2012). In particular, diet-induced metabolic alterations in mammals are passed from father to offspring (Q. Chen et al., 2016; Grandjean et al., 2015), suggesting sperm-mediated epigenetic inheritance (Rando, 2012).

2.1.1 Metabolic Syndrome (MS)

Metabolic disorders usually evolve from MS which is also named as *syndrome X* or *dysmetabolic syndrome*. MS refers to a set of pathological conditions characterized by hypertension, hyperlipidemia, insulin resistance, and abdominal obesity; and frequently named as a *precursor state* for cardiovascular diseases (CVDs) and T2D (Wilson et al., 2005).

The first descriptions of the clustering of various components of the MS has a century of history (Kylin, 1923). However, what we now call the MS was the definition of *syndrome X* suggested by Reaven in 1988 (Reaven, 1988). In the last two decades, global organizations have made a number of attempts to introduce a unified definition for MS. As a result, four slightly different definitions became popular worldwide introduced by:

- WHO (Alberti & Zimmet, 1998; World Health Organization, 1999)
- EGIR (Balkau & Charles, 1999)
- NCEP/ATP III (National Cholesterol Education Programme/Adult Treatment Panel III, 2002)
- IDF (Alberti, Zimmet, & Shaw, 2005; International Diabetes Federation, 2006)

IDF reports that nearly one fourth of the global population has MS (O'Neill & O'Driscoll, 2015) although this estimate varies widely due to the MS definition used as well as the ethnicity, sex, and age of the population under investigation (Kaur, 2014).

2.1.2 Type 2 Diabetes Mellitus (T2D)

Diabetes mellitus is a chronic condition in which blood sugar exceeds normal levels as a result of inability of the body to produce sufficient amount of insulin hormone or to use insulin effectively; and patients with MS are five fold more likely to develop diabetes (Eckel et al., 2005; Zimmet et al., 2005). Diabetes is divided into three main classes including type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes (GDM). However, T2D accounts for ~95% of all cases (Centers for Disease Control and Prevention, 2011). According to IDF diabetes atlas, global prevalence of diabetes is 8.8% (425 m) and total healthcare expenditures for diabetes is above USD 727 billion as of 2017. By 2045, this rate is expected to increase to 9.9% (629 m) while the cost reaches to USD 776 billion (International Diabetes Federation, 2017).

2.1.3 Coronary Artery Disease (CAD)

People with MS or T2D has two to three times increased risk of cardiovascular diseases (CVDs) (Eckel et al., 2005; Zimmet et al., 2005) those are a cluster of disorders in blood vessels and heart including: Renal artery stenosis (RAS), cerebrovascular disease, peripheral arterial disease (PAD), Coronary artery disease (CAD), and aortic aneurysm (AOA). CVDs are known to be the leading source of death globally that cause one in every three deaths (Mendis et al., 2011). On the other hand, CAD, also known as *ischemic heart disease* (IHD) or *coronary heart disease* (CHD), accounts for 75% of cardiovascular deaths in females and 80% of cardiovascular deaths in males (Mendis et al., 2011).

2.1.4 Medical Studies Associating MS, T2D, and CAD

Over the past three decades, many medical research papers have been published that clearly demonstrate the relation between MS, T2D, and CAD. In 2001, Isomaa *et al.* found that MS was present in ~80% of subjects with T2D and the presence of MS increased the risk of CHD three-fold and increased the risk of cardiovascular mortality and morbidity by 1.8-fold. (Isomaa et al., 2001). In 2004, Grundy described obesity-induced MS as a multidimensional risk factor for atherosclerotic cardiovascular disease (ASCVD) and T2D (Grundy, 2004). Again in 2004, Grundy *et al.* reported that in patients with MS, the risk of developing ASCVD increases at least twice, and the risk of developing T2D increases five times, regardless of gender (Grundy et al., 2004). In 2005, Wilson *et al.* observed that MS accounts for up to one third of CVD in men and approximately half of new T2D over 8 years of follow-up (Wilson et al., 2005). Many other studies demonstrated the association and parallel incidence of MS, T2D, and CAD (Table 2.1).

Table 2.1 Medical and computational studies on MS, T2D, and CAD

Publication	Method / Key Findings
(Isomaa et al., 2001)	MS is present in 80% of subjects with T2D. MS increases the risk of CHD by 3-fold, and CVD-caused mortality by 1.8-fold
(Laaksonen et al., 2002)	Persons with MS are at high risk for developing T2D during the 4-year follow-up.
(Hanson et al., 2002)	Among 890 originally non-diabetic participants with MS, 144 developed T2D in a follow-up of 4 years.
(S. M. Grundy, 2004)	Obesity-induced MS is a multidimensional risk factor for ASCVD and T2D.
(Grundy et al., 2004)	MS increases the risk of developing ASCVD twice and T2D 5-times, regardless of gender.
(Hu et al., 2004)	432 out of 1,119 deaths are caused by CVD. The overall hazard ratios of persons with MS are 1.44 times higher.
(Carr & Brunzell, 2004)	Persons with MS are at particularly high risk (20-30%) for premature CAD if they also have T2D.
(Wilson et al., 2005)	MS accounts for up to 1/3 of CVD in men and 1/2 of new T2D over 8 years of follow-up.
(Galassi, et al., 2006)	MS increases all-causes and CVD-caused mortality, as well as CVD, CHD, and stroke incidences.
(Skov, et al., 2012)	A pathway and network analysis. Displayed a statistically significant cluster of dysregulated genes in the arteries of diabetic patients.
(Chan et al., 2014)	An integrative pathway and network analysis. Identified multiple biological pathways and key regulatory genes involved in CVD and T2D.
(Dong et al., 2014)	An integrative network analysis identified four common pathways in T2D and CAD.
(Ko et al., 2016)	A novel approach that utilizes underlying molecular pathways and common disease-related genes to identify comorbid diseases through molecular interaction networks.
(Liu et al., 2016)	A WGCNA to identify specific hub genes and modules in CAD. 3711 genes and 21 modules associated with CAD.
(Shu et al., 2017)	An integrative analysis based on five multi-ethnic GWAS. Identified common disease sub-networks and metabolic pathways in T2D and CVD.
(Wei Zhao et al., 2017)	A genome-wide study on multiple ancestry groups including 265,678 T2D and 260,365 CHD subjects. Reported new genetic loci that are shared by CHD and T2D.

2.1.5 Computational Studies Associating MS, T2D, and CAD

Although there are many medical studies conducted to understand the relationship between MS, T2D, and CAD; the genome studies that aimed to reveal underlying mechanisms remained limited. The relationship in molecular level is not well understood since the connection between metabolic disorders is very complicated due to the fact that MS affects numerous metabolic pathways that produce a large number of potential risk factors, thus it is extremely difficult to distinguish the important ones from the insignificant ones.

Nevertheless, some promising discoveries have been made recently on these complex diseases with the help of recent developments in computational biology (Table 2.1). In 2014, with an integrative analysis on biological pathways and networks, Chan *et al.* discovered multiple biological pathways and key regulatory genes involved in CVD and T2D development (Chan et al., 2014). In 2016, Ko *et al.* proposed a novel approach that utilizes underlying molecular pathways and common disease-related genes to identify comorbid diseases through molecular interaction networks (Ko et al., 2016). In the same year, Liu *et al.* performed a weighted gene co-expression network analysis (WGCNA) to identify specific hub genes and modules associated with CAD; and they associated 3711 genes in 21 modules with CAD (Liu et al., 2016). In 2017, Shu *et al.* conducted a broad integrative analysis based on five multi-ethnic genome-wide association studies; and they identified the common disease sub-networks and metabolic pathways in T2D and CVD (Shu et al., 2017). In the same year, Zhao *et al.* performed a genome-wide study on multiple ancestry groups including 265,678 T2D and 260,365 CHD subjects; and they reported new genetic loci that are shared by CHD and T2D (Wei Zhao et al., 2017).

2.2 Microarray Data and Analysis

Gene expression is the process of synthesizing functional gene products (i.e. *phenotype*) using the gene information (i.e. *genotype*). The gene products are either proteins or functional RNAs, depending on whether the gene is coding or noncoding. On the other hand, gene expression profiling is the measurement of the activity (*expression level*) of thousands of genes at the same time to monitor cellular function at a global level. By examining these profiles, we can distinguish between diseased and healthy cells, or observe how cells respond to a specific treatment. Various transcriptomic technologies may be used to generate analytical data.

DNA microarrays, also called as *biochips* or *DNA chips*, are one of the most popular transcriptomic technologies. They are solid surfaces consisting of microscopic DNA spots. Each spot involves 10^{-12} moles (i.e. *picomoles*) of DNA sequence that is named as *probes*. A probe can be a short section of a gene or other DNA element that are utilized in probe-target hybridization.

DNA microarrays provide a picture of all transcriptional activity in a biological sample. Different from most conventional molecular biology tools, that usually allow a single or very small number of genes to run, microarrays facilitate the discovery of completely new and unexpected functional roles of gene. The power of these tools has been applied to a variety of applications, such as exploring new disease sub-types, identifying underlying disease or drug response mechanisms, and developing new diagnostic tools. However, DNA microarray technology produces a large amount of data that forces us to analyze using modern statistical and computational tools.

2.2.1 Differential Gene Expression Analysis

Because of the fact that the microarray data sets are commonly very large to perform computational analyses on, dimensional reduction has become one of the primary tasks in any bioinformatics study, regardless of the computational method that will be used. By performing some statistical analyses, one can observe the quantitative changes in gene expression levels among two or more sample groups and eliminate the insignificant features (genes) to reduce the dimension of data; that is called differential gene expression analysis.

However, detecting the differentially expressed genes (***DEGs***) is a challenging task as it aims to eliminate as much data as possible while minimizing the loss of significant genes. There are different methods for differential expression analysis. such as *DESeq* (Anders & Huber, 2010) and *edgeR* (Robinson, McCarthy, & Smyth, 2010) based on negative binomial distributions; or *EBSeq* (Leng et al., 2013) and *baySeq* (Hardcastle & Kelly, 2010) which are Bayesian approaches using a negative binomial model. However, it is very important to take the experimental design into consideration when choosing an analysis method. While some of the tools can only perform pair-wise comparison, others such as *edgeR*, *limma* (Smyth, 2005), and *DESeq* can perform multiple comparisons.

2.2.1.1 Fold Change

Using fold-changes (***FC***) is one of the simplest and most popular methods that is used to discover DEGs. A FC can be described as the ratio between two samples (e.g., for given samples *I* and *J*, the FC of *J* with respect to *I* is computed as *J/I*). However, it is more common to use logarithmic fold change ratios, also named as *logFC*, *log2FC* or *loget* (Pacholewska, 2017), in microarray experiments because proportional changes are more biologically meaningful than additive chances.

The log2FC can be defined as $log_2(E_i / E_j)$ where E_i and E_j are gene expression values for two samples *I* and *J* (e.g. different subjects or conditions). To obtain DEGs using the log2FC, an arbitrary threshold (τ) value is selected and all genes that differ by more than τ are considered as DEGs. Since both down-regulated (*log2FC* ≤ -*τ*) and up-regulated (*log2FC* ≥ *τ*) genes are considered to be differentially expressed, the absolute values can be used in DEG detection (|*log2FC*| ≥ *τ*). Additionally, *τ* = *1* is a very common definition due to the fact that the FC with 2-fold usually accepted a sufficient cutoff.

2.2.1.2 The t-test

Although the FC cutoffs are very useful in producing biologically meaningful results, they have some limitations such as not taking variability into account or not guaranteeing reproducibility (Y. Chen, Dougherty, & Bittner, 1997). Therefore, it has become very common to benefit from the traditional statistical tests.

Two-sample *t*-test is a straightforward method to use (Peck & Devore, 2011). However, two-sample *t*-test descriptions differentiate according to two conditions: 1) whether it is logical to assume that gene expression levels show an equal variance under the compared conditions, 2) whether the sample size (K_1 and K_2) is large. Since typically both K_1 and K_2 are small and variances are unequal in gene expression data (Thomas, Olson, Tapscott, & Zhao, 2001), it will be relevant to use the *t*-test with two normally-distributed, independent, small samples with unequal variances.

Let E_{jk} be the expression level of gene *j* with under condition *k*. For two conditions, *k=1* and *k=2*, If the sample means are:

$$\bar{E}_{A(1)}=\frac{\sum_{k=1}^{K_1} E_{Ak}}{K_1}\ ,\ \bar{E}_{A(2)}=\frac{\sum_{k=K_1+1}^{K_1+K_2} E_{Ak}}{K_2} \tag{2.1}$$

and the variances are:

$$s_{A(1)}^2=\frac{\sum_{k=1}^{K_1}(E_{Ak}-\bar{E}_{A(1)})^2}{K_1-1}\ ,\ s_{A(2)}^2=\frac{\sum_{k=K_1+1}^{K_1+K_2}(E_{Ak}-\bar{E}_{A(2)})^2}{K_2-1} \tag{2.2}$$

Then the *t*-statistic will be:

$$T_A=\frac{\bar{E}_{A(1)}-\bar{E}_{A(2)}}{\sqrt{s_{A(1)}^2/K_1+s_{A(2)}^2/K_2}} \tag{2.3}$$

By comparing the T_j with the critical value that is corresponding to the sample size (degree of freedom) and the desired confidence level (*p*-value), one can determine if a gene is differentially expressed or not through *t*-statistics. Although the significance thresholds can be chosen arbitrarily, it is very common to set *p*-value ≤ 0.01 or *p*-value ≤ 0.05.

2.2.1.3 Combined Methods

Although the conventional statistical methods seemed to be an alternative to FC at the first place, they were soon found to have other limitations such as giving high false discovery rates (*FDRs*) in small samples, or being weakly related with FC (McCarthy & Smyth, 2009). Therefore, meeting both FC cutoff and *p*-value criteria together has become a general opinion in DEG detection.

Patterson *et al.* applied statistical comparison cutoffs (*p*-value < 0.01 or *p*-value < 0.05) on different FC values (FC > 1.5, FC > 2 or FC > 4) to identify the significant DEGs (Patterson et al., 2006). They presented that the sets of DEGs obtained using the combinations of *p*-value and FC are more concordant with microarray platforms in comparison to the ones obtained using *p*-value or FC alone. Similarly, Huggins *et al.* considered DEGs significant if they satisfy a FC of at least 1.3 (FC > 1.3) and a statistical significance criteria (*p*-value < 0.2) simultaneously (Huggins et al., 2008). As a result, they showed that the DEG lists generated using this combination are biologically more significant than the ones generated using *p*-values alone. On the other hand, Peart *et al.* and Raouf *et al.* required genes to satisfy a maximum *p*-value, which is adjusted for multiple-testing (*p*-value < 0.05), and a minimum fold-chage condition (FC > 1.5) (Peart et al., 2005; Raouf et al., 2008).

2.3 Gene Ontology (GO)

The GO platform is a comprehensive bioinformatics knowledge-base that includes controlled and structured (i.e. machine-readable and human-readable) vocabulary of terms, providing a uniformed annotation for attributes of genes and functional gene products in a wide variety of species (The Gene Ontology Consortium, 2008). The GO project emerged as a result of collaborative efforts and contributions of several bioinformatics resource centers and major model-organism databases; and assist biomedical researchers in the annotation phase of large-scale molecular biology and genetics experiments.

An ontology in GO is a hierarchy of terms in three key biological domains that are common to all organisms: cellular component (CC), biological process (BP), and molecular function (MF). Cellular components consist of the sections of a cell and its extracellular elements. Biological processes represent all biological operations and elemental activities that are related to the functionality of integrated living units such as cells, tissues, organs, and organisms. On the other hand, molecular functions correspond to the molecular level activities of gene products (i.e. protein or RNA), like catalysis or binding.

2.3.1 GO Enrichment Analysis

A main practice of the GO is to conduct term enrichment analysis within a subset of genes. In particular, one can identify the GO terms that are significantly over-represented or under-represented under specific conditions, by performing term enrichment analysis on a differentially expressed gene list.

There are a several methods and tools to realize term enrichment using GO, like DAVID (Dennis et al., 2003), GO::TermFinder (Boyle et al., 2004), Blast2GO (Conesa et al., 2005), g:Profiler (Reimand, Kull, Peterson, Hansen, & Vilo, 2007), GSEA (Subramanian, Kuehn, Gould, Tamayo, & Mesirov, 2007), ToppGene (J. Chen, Bardes, Aronow, & Jegga, 2009), GOrilla (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009), and topGO (Alexa, Rahnenführer, & Lengauer, 2006).

These tools and methods may be diversified according to the input type, type of correction method applied, or type of the statistical tests applied. Some use ranked gene lists as input, while others use unranked ones. There are also more complicated methods that avoid arbitrary cutoffs by enabling each gene to be associated with an expression level. In addition, the most widely used statistical significance tests are *Hypergeometric test* and *Fisher's exact test* (D. W. Huang, Sherman, & Lempicki, 2009; Rivals, Personnaz, Taing, & Potier, 2007); and the most common correction methods are *Bonferroni* and *FDR*.

2.3.2 GO Similarity Analysis

It is possible to measure the level of functional similarity of genes and proteins by the semantic similarity of their GO annotations. Thus, the GO similarities have been commonly used in computational biology applications, especially in the studies that perform gene clustering (Bolshakova, Azuaje, & Cunningham, 2005; Wolting, McGlade, & Tritchler, 2006), gene function prediction (Tao, Sam, Li, Friedman, & Lussier, 2007), and protein localization (Lei & Dai, 2006).

To measure the semantic similarity among multiple GO terms, various strategies have been developed; and their correlation with protein-protein interactions (Xu, Du, & Zhou, 2008), gene expressions (Sevilla et al., 2005), and sequence similarities (Lord, Stevens, Brass, & Goble, 2003) have been confirmed.

Some methods (Dennis et al., 2003; Gunther et al., 2005) only consider the functional similarity of genes and use the kappa statistics of similar annotations or the frequency of incidence of GO terms when calculating the similarities. However, the main shortcoming of these approaches is that they ignore semantic relationships, such as *is-a* and *part-of*, between the terms. Other methods developed for natural language taxonomies (J. J. Jiang & Conrath, 1997; Lin, 1998; Resnik, 1999), measure the semantic similarity by considering the distance of two terms to the closest common ancestor term. The comparative evaluations (Guo, Liu, Shriver, Hu, & Liebman, 2006; Sevilla et al., 2005; H. Wang, Azuaje, Bodenreider, & Dopazo, 2004) have shown that each approach has its own advantages and drawbacks, but that Resnik's method provides a higher correlation with gene expression and sequence similarity.

2.4 Protein–Protein Interaction Networks (PPINs)

PPINs are complex topological architectures that fulfill a specific biological function through electrostatic forces and biochemical events. On the other hand, a *protein interactome* represents the whole set of protein-protein interactions partaking within a biological system.

A PPIN is commonly modeled as an undirected graph, $G(V, E)$, where V denotes the nodes or vertices (i.e. proteins), and E denotes the edges (i.e. pairwise protein-protein interactions). An interaction between two proteins may represent the information of physical association, functional association, co-localization, or direct interaction. Edges are usually undirected and weighted, connect pairs of interacting proteins and the edge weights represent how strongly interacting two proteins. Sometimes, edge weights denote the reliability information associated to the corresponding interactions. PPINs have similar topological features with the real-world networks, like communication networks and social networks. Typically, PPINs are scale-free networks with high-degree of clustering and small-world property (X.-F. Zhang, Dai, Ou-Yang, & Yan, 2014).

PPIs serve in almost every cellular process, so understanding the main mechanism of these interactions is crucial to infer novel functions of gene products, to support predictions in pathogenesis studies, or to detect the alterations in cell physiology of diseased samples. It is also important for drug development (Hopkins, 2008), since drugs that bind to proteins change some functions in a corrupted PPIN.

Discovering PPINs has been a major challenge in computational biology over the last decade. Consequently, the bioinformaticians developed several strategies and methods to identify PPINs. These strategies can be separated by each other according to the experimental and computational methods they use (Jordán, Nguyen, & Liu, 2012).

There are two types of experimental approaches, the traditional (low-throughput) and the high-throughput ones. The traditional approaches provide high resolution (atomic level) outcomes and are thus very detailed and informative. They supply essential information to identify disease related protein complexes and design target oriented molecular therapies. However, they are low-throughput because of the high demands in sample quality and amounts of material needed for structural determination. On the other hand, high-throughput methods report interactions at a larger scale and assess interactions globally at cellular level. They offer information at a relatively low resolution and just report the existence of particular inter-molecular interactions. Still, they provide a useful basis for further experimentation and analysis of molecular networks within cells or organelles. The main disadvantages of high-throughput methods are being labour-intensive and having high false negative and false positive rates (Podobnik et al., 2016).

On the other hand, the computational methods can be divided into three groups according to the source of information they use: genome-based, sequence-based, or structure-based (Schuster-Böckler & Bateman, 2008). In addition, it is also possible to integrate multiple data sources by Bayesian network approach (Jansen et al., 2003), probabilistic decision tree approach (L. V. Zhang, Wong, King, & Roth, 2004), kernels methods (Ben-Hur & Noble, 2005), or the hybrid approach (Bui, Katrenko, & Sloot, 2011).

The rapid advance in computational methods in the recent years have provided bioinformaticians an opportunity to develop massive PPI databases (Table 2.2). Some of these are HPRD (Keshava Prasad et al., 2009), MINT (Licata et al., 2012), BioGRID (Chatr-aryamontri et al., 2017), HIPPIE (Alanis-Lobato, Andrade-Navarro, & Schaefer, 2017), FunCoup (Ogris, Guala, Kaduk, & Sonnhammer, 2018), HumanNet (Hwang et al., 2019), and STRING (Szklarczyk et al., 2019).

Table 2.2 Number of unique nodes and edges of popular PPI databases for Homo Sapiens (Human). Unique nodes show non-redundant proteins or genes. Unique edges represent non-redundant undirected interactions between two interactors (proteins or genes)

Database	Unique Nodes	Unique Edges	Version
HPRD	30,047	41,327	Release 9 (April 2010)
MINT	7,411	51,886	2012 Update (Jan 2012)
BioGRID	23,505	373,866	3.5.171 (April 2019)
HIPPIE	20,781	411,430	V2.2 (Feb 2019)
FunCoup	18,355	6,403,719	V4 (Jan 2018)
HumanNet	16,243	476,399	V1 (2011)
STRING	19,257	5,879,727	V11.0 (Jan 2019)

On the other hand, researchers may have trouble while choosing the best PPIN for a biological application because of the plethora of databases. Fortunately, there are many comparative studies assessing PPIN databases based on both topological and biological features. Mathivanan *et al.* analyzed eight public PPIN databases including Reactome, PDZBase, MIPS, MINT, IntAct, HPRD, DIP, and BIND by literature citations, protein coverage, network size, overlapping, and other topological features (Mathivanan et al., 2006). Similarly, Lehne *et al.* compared six major PPIN databases including MINT, IntAct, HPRD, DIP, BioGRID, and BIND by considering topological features and overlapping (Lehne & Schlitt, 2009). Turinsky *et al.* systematically compared the interaction and protein agreement of ten PPIN databases including OPHID, MPPI, MPact, MINT, IntAct, HPRD, DIP, CORUM, BioGRID, and BIND (Turinsky, Razick, Turner, Donaldson, & Wodak, 2011). Huang *et al.* evaluated 21 different human genome-wide PPINs for their coverage capability on 446 disease gene sets by comparing their construction method, molecular interaction types, network similarities and size (J. K. Huang et al., 2018).

2.4.1 STRING

STRING (https://string-db.org) is a popular protein association networks (*PANs*) database that is periodically improved and updated since the initial version, which is v3.0 including 261,033 proteins in 89 organisms, has been published in 2003. It collects and arranges PPI data by integrating known and predicted protein-protein associations for 5090 different organisms (STRING v11) including homo-sapiens (Szklarczyk et al., 2019).

The protein associations in STRING consist of *functional* (indirect) and *physical* (direct) interactions that are known to be specific and biologically significant. The strength of each interaction is determined by collecting and re-evaluating the experimental data available in the PPINs, and retrieving the known protein complexes and pathways from curated databases. In addition, STRING combines four different sources to predict the strength of interactions including: (1) co-expression analysis, (2) transfer of PPI information between different organisms through gene orthology, (3) identification of shared selective signals across genomes, and (4) text-mining of the scientific literature (Szklarczyk et al., 2017).

An edge weight (*score*) is assigned for every interaction listed in STRING. These scores are normalized to the range of [0,1] and denote the confidence level of the interactions. A higher confidence denotes that the given interaction is more specific, biologically more significant, and easier to reproduce. For each PPI within STRING, there are seven different channels of supporting evidence, that are separated by type and source of the evidence. Each of these channels are collected, scored, and compared individually. A final confidence score, which is named as ***combined score***, is computed for each PPI based on the seven channels. The combined scores are preferably used to sort and filter the interactions while building PPINs. STRING defined the typical confidence thresholds for the combined scores as follows: 0.90 = highest confidence, 0.70 = high confidence, 0.40 = medium confidence, and 0.15 = low confidence.

2.4.2 INet

Because it has large number of entries (proteins and interactions) and high coverage and integration, STRING became a prominent database and had been used in many successful bioinformatics applications focused on gene-disease association

(Lan, Wang, Li, Peng, & Wu, 2015; Moreau & Tranchevent, 2012; X. Wang, Gulbahce, & Yu, 2011). On the other hand, in 2017, Yang *et al.* proposed INet (F. Yang et al., 2017), as an integrated network model and suggested it as an equally or better performing alternative to STRING for disease gene prediction.

By definition, INet is a weighted human gene association network (WGAN) that integrates four well-known WGANs (FunCoup, HumanNet, HIPPIE, and STRING) by utilization of *information entropy*. As well as INet covers all nodes of the four original networks, it calculates the combined edge weights for all overlapping edges by use of an *information entropy* algorithm. Since the overlapping edges of the four existing networks are very limited (common nodes > 72% and common edges < 12%) (Table 2.3), the INet is expected to be a much larger network including richer biological information and high functional relevance between strongly interacting gene (or protein) pairs.

Yang *et al.* made several assessments on INet, STRING, FunCoup, HIPPIE, and HumanNet; and they evaluated the performance in terms of the percentage of actual disease genes in predicted disease genes and the ratio of true positives (*TPR*) to false positives (*FPR*). As a result, they suggested that INet and STRING show very similar performances and both outperform FunCoup, HIPPIE, and HumanNet in disease gene prediction.

Table 2.3 The common nodes and edges of four networks (F. Yang et al., 2017)

	HIPPIE	**HumanNET**	**FunCoup**	**STRING**
Proportion of common nodes occupied in other networks	73.43%	74.66%	72.94%	74.80%
Proportion of common edges occupied in other networks	11.24%	4.05%	0.65%	0.83%

2.5 Clustering Algorithms

PPIN clustering, also known as ***module detection***, is the process of analyzing the functional and topological features of a PPIN to find out the groups of interacting proteins that serve together in particular biological functions or that participate in the same biological processes. Many recent studies have indicated that PPIN clustering is an effective approach to discover functions of novel proteins or identify functional modules, thus it has become a hot research topic in systems biology (King, Przulj, & Jurisica, 2004; Shih & Parthasarathy, 2012; J. Wang, Li, Chen, & Pan, 2011; S. Zhang, Ning, & Zhang, 2006).

Although identification of such modules is a computationally complex problem, many approaches utilizing different computational strategies, such as community detection and graph clustering have emerged in the last decade. Besides, traditional clustering algorithms, such as RNSC (King et al., 2004), MCODE (Bader & Hogue, 2003), and MCL (Dongen, 2000) have been successfully adapted for PPIN clustering. Many of the clustering methods have been extensively compared in the surveys (Bhowmick & Seah, 2016; Pizzuti & Rombo, 2014; J. Wang, Li, Deng, & Pan, 2010; X. Wang et al., 2011; X.-F. Zhang et al., 2014) and they typically classified depending on four main characteristics: (i) whether the graph is weighted, (ii) whether the clusters are overlapping, (iii) whether the method is providing complete coverage, (iv) computational method used.

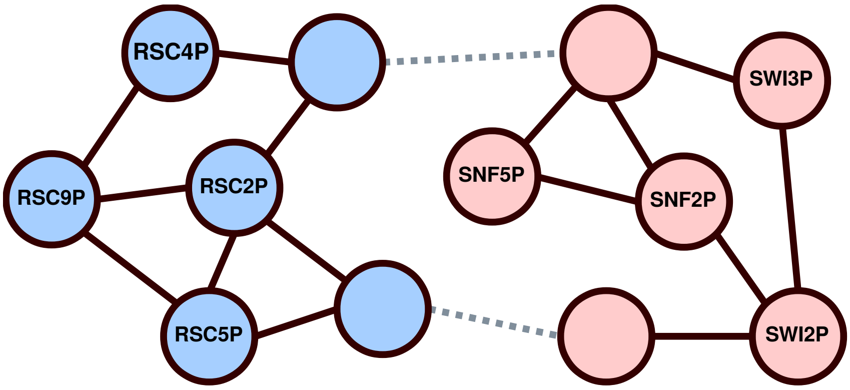

Figure 2.1 An example of PPIN clustering

A clustering algorithm may or may not support the weighted graphs. On the other hand, the clusters generated by a clustering algorithm may be overlapping or disjoint (non-overlapping). In networks with overlapping clusters, a protein can be a member of multiple clusters. A clustering method providing complete coverage assigns a cluster for all nodes (i.e. there is not any cluster-less proteins in the PPIN). On the other hand, by computational method used, a clustering algorithm can be classified in seven categories (Bhowmick & Seah, 2016):

- Genomic Data Driven Algorithms
- Multiple Clustering based Algorithms
- Hierarchical (Graph Cut) Algorithms
- Random Walk based (Message Passing) Algorithms
- Complete Enumeration Algorithms
- Flow-based Algorithms
- Heuristic-based Algorithms

Genomic Data Driven methods integrate PPI data and genomic data to eliminate the noise problems in PPINs. ***Multiple Clustering based*** methods perform multiple clustering instead of a single clustering and combine the generated clusters to achieve the final clustering. ***Hierarchical*** algorithms utilize graph-theoretic (i.e. topological) properties of the PPINs to generate clusters. ***Random Walk based*** approaches utilize the stationary distribution of the Markov chain to solve the graph clustering problem. ***Complete Enumeration*** algorithms apply enumeration on all possible sub-graphs of which density exceed a particular threshold. On the other hand, ***Flow-based*** methods distinguish clusters with weak inter-cluster flow and high intra-cluster flow using a series of flow expansions and contraction. Lastly, in ***Heuristic-based*** methods, the clustering is achieved by a greedy heuristic approach that is based on measurements of similarity and dissimilarity between nodes.

Table 2.4 Main features of three clustering algorithms for extracting clusters from weighted PPINs

Algorithm	**Computational Method**	**Weighted**	**Overlapping**	**Full Coverage**
MCL	Flow-based	Yes	No	Yes
SPICi	Heuristic-based	Yes	No	No
Linkcomm	Graph-cut & Hierarchical	Yes	Yes	No

Selecting the right algorithm is another challenge in PPIN clustering. Several recent studies evaluated more than 50 methods proposed for PPIN clustering (Bhowmick & Seah, 2016; Ji, Zhang, Liu, Quan, & Liu, 2014; Pizzuti & Rombo, 2014; J. Wang et al., 2010; X.-F. Zhang et al., 2014). However, less than half of these approaches support weighted graphs and the results vary widely, depending on overlap and coverage characteristics. Therefore, it would be wise to apply multiple clustering methods with different characteristics as listed in Table 2.4 instead of considering the results of a single clustering algorithm.

2.5.1 MCL

MCL (Markov Clustering, *https://micans.org/mcl*) is a well-kown stochastic flow-based graph clustering algorithm (Enright, Van Dongen, & Ouzounis, 2002). To partition a given PPIN *G = (V, E)* into sub-graphs, MCL first gives a similarity score (e.g. BLAST *E*-value or GO similarity score) to each edge *(υ, ν)* between nodes *υ* and *ν* using a function *f : E* → ℝ, then defines a *weight transition matrix* (*W*) given by:

$$W[\upsilon,\nu] = I((\upsilon,\nu))\, f(\upsilon,\nu) \tag{2.4}$$

where the indicator function *I((υ, ν)) = 1* if *(υ, ν)* ∈ *E* and *I((υ, ν)) = 0* otherwise. Then, MCL performs a normalization based on the *weight transition matrix W* and constructs the *column-wise transition probability matrix* (*M*):

$$M[\upsilon,\nu] = \frac{W[\upsilon,\nu]}{\sum_{x} W[\upsilon,x]} \tag{2.5}$$

To separate the graph into different segments, MCL simulates random walks by iteratively alternating two operations that are called *inflation* and *expansion*, until the convergence. In expansion, the transition matrix *M* is raised to the power of *p*:

$$M_t[\upsilon,\nu] = (M_{t-1}[\upsilon,\nu])^p \tag{2.6}$$

This operation actually represents the transformation of M_{t-1} into a *transition probability matrix* of all possible random walks over *p* steps. In the inflation, the transition matrix *M* is raised to the power of $h > 1$ (Hadamard power) followed by re-normalization. This corresponds to an entry-wise exponentiation and normalization:

$$\Gamma_r M_t[\upsilon, v] = \frac{M_{t-1}[\upsilon, v]^h}{\sum_x M_{t-1}[\upsilon, x]^h} \tag{2.7}$$

Since the exponent $h > 1$, the entries with low transition probabilities are suppressed, while the entries with high transition probabilities are inflated (i.e., favored), thus favoring densely connected regions.

However, the MCL algorithm may reveal imbalanced clusters of which sizes are significantly different. A side effect of having imbalanced clusters is the formation of Singleton clusters. In order to avoid the fragmentation and scalability problems, the Multi-Level Regularized MCL algorithm, MLR-MCL (Satuluri & Parthasarathy, 2009) is proposed.

On the other hand, the original implementation of MCL algorithm ignores the overlapping clusters which may be useful in network analysis. To this end, SR-MCL (Shih & Parthasarathy, 2012), which is another MCL-based clustering algorithm creating overlapping clusters, is proposed. Essentially, SR-MCL is an augmented version of MCL that iteratively executes the conventional MCL clustering process to generate different clusterings on the same network. It keeps generating clusters until the resulting clusters are always the same, then it applies a post-processing to remove redundant clusters and obtains the final set of overlapping clusters.

2.5.2 SPICi

Another important feature that a clustering tool necessarily should have is *scalability* that denotes the ability of handling computational complexity of clustering large PPINs. Jiang *et al.* proposed a heuristic local clustering algorithm, SPICi (‘spicy’, *http://compbio.cs.princeton.edu/spici*, Speed and Performance In Clustering) to handle *scalability* problem (P. Jiang & Singh, 2010).

Given a undirected weighted graph *G = (V, E)*, SPICi aims to generate a set of disjoint dense sub-graphs. In *G*, the edge weight $w_{\upsilon,v}$ for every edge *(υ, v)* ∈ *E* is in the range of (0, 1]. On the other hand, $w_{\upsilon,v} = 0$ if the nodes *υ* and *v* are unconnected. For each node *υ*, SPICi defines the *weighted degree*, $d_w(\upsilon)$, as the total confidence value of all of its incident edges:

$$d_w(\upsilon) = \sum_{v:(\upsilon,\, v)\in \mathrm{E}} w_{\upsilon,v} \tag{2.8}$$

Then, for each set of nodes $U \subset V$, SPICi defines a *density* value that varies in the range of [0, 1] and indicates how close the induced subgraph is to a clique. The *density* is calculated by dividing the sum of the edge weights among all nodes of the subgraph by the total number of possible edges:

$$density(U) = \frac{\sum_{\upsilon,\, v\in U} w_{\upsilon,\, v}}{|U|(|U|-1)/2} \tag{2.9}$$

Finally, for each node *υ* and node set $U \subset V$, SPICi defines the *support* of *υ* by *U* as the sum of the confidence values of *υ*’s edges that are incident to nodes in *U*:

$$support(\upsilon, U) = \sum_{v\in U} w_{\upsilon,\, v} \tag{2.10}$$

SPICi forms one cluster at a time by executing two steps: 1) selecting seeds by a heuristic approach, 2) building or expanding clusters using the obtained seeds. Each cluster is expanded from an original seed pair of nodes. To select the seed nodes, SPICi first finds the node with the largest weighted degree. Then, it identifies the best pair of nodes as *seed* by following a binned selection process.

After obtaining two seed nodes with an edge between them, SPICi forms a cluster by iteratively adding the unclustered (adjacent) node with the highest *support* for the cluster. Nodes are added as long as the *support* is higher than a user-defined support threshold, T_s and the overall cluster *density* remains higher than a user-defined density threshold, T_d. Once the *support* among all the unclustered nodes drops below T_s or *density* of the formed cluster drops below T_d, SPICi returns the cluster as an output and removes the nodes of the returned cluster from the network. The same procedure is repeated until all nodes in the network are clustered.

Jiang *et al.* evaluated the performance of SPICi and other nine clustering algorithms in terms of memory usage and running time on five different biological networks (Table 2.5). The results showed that SPICi, which has a time complexity of $O(V \log V+E)$ and space complexity of $O(E)$, is significantly fast and memory-efficient; and it can be considered as a highly scalable tool for large PPI networks.

2.5.3 Linkcomm

While MCL provides complete coverage and assigns a cluster to each protein in a network, SPICi may output some unclustered nodes. On the other hand, both methods performs a *node-oriented* clustering and place each protein in a maximum of one community. However, such an approach in which each node can only be included in a single community is unsuitably restrictive for densely connected graphs where sub-networks often overlap. To overcome such a problem Ahn *et al.* proposed a link-similarity based community detection algorithm (Ahn, Bagrow, & Lehmann, 2010) which has been later released as an R library, Linkcomm (Kalinka & Tomancak, 2011).

Table 2.5 Execution time (sec) and peak memory consumption (MBs) of clustering algorithms (Jiang & Singh, 2010). Memory consumption of MCUPGMA is fixed for these networks since it is pre-allocated by a default limit. The algorithms that cannot cluster the network within 12 hours are shown in the table with N/A

Running Time (sec)	**Bayesian Human**	**STRING Human**	**STRING Yeast**	**BioGRID Human**	**BioGRID Yeast**
SPICi	**1111**	**7**	**2**	**1**	**1**
MCUPGMA	N/A	33	9	4	5
MCL	N/A	4926	645	114	336
SPC	N/A	N/A	219	215	183
RNSC	N/A	N/A	1325	17	172
MCODE	N/A	N/A	7848	49	101
NetworkBLAST	N/A	N/A	7848	427	1904
DPClus	N/A	N/A	N/A	2113	1602
CFinder	N/A	N/A	N/A	25	N/A
DME	N/A	N/A	N/A	N/A	N/A

Memory Usage (MBs)	**Bayesian Human**	**STRING Human**	**STRING Yeast**	**BioGRID Human**	**BioGRID Yeast**
SPICi	**1143.0**	**90.5**	**15.1**	**1.5**	**1.2**
MCUPGMA	N/A	259.1	259.1	259.1	259.1
MCL	N/A	357.0	111.7	24.9	73.3
SPC	N/A	N/A	311.0	430.3	220.5
RNSC	N/A	349.4	82.3	9.8	25.9
MCODE	N/A	N/A	606.9	306.1	375.6
NetworkBLAST	N/A	N/A	72.8	60.5	61.9
DPClus	N/A	N/A	N/A	202.1	140.2
CFinder	N/A	N/A	N/A	23.0	N/A
DME	N/A	N/A	N/A	N/A	N/A

For a given unweighted, undirected graph, the set of node i and its neighbors are denoted as $n_+(i)$. To obtain link communities in such networks, first the similarity S, between link pairs (e_{ik}, e_{jk}) that share a node is calculated by:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \tag{2.11}$$

Then, by agglomerating ties in *S* simultaneously, a link dendrogram is built by use of a single-linkage hierarchical clustering approach. Cutting this dendrogram with certain threshold values yields link communities.

The threshold value is defined as ***maximum partition density***. For given graph *G* with *N* nodes and *L* links, $P = \{P_1, P_2, ..., P_s\}$ represents partition of the links into *s* subsets. The number of connected nodes in subset P_s is $n_s = |U_{e_{ij} \in P_s} \{i, j\}|$ and the number of links is $m_s = |P_s|$. The *link density*, D_s, of community *s* is obtained by normalizing m_s by the maximum and minimum numbers of edges possible among n_s linked nodes:

$$D_s = \frac{m_s - (n_s - 1)}{n_s(n_s - 1)/2 - (n_s - 1)} \tag{2.12}$$

and the average of the *link density* weighted by the fraction of present links, gives the *partition density*, *D*:

$$D = \frac{1}{L} \sum_s m_s D_s = \frac{2}{L} \sum_s m_s \frac{m_s - (n_s - 1)}{(n_s - 2) - (n_s - 1)} \tag{2.13}$$

Zhang *et al.* compared ten different clustering algorithms in terms of accuracy (ACC) (X. Li, Wu, Kwoh, & Ng, 2010), maximum matching ratio (MMR) (Nepusz, Yu, & Paccanaro, 2012), fraction of matched complexes (FRAC), and precision-recall score (PR) (Song & Singh, 2009). The performance of each algorithm has been evaluated on six different biological networks by considering how well the gold standards are recovered by the predicted complexes. The results in Table 2.6 show that Linkcomm and SR-MCL generate more clusters with high accuracy in comparison to other approaches (X.-F. Zhang et al., 2014).

Table 2.6 Average benchmark results for six unweighted PPI networks (Collins, Gavin, Krogan-Core, Krogan-Extended, DIP, and BioGRID) with respect to the CYC2008 gold standard (X.-F. Zhang et al., 2014). The highest two values in each measurement are highlighted

Algorithm	Coverage	# of Clusters	ACC	FRAC	MMR	PR
GMFTP	1539	299	0.731	0.438	**0.714**	**0.453**
AP	**3250**	343	0.526	0.399	0.22	0.194
CFinder	1123	107	0.416	0.253	0.576	0.315
ClusterONE	1676	349	0.645	0.382	0.68	0.382
Linkcomm	2152	**1452**	**0.739**	**0.487**	**0.69**	0.337
MCL	2507	382	0.579	0.33	0.661	0.29
MCODE	773	113	0.461	0.275	0.57	**0.39**
MINE	1289	223	0.645	0.367	0.671	0.42
SPICi	1610	289	0.591	0.331	0.666	0.382
SR-MCL	**3282**	**1645**	**0.742**	**0.47**	0.656	0.241

2.6 Integrated Methods for Gene-Disease Association

Since relying upon a single type of biological information does not provide reliable results, integrating multiple data sources such as protein-protein interaction / association network, gene expression, Gene Ontology, functional annotation, and DNA sequence have become essential in the post-genomic era.

Hubner *et al.* discovered 73 new genes and regulatory pathways underlying MS and CVD phenotypes by integrating linkage analysis with genome-wide expression profiling (Hubner et al., 2005). Presson *et al.* presented the IWGCNA method that integrates weighted gene co-expression network analysis (WGCNA) with genetic marker data (SNP) to detect disease related modules in Chronic Fatigue Syndrome (Presson et al., 2008). Radivojac *et al.* proposed an approach to predict gene–disease associations (GDAs) based on a PPIN as well as functional annotation and protein sequence (Radivojac et al., 2008). Similarly, Wu *et al.* integrated the gene expressions data with PPIN to prioritize genes associated with cancer (C. Wu, Zhu, & Zhang, 2012). Magger *et al.* integrated PPIN data with tissue specific gene expression data in order to build tissue-specific PPINs for 60 tissues; and they performed disease gene prioritization on them (Magger, Waldman, Ruppin, & Sharan, 2012). Liu *et al.* combined WGCNA with functional and pathway enrichment analyses to discover hub genes and particular modules associated with CAD (Liu et al., 2016).

2.7 Integrated Databases for Gene-Disease Association

In the recent years, many integrative databases are founded to collect the results of the GDA studies. These databases can integrate data from animal models, GWAS catalogs, or expert curated repositories as well as they can obtain GDAs from scientific literature by text mining. They generally diverge by the data sources they scan and the scoring method they use to assign confidence levels of the GDAs.

In Table 2.7, we present some statistical features and the main differences of three largest GDA databases: CTD (Davis et al., 2019), DISEASES (Pletscher-Frankild, Pallejà, Tsafou, Binder, & Jensen, 2015), DisGeNET (Piñero et al., 2017).

Table 2.7 The number of curated genes (*G*), the number of curated diseases (*D*), the number of curated gene-disease associations (*GDA*), and types of integrated data sources for the three largest GDA databases: CTD, DISEASES, and DisGeNET. Types of data sources are: Curated Knowledge (C), Animal Model (M), Experimental (E), Inferred (I), and Literature (L)

Database	*G*	*D*	*GDA*	Types of Data Sources
CTD	8,572	5,790	38,928	C + L
DISEASES	2,001	735	15,231	C + E + L
DisGeNET	**9,413**	**10,370**	**81,746**	C + M + I + L

2.7.1 DisGeNET

DisGeNET database integrates information of human gene-disease associations (GDAs) and variant-disease associations (VDAs) from various repositories including Mendelian, complex and environmental diseases. DisGeNET (v6.0) contains 628,685

GDAs, between 17,549 genes and 24,166 diseases. The GDAs in DisGeNET are organized based on the types of source databases:

- **Curated (C):** GDAs from UniProt (UniProt Consortium, 2018), PsyGeNET (Gutiérrez-Sacristán et al., 2015), Orphanet (Weinreich, Mangon, Sikkens, Teeuw, & Cornel, 2008), the CGI (Tamborero et al., 2018), CTD (Davis et al., 2019), ClinGen (Rehm et al., 2015), and the Genomics England PanelApp (Genomics England PanelApp, 2019).
- **Animal Models (M):** GDAs from RGD (Laulederkind et al., 2018), MGD (Smith et al., 2018), and CTD (mouse and rat).
- **Inferred (I):** GDAs from the HPO (Köhler et al., 2019), and VDAs reported by Clinvar (Landrum & Kattman, 2018), the GWAS db (M. J. Li et al., 2016), and GWAS catalog (MacArthur et al., 2017).
- **Literature (L):** GDAs from BeFree (Bravo, Cases, Queralt-Rosinach, Sanz, & Furlong, 2014; Bravo, Piñero, Queralt-Rosinach, Rautschka, & Furlong, 2015) and LHGDN (Bundschus, Dejori, Stetter, Tresp, & Kriegel, 2008).

Curated data (*C*) contain GDAs reported by the expert curated resources. Animal model data (*M*) include GDAs provided by the resources containing information about animal models (rat and mouse) of disease. DisGeNET uses orthology information to map the associations to the human genes. Inferred data (*I*) refer to GDAs inferred from *HPO* and *VDAs*. In the case of *HPO*, GDAs are inferred from phenotype-disease via triangulation. In the case of *VDAs*, a GDA is created for each gene annotated to the variant and the disease annotated to the variant. Literature data (*L*) denote GDAs extracted by text-mining of *LHGDN* and *MEDLINE* abstracts via *BeFree* system.

2.7.1.1 Calculating DisGeNET Scores

DisGeNET uses scores to rank the GDAs according to their level of evidence. These scores varies between 0 and 1 depending on the number of data sources and publications reporting the association, and type (level of curation) of these sources. The score (*S*) for GDAs is computed according to:

$$S = C + M + I + L \tag{2.14}$$

where:

$$C = \begin{cases} 0.6 & if\ N_{sources(i)} > 2 \\ 0.5 & if\ N_{sources(i)} = 2 \\ 0.4 & if\ N_{sources(i)} = 1 \\ 0 & otherwise \end{cases} \qquad \begin{aligned} M &= \begin{cases} 0.2 & if\ N_{sources(j)} > 0 \\ 0 & otherwise \end{cases} \\ I &= \begin{cases} 0.1 & if\ N_{sources(k)} > 0 \\ 0 & otherwise \end{cases} \\ L &= \begin{cases} 0.1 & if\ N_p > 9 \\ N_p * 0.01 & if\ N_p \leqslant 9 \end{cases} \end{aligned} \tag{2.15}$$

Here, $N_{sources(i)}$ is the number of curated sources, $N_{sources(j)}$ is the number of model organisms, $N_{sources(k)}$ is the number of inferential sources, and N_p is the number of publications supporting the GDA.

2.8 Novelty of the Proposed Study

We mentioned several medical and bioinformatics applications that aim to identify underlying molecular mechanisms of metabolic disorders by use of different biological data sources and computational approaches. However, most of these studies either focused on a single type of metabolic disorder or did not diversify the integrated data sources and/or computational methods that are used in analyses.

In our study, we constructed three functional interaction networks for MS, T2D, and CAD disease by integration of multiple biological data sources such as protein-protein interactions, gene expressions, and gene ontologies. Then, we detected the disease related protein complexes using different clustering approaches. By analyzing the overlapping sub-graphs, we obtained shared disease genes for metabolic disorders in question.

As well as presenting a comparative performance evaluation of different PPIN databases and clustering algorithms in disease gene prediction, this study is novel in virtue of being the initial effort to collectively analyze MS, T2D, and CAD in bioinformatics perspective.

CHAPTER THREE
METHOD

3.1 System Overview

We present a general overview of our system in four main steps in Figure 3.1. In step A, we reduced the size of STRING and INet networks. We first mapped the proteins in both networks to the gene symbols in our gene expression data set. Then, by filtering out the unmapped nodes, we reduced the density of both networks approximately 34%. The main reduction in the number of edges has happened when the insignificant interactions were removed. By using the medium-confidence threshold (0.4) that is stated in Section 2.4.1, we shrank the STRING network to the 10% of its initial size. Similarly, we selected 0.175 as the confidence cutoff for INet to get same reduction ratio with STRING in terms of network size.

Table 3.1 Number of nodes and edges of the STRING and INet topologies before and after data reduction

	Before Reduction		After Reduction		% of Reduction	
Network	**Nodes**	**Edges**	**Nodes**	**Edges**	**Nodes**	**Edges**
STRING	19,576	5,676,528	13,969	568,020	29	90
INet	19,290	7,077,509	12,264	710,660	36	90

In step B, we executed three procedures to construct disease networks for MS, CAD, and T2D: (1) preprocessing of microarray data, (2) detection of DEGs, (3) integration of PPINs with the DEGs. In procedure 1, we first removed invalid and null rows. Then, we filtered out the duplicated rows (probes) by mapping each probe identifier to its gene symbol and aggregating the probes corresponding to the same gene symbol. Thus, we obtained 24,279 unique genes from 50,400 probes. In procedure 2, we detected DEGs by considering both fold-change values and significance score of *t*-test. In procedure 3, we constructed disease networks by placing each DEG to the STRING and the INet topologies.

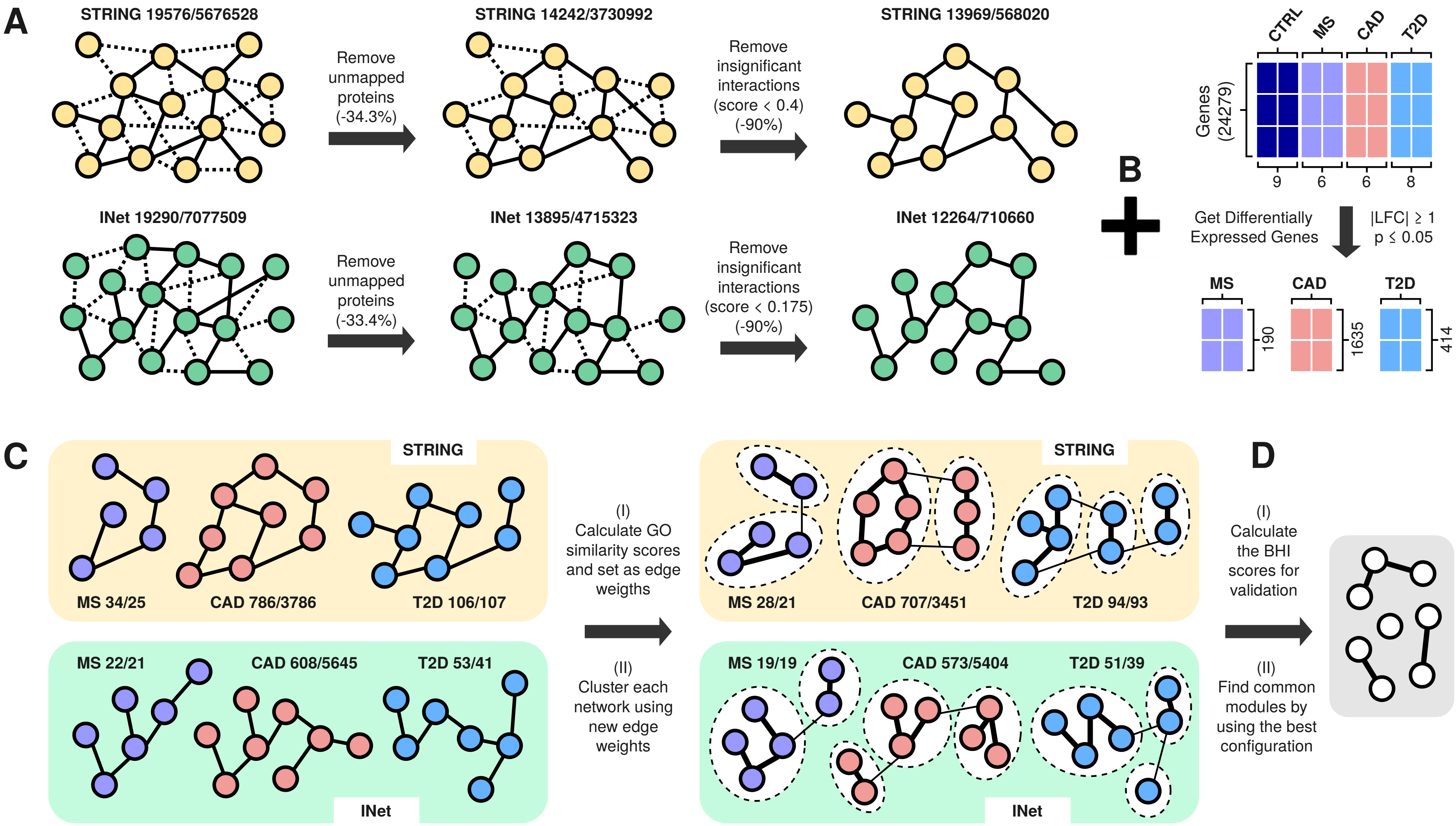
A
STRING 19576/5676528
Remove unmapped proteins (-34.3%)
STRING 14242/3730992
Remove insignificant interactions (score < 0.4) (-90%)
STRING 13969/568020
INet 19290/7077509
Remove unmapped proteins (-33.4%)
INet 13895/4715323
Remove insignificant interactions (score < 0.175) (-90%)
INet 12264/710660
B
CTRL
MS
CAD
T2D
Genes (24279)
9
6
6
8
Get Differentially Expressed Genes
|LFC| ≥ 1
p ≤ 0.05
MS
CAD
T2D
190
1635
414
C
STRING
MS 34/25
CAD 786/3786
T2D 106/107
MS 22/21
CAD 608/5645
T2D 53/41
INet
(I) Calculate GO similarity scores and set as edge weigths
(II) Cluster each network using new edge weights
STRING
MS 28/21
CAD 707/3451
T2D 94/93
MS 19/19
CAD 573/5404
T2D 51/39
INet
D
(I) Calculate the BHI scores for validation
(II) Find common modules by using the best configuration

Figure 3.1 The system overview

In step C, we clustered the disease networks to identify the significant functional modules. To be able to generate biologically meaningful clusters, we assigned GO semantic similarity scores as edge weights. Then, we executed three clustering algorithms (MCL, SPICi, and Linkcomm) on each network and generated the disease-related protein complexes. Before proceeding to the last step, we performed cluster validation by biological homogeneity index (*BHI*) for each disease network. By this means, we selected the best clustering algorithm and network construction parameters to be used in the extraction of overlapping disease modules. Finally, in step D, we detected the overlapping modules in MS, T2D, and CAD disease networks to obtain common sub-modules that are expected to include shared disease genes for metabolic disorders in question.

3.2 Peripheral Blood Gene Expression Data

The microarray experiments has shown that peripheral blood gene expression profiling is an effective way to distinguish the phenotypically unique cohorts of patients suffering from a wide variety of diseases (Aune, Maas, Moore, & Olsen, 2003; Baechler et al., 2003; Bomprezzi et al., 2003). Based on this claim, Grayson *et al.* comparative evaluation on peripheral blood transcript levels of patients with CAD, T2D, MS, and RA (rheumatoid arthritis) in order to determine if patients with metabolic disorders own distinct gene expression profiles (Grayson, Wang, & Aune, 2011). To do that, they recruited subjects with CAD (n=6), T2D (n=8), MS (n=6), RA (n=6), and 9 individuals who were not currently taking medications for any disease state, and had never been diagnosed with a chronic illness, to present as the control (CTRL) cohort. After analyzing the peripheral blood samples of 35 subjects and profiling the gene expressions by use of the human exonic evidence based oligonucleotide (HEEBO) array, they deployed the resulting data set to public in National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) under the GSE23561 identifier.

3.2.1 Preprocessing

Each series matrix file provided under GSE23561 data set includes 50400 oligonucleotide probes and their expression values that are normalized to whole array intensity sum of 10,000 that is giving an average intensity per probe of 0.2. However, we used the raw (*F635 median*) values instead of normalized series matrices to apply a logarithmic transformation.

We normalized the matrices by transforming the *F635 median* value of each probe to *log_2(F635 median)*. Then, we removed the invalid and null rows in the data set and aggregated multiple probes corresponding to the same gene by mapping probe identifiers to gene symbols (Symbol v12) using GPL10775 platform. Here, we applied a median based aggregation. As a result, we obtained 24,279 log transformed gene expression values for each one of the 29 samples. The distribution of gene expression values for each subject is shown in Figure 3.2.

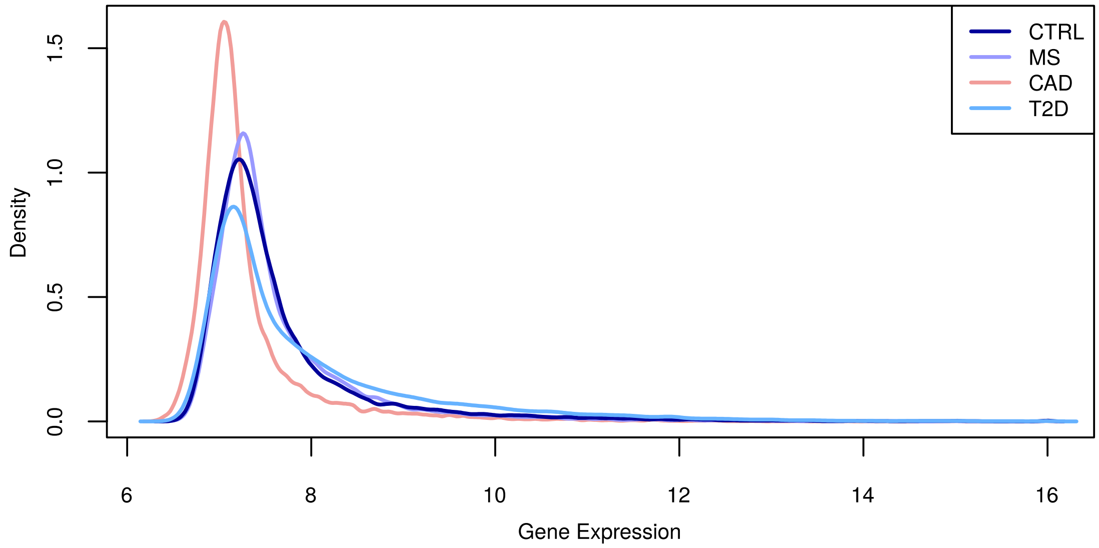

Figure 3.2 Distribution of log-normalized gene expression values

3.2.2 Detecting DEGs

We first applied a *t*-test to identify the significantly differentiated genes between control and disease groups. By filtering out the genes with *p*-value > 0.05, we obtained the significant genes in the following numbers: 307 (MS), 435 (T2D), and 2679 (CAD). Then, we calculated the fold-change values by:

$$FC_{(i)} = \left| m_{CTRL(i)} - m_{DISEASE(i)} \right| \tag{3.1}$$

where $m_{CTRL(i)}$ denotes the mean of log-normalized expression values of gene i in control (CTRL) group and $m_{DISEASE(i)}$ denotes the mean of log-normalized expression values of gene i in a disease group (i.e., MS, T2D, or CAD). After applying the fold-change cutoff, FC $\geq$ 1, we concluded the analysis with the following number of DEGs: 190 (MS), 414 (T2D), and 1635 (CAD).

3.3 PPIN Data

All versions of STRING topologies for different types of species are available in the STRINGdb website (*https://string-db.org*). We made use of human PPIN v10.5 (*9606.protein.links.v10.5*) which consists of 19,576 unique proteins with 11,353,056 directed and reversely-duplicated interactions. By merging the directed interactions between the same proteins, we obtained 5,676,528 unique (non-redundant) and undirected edges (Figure 3.1 A). In STRING, nodes are named by *ensemble protein identifiers* and edge weights are represented by *combined scores* in the range of [0, 1000].

On the other hand, we obtained INet topology from Yang *et al.*'s study (F. Yang et al., 2017). INet covers 19,290 unique genes and there are 7,077,509 undirected and non-redundant interactions between them (Figure 3.1 A). The nodes are presented by *ensemble identifiers* and the weight of each edge is in the range of [0, 1].

3.3.1 Protein-Gene Mapping

In order to place the genes in our gene expression data set into the STRING and INet topologies, we mapped the *ensemble identifiers* to the *official gene symbols*. We performed this mapping in R by use of *STRINGdb* (v1.22.0) and *org.Hs.eg.db* packages (Carlson, 2018) available in Bioconductor (*https://bioconductor.org*). Then, we eliminated the nodes that could not be mapped to the genes in our data set and the edges between them. The eliminations of non-available genes correspond to reduction in 27% of nodes and 34% of edges for the STRING; and 28% of nodes and 33% of edges for the INet network (Figure 3.1 A).

3.3.2 Filtering Insignificant PPIs

In the STRING network, we first normalized the *combined scores* to the range of [0, 1]. Then, we filtered out the interactions below the medium-confidence (i.e., *combined score* < *0.4*). Compared to the initial topology, the resulting network was including only 10% of the edges while keeping 71% of the nodes (Table 3.1). Because of the absence of a *de facto* confidence threshold for INet, we set the cutoff as 0.175 providing 10% reduction obtained as in the STRING network. Unlike STRING, the remained nodes accounted for 64% of the initial nodes in the INet network (Table 3.1). On the other hand, 12,109 of the nodes and 188,188 of the edges are intersecting in the final versions of two networks (Table 3.2).

Table 3.2 Number of overlapping nodes and edges for the STRING and INet topologies

	# after reduction		**# of overlap**		**% of overlap**	
Network	**Nodes**	**Edges**	**Nodes**	**Edges**	**Nodes**	**Edges**
STRING	13,969	568,020	12,109	188,188	86.6	33.1
INet	12,264	710,660	12,109	188,188	98.7	26.5

3.4 Integrating DEGs with PPINs

For both the STRING and INet topologies, we filtered out all of the genes that are not represented in the DEG set for any disease subject (Figure 3.1 B). As a result, we obtained three separate differential co-expression networks for two topologies (Figure 3.1 C). The numbers of nodes/edges in the resulting STRING-based disease networks were as follows: 34/25 (MS), 786/3786 (CAD), and 106/107 (T2D). The numbers of nodes/edges in the resulting INet-based disease networks were as follows: 22/21 (MS), 608/5645 (CAD), and 53/41 (T2D).

3.5 Generating GO Similarity Scores

We obtained GO semantic similarity scores for each connected gene pair using *GoSemSim* R package (Yu et al., 2010). *GOSemSim* generates GO semantic similarity scores by using different *similarity measures*, different *combination methods*, and different *orthogonal ontologies*. The *similarity measure* is either one of four information content (*IC*) based method (Resnik, Lin, Rel, Jiang) or a graph-based method (Wang) that are used in determination of the semantic similarity of two GO terms. The *combination strategy* is one of the *max*, *avg*, *rcmax*, or *best-match average* (*BMA*) and needed to calculate overall semantic similarity score on all pairs of two GO term sets. On the other hand, the reference *orthogonal ontologies* can be *biological process* (BP), *cellular component* (CC), or *molecular function* (MF) and used to specify which ontology will be considered while generating similarity scores.

To combine GO terms, we selected *BMA* which has been suggested as the best combination method in the previous studies (Pesquita et al., 2008; X. Wu, Pang, Lin, & Pei, 2013). On the other hand, we repeated our analyses for each type of *similarity measure* (Resnik, Lin, Rel, Jiang, Wang) and each type of *orthogonal ontology* (BP, CC, MF) to perform a comparative evaluation on our data set (Figure 3.3).

3.5.1 Information Content Based Similarity Methods

The IC-based approaches use the annotation statistics (i.e. information content) of the common ancestor terms to measure the semantic similarity of two GO terms. The semantic similarity score depends on the frequencies of two GO terms involved and that of their closest common ancestor term (i.e. the *most informative common ancestor*) within a given corpus of GO annotations. On the other hand, the IC of a GO concept is calculated by taking the negative logarithm of the probability that the term being included in the respective GO corpus. In this regard, rarely used GO terms are richer in the information they contain.

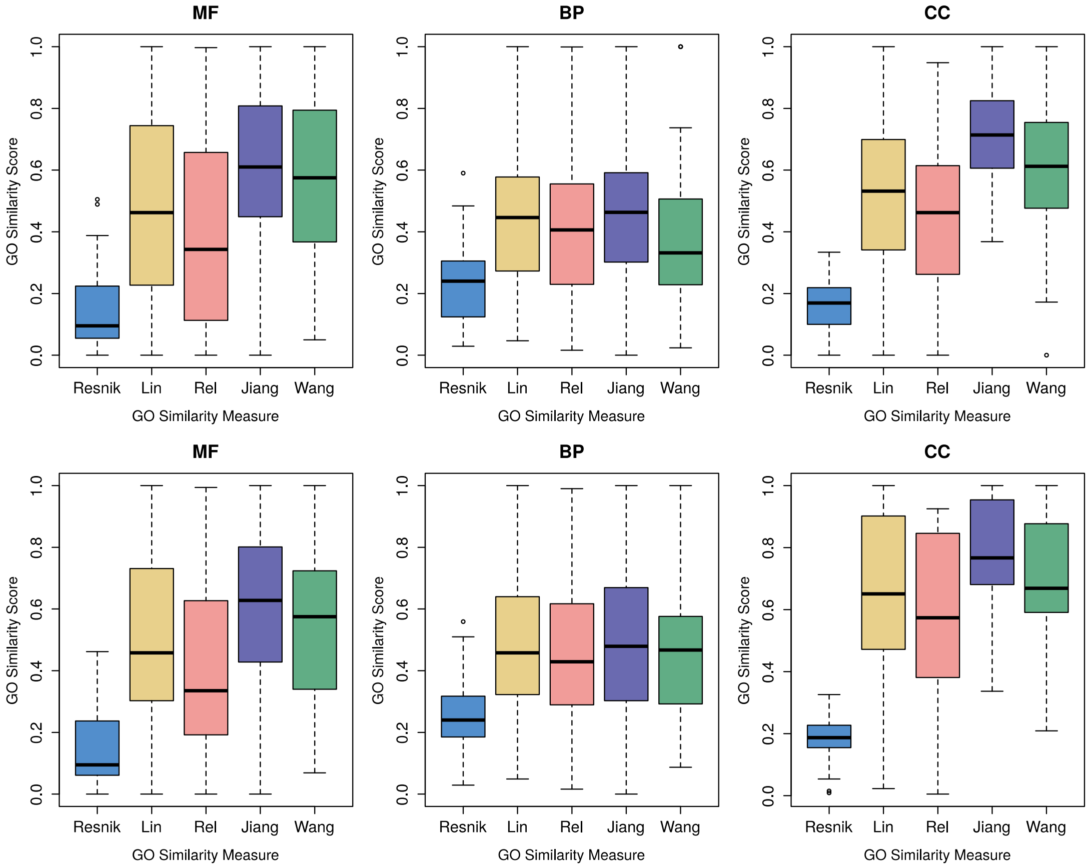

Figure 3.3 Distribution of GO semantic similarity scores in T2D subject. Three plots on top are generated using STRING. Three plots on bottom are generated using INet

For a given GO term θ and a set T consisting of θ's children terms, the frequency $f(\theta)$ and the information content $IC(\theta)$ are defined as:

$$f(\theta) = \frac{|T|}{N}, \quad IC(\theta) = -\log(f(\theta)) \tag{3.2}$$

where N is the total number of terms in the GO corpus. Since GO allow each term to have multiple parents, any two terms can be linked to a parent through multiple paths. The similarity of terms θ_1 and θ_2 is calculated by use of the IC of each term and the IC of their *most informative common ancestor* (θ_A).

The Resnik's similarity (Resnik, 1999) of two GO terms is defined as:

$$sim_{Resnik}(\theta_1, \theta_2) = IC(\theta_A) \tag{3.3}$$

The Lin's measure (Lin, 1998) is defined as:

$$sim_{Lin}(\theta_1, \theta_2) = \frac{2\,IC(\theta_A)}{IC(\theta_1)+IC(\theta_2)} \tag{3.4}$$

The Relevance (Rel) method (Schlicker, Domingues, Rahnenführer, & Lengauer, 2006) is a combination of Resnik's and Lin's method:

$$sim_{Rel}(\theta_1, \theta_2) = \frac{2\,IC(\theta_A)(1-f(\theta_A))}{IC(\theta_1)+IC(\theta_2)} \tag{3.5}$$

And Jiang's similarity (J. J. Jiang & Conrath, 1997) is defined as:

$$sim_{Jiang}(\theta_1, \theta_2) = 1-min(1,\ IC(\theta_1)+IC(\theta_2)-2\,IC(\theta_A)) \tag{3.6}$$

3.5.2 Graph Based Similarity Methods

The graph-based methods compute the semantic similarity using the topological properties of GO. In GO database, the ontologies are formed as a *directed acyclic graph* (*DAG*) in which the nodes denote concepts (i.e. terms) and the edges denote two types of semantic relations (*'is-a'* and *'part-of'*). In such a structure, a GO term *i* can be defined as $DAG_i = (i, T_i, E_i)$ where T_i is the corpus of GO terms in DAG_i, consisting of term *i* and all of its ancestor terms, and E_i is the set of edges connecting the GO terms in DAG_i.

The Wang's method (J. Z. Wang, Du, Payattakool, Yu, & Chen, 2007) computes the GO semantic similarity between two GO terms by taking into account their distance in the GO graph and their connections with the common ancestor terms. It first sets the semantic value (*SV*) of GO term *i* as the cumulative contribution of all terms in *DAG$_i$* to the semantics of term *i*. In *DAG$_i$*, the terms closest to term *i* provide the highest semantic contribution to it. Therefore, the contribution of a GO term *j* to the semantic of GO term *i* is defined as the *S-value* of term *j* related to term *i*. For any *j* in *DAG$_i$*, *S$_i$(j)* denotes its *S-value* related to the GO term *i*, and can be defined as:

$$\begin{aligned} &S_i(i) = 1 \\ &S_i(j) = max\{w_e \times S_i(j') \mid j' \in children\,of\,(j)\}\,if\;j \neq i \end{aligned} \tag{3.7}$$

where w_e represents the *semantic contribution factor* for edge $e \in E_i$ connecting term *j* and its child *j'*. Contribution of term *i* to its own is always defined as 1. After calculating the contributions (i.e. *S-values*) of all terms in *DAG$_i$*, the semantic value (*SV*) of *i*, which the sum of the *S-values*, is obtained as:

$$SV(i) = \sum_{j \in T_i} S_i(j) \tag{3.8}$$

Based on the semantic values and contribution of the common ancestor term *j*, the semantic similarity between terms *i* and *k* is defined as:

$$sim_{Wang}(i,k) = \frac{\sum_{j \in T_i \cap T_k} S_i(j) + S_k(j)}{SV(i) + SV(k)} \tag{3.9}$$

where *S$_i$(j)* and *S$_k$(j)* are the contributions of the GO term *j* to the GO term *i* and GO term *k*, respectively.

3.5.3 Combination Methods

For given two genes g_1 and g_2 annotated by two sets of GO terms $T_1 = \{t_{11}, t_{12} \ldots t_{1n}\}$ and $T_2 = \{t_{21}, t_{22} \ldots t_{2k}\}$ respectively, *GOSemSim* computes the semantic similarity between g_1 and g_2 by combining T_1 and T_2. To that end, it utilizes one of the four combination methods called *max*, *rcmax*, *avg*, and *BMA*.

The *max* technique sets the semantic similarity score between g_1 and g_2 as the maximum semantic similarity over all pairs of GO terms between T_1 and T_2.

$$sim_{max}(g_1, g_2) = \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant k} sim(t_{1i}, t_{2j}) \tag{3.10}$$

The *avg* sets the similarity score by taking an average over all pairs of GO terms.

$$sim_{avg}(g_1, g_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} sim(t_{1i}, t_{2j})}{n \times k} \tag{3.11}$$

Similarities between T_1 and T_2 form a matrix. The *rcmax* method uses the maximum of *ColumnScore* (average of maximum similarities on each column) and *RowScore* (average of maximum similarities on each row) to calculate *sim(g₁, g₂)*.

$$sim_{rcmax}(g_1, g_2) = max\left(\frac{\sum_{i=1}^{n} \max_{1\leqslant j\leqslant k} sim(t_{1i}, t_{2j})}{n}, \frac{\sum_{j=1}^{k} \max_{1\leqslant i\leqslant n} sim(t_{1i}, t_{2j})}{k}\right) \tag{3.12}$$

The *BMA* (*Best-Match Average*) method, utilizes the same similarity matrix as in *rcmax*, but it takes the average of all maximum similarities on each column and row.

$$sim_{BMA}(g_1, g_2) = \frac{\sum_{i=1}^{n} \max_{1\leqslant j\leqslant k} sim(t_{1i}, t_{2j}) + \sum_{j=1}^{k} \max_{1\leqslant i\leqslant n} sim(t_{1i}, t_{2j})}{n+k} \tag{3.13}$$

3.6 Implementation of the Clustering Algorithms

We executed three different clustering algorithms, MCL (Enright et al., 2002), SPICi (P. Jiang & Singh, 2010), and Linkcomm (Ahn et al., 2010; Kalinka & Tomancak, 2011), on ninety different combinations of disease networks constructed for MS, CAD, and T2D by using two different PPI topologies (STRING and INet), three different orthogonal ontologies (BP, CC, MF), and five different GO similarity measures (Resnik, Lin, Rel, Jiang, Wang). We summarize the clustering, evaluation, and validation steps in Figure 3.4. All datasets, source code, and analysis results are available on GitHub (*https://github.com/smtnkc/go-cluster*).

3.6.1 Libraries, Functions, Parameters, and Running Environment

We used MCL (Jäger, 2015) and Linkcomm (Kalinka & Guenoche, 2014) R packages, and SPICi (Peng & Singh, 2010) python library to generate clusters. We executed *mcl*, *getLinkCommunities*, and *spici* functions for MCL, Linkcomm, and SPICi, respectively. For all methods, we set the input as an undirected and weighted graph without self loops. In SPICi, we set the minimum cluster density (*d*) to 0.5, the minimum support threshold (*g*) to 0.5, the minimum cluster size (*s*) to 2, and graph mode to *0* (*sparse graph*). On the other hand, we used the default parameters in MCL and Linkcomm. All executions have been performed on a computer with Intel Core i5-4200U processor, 8 GB of RAM, and Ubuntu 18.04 operating system.

3.6.2 Execution Time and Memory Consumption

We present the average execution time and maximum memory consumption of each algorithm in Table 3.3 and Table 3.4, respectively. As it is expected, SPICi is significantly faster and memory efficient (i.e. scalable), especially in large networks (i.e. CAD).

Table 3.3 Average execution time (seconds) for MCL, Linkcomm, and SPICi

Topology	Subject	MCL	Linkcomm	SPICi
STRING	MS	0.2	1.6	0.006
	T2D	1.2	2.1	0.005
	CAD	37.4	32.6	0.02
INet	MS	0.8	1.5	0.006
	T2D	1.2	1.5	0.006
	CAD	28.5	330.6	0.02

Table 3.4 Average memory consumption (MBs) of MCL, Linkcomm, and SPICi

Topology	Subject	MCL	Linkcomm	SPICi
STRING	MS	1.2	0.1	3.7
	T2D	19.2	2.2	3.8
	CAD	2,298	4,289	4.8
INet	MS	0.9	0.2	3.7
	T2D	4.6	0.3	3.7
	CAD	1,642	608.7	5.2

3.7 Post-clustering Validation

To validate the biological significance of the produced modules, we calculated the biological homogeneity index (*BHI*) scores for each disease network that are constructed by different configurations and clustered by different algorithms. Basically, BHI quantifies how biologically homogeneous the clusters are; and it controls to what extent the genes incorporated into the same cluster by statistical methods, belong to the same functional classes. Thus, it is a useful metric to evaluate consistency and performance of the clustering algorithms. The *BHI* scores vary in the range of [0, 1], and larger values represent biologically more homogeneous clusters.

We generated the *BHI* scores using *clValid* R package (Brock, Pihur, Datta, & Datta, 2008) (Figure 3.4). The *BHI* function in *clValid* takes three important arguments including a clustered network, a *Bioconductor* annotation list, and a category parameter indicating the GO categories to use for biological validation (BP, CC, MF, or ALL). As annotation reference, we used *hgu133a.db* (Carlson, 2016) annotation database available in Bioconductor. We mapped the official gene symbols in our networks to the probe identifiers available in *hgu133a.db*.

3.8 Overlapping Disease Modules

To detect the shared disease related modules for MS, T2D, and CAD, we analyzed the overlapping clusters of the disease networks. We first compared different configurations for network construction and clustering algorithms to find out the parameters that are producing the most biologically homogeneous clusters (i.e., the clusters with the largest *BHI* score). After the comparative evaluation of 270 different configuration, we decided to construct the disease networks on INet topology. We selected Wang measure and MF ontology to calculate GO semantic similarity scores. For clustering, we ran both Linkcomm and SPICi algorithms, since they achieved to almost identical *BHI* scores (Figure 3.4).

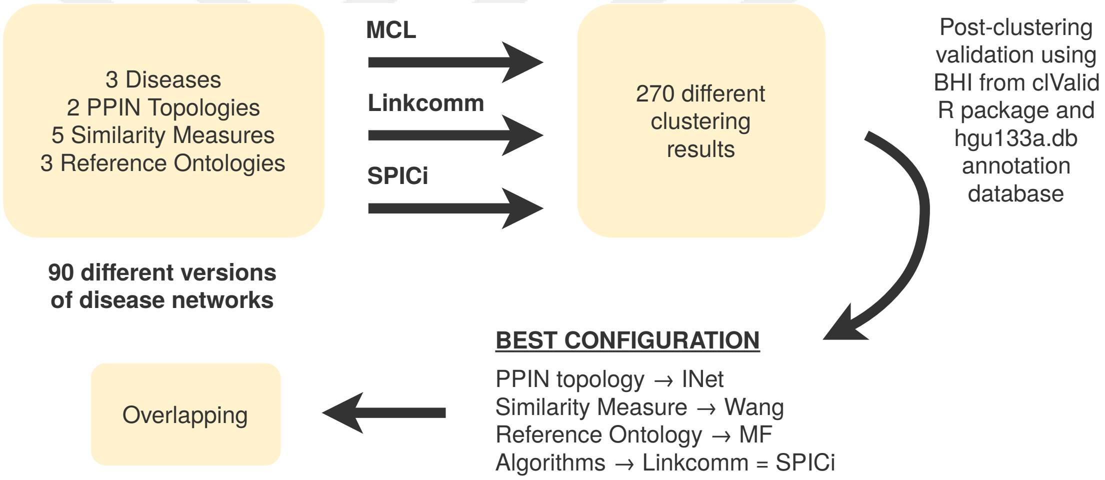
3 Diseases
2 PPIN Topologies
5 Similarity Measures
3 Reference Ontologies
MCL
Linkcomm
SPICi
270 different clustering results
Post-clustering validation using BHI from clValid R package and hgu133a.db annotation database
90 different versions of disease networks
BEST CONFIGURATION
PPIN topology → INet
Similarity Measure → Wang
Reference Ontology → MF
Algorithms → Linkcomm = SPICi
Overlapping

Figure 3.4 The summary of clustering, validation, and overlapping steps

In the results and discussions chapter, we will explicitly present the revealed clusters for each disease network and the common modules detected for each pair of diseases alongside a comparative evaluation of network construction and clustering methods. Additionally, we perform a biological assessment for our findings.

CHAPTER FOUR
RESULTS AND DISCUSSIONS

We have followed an integrative approach that aims to combine multiple biological data sources and computational methods to identify common disease related protein complexes in MS, T2D, and CAD. To obtain the best configuration for clustering of the disease networks, we comparatively evaluated the biological significance of 270 networks clustered using various parameters. These parameters include: three gene expression matrices (MS, T2D, CAD), two protein-protein interaction networks (STRING, INet), five GO similarity measures (Resnik, Rel, Lin, Jiang, Wang), three orthogonal ontologies (MF, BP, CC), and three clustering algorithms (MCL, SPICi, Linkcomm). As a significance indicator, we used the biological homogeneity index (*BHI*) and after we obtained the most accurate clustering, we overlapped the revealed clusters to identify the shared disease-related genes.

4.1 Comparison of PPI Networks

For each PPIN topology, we evaluated the biological homogeneity of clusters produced by 135 different configurations for three disease networks, MS, T2D, and CAD. For 92 out of 135 configurations (68%) the INet network outperformed the STRING network. On the other hand, the STRING performed better in 19 cases (14%), while it showed equal success with INet in 24 cases (18%). The mean *BHI* scores achieved on MS, T2D, and CAD are 0.449 and 0.408 for INet and STRING respectively. We present the average and disease subject specific *BHI* scores in Table 4.1 where *BHI* scores vary in the range of [0, 1] and larger values correspond to biologically more homogeneous clusters. The results show that constructing the disease networks on INet topology would provide more homogeneous clusters for our data set.

Table 4.1 The average *BHI* scores for STRING and INet topologies. The scores are produced using 45 different configurations (five GO similarity measures, three orthogonal ontologies, and three clustering algorithm) for each disease network (MS, T2D, and CAD). The largest value in each column is highlighted and corresponds to most biologically homogeneous clustering

	MS	T2D	CAD	AVERAGE
BHI_{STRING}	0.308	0.449	0.467	0.408
BHI_{INet}	**0.407**	**0.458**	**0.481**	**0.449**

4.2 Comparison of GO Similarity Measures

To select the most accurate GO similarity measure to calculate edge weights on the disease networks based on semantic similarity, we performed clustering procedure by using each of five similarity measures (Jiang, Lin, Rel, Resnik, and Wang) on the INet topology. We generated the BHI scores using three types of orthogonal ontologies (BP, CC, and MF). We present the average and disease subject specific *BHI* scores for each measure in Table 4.2.

Table 4.2 The average *BHI* scores produced by nine different configurations (three orthogonal ontology and three clustering algorithm) for MS, T2D, and CAD. The disease networks are constructed using the INet topology. The largest value in each column is highlighted and corresponds to most biologically homogeneous clustering

	MS	T2D	CAD	AVG
BHI_{Jiang}	0.425	**0.492**	0.477	0.465
BHI_{Lin}	0.415	0.491	0.482	0.463
BHI_{Rel}	0.441	0.491	0.481	0.471
BHI_{Resnik}	0.300	0.326	**0.484**	0.370
BHI_{Wang}	**0.456**	**0.492**	0.481	**0.476**

On disease networks of MS and T2D, Wang similarity measure outperformed the others. Although Resnik measure generated slightly more homogeneous clusters on CAD, its performance is significantly worse on MS and T2D. Therefore, the Wang, which had a more successful average score than the others, stepped forward as the best similarity measure for our analyses.

4.3 Comparison of Orthogonal Ontologies

After selecting the INet topology and the Wang measure, we need to choose the most efficient orthogonal ontology between BP, CC, and MF to generate edge weights based on GO semantic similarity and to perform clustering procedure on each disease network. By following a similar approach, we compared the average *BHI* scores achieved by each ontology on each disease network constructed by INet. In Table 4.3, we present the average *BHI* scores generated on the INet topology using Wang similarity measure and different clustering algorithms. On each disease subject, MF outperformed the BP and CC in terms of creating more homogeneous clusters.

Table 4.3 The average *BHI* scores produced by three clustering algorithms for MS, T2D, and CAD. The disease networks are constructed using the INet topology. The GO similarity scores are generated using the Wang measure. The largest value in each column is highlighted and corresponds to most biologically homogeneous clustering

	MS	T2D	CAD	AVERAGE
BHI_{BP}	0.442	0.486	0.478	0.468
BHI_{CC}	0.444	0.490	0.481	0.472
BHI_{MF}	**0.480**	**0.498**	**0.483**	**0.487**

4.4 Comparison of Clustering Algorithms

We ran each clustering algorithm on the disease networks built by use of the INet topology, the Wang measure, and MF ontology. We present the number of clustered nodes (m_n), the number of clusters produced (m_c), and the average cluster size (m_n / m_c) for each clustering algorithm in Table 4.4. We also show the coverage rate ($100 \times m_n / m_a$) and the number of clusters by the network size (m_a / m_c) in Figure 4.1.

The results show that MCL outperforms the other algorithms in terms of coverage rate (*100 x* m_n / m_a) and the average cluster size (m_n / m_c). Similarly, MCL produces significantly more clusters (m_c) than SPICi and Linkcomm for T2D and MS networks. However, the biological homogeneity (*BHI*) achieved by MCL is not as high as SPICi and Linkcomm (Table 4.5). Since *BHI* scores obtained by SPICi and Linkcomm algorithms are very close, we decided to apply clustering by using both algorithms in order to verify the results and not to miss out any significant clusters revealed by one of the algorithms.

Table 4.4 The number of clustered nodes (m_n), the number of clusters produced (m_c), and the average cluster size (m_n / m_c) for MCL, SPICi, and Linkcomm algorithms. Each value is generated by using the INet topology, the Wang measure, and MF ontology

		MCL			**SPICi**			**Linkcomm**		
Topology	**Subject**	m_n	m_c	m_n / m_c	m_n	m_c	m_n / m_c	m_n	m_c	m_n / m_c
STRING	MS	29	7	4.1	6	2	3	3	1	3
	T2D	93	24	3.9	37	15	2.5	31	9	3.4
	CAD	716	111	6.5	354	84	4.2	635	235	2.7
INet	MS	19	3	6.3	3	1	3	7	2	3.5
	T2D	50	13	3.8	11	5	2.2	6	2	3
	CAD	580	76	7.6	296	68	4.4	532	137	3.9

Table 4.5 The *BHI* scores produced by three clustering algorithms for MS, T2D, and CAD. The disease networks are constructed using the INet topology. The GO similarity scores are generated using the Wang measure and MF ontology. The largest value in each column is highlighted and corresponds to most biologically homogeneous clustering

	MS	**T2D**	**CAD**	**AVG**
$\boldsymbol{BHI_{MCL}}$	0.441	0.493	0.467	0.467
$\boldsymbol{BHI_{SPICi}}$	**0.500**	**0.500**	**0.492**	**0.497**
$\boldsymbol{BHI_{Linkcomm}}$	**0.500**	**0.500**	0.490	**0.497**

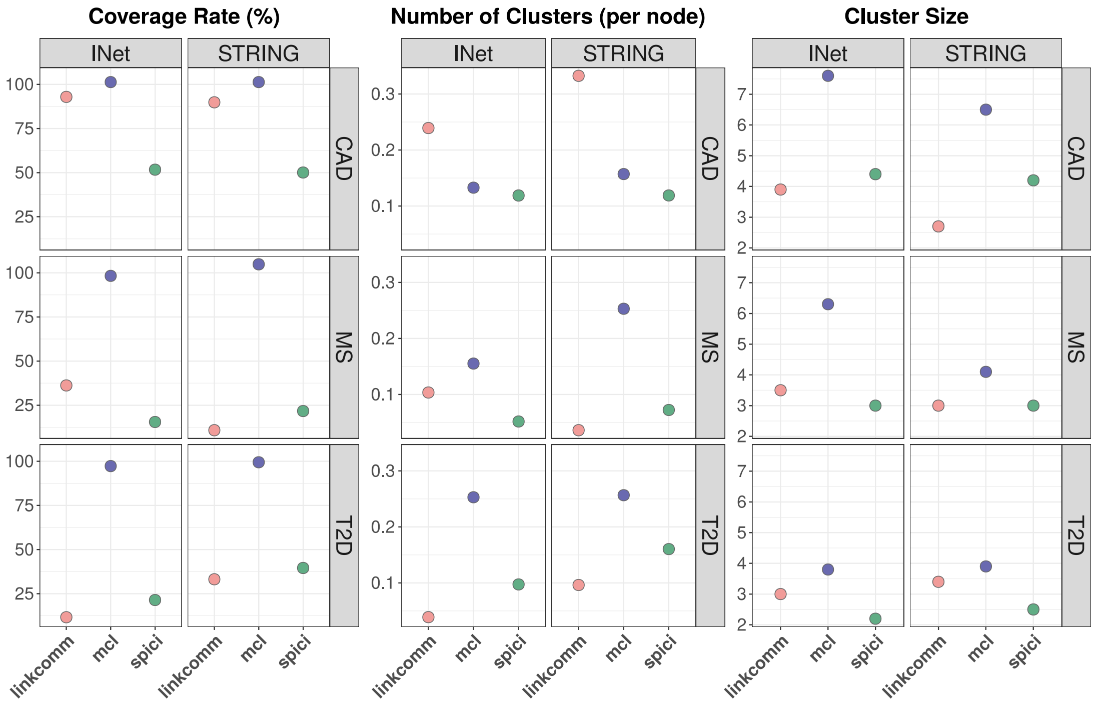

Figure 4.1 Coverage rate (%), number of clusters per node, and average cluster size for Linkcomm, MCL, and SPICi algorithms. Coverage rate is *100 x* (m_n / m_a), number of clusters per node is m_a / m_c, the cluster size is m_n / m_c where m_a is the number of all nodes, m_n is the number of clustered nodes, and m_c is the number of clusters. The values are generated by using the STRING and INet topologies, the Wang measure, and MF ontology

4.5 Discovered Disease Modules

Using the Linkcomm clustering algorithms, we revealed two clusters with seven unique genes for MS, four clusters with ten unique genes for T2D, and 137 clusters with 532 unique genes for CAD. Using the SPICi clustering algorithm, we detected one cluster with four genes for MS, six clusters with 13 genes for T2D, and 68 clusters with 296 genes for CAD.

4.5.1 Discovered Disease Modules for MS

We identified two clusters, {*RSP9, ARPC1B, POLR2L, ANAPC2*} and {*ANAPC2, VPS28, PCGF6, HCFC1*}, in the MS network by using the Linkcomm algorithm. On the other hand, we detected one cluster, {*GSPT2, RPS9, POLR2L, ANAPC2*} by using the SPICi algorithm. We show the discovered clusters in Figure 4.2.

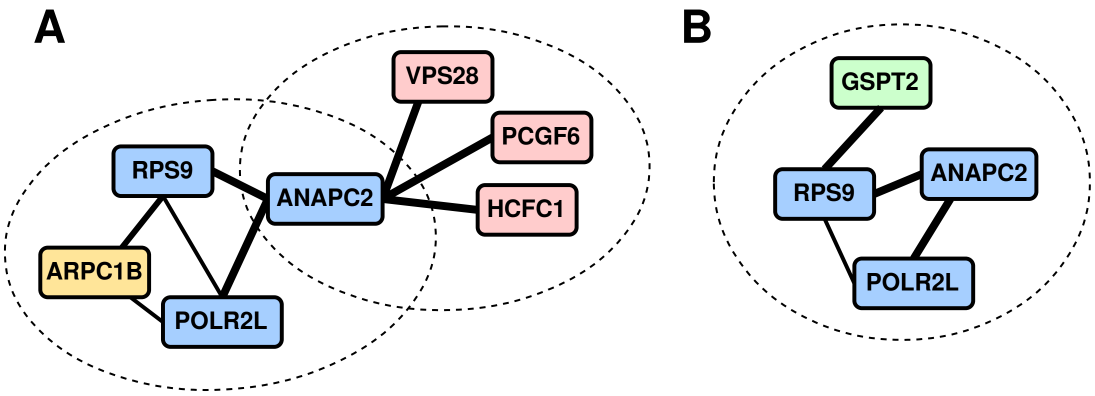

Figure 4.2 Discovered clusters in MS network. (A) Two modules with seven unique genes are detected by using the Linkcomm algorithm. (B) One module with four unique genes are detected by using the SPICi algorithm. The edge thicknesses represent the GO similarity value of two genes. The common nodes are shown in the same color

4.5.2 Discovered Disease Modules for T2D

We identified four disease modules, {*SP1, POU2F1*}, {*HBD, ALAS2*}, {*PLEKHG5, RAC3, ARHGAP10*} and {*DYNC1I1, SPTBN2, RILP*}, in T2D network by using the Linkcomm algorithm. On the other hand, we detected five modules, {S*P1, POU2F1*}, {*HBD, ALAS2*}, {*PELI3, MAP3K14*}, {*RAC3, NRBP1*} and {*TXNRD2, SAMM50, BCS1L*} in the same network by using the SPICi algorithm. We show the revealed clusters in Figure 4.3.

4.5.3 Discovered Disease Modules for CAD

We identified 68 clusters with 296 genes for the CAD network by using the SPICi algorithm. On the other hand, we detected 137 clusters with 532 unique genes by using the Linkcomm algorithm. However, total number of genes in 137 clusters reached up to 2,907 since the Linkcomm algorithm produce overlapping clusters. Moreover, the number of edges between the clustered nodes were 5,461 and 3,948 for Linkcomm and SPICi, respectively. Because of the huge size and complexity of these networks, we summarize the revealed clusters in the supplementary materials.

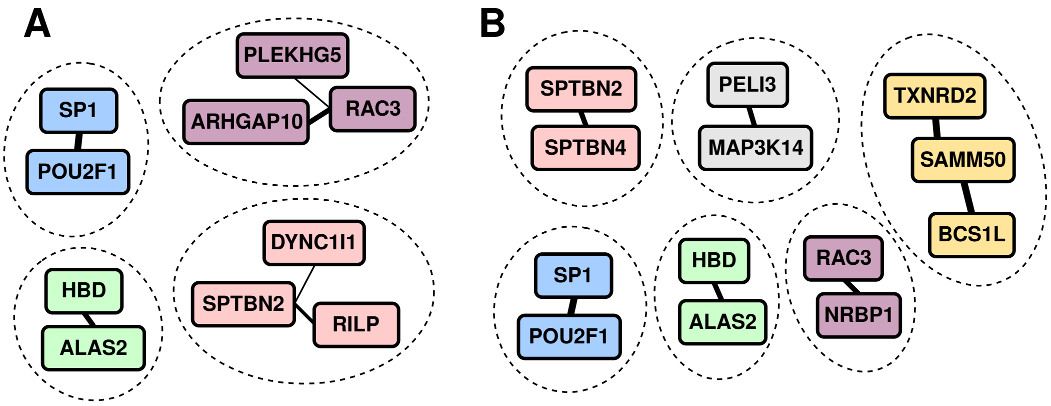

Figure 4.3 Discovered clusters in T2D network. (A) Four modules with ten unique genes are detected by using the Linkcomm algorithm. (B) Six modules with 13 unique genes are detected by using the SPICi algorithm. The edge thicknesses represent the GO similarity value of two genes. The modules with common nodes are shown in the same color

4.6 Common Disease Modules

By overlapping the clusters of the Linkcomm algorithm, we detected two modules, {*ARPC1B, ANAPC2, RPS9, POLR2L*} and {*VPS28, PCGF6, HCFC1, ANAPC2*}, and 12 individual genes, {*TNFSF13, TNFRSF13B, NFKBIB, FRG1, S100A8, ENTPD2, SIX3, LHX2, GSPT2, ISYNA1, COX5A, GLUD2*}, shared by MS–CAD pair (Figure 4.4). On the other hand, we found a single shared gene, *SP1*, alongside a shared module, {*ALAS2, HBD*}, for T2D–CAD (Figure 4.4).

By overlapping the clusters of the SPICi algorithm, we identified one module, {GSPT2, *ANAPC2, RPS9, POLR2L*}, and 15 individual genes, {*TNFSF13, TNFRSF13B, NFKBIB, FRG1, S100A8, ENTPD2, SIX3, LHX2, ISYNA1, COX5A, GLUD2, VPS28, PCGF6, HCFC1, ANAPC2*}, shared by MS–CAD (Figure 4.5). As same as Linkcomm, one common module, {*ALAS2, HBD*}, and one common gene, *SP1*, revealed for T2D–CAD pair (Figure 4.5).

Unlike MS–CAD and T2D–CAD, we could not identify an intersecting module for the MS–T2D using neither the SPICi nor the Linkcomm clustering. Therefore, our analyses could not reveal a disease module shared by all disease groups.

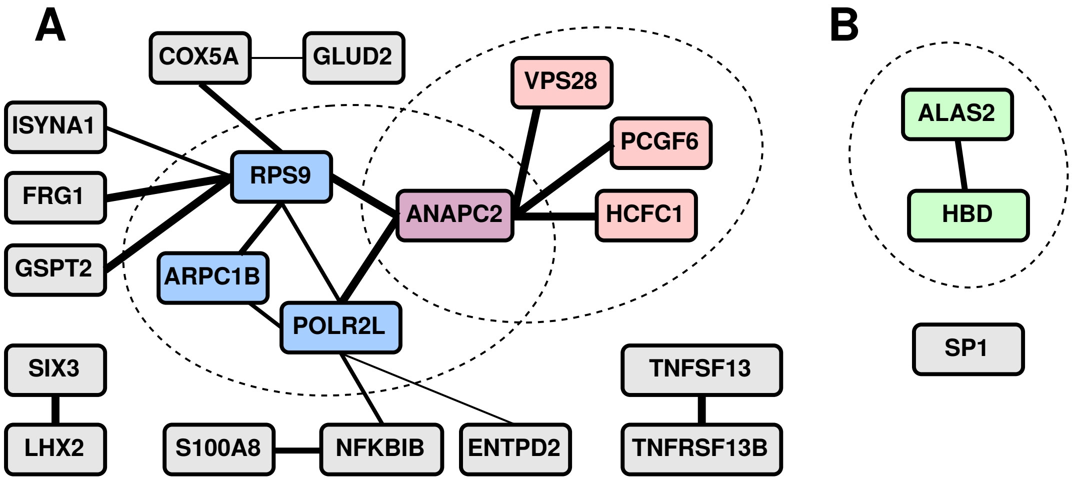

Figure 4.4 Overlapping modules and genes of the Linkcomm clustering. (A) Two modules with seven genes and 12 individual genes are shared by MS-CAD pair. (B) One module with two genes (ALAS2, HBD) and the SP1 gene are shared by T2D-CAD pair. The gray nodes represent the common genes that are either unclustered or diversely clustered in each network. The edge thicknesses denote the GO similarity value of two genes. The modules with common nodes are shown in the same color

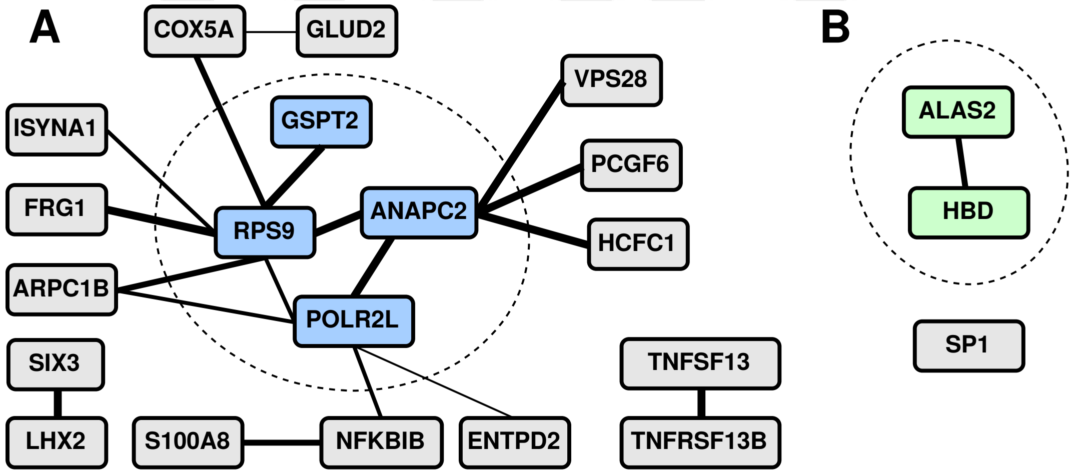

Figure 4.5 Overlapping modules and genes of the SPICi clustering. (A) One module with four genes and 15 individual genes are shared by MS-CAD pair. (B) One module with two genes (ALAS2, HBD) and the SP1 gene are shared by T2D-CAD pair. The gray nodes represent the common genes that are either unclustered or diversely clustered in each network. The edge thicknesses denote the GO similarity value of two genes. The modules with common nodes are shown in the same color

4.7 Evaluation of the Gene-Disease Associations

In total, SPICi and Linkcomm algorithms revealed 19 unique genes (*ANAPC2, ARPC1B, COX5A, ENTPD2, FRG1, GLUD2, GSPT2, HCFC1, ISYNA1, LHX2, NFKBIB, PCGF6, POLR2L, RPS9, S100A8, SIX3, TNFRSF13B, TNFSF13, VPS28*) shared by MS–CAD pair and three unique genes (*ALAS2, HBD, SP1*) shared by T2D–CAD pair. By using the DisGeNET (Piñero et al., 2017), we identified 1,136 gene-disease associations (GDAs) for 22 genes in question. We present the number of GDAs reported for each gene (N_{gda}) in Table 4.6.

Then, we obtained the GDAs that are related with our diseases (MS, T2D, and CAD) by filtering out the associations reported for irrelevant disease classes. We selected three disease classes: nutritional and metabolic diseases (NMD), endocrine system diseases (ESD), and cardiovascular diseases (CVD). In ESD class, we also applied keyword filtering to obtain GDAs that are related with only diabetes and diabetic complications (i.e., DIAB sub class).

In NMD class, we identified 11 genes (*APC2, COX5A, FRG1, HCFC1, ISYNA1, NFKBIB, S100A8, TNFSF13, ALAS2, HBD, SP1*) with 57 GDAs reported by 82 articles (Table 4.7). In DIAB sub class, we identified 4 genes (*ISYNA1, S100A8, SIX3, SP1*) with 12 GDAs reported by 15 articles (Table 4.8). In CVD class, we found 11 genes (*ALAS2, APC2, COX5A, ENTPD2, ISYNA1, RPS9, S100A8, SIX3, SP1, TNFRSF13B, TNFSF13*) with 60 GDAs reported by 88 articles (Table 4.9).

Five out of 19 genes (*APC2, COX5A, ISYNA1, S100A8, TNFSF13*) identified for MS–CAD pair were already associated with both NMD and CVD class diseases. On the other hand, three genes (*FRG1, HCFC1, NFKBIB*) were only associated with NMD, three genes (*ENTPD2, RPS9, TNFRSF13B*) were only associated with CVD, and one gene (*SIX3*) was associated with DIAB and CVD. The remaining genes (*ARPC1B, GLUD2, GSPT2, LHX2, PCGF6, POLR2L, VPS28*) are novel for MS and CAD, since they were not previously associated with NMD or CVD (Table 4.10).

Among three disease genes (*ALAS2, HBD, SP1*) identified for T2D–CAD pair, only *SP1* was previously associated with DIAB and CVD. *ALAS2* was associated with NMD and CVD, while *HBD* was only associated with NMD (Table 4.10). In this respect, we suggest only the *HBD* as a novel gene shared by T2D–CAD.

Table 4.6 Gene symbol, entrez identifier, description, and the number of GDAs (N_{gda}) reported for each gene according to the DisGeNET. * ANAPC2 is replaced by APC2 in the DisGeNET

Symbol	**Entrez**	**Description**	N_{gda}
APC2*	10297	APC2, WNT signaling pathway regulator	82
ARPC1B	10095	actin related protein 2/3 complex subunit 1B	6
COX5A	9377	cytochrome c oxidase subunit 5A	75
ENTPD2	954	ectonucleoside triphosphate diphosphohydrolase 2	12
FRG1	2483	FSHD region gene 1	47
GLUD2	2747	glutamate dehydrogenase 2	5
GSPT2	23708	G1 to S phase transition 2	4
HCFC1	3054	host cell factor C1	43
ISYNA1	51477	inositol-3-phosphate synthase 1	95
LHX2	9355	LIM homeobox 2	16
NFKBIB	4793	NFKB inhibitor beta	12
PCGF6	84108	polycomb group ring finger 6	1
POLR2L	5441	RNA Polymerase II Subunit L	0
RPS9	6203	ribosomal protein S9	2
S100A8	6279	S100 calcium binding protein A8	199
SIX3	6496	SIX homeobox 3	84
TNFRSF13B	23495	TNF receptor superfamily member 13B	118
TNFSF13	8741	TNF superfamily member 13	90
VPS28	51160	VPS28, ESCRT-I subunit	1
ALAS2	212	5'-aminolevulinate synthase 2	59
HBD	3045	hemoglobin subunit delta	26
SP1	6667	Sp1 transcription factor	159
TOTAL	–	–	**1136**

Table 4.7 Gene symbol, entrez identifier, disease pair, the number of GDAs reported for ***nutritional and metabolic diseases*** ($N_{gda(NMD)}$), the number of publications supporting the corresponding associations (N_P), and the maximum GDA score (S_{max}) for each gene according to the DisGeNET. * ANAPC2 is replaced by APC2 in the DisGeNET

Symbol	**Entrez**	**Disease Pair**	$N_{gda(NMD)}$	N_P	S_{max}
APC2*	10297	MS–CAD	2	0	0.100
COX5A	9377	MS–CAD	9	12	0.340
FRG1	2483	MS–CAD	1	0	0.100
HCFC1	3054	MS–CAD	4	2	0.100
ISYNA1	51477	MS–CAD	6	9	0.030
NFKBIB	4793	MS–CAD	1	1	0.010
S100A8	6279	MS–CAD	14	18	0.030
TNFSF13	8741	MS–CAD	2	2	0.010
ALAS2	212	T2D–CAD	13	32	0.600
HBD	3045	T2D–CAD	1	0	0.100
SP1	6667	T2D–CAD	4	6	0.320
TOTAL	–	–	**57**	**82**	–

Table 4.8 Gene symbol, entrez identifier, disease pair, the number of GDAs reported for ***diabetes and diabetic complications*** ($N_{gda(DIAB)}$), the number of publications supporting the corresponding associations (N_P), and the maximum GDA score (S_{max}) for each gene according to the DisGeNET

Symbol	**Entrez**	**Disease Pair**	$N_{gda(DIAB)}$	N_P	S_{max}
ISYNA1	51477	MS–CAD	6	9	0.010
S100A8	6279	MS–CAD	4	4	0.030
SIX3	6496	MS–CAD	1	1	0.100
SP1	6667	T2D–CAD	1	1	0.010
TOTAL	–	–	**12**	**15**	–

Table 4.9 Gene symbol, entrez identifier, disease pair, the number of GDAs reported for ***cardiovascular diseases*** ($N_{gda(CVD))}$), the number of publications supporting the corresponding associations (N_P), and the maximum GDA score (S_{max}) for each gene according to the DisGeNET.
* ANAPC2 is replaced by APC2 in the DisGeNET

Symbol	**Entrez**	**Disease Pair**	$N_{gda(CVD)}$	N_P	S_{max}
APC2*	10297	MS–CAD	1	1	0.010
COX5A	9377	MS–CAD	4	4	0.010
ENTPD2	954	MS–CAD	1	1	0.200
ISYNA1	51477	MS–CAD	13	16	0.020
RPS9	6203	MS–CAD	1	1	0.010
S100A8	6279	MS–CAD	13	20	0.040
SIX3	6496	MS–CAD	2	2	0.010
TNFRSF13B	23495	MS–CAD	6	12	0.450
TNFSF13	8741	MS–CAD	9	15	0.060
ALAS2	212	T2D–CAD	3	5	0.020
SP1	6667	T2D–CAD	7	11	0.030
TOTAL	–	–	**60**	**88**	–

Another important point is that although our clustering algorithms could not reveal a disease module shared by all disease groups, three genes (*ISYNA1, S100A8, SP1*) that we identified were associated with NMD, DIAB, and CVD disease classes. Additionally, three genes (*ALAS2, HBD, SIX3*) are potentially shared by all disease groups, since they were associated with a complementary disease class in the DisGeNET (e.g., DIAB is a complementary class for a gene shared by MS–CAD).

In a nutshell, by using the DisGeNET, we biologically evaluated our GDA results for 22 genes identified as shared disease genes for MS–CAD and T2D–CAD pairs in our analyses. In pairwise associations, we obtained *full-matching* (i.e., previous association with both disease classes) for six genes and *half-matching* (i.e., previous association with one of the disease classes) for eight genes. In addition, we identified eight novel genes that have no previous association with any of the disease classes.

Table 4.10 Gene symbol, entrez identifier, disease pair, GDA reported disease classes, and matching result (full, half, or none) according to DisGeNET. * ANAPC2 is replaced by APC2 in DisGeNET

Symbol	Entrez	Disease Pair	GDA reported class(es)	Matching
APC2*	10297	MS–CAD	NMD + CVD	Full
ARPC1B	10095	MS–CAD	–	None
COX5A	9377	MS–CAD	NMD + CVD	Full
ENTPD2	954	MS–CAD	CVD	Half
FRG1	2483	MS–CAD	NMD	Half
GLUD2	2747	MS–CAD	–	None
GSPT2	23708	MS–CAD	–	None
HCFC1	3054	MS–CAD	NMD	Half
ISYNA1	51477	MS–CAD	NMD + DIAB + CVD	Full
LHX2	9355	MS–CAD	–	None
NFKBIB	4793	MS–CAD	NMD	Half
PCGF6	84108	MS–CAD	–	None
POLR2L	5441	MS–CAD	–	None
RPS9	6203	MS–CAD	CVD	Half
S100A8	6279	MS–CAD	NMD + DIAB + CVD	Full
SIX3	6496	MS–CAD	DIAB + CVD	Half
TNFRSF13B	23495	MS–CAD	CVD	Half
TNFSF13	8741	MS–CAD	NMD + CVD	Full
VPS28	51160	MS–CAD	–	None
ALAS2	212	T2D–CAD	NMD + CVD	Half
HBD	3045	T2D–CAD	NMD	None
SP1	6667	T2D–CAD	NMD + DIAB + CVD	Full

4.8 Functional and Relational Evaluation of the Novel Disease Genes

To gain an insight into the functional and biological features of the eight novel disease genes (*ARPC1B, GLUD2, GSPT2, HBD, LHX2, PCGF6, POLR2L, VPS28*), we investigated their associations with non-metabolic disorders and their interactions with other genes associated with metabolic disorders.

ARPC1B encodes one of seven subunits of the human Arp2/3 protein complex which has been implicated in the control of actin polymerization in cells. It is known that *ARPC1B* plays a major role in the regulation of the actin cytoskeleton and its deficiency causes platelet and immune system abnormalities (Kahr et al., 2017).

GLUD2 encodes an enzyme localized to the mitochondrion and acts as a homohexamer to recycle glutamate during neurotransmission. *GLUD2* is associated with Parkinson disease (Plaitakis et al., 2010; Plaitakis, Latsoudis, & Spanaki, 2011; Plaitakis, Zaganas, & Spanaki, 2013). More importantly, its housekeeping isoform *GLUD1* is clearly associated with several metabolic disorders and diabetic conditions such as hypoglycemia, hyperinsulinism, and hyperammonemia (Balasubramaniam et al., 2011; Chik, Chan, Lam, & Ng, 2008; Corrêa-Giannella et al., 2012; Darendeliler & Bas, 2006; Flanagan et al., 2010; Ihara et al., 2005; MacMullen et al., 2001; Meissner, Mayatepek, Kinner, & Santer, 2004; Santer et al., 2001; C. A. Stanley et al., 1998; Charles A. Stanley, 2011; Tran et al., 2015).

GSPT2 encodes a GTP-binding protein which has an essential role at the G1 to S-phase transition of the cell cycle in human and yeast cells. *GSPT2* is associated with intellectual disability (Grau et al., 2017). It is closely related to *GSPT1* and shown to interact with *PABPC1* (Hoshino, Imai, Kobayashi, Uchida, & Katada, 1999). However, none of these genes are previously associated with metabolic disorders.

HBD and *HBB* genes are normally expressed in the adults and responsible from constitution of the hemoglobin. Mutations in the *HBD* are associated with Delta-thalassemia, an inherited blood disorder characterized by abnormal hemoglobin production. (Matsunaga, Kimura, Yamada, Fukumaki, & Takagi, 1985; Vives-Corrons, Pujades, Miguel-García, Miguel-Sosa, & Cambiazzo, 1992; J. W. Zhang, Stamatoyannopoulos, & Anagnou, 1988). On the other hand, *HBB* is associated with several CVDs (Bender, 1993; Dinakaran et al., 2014; Makani et al., 2011).

LHX2 encodes a protein that belongs to a large protein family, members of which carry the LIM domain and function as a transcriptional regulator. It is associated with neoplastic process, digestive system diseases, and rheumatoid arthritis (Galligan, Baig, Bykerk, Keystone, & Fish, 2007; Kuzmanov et al., 2014; Shi et al., 2015). *LHX2* is shown to interact with *CITED2* (Glenn & Maurer, 1999) which is strongly associated with several heart diseases (Sperling et al., 2005; Su et al., 2013; Sun et al., 2006; Yin et al., 2002).

PCGF6 encodes a Polycomb group (PcG) protein, which acts as a master regulator to ensure embryonic stem cell development and differentiation (C.-S. Yang, Chang, Dang, & Rana, 2016; Wukui Zhao et al., 2017). *PCGF6* is most closely related to *PCGF2* (i.e. *MEL-18*) that is known as a marker in breast cancer (Lee et al., 2014; Park et al., 2011; Riis et al., 2010).

POLR2L encodes a subunit of RNA polymerase II that is the polymerase responsible for synthesizing messenger RNA in eukaryotes, and it is shown to interact with *POLR2A* (Acker et al., 1997) which is associated with heart failure and cardiomyopathy (Brattelid et al., 2010).

VPS28 encodes a protein subunit of the ESCRT-I complex, which functions in the transportation and sorting of proteins into subcellular vesicles. Although there is not a GDA reported for *VPS28* in the literature, *VPS37C* in the same subunit (ESCRT-I) is associated with rheumatoid arthritis and cardiometabolic disorders (Eyre et al., 2012; Kettunen et al., 2012). Additionally, *VPS37A* in the same subunit has a strong association with hereditary spastic paraplegia (Zivony-Elboum et al., 2012).

As a conclusion, we discovered that five out of eight genes (*GLUD2, HBD, LHX2, POLR2L, VPS28*) that we identified as novel disease genes for MS–CAD and T2D–CAD have some indirect associations with diseases in NMD and CVD class. On the other hand, there is no previous metabolic disorder association reported for the remaining three genes (*ARPC1B, GSPT2, PCGF6*) in the literature.

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

Identifying the shared disease-genes and protein complexes for multiple metabolic disorders is crucial to enable accurate prognoses and to design targeted drug therapies. In this respect, we have suggested an integrative bioinformatics model that aims to combine multiple biological data sources and computational methods to identify common disease related genes and modules in MS, T2D, and CAD.

We constructed 90 disease networks for each disease group by using of different protein-protein interaction network topologies, orthogonal ontologies, GO similarity measures, and clustering algorithms. Then, we evaluated the performance of each configuration by considering the biological homogeneity achieved in the generated clusters. The networks constructed using the INet topology provided higher BHI scores in comparison to networks constructed using the STRING. In calculation of GO semantic similarity scores, the only graph-based similarity measure, Wang, performed better than all of the information content based similarity measures (Jiang, Lin, Rel, and Resnik). Especially, the BHI scores obtained by combining the Wang measure and MF ontology were significantly higher in comparison to any other measure. For clustering, performance of the Linkcomm and the SPICi algorithms was almost identical and better than MCL. Therefore, we executed both algorithms to generate the clusters.

By overlapping the generated clusters, we identified 22 genes (*ANAPC2, ARPC1B, COX5A, ENTPD2, FRG1, GLUD2, GSPT2, HCFC1, ISYNA1, LHX2, NFKBIB, PCGF6, POLR2L, RPS9, S100A8, SIX3, TNFRSF13B, TNFSF13, VPS28, ALAS2, HBD, SP1*) shared by MS–CAD and T2D–CAD pairs. Three of these genes (*ISYNA1, S100A8, SP1*) were previously associated with all of the nutritional and metabolic diseases (NMD), diabetes and diabetic complications (DIAB), and cardiovascular diseases (CVD). Four of them (*ALAS2, APC2, COX5A, TNFSF13*) were associated with NMD and CVD class diseases, and one gene (*SIX3*) was associated with DIAB and CVD class diseases. Four genes (*FRG1, HCFC1,* *NFKBIB, HBD*) were associated with only NMD, and three genes (*ENTPD2, RPS9, TNFRSF13B*) were associated with only CVD. The remaining seven genes (*ARPC1B, GLUD2, GSPT2, LHX2, PCGF6, POLR2, VPS28*) were not associated with any of the NMD, DIAB, or CVD class diseases.

We performed a functional and relational analysis for these seven genes and *HBD*, which revealed as a shared disease gene for T2D–CAD pair in our analyses but is not associated with neither DIAB nor CVD class diseases before. We investigated the associations with non-metabolic disorders and the gene-gene interactions with external genes, which have associations with NMD, DIAB, or CVD class diseases.

We found that five out of eight novel genes (*GLUD2, HBD, LHX2, POLR2L, VPS28*) that we identified as novel disease genes for MS–CAD and T2D–CAD have some indirect associations with diseases in NMD and CVD class. On the other hand, *ARPC1B* was associated with platelet and immune system abnormalities, *GSPT2* was associated with intellectual disability, and *PCGF6* was associated with breast cancer. However, no previous association with NMD, DIAB, or CVD class diseases reported for these three genes.

Our study provided some strong evidence that there are common disease genes underlying the MS, T2D, and CAD. Nevertheless, further investigation with different data sets is required to validate these new findings. In the future studies, the integrated biological data sources may be diversified by use of human disease networks (HDNs), and some machine learning strategies may be executed to select the best network construction parameters. Additionally, different biological validation metrics such as Biological Stability Index (BSI) (Datta & Datta, 2006) and Clustering Quality Score (CQS) (Gat-Viks, Sharan, & Shamir, 2003) may be combined with the Biological Homogeneity Index (BHI) in the post-clustering validation.

REFERENCES

Acker, J., Graaff, M. de, Cheynel, I., Khazak, V., Kedinger, C., & Vigneron, M. (1997). Interactions between the Human RNA Polymerase II Subunits. *Journal of Biological Chemistry*, *272*(27), 16815–16821.

Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–764.

Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414.

Alberti, K. G. M. M., Zimmet, P., & Shaw, J. (2005). The metabolic syndrome—A new worldwide definition. *The Lancet*, *366*(9491), 1059–1062.

Alberti, K. G. M. M., & Zimmet, P. Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabetic Medicine: A Journal of the British Diabetic Association*, *15*(7), 539–553.

Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, *22*(13), 1600–1607.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Aune, T. M., Maas, K., Moore, J. H., & Olsen, N. J. (2003). Gene expression profiles in human autoimmune disease. *Current Pharmaceutical Design*, *9*(23), 1905–1917.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*(1), 2.

Baechler, E. C., Batliwalla, F. M., Karypis, G., Gaffney, P. M., Ortmann, W. A., Espe, K. J., … Behrens, T. W. (2003). Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(5), 2610–2615.

Balasubramaniam, S., Kapoor, R., Yeow, J. H. H., Lim, P. G., Flanagan, S., Ellard, S., & Hussain, K. (2011). Biochemical evaluation of an infant with hypoglycemia resulting from a novel de novo mutation of the GLUD1 gene and hyperinsulinism-hyperammonemia syndrome. *Journal of Pediatric Endocrinology & Metabolism: JPEM*, *24*(7–8), 573–577.

Balkau, B., & Charles, M. A. (1999). Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). *Diabetic Medicine : A Journal of the British Diabetic Association*, *16*(5), 442–443.

Bender, M. A. (1993). Sickle Cell Disease. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, & A. Amemiya (Eds.), *GeneReviews*. Seattle (WA): University of Washington, Seattle.

Ben-Hur, A., & Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, *21*(suppl_1), i38–i46.

Bhowmick, S. S., & Seah, B. S. (2016). Clustering and summarizing protein-protein interaction networks: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 638–658.

Bolshakova, N., Azuaje, F., & Cunningham, P. (2005). A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, *21*(10), 2546–2547.

Bomprezzi, R., Ringnér, M., Kim, S., Bittner, M. L., Khan, J., Chen, Y., … Trent, J. M. (2003). Gene expression profile in multiple sclerosis patients and healthy controls: Identifying pathways relevant to disease. *Human Molecular Genetics*, *12*(17), 2191–2199.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715.

Brattelid, T., Winer, L. H., Levy, F. O., Liestøl, K., Sejersted, O. M., & Andersson, K. B. (2010). Reference gene alternatives to Gapdh in rodent and human heart failure gene expression studies. *BMC Molecular Biology*, *11*, 22.

Bravo, À., Cases, M., Queralt-Rosinach, N., Sanz, F., & Furlong, L. I. (2014). A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed Research International*, *2014*, 1–11.

Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinformatics*, *16*, 55.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, *25*(1), 1–22.

Bui, Q.-C., Katrenko, S., & Sloot, P. M. A. (2011). A hybrid approach to extract protein–protein interactions. *Bioinformatics*, *27*(2), 259–265.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., & Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, *9*, 207.

Carlson, M. (2016). *Hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a) R package version 3.2.3*. Retrieved July 1, 2019, from http://bioconductor.org/packages/hgu133a.db

Carlson, M. (2018). *Org.Hs.eg.db: Genome wide annotation for Human R package version 3.7.0*. Retrieved July 1, 2019, from https://bioconductor.org/packages/org.Hs.eg.db

Carone, B. R., Fauquier, L., Habib, N., Shea, J. M., Hart, C. E., Li, R., … Rando, O. J. (2010). Paternally Induced Transgenerational Environmental Reprogramming of Metabolic Gene Expression in Mammals. *Cell*, *143*(7), 1084–1096.

Carr, M. C., & Brunzell, J. D. (2004). Abdominal Obesity and Dyslipidemia in the Metabolic Syndrome: Importance of Type 2 Diabetes and Familial Combined Hyperlipidemia in Coronary Artery Disease Risk. *The Journal of Clinical Endocrinology & Metabolism*, *89*(6), 2601–2607.

Centers for Disease Control and Prevention. (2011). *National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011*. Retrieved March 15, 2019, from https://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf

Chan, K. H. K., Huang, Y.-T., Meng, Q., Wu, C., Reiner, A., Sobel, E. M., … Liu, S. (2014). Shared molecular pathways and gene networks for cardiovascular disease and type 2 diabetes mellitus in women across diverse ethnicities. *Circulation. Cardiovascular Genetics*, *7*(6), 911–919.

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., … Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, *45*(D1), D369–D379.

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(suppl_2), W305–W311.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., … Zhou, Q. (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science*, *351*(6271), 397–400.

Chen, Y., Dougherty, E. R., & Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, *2*(4), 364–375.

Chik, K.-K., Chan, C.-W., Lam, C.-W., & Ng, K.-L. (2008). Hyperinsulinism and hyperammonaemia syndrome due to a novel missense mutation in the allosteric domain of the glutamate dehydrogenase 1 gene. *Journal of Paediatrics and Child Health*, *44*(9), 517–519.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676.

Corrêa-Giannella, M. L., Freire, D. S., Cavaleiro, A. M., Fortes, M. A. Z., Giorgi, R. R., & Pereira, M. A. A. (2012). Hyperinsulinism/hyperammonemia (HI/HA) syndrome due to a mutation in the glutamate dehydrogenase gene. *Arquivos Brasileiros De Endocrinologia E Metabologia*, *56*(8), 485–489.

Darendeliler, F., & Bas, F. (2006). Hyperinsulinism in infancy—Genetic aspects. *Pediatric Endocrinology Reviews: PER*, *3*(Suppl 3), 521–526.

Datta, S., & Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, *7*(1), 397.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., … Mattingly, C. J. (2019). The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Research*, *47*(D1), D948–D954.

Daxinger, L., & Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature Reviews Genetics*, *13*(3), 153–162.

De Rosa, S., Arcidiacono, B., Chiefari, E., Brunetti, A., Indolfi, C., & Foti, D. P. (2018). Type 2 Diabetes Mellitus and Cardiovascular Disease: Genetic and Epigenetic Links. *Frontiers in Endocrinology*, *9*, 2.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(9), R60.

Dinakaran, V., Rathinavel, A., Pushpanathan, M., Sivakumar, R., Gunasekaran, P., & Rajendhran, J. (2014). Elevated levels of circulating DNA in cardiovascular disease patients: Metagenomic profiling of microbiome in the circulation. *PloS One*, *9*(8), e105221.

Dong, C., Tang, L., Liu, Z., Bu, S., Liu, Q., Wang, Q., … Duan, S. (2014). Landscape of the relationship between type 2 diabetes and coronary heart disease through an integrated gene network analysis. *Gene*, *539*(1), 30–36.

Dongen, S. (2000). *A Cluster Algorithm for Graphs*. PhD Thesis, University of Utrecht, Utrecht.

Eckel, R. H., Grundy, S. M., & Zimmet, P. Z. (2005). The metabolic syndrome. *The Lancet*, *365*(9468), 1415–1428.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., … Worthington, J. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, *44*(12), 1336–1340.

Flanagan, S. E., Kapoor, R. R., Mali, G., Cody, D., Murphy, N., Schwahn, B., … Ellard, S. (2010). Diazoxide-responsive hyperinsulinemic hypoglycemia caused by HNF4A gene mutations. *European Journal of Endocrinology*, *162*(5), 987–992.

Galassi, A., Reynolds, K., & He, J. (2006). Metabolic Syndrome and Risk of Cardiovascular Disease: A Meta-Analysis. *The American Journal of Medicine*, *119*(10), 812–819.

Galligan, C. L., Baig, E., Bykerk, V., Keystone, E. C., & Fish, E. N. (2007). Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: Correlates with disease activity. *Genes and Immunity*, *8*(6), 480–491.

Garrod, A. E. (1908). The croonian lectures in inborn errors of metabolism. *The Lancet*, *172*(4427), 1–7.

Gat-Viks, I., Sharan, R., & Shamir, R. (2003). Scoring clustering solutions by their biological relevance. *Bioinformatics*, *19*(18), 2381–2389.

Genomics England PanelApp. (2019). *A crowdsourcing tool to allow gene panels to be shared, downloaded, viewed and evaluated by the scientific community*. Retrieved May 23, 2019, from https://panelapp.genomicsengland.co.uk/

Glenn, D. J., & Maurer, R. A. (1999). MRG1 Binds to the LIM Domain of Lhx2 and May Function as a Coactivator to Stimulate Glycoprotein Hormone α-Subunit Gene Expression. *Journal of Biological Chemistry*, *274*(51), 36159–36167.

Grandjean, V., Fourré, S., De Abreu, D. A. F., Derieppe, M.-A., Remy, J.-J., & Rassoulzadegan, M. (2015). RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders. *Scientific Reports*, *5*, 18193.

Grau, C., Starkovich, M., Azamian, M. S., Xia, F., Cheung, S. W., Evans, P., … Scott, D. A. (2017). Xp11.22 deletions encompassing CENPVL1, CENPVL2, MAGED1 and GSPT2 as a cause of syndromic X-linked intellectual disability. *PloS One*, *12*(4), e0175962.

Grayson, B. L., Wang, L., & Aune, T. M. (2011). Peripheral blood gene expression profiles in metabolic syndrome, coronary artery disease and type 2 diabetes. *Genes & Immunity*, *12*(5), 341–351.

Grundy, S. M., Hansen, B., Smith, S. C., Cleeman, J. I., & Kahn, R. A. (2004). Clinical Management of Metabolic Syndrome. *Circulation*, *109*(4), 551–556.

Grundy, Scott M. (2004). Obesity, Metabolic Syndrome, and Cardiovascular Disease. *The Journal of Clinical Endocrinology & Metabolism*, *89*(6), 2595–2600.

Gunther, C. C., Langaas, M., Lydersen, S., Beisvåg, V., Junge, F. R. K., Bergum, H., & Lægreid, A. (2005). Statistical hypothesis testing of association between two reporter lists within the GO-hierarchy. *In European Meeting of Statisticians*, *134*.

Guo, X., Liu, R., Shriver, C. D., Hu, H., & Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, *22*(8), 967–973.

Gutiérrez-Sacristán, A., Grosdidier, S., Valverde, O., Torrens, M., Bravo, À., Piñero, J., … Furlong, L. I. (2015). PsyGeNET: A knowledge platform on psychiatric disorders and their genes. *Bioinformatics*, *31*(18), 3075–3077.

Hanson, R. L., Imperatore, G., Bennett, P. H., & Knowler, W. C. (2002). Components of the "Metabolic Syndrome" and Incidence of Type 2 Diabetes. *Diabetes*, *51*(10), 3120–3127.

Hardcastle, T. J., & Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*(1), 422.

Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology*, *4*(11), 682–690.

Hoshino, S., Imai, M., Kobayashi, T., Uchida, N., & Katada, T. (1999). The Eukaryotic Polypeptide Chain Releasing Factor (eRF3/GSPT) Carrying the translation termination signal to the 3′-Poly(A) tail of mRNA direct association of eRF3/GSPT with polyadenylate-binding protein. *Journal of Biological Chemistry*, *274*(24), 16677–16680.

Hu, G., Qiao, Q., Tuomilehto, J., Balkau, B., Borch-Johnsen, K., Pyorala, K., & DECODE Study Group. (2004). Prevalence of the metabolic syndrome and its relation to all-cause and cardiovascular mortality in nondiabetic European men and women. *Archives of Internal Medicine*, *164*(10), 1066–1076.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13.

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., & Ideker, T. (2018). Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Systems*, *6*(4), 484-495.e5.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., … Aitman, T. J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, *37*(3), 243–253.

Huggins, C. E., Domenighetti, A. A., Ritchie, M. E., Khalil, N., Favaloro, J. M., Proietto, J., … Delbridge, L. M. D. (2008). Functional and metabolic remodelling in GLUT4-deficient hearts confers hyper-responsiveness to substrate intervention. *Journal of Molecular and Cellular Cardiology*, *44*(2), 270–280.

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., & Lee, I. (2019). HumanNet v2: Human gene networks for disease research. *Nucleic Acids Research*, *47*(D1), D573–D580.

Ihara, K., Miyako, K., Ishimura, M., Kuromaru, R., Wang, H.-Y., Yasuda, K., & Hara, T. (2005). A case of hyperinsulinism/hyperammonaemia syndrome with reduced carbamoyl-phosphate synthetase-1 activity in liver: A pitfall in enzymatic diagnosis for hyperammonaemia. *Journal of Inherited Metabolic Disease*, *28*(5), 681–687.

International Diabetes Federation. (2006). *The IDF consensus worldwide definition of the metabolic syndrome*. Retrieved March 15, 2019, from https://www.idf.org/e-library/consensus-statements/60-idfconsensus-worldwide-definitionof-the-metabolic-syndrome.html

International Diabetes Federation. (2017). *IDF Diabetes Atlas 8th edition*. Retrieved March 15, 2019, from https://diabetesatlas.org/resources/2017-atlas.html

Isomaa, B., Almgren, P., Tuomi, T., Forsen, B., Lahti, K., Nissen, M., … Groop, L. (2001). Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome. *Diabetes Care*, *24*(4), 683–689.

Jäger, M. L. (2015, March 11). *MCL: Markov Cluster Algorithm*. Retrieved August 20, 2019, from https://cran.r-project.org/package=MCL

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., … Gerstein, M. (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, *302*(5644), 449–453.

Ji, J., Zhang, A., Liu, C., Quan, X., & Liu, Z. (2014). Survey: Functional Module Detection from Protein-Protein Interaction Networks. *IEEE Transactions on Knowledge and Data Engineering*, *26*(2), 261–277.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of 10th International Conference Research on Computational Linguistics*, 1–15.

Jiang, P., & Singh, M. (2010). SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics (Oxford, England)*, *26*(8), 1105–1111.

Jordán, F., Nguyen, T.-P., & Liu, W. (2012). Studying protein–protein interaction networks: A systems view on diseases. *Briefings in Functional Genomics*, *11*(6), 497–504.

Kahr, W. H. A., Pluthero, F. G., Elkadri, A., Warner, N., Drobac, M., Chen, C. H., … Muise, A. M. (2017). Loss of the Arp2/3 complex component ARPC1B causes platelet abnormalities and predisposes to inflammatory disease. *Nature Communications*, *8*, 14816.

Kalinka, A. T., & Guenoche, A. (2014). *linkcomm: Tools for Generating, Visualizing, and Analysing Link Communities in Networks*. Retrieved August 20, 2019, from https://cran.r-project.org/package=linkcomm

Kalinka, A. T., & Tomancak, P. (2011). linkcomm: An R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*, *27*(14), 2011–2012.

Kaur, J. (2014). A comprehensive review on metabolic syndrome. *Cardiology Research and Practice*, *2014*, 943162.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., … Pandey, A. (2009). Human Protein Reference Database—2009 update. *Nucleic Acids Research*, *37*(suppl_1), D767–D772.

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., … Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, *44*(3), 269–276.

King, A. D., Przulj, N., & Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics (Oxford, England)*, *20*(17), 3013–3020.

Ko, Y., Cho, M., Lee, J.-S., & Kim, J. (2016). Identification of disease comorbidity through hidden molecular mechanisms. *Scientific Reports*, *6*, 39433.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., … Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, *47*(D1), D1018–D1027.

Kuzmanov, A., Hopfer, U., Marti, P., Meyer-Schaller, N., Yilmaz, M., & Christofori, G. (2014). LIM-homeobox gene 2 promotes tumor growth and metastasis by inducing autocrine and paracrine PDGF-B signaling. *Molecular Oncology*, *8*(2), 401–416.

Kylin, E. (1923). Studien ueber das Hypertonie-Hyperglyka. *Zentralblatt Fuer Innere Medizin*, *44*, 105–127.

Laaksonen, D. E., Lakka, H.-M., Niskanen, L. K., Kaplan, G. A., Salonen, J. T., & Lakka, T. A. (2002). Metabolic syndrome and development of diabetes mellitus: Application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study. *American Journal of Epidemiology*, *156*(11), 1070–1077.

Lan, W., Wang, J., Li, M., Peng, W., & Wu, F. (2015). Computational approaches for prioritizing candidate disease genes based on PPI networks. *Tsinghua Science and Technology*, *20*(5), 500–512.

Landrum, M. J., & Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. *Human Mutation*, *39*(11), 1623–1630.

Laulederkind, S. J. F., Hayman, G. T., Wang, S.-J., Smith, J. R., Petri, V., Hoffman, M. J., … Shimoyama, M. (2018). A Primer for the Rat Genome Database (RGD). *Methods in Molecular Biology*, *1757*, 163–209.

Lee, J.-Y., Park, M. K., Park, J.-H., Lee, H. J., Shin, D. H., Kang, Y., … Kong, G. (2014). Loss of the polycomb protein Mel-18 enhances the epithelial-mesenchymal transition by ZEB1 and ZEB2 expression through the downregulation of miR-205 in breast cancer. *Oncogene*, *33*(10), 1325–1335.

Lehne, B., & Schlitt, T. (2009). Protein-protein interaction databases: Keeping up with growing interactomes. *Human Genomics*, *3*(3), 291.

Lei, Z., & Dai, Y. (2006). Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, *7*(1), 491.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., … Kendziorski, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, *29*(8), 1035–1043.

Li, M. J., Liu, Z., Wang, P., Wong, M. P., Nelson, M. R., Kocher, J.-P. A., … Wang, J. (2016). GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Research*, *44*(D1), D869-876.

Li, X., Wu, M., Kwoh, C.-K., & Ng, S.-K. (2010). Computational approaches for detecting protein complexes from protein interaction networks: A survey. *BMC Genomics*, *11*(1), S3.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., … Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, *40*(D1), D857–D861.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304.

Liu, J., Jing, L., & Tu, X. (2016). Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC Cardiovascular Disorders*, *16*(1), 54.

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics*, *19*(10), 1275–1283.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901.

MacMullen, C., Fang, J., Hsu, B. Y., Kelly, A., de Lonlay-Debeney, P., Saudubray, J. M., … Hyperinsulinism/hyperammonemia Contributing Investigators. (2001). Hyperinsulinism/hyperammonemia syndrome in children with regulatory mutations in the inhibitory guanosine triphosphate-binding domain of glutamate dehydrogenase. *The Journal of Clinical Endocrinology and Metabolism*, *86*(4), 1782–1787.

Magger, O., Waldman, Y. Y., Ruppin, E., & Sharan, R. (2012). Enhancing the Prioritization of Disease-Causing Genes through Tissue Specific Protein Interaction Networks. *PLOS Computational Biology*, *8*(9), e1002690.

Makani, J., Menzel, S., Nkya, S., Cox, S. E., Drasar, E., Soka, D., … Thein, S. L. (2011). Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood*, *117*(4), 1390–1392.

Mathivanan, S., Periaswamy, B., Gandhi, T., Kandasamy, K., Suresh, S., Mohmood, R., … Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, *7*(5), S19.

Matsunaga, E., Kimura, A., Yamada, H., Fukumaki, Y., & Takagi, Y. (1985). A novel deletion in delta beta-thalassemia found in Japan. *Biochemical and Biophysical Research Communications*, *126*(1), 185–191.

McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics (Oxford, England)*, *25*(6), 765–771.

Meissner, T., Mayatepek, E., Kinner, M., & Santer, R. (2004). Urinary alpha-ketoglutarate is elevated in patients with hyperinsulinism-hyperammonemia syndrome. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *341*(1–2), 23–26.

Mendis, S., Puska, P., Norrving, B., & World Health Organization. (2011). *Global atlas on cardiovascular disease prevention and control*. Retrieved March 15, 2019, from https://who.int/cardiovascular_diseases/publications/atlas_cvd/en/

Moreau, Y., & Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nature Reviews Genetics*, *13*(8), 523–536.

National Cholesterol Education Programme/Adult Treatment Panel III. (2002). *Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult*

Treatment Panel III) final report. Retrieved March 15, 2019, from https://www.nhlbi.nih.gov/files/docs/resources/heart/atp-3-cholesterol-full-report.pdf

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, *9*(5), 471–472.

Ogris, C., Guala, D., Kaduk, M., & Sonnhammer, E. L. L. (2018). FunCoup 4: New species, data, and visualization. *Nucleic Acids Research*, *46*(D1), D601–D607.

O'Neill, S., & O'Driscoll, L. (2015). Metabolic syndrome: A closer look at the growing epidemic and its associated pathologies. *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity*, *16*(1), 1–12.

Pacholewska, A. (2017). "Loget"—A Uniform Differential Expression Unit to Replace "logFC" and "log2FC." *Matters*, *3*(10), e201706000011.

Park, J. H., Lee, J. Y., Shin, D. H., Jang, K. S., Kim, H. J., & Kong, G. (2011). Loss of Mel-18 induces tumor angiogenesis through enhancing the activity and expression of HIF-1α mediated by the PTEN/PI3K/Akt pathway. *Oncogene*, *30*(45), 4578–4589.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., … Wolfinger, R. D. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology*, *24*(9), 1140–1150.

Peart, M. J., Smyth, G. K., Laar, R. K. van, Bowtell, D. D., Richon, V. M., Marks, P. A., … Johnstone, R. W. (2005). Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proceedings of the National Academy of Sciences*, *102*(10), 3697–3702.

Peck, R., & Devore, J. L. (2011). *Statistics: The Exploration & Analysis of Data*. Cengage Learning.

Peng, J., & Singh, M. (2010). *SPICi: Speed and Performance In Clustering*. Retrieved August 20, 2019, from https://compbio.cs.princeton.edu/spici/

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., & Couto, F. M. (2008). Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics*, *9*(Suppl 5), S4.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., … Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839.

Pizzuti, C., & Rombo, S. E. (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, *30*(10), 1343–1352.

Plaitakis, A., Latsoudis, H., Kanavouras, K., Ritz, B., Bronstein, J. M., Skoula, I., … Spanaki, C. (2010). Gain-of-function variant in GLUD2 glutamate dehydrogenase modifies Parkinson's disease onset. *European Journal of Human Genetics: EJHG*, *18*(3), 336–341.

Plaitakis, A., Latsoudis, H., & Spanaki, C. (2011). The human GLUD2 glutamate dehydrogenase and its regulation in health and disease. *Neurochemistry International*, *59*(4), 495–509.

Plaitakis, A., Zaganas, I., & Spanaki, C. (2013). Deregulation of glutamate dehydrogenase in human neurologic disorders. *Journal of Neuroscience Research*, *91*(8), 1007–1017.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). DISEASES: Text mining and data integration of disease–gene associations. *Methods*, *74*, 83–89.

Podobnik, M., Kraševec, N., Zavec, A. B., Naneh, O., Flašker, A., Caserman, S., … Anderluh, G. (2016). How to study protein-protein interactions. *Acta Chimica Slovenica*, *63*(3), 424–439.

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., … Horvath, S. (2008). Integrated Weighted Gene Co-expression Network

Analysis with an Application to Chronic Fatigue Syndrome. *BMC Systems Biology*, *2*(1), 95.

Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., & Mooney, S. D. (2008). An integrated approach to inferring gene–disease associations in humans. *Proteins: Structure, Function, and Bioinformatics*, *72*(3), 1030–1037.

Rando, O. J. (2012). Daddy Issues: Paternal Effects on Phenotype. *Cell*, *151*(4), 702–708.

Raouf, A., Zhao, Y., To, K., Stingl, J., Delaney, A., Barbara, M., … Eaves, C. (2008). Transcriptome Analysis of the Normal Human Mammary Cell Commitment and Differentiation Process. *Cell Stem Cell*, *3*(1), 109–118.

Reaven, G. M. (1988). Role of Insulin Resistance in Human Disease. *Diabetes*, *37*(12), 1595–1607.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., … Watson, M. S. (2015). ClinGen—The Clinical Genome Resource. *The New England Journal of Medicine*, *372*(23), 2235–2242.

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:Profiler—A web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(suppl_2), W193–W200.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Riis, M. L. H., Lüders, T., Nesbakken, A.-J., Vollan, H. S., Kristensen, V., & Bukholm, I. R. K. (2010). Expression of BMI-1 and Mel-18 in breast tissue—A diagnostic marker in patients with breast cancer. *BMC Cancer*, *10*, 686.

Rivals, I., Personnaz, L., Taing, L., & Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics*, *23*(4), 401–407.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Santer, R., Kinner, M., Passarge, M., Superti-Furga, A., Mayatepek, E., Meissner, T., … Schaub, J. (2001). Novel missense mutations outside the allosteric domain of glutamate dehydrogenase are prevalent in European patients with the congenital hyperinsulinism-hyperammonemia syndrome. *Human Genetics*, *108*(1), 66–71.

Satuluri, V., & Parthasarathy, S. (2009). Scalable Graph Clustering Using Stochastic Flows: Applications to Community Discovery. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 737–746.

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, *7*(1), 302.

Schuster-Böckler, B., & Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biology*, *9*(1), R9.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., … Rubio, A. (2005). Correlation Between Gene Expression and GO Semantic Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(4), 330–338.

Shi, X., Zhan, L., Xiao, C., Lei, Z., Yang, H., Wang, L., … Zhang, H.-T. (2015). MiR-1238 inhibits cell proliferation by targeting LHX2 in non-small cell lung cancer. *Oncotarget*, *6*(22), 19043–19054.

Shih, Y.-K., & Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics*, *28*(18), i473–i479.

Shu, L., Chan, K. H. K., Zhang, G., Huan, T., Kurt, Z., Zhao, Y., … Yang, X. (2017). Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLOS Genetics*, *13*(9), 1–25.

Skov, V., Knudsen, S., Olesen, M., Hansen, M. L., & Rasmussen, L. M. (2012). Global gene expression profiling displays a network of dysregulated genes in non-atherosclerotic arterial tissue from patients with type 2 diabetes. *Cardiovascular Diabetology*, *11*, 15.

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J., & Mouse Genome Database Group. (2018). Mouse Genome Database (MGD)-2018: Knowledgebase for the laboratory mouse. *Nucleic Acids Research*, *46*(D1), D836–D842.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 397–420.

Song, J., & Singh, M. (2009). How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, *25*(23), 3143–3150.

Sperling, S., Grimm, C. H., Dunkel, I., Mebus, S., Sperling, H.-P., Ebner, A., … Hammer, S. (2005). Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Human Mutation*, *26*(6), 575–582.

Stanley, C. A., Lieu, Y. K., Hsu, B. Y., Burlina, A. B., Greenberg, C. R., Hopwood, N. J., … Poncz, M. (1998). Hyperinsulinism and hyperammonemia in infants with regulatory mutations of the glutamate dehydrogenase gene. *The New England Journal of Medicine*, *338*(19), 1352–1357.

Stanley, Charles A. (2011). Two genetic forms of hyperinsulinemic hypoglycemia caused by dysregulation of glutamate dehydrogenase. *Neurochemistry International*, *59*(4), 465–472.

Su, D., Li, Q., Guan, L., Gao, X., Zhang, H., Dandan, E., … Ma, X. (2013). Down-regulation of EBAF in the heart with ventricular septal defects and its regulation by histone acetyltransferase p300 and transcription factors smad2 and cited2. *Biochimica Et Biophysica Acta*, *1832*(12), 2145–2152.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., & Mesirov, J. P. (2007). GSEA-P: A desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, *23*(23), 3251–3253.

Sun, W., Kim, K.-H., Noh, M., Hong, S., Huh, P. W., Kim, Y., & Kim, H. (2006). Induction of CITED2 expression in the rat hippocampus following transient global ischemia. *Brain Research*, *1072*(1), 15–18.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. von. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., … von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, *45*(D1), D362–D368.

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., … Lopez-Bigas, N. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, *10*(1), 25.

Tao, Y., Sam, L., Li, J., Friedman, C., & Lussier, Y. A. (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, *23*(13), i529–i538.

Tenekeci, S., & Isik, Z. (2018). Gene co-expression network analysis to identify shared disease genes in metabolic syndrome, type 2 diabetes, and coronary artery disease. *The International Symposium on Health Informatics and Bioinformatics*.

The Gene Ontology Consortium. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research*, *36*(Database issue), D440–D444.

Thomas, J. G., Olson, J. M., Tapscott, S. J., & Zhao, L. P. (2001). An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, *11*(7), 1227–1236.

Tran, C., Konstantopoulou, V., Mecjia, M., Perlman, K., Mercimek-Mahmutoglu, S., & Kronick, J. B. (2015). Hyperinsulinemic hypoglycemia: Think of hyperinsulinism/hyperammonemia (HI/HA) syndrome caused by mutations in the GLUD1 gene. *Journal of Pediatric Endocrinology & Metabolism: JPEM*, *28*(7–8), 873–876.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., & Wodak, S. J. (2011). Interaction databases on the same page. *Nature Biotechnology*, *29*, 391–393.

UniProt Consortium, T. (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *46*(5), 2699.

Vives-Corrons, J. L., Pujades, M. A., Miguel-García, A., Miguel-Sosa, A., & Cambiazzo, S. (1992). Rapid detection of Spanish (delta beta)zero-thalassemia deletion by polymerase chain reaction. *Blood*, *80*(6), 1582–1585.

Wang, H., Azuaje, F., Bodenreider, O., & Dopazo, J. (2004). Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, *2004*, 25–31.

Wang, J., Li, M., Chen, J., & Pan, Y. (2011). A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(3), 607–620.

Wang, J., Li, M., Deng, Y., & Pan, Y. (2010). Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, *11*(Suppl 3), S10.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274–1281.

Wang, X., Gulbahce, N., & Yu, H. (2011). Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, *10*(5), 280–293.

Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E., & Cornel, M. C. (2008). Orphanet: A European database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, *152*(9), 518–519.

Wilson, P. W. F., D'Agostino, R. B., Parise, H., Sullivan, L., & Meigs, J. B. (2005). Metabolic Syndrome as a Precursor of Cardiovascular Disease and Type 2 Diabetes Mellitus. *Circulation*, *112*(20), 3066–3072.

Wolting, C., McGlade, C. J., & Tritchler, D. (2006). Cluster analysis of protein array results via similarity of Gene Ontology annotation. *BMC Bioinformatics*, *7*(1), 338.

World Health Organization. (1999). *Definition, diagnosis and classification of diabetes mellitus and its complications: Report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus*. Retrieved March 15, 2019, from https://apps.who.int/iris/handle/10665/66040

Wu, C., Zhu, J., & Zhang, X. (2012). Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, *13*(1), 182.

Wu, X., Pang, E., Lin, K., & Pei, Z.-M. (2013). Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method. *PLoS ONE*, *8*(5).

Xu, T., Du, L., & Zhou, Y. (2008). Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, *9*(1), 472.

Yang, C.-S., Chang, K.-Y., Dang, J., & Rana, T. M. (2016). Polycomb Group Protein Pcgf6 Acts as a Master Regulator to Maintain Embryonic Stem Cell Identity. *Scientific Reports*, *6*, 26899.

Yang, F., Wu, D., Lin, L., Yang, J., Yang, T., & Zhao, J. (2017). The integration of weighted gene association networks based on information entropy. *PLOS ONE*, *12*(12), e0190029.

Yin, Z., Haynie, J., Yang, X., Han, B., Kiatchoosakun, S., Restivo, J., … Yang, Y.-C. (2002). The essential role of Cited2, a negative regulator for HIF-1alpha, in heart development and neurulation. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(16), 10488–10493.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, *26*(7), 976–978.

Zhang, J. W., Stamatoyannopoulos, G., & Anagnou, N. P. (1988). Laotian (delta beta) (0)-thalassemia: Molecular characterization of a novel deletion associated with increased production of fetal hemoglobin. *Blood*, *72*(3), 983–988.

Zhang, L. V., Wong, S. L., King, O. D., & Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, *5*(1), 38.

Zhang, S., Ning, X., & Zhang, X.-S. (2006). Identification of functional modules in a PPI network by clique percolation clustering. *Computational Biology and Chemistry*, *30*(6), 445–451.

Zhang, X.-F., Dai, D.-Q., Ou-Yang, L., & Yan, H. (2014). Detecting overlapping protein complexes based on a generative model with functional and topological properties. *BMC Bioinformatics*, *15*(1), 186.

Zhao, Wei, Rasheed, A., Tikkanen, E., Lee, J.-J., Butterworth, A. S., Howson, J. M. M., … Saleheen, D. (2017). Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics*, *49*(10), 1450–1457.

Zhao, Wukui, Tong, H., Huang, Y., Yan, Y., Teng, H., Xia, Y., … Qin, J. (2017). Essential Role for Polycomb Group Protein Pcgf6 in Embryonic Stem Cell

Maintenance and a Noncanonical Polycomb Repressive Complex 1 (PRC1) Integrity. *The Journal of Biological Chemistry*, *292*(7), 2773–2784.

Zimmet, P. Z., Shaw, J. E., & Alberti, K. G. M. M. (2005). Mainstreaming the metabolic syndrome: A definitive definition. *The Medical Journal of Australia*, *183*(4), 175–176.

Zivony-Elboum, Y., Westbroek, W., Kfir, N., Savitzki, D., Shoval, Y., Bloom, A., … Falik-Zaccai, T. C. (2012). A founder mutation in Vps37A causes autosomal recessive complex hereditary spastic paraparesis. *Journal of Medical Genetics*, *49*(7), 462–472.

APPENDICES

APPENDIX 1: LIST OF ACRONYMS

ACC	Accuracy
AOA	Aortic aneurysm
ASCVD	Atherosclerotic cardiovascular disease
BHI	Biological homogeneity index
BMA	Best-match average
BP	Biological process
BSI	Biological stability index
CAD	Coronary artery disease
CC	Cellular component
CHD	Coronary heart disease
CQS	Clustering quality score
CTRL	Control
CVD	Cardiovascular disease
DAG	Directed acyclic graph
DEG	Differentially expressed gene
DIAB	Diabetes
EGIR	European Group for the Study of Insulin Resistance
ESD	Endocrine system disease
FC	Fold-change
FDR	False discovery rate
FPR	False positive rate
FRAC	Fraction
GDA	Gene-disease association
GDM	Gestational diabetes mellitus
GEO	Gene Expression Omnibus
GO	Gene Ontology

HDN	Human disease network
HEEBO	Human exonic evidence-based oligonucleotide
HPO	Human Phenotype Ontology
IC	Information content
IDF	International Diabetes Federation
LHGDN	Literature-derived Human Gene-Disease Network
logFC	Logarithmic fold-change
MF	Molecular function
MICA	The most informative information ancestor
MMR	Maximum matching ratio
MS	Metabolic syndrome
NB	Negative binomial
NCBI	National Center for Biotechnology Information
NMD	Nutritional and metabolic diseases
PAD	Peripheral arterial disease
PAN	Protein association network
PPI	Protein-protein interaction
PPIN	Protein-protein interaction network
PR	Precision-recall score
RA	Rheumatoid arthritis
RAS	Renal artery stenosis
T1D	Type-1 diabetes mellitus
T2D	Type-2 diabetes mellitus
TPR	True positive rate
VDA	Variant-disease association
WGAN	Weighted Gene Association Network
WGCN	Weighted Gene Co-expression Network
WGCNA	Weighted Gene Co-expression Network Analysis
WHO	World Health Organization