Identification of Differential Gene Pathways in Microarray Data

A Dissertation Presented

by

Qiao Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Department of Applied Mathematics and Statistics

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Qiao Zhang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Song Wu - Dissertation Advisor
Assistant Professor, Department of Applied Mathematics and Statistics

Wei Zhu - Chairperson of Defense
Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Jie Yang - Dissertation Co-Advisor
Assistant Professor, Department of Preventive Medicine

Jian Cao - Outside Member
Associate Professor of Research Medicine, Stony Brook University Medical Center

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Identification of Differential Gene Pathways in Microarray Data

by

Qiao Zhang

Doctor of Philosophy

in

Department of Applied Mathematics and Statistics

Stony Brook University

2014

A gene pathway typically refers to a group of genes and small molecules that work together to control one or more cell functions. In systems biology, pathway analysis is of paramount biological importance, and recent studies revealed that malfunction of gene pathways could induce disease manifestations, such as cancer. Usually, a gene pathway consists of two components: the upstream factors, which are signaling molecules transmitting stimulus from cell surface to nucleus, and the downstream factors, which respond to cell signaling through changes of their expression levels. Although several methods have been reported for analysis of gene pathways, almost all of them focus on the upstream factors of a pathway, ignoring the rich information from the downstream factors.

In this thesis work, we first investigated and compared the existing gene pathway analysis methods, particularly on three most popular ones: Gene Set Enrichment Analysis (GSEA), Principal Component Analysis (PCA), and Canonical Discriminant Analysis (CDA). We then proposed an innovative method based on the concept of integrating the statistical information from both upstream and downstream factors to infer differential gene pathways. More specifically, the Relax Intersection-Union Test (RIUT) framework was employed to combine evidences from upstream and downstream factors.

We performed intensive simulation studies with GSEA, PCA and CDA. We found out both the limitations and strengths of these methods under various data structures, and we identified scenarios in which each method can outperform the others. Furthermore, we demonstrated that our proposed combining method outperforms the above existing methods in terms of both power and interpretability in biology.

We applied the combining method to two real data sets: the p53 data set and *Essential thrombocythaemia* data set. The results suggest that in the combining method, GSEA is more appropriate for the upstream subgroup and CDA is more powerful for the downstream subgroup due to their distinct data structures.

Table of Contents

Chapter 1 Biology Fundamentals 1

1.1 The Central Dogma 1

1.2 Gene Expression and Regulation 2

1.3 Signal Transduction Pathway 4

1.4 Transcriptional Factors 8

1.5 DNA Microarray 11

1.6 Database 13

Chapter 2 Review of Previous Studies on Gene Set Analysis 19

2.1 Two Types of Hypotheses 19

2.2 Overview of Existing Gene Set Analysis 22

2.2.1 Over-Representation Analysis 23

2.2.2 Functional Class Scoring (FCS) 23

2.2.2.1 Model-based Method 26

2.2.2.2 Non-model-based Method 28

2.2.3 Pathway Topology-based Approaches 30

2.3 Permutation Test 31

2.4 False Discovery Rate 33

Chapter 3 Existing Statistical Methods on Gene Set Analysis 38

3.1 Gene Set Enrichment Analysis 38

3.1.1 Introduction 38

3.1.2 Mathematical Description of GSEA 39

3.1.3 Limitation 42

3.2 Gene Pathway Analysis by Principal Component Analysis 43

3.2.1 Introduction 43

3.2.2 Eigenvalue Decomposition (Johnson & Wichern, 2007) 44

3.2.3 Singular Value Decomposition (Berrar, 2002) 46

3.2.4 Application in Gene Pathway Analysis 48

3.2.5 Limitation 50

3.3 Canonical Discriminant Analysis/MANOVA 51

3.3.1 Introduction 51

3.3.2 Mathematical Implementation 52
3.3.3 Comparison with MANOVA/Hotelling’s T Test 53
3.3.4 Singular Matrix Problem 55
3.4 Intersection-Union Test and Union-Intersection Test 57
3.4.1 Definition 57
3.4.2 Method 59
3.4.2.1 Fisher’s Method 60
3.4.3 Satterwhite’s Approximation for Dependent Hypotheses 61
3.4.4 Z-transform Method 63
3.4.5 Minimum P Value and Maximum P Value 64
3.4.5.1 Methods to Compute Maximum P Value 66
Chapter 4 The Combined Method 72
4.1 Main Goal and Significance of the Study 72
4.2 Method 74
4.3 Simulations 75
4.3.1 Simulation for PCA 75
4.3.2 Simulation for GSEA 77
4.3.3 Study of CDA 84
4.3.4 Comparison of PCA, GSEA and CDA 85
4.3.5 Combining Upstream and Downstream Factors 90
4.3.6 Summary of the Simulation Studies 93
Chapter 5 Real Data Analysis 95
5.1 Materials 95
5.2 Database 96
5.3 Upstream Factors and Downstream Factors 96
5.4 Real Data Analysis 108
5.4.1 P53 Data Set 108
5.4.2 ET Data Set 111
5.5 Conclusion 113
Chapter 6 Discussion and Future Work 115

List of Figures

Figure 1-1: The central dogma of molecular biology. 1

Figure 1-2: Structure of gene. 2

Figure 1-3: Signal transduction pathway (Gene VIII). 5

Figure 1-4: JAK/STAT pathway. 6

Figure 1-5: Activator stimulates basal transcriptional complex. 9

Figure 1-6: Hybridization of the targets to probes in microarray 11

Figure 1-7: The database structure of TRANSFAC. 15

Figure 1-8: A summary of BRCA1 transcription factor in JASPAR CORE database. 18

Figure 2-1: Overview of 3 generations of gene set analysis: ORA, FCS and PT. 22

Figure 2-2: Overview of the strategies for existing FCS approaches. 25

Figure 2-3: A pathway with topology. 31

Figure 2-4: A schematic diagram for permutation test based on subject sampling. 32

Figure 2-5: A density histogram of 3170 p values from Hedenfalk et al. data. 36

Figure 3-1: Enrichment score (Aravind Subramaniana, 2005). 40

Figure 3-2: Asymmetry of GSEA results due to unbalanced global phenotype expression and gene set collection bias. (Aravind Subramaniana, 2005) 41

Figure 3-3: Flowchart of gene pathway analysis by PCA. 49

Figure 3-4: Histogram of maximum p value based on 1000 simulations. 68

Figure 4-1: Example of complex gene pathway. 74

Figure 4-2: Flowchart of our proposed combined method of gene set analysis. 74

Figure 4-3: Result of simulations comparing PCA under two different correlation structures. ... 76

Figure 4-4: Simulation results of GSEA on gene set with 10% of DEGs (a) and 50% of DEGs (b). 80

Figure 4-5: Simulation results of GSEA on gene set with 10% of DEGs (a, c) and 50% of DEGs (b, d). 81

Figure 4-6: Simulation results of GSEA and Quadratic GSEA on gene sets of both up- and down-regulated genes 83

Figure 4-7: Simulation results of CDA on gene set with 10% of DEGs (a) and 50% of DEGs (b). 84

Figure 4-8: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with 10% DEGs. 86

Figure 4-9: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with 50% DEGs. 88

Figure 4-10: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with mixture of up-regulated and down-regulated DEGs. 89

Figure 5-1: A density histogram of the p values from p53 dataset (a) and ET dataset (b). 98

Figure 5-2: Gene set size of p53 dataset and ET dataset 100

Figure 5-3: Mean pairwise correlation for each gene set in p53 dataset and ET dataset. 101

Figure 5-4: Mean absolute pairwise correlation for each gene set in p53 dataset and ET dataset. 102

Figure 5-5: Percent of DEGs ($p < 0.1$) in p53 dataset and ET dataset. 103

Figure 5-6: Percent of up-regulated genes out of DEGs in p53 dataset and ET dataset. 104

Figure 5-7: Mean absolute difference of DEGs in p53 dataset and ET dataset. 105

Figure 5-8: Maximum absolute difference of DEGs in p53 dataset and ET dataset. 106

Figure 5-9: Mean absolute t score of DEGs in p53 dataset and ET dataset................................ 107

Figure 5-10: Maximum absolute t score of DEGs in p53 dataset and ET dataset. 108

List of Tables

Table 1-1: Summary of the content and growth of the JASPAR database 16

Table 1-2: Summary of content of JASPAR CORE database updated in 2013 17

Table 2-1: Number of errors committed when testing m hypothesis. 34

Table 4-1: Parameters for simulations of illustrating the limitation of PCA. 75

Table 4-2: Parameters of simulations for GSEA and CDA. 79

Table 4-3: Simulation results of power in RIUT and overall test. 91

Table 4-4: Simulation results of power in RIUT in upstream factors with 10% DEGs. 92

Table 4-5: Simulation results of power in RIUT in upstream factors with 50% DEGs. 93

Table 5-1: Summary of p53 dataset and ET dataset. 96

Table 5-2: Summary of gene set databases: C2 and JASPAR. 96

Table 5-3: Summary of upstream and downstream gene sets in p53 and ET. 97

Table 5-4: The number of gene sets with p value <0.01, 0.05 and q value<0.25 in p53 data set. 109

Table 5-5: Top ten ranked gene sets by GSEA/CDA method in p53 data set. 110

Table 5-6: Top ten ranked gene sets from overall CDA test in p53 data set. 111

Table 5-7: The number of gene sets with p value <0.01, 0.05 and q value<0.25 in ET data set. 112

Table 5-8: Top ten ranked gene sets by GSEA/CDA method in ET data set. 113

Table 5-9: Top ten ranked gene sets from overall CDA test in ET data set. 113

Acknowledgments

I would never have been able to finish my dissertation without the guidance of my committee members, support from my family, and help from group members.

I would like to thank my committee members, who kindly helped me through my PhD study. Especially I would like to express my gratitude to Professor Yang, my dissertation co-advisor and supervisor of my research assistant job. Professor Yang provided me with financial support throughout my PhD study and guided me with knowledge and practical skills beyond the text books. My ability of handling real world data was built up on Professor Yang’s patient and kind trainings.

I would like to thank my husband, Jianfei Wang, for love and support through my study. Especially after our daughter was born, my husband did his best to relieve my obligation of housework and soothe me whenever I felt depressed. Without him, I would not be able to finish it. I would also like to thank my parents, who always believe in me and love me. Their love and support are light shredding into the occasional dark moments of my life.

Finally, I would like to thank all the members in our group, especially the girls who worked with me in BCC. Thank you for listening to my complaints and offering all kinds of excellent ideas and suggestions. Especially Jiawen, you were my best friend for the past three years, and will be my beloved friend for good. Thank Hao for his technical support in R programming. You guys gave me a sweet and precious memory and made me feel young again in my early 30s. How lucky I met all of you!

Chapter 1 Biology Fundamentals

Many diseases have complex genetic causes. While some are monogenic (caused by defects in only one gene, e.g. Huntington's disease), most common diseases are polygenic (under the influence of multiple genes, e.g. cancer, diabetes, and heart diseases, etc.). The reason is that in cells genes do not act by themselves, but rather interact with each other through their RNAs and protein expression products and assemble into functional networks. These networks of genes are so-called *signal transduction pathways*, which are working modules of a cell system regulated by development and environment stimuli. Malfunction of the signal transduction pathways may induce disease manifestations, such as cancers. In this chapter, we introduce essential concepts of gene expression, transcription factor, signal transduction pathway, and microarray.

1.1 The Central Dogma

The central dogma of molecular biology is a framework for understanding the relationship among DNA, RNA and proteins (Crick, 1970). In general, it can be summarized as DNA makes RNA that in turns makes proteins (Figure 1-1).

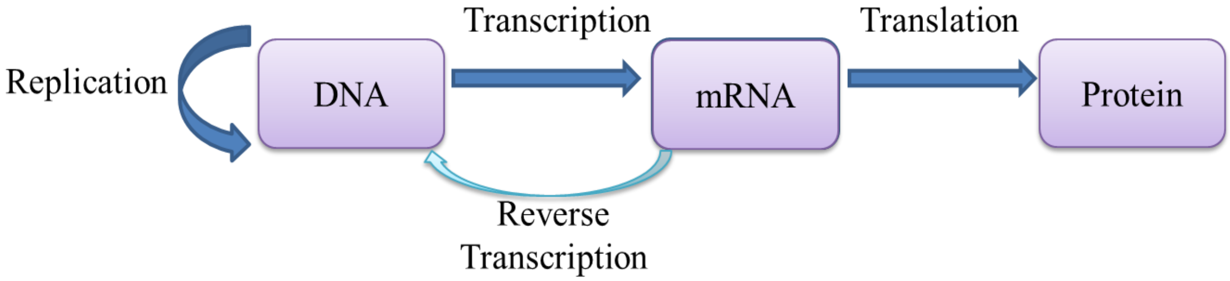


Figure 1-1: The central dogma of molecular biology. Transcription is the first step of gene expression, and the consequent change in the concentration level of messenger RNA (mRNA) is detectable by modern biotechnology such as microarray.

The process that the information contained in a section of DNA is transferred to messenger RNA (mRNA) is called transcription. In eukaryotes, transcription is performed by an enzyme called RNA polymerase in the assistant of promoter and transcription factors (details in 1.2 and 1.4, respectively). The mRNA will encounter a series of modifications and splicing, and be translated into proteins. In most organisms, this process is irreversible; however, in retrovirus (e.g. HIV), mRNA can be reversely transcribed back to DNA, known as complementary DNA for it is complementary copy of mRNA, by reverse transcriptase. This technique is now widely used in expression microarray (details in 1.5).

1.2 Gene Expression and Regulation

A gene is a segment of DNA specifying production of a polypeptide chain (Lewin, 2004); it occupies a specific location on a chromosome and determines a particular characteristic in an organism. A typical eukaryotic protein-coding gene includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons) (Figure 1-2).

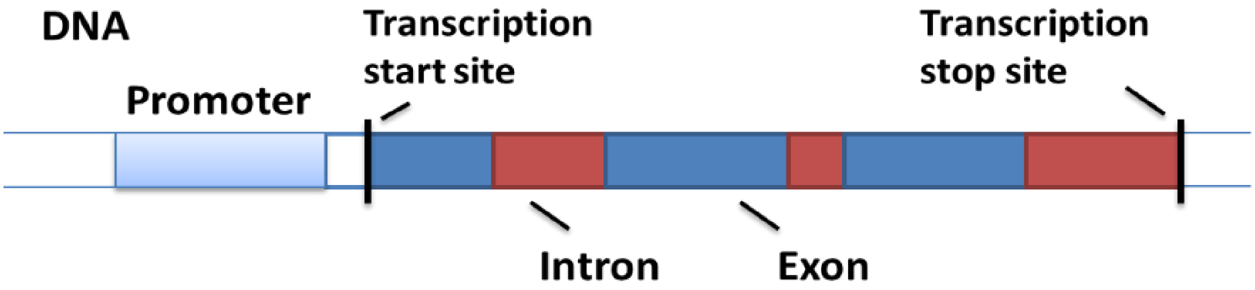


Figure 1-2: Structure of gene. This is the structure of typical Eukaryotic gene, including uncoding promoter region, and coding region with introns and exons.

In eukaryotes, the transcription process is very complicated, and it contains important protein-DNA binding procedures. One essential DNA element for initiating transcription is promoter. Eukaryotic promoters are regions of DNAs that are found typically at -30, -75, and -90 base pairs upstream from the *transcription start site* and can facilitate the transcription of the gene. Promoters contain specific DNA sequences and response elements that provide a secure initial binding site for necessary proteins. For transcription to take place, the enzyme that synthesizes RNA, known as RNA polymerase, must attach to promoter. However, eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Instead, a collection of proteins called *transcription factors* recruit RNA polymerase and mediate the initiation of transcription.

Regulation of gene expression is essential for organisms as it increases the versatility and adaptability of an organism by controlling of the amount and timing of appearance of the functional product of a gene, such as proteins, RNAs and etc. Furthermore, in human beings and other multicellular organisms, gene regulation drives the processes of cell differentiation and morphogenesis.

Gene expression can be regulated at several levels: transcriptional regulation, post-transcriptional regulation, translational regulation, and post-translational regulation. Transcriptional regulation is one of the most important and widely-used strategies, which is usually mediated through activation or inhibition of signal transduction pathway (details in 1.3). Change of activity in transcription factors is an essential mechanism in gene expression regulation (details in 1.4). Transcriptional regulation usually triggers dramatic change in gene expression level, which can be detected by modern biotechnology like microarray, RNA-Seq and etc.

1.3 Signal Transduction Pathway

Signal transduction pathway describes a group of proteins and small molecules that work together to control one or more cell functions, such as cell division or cell death, when extracellular signaling molecules are present. After the first molecule in a pathway receives a signal, it activates another downstream molecule. This process is repeated until the last molecule is activated, which is usually a transcription factor. Then the transcription of target genes will be activated and the corresponding cell function is carried out. Abnormal activation of signal pathways can lead to various diseases including cancer. Therefore, identifying differentially expressed signal pathways in a certain disease helps understand the mechanism of the disease and provides potential targets for drug development.

A signal transduction pathway typically consists of receptors, intermediate enzymes, and effectors (Figure 1-3). Receptors can be categorized into two types: extracellular receptors and intracellular receptors. Extracellular receptors are most common ones, and are transmembrane proteins embedded in the cell membrane with one part of the receptor on the outside of the cell and the other in the inside. When stimulus occurs, certain messenger molecules called ligand will bind to the outside part of receptor and induce conformational change in the inside part. Different receptors only bind to their own specific ligands and stimulate particular signal pathways. A receptor typically is activated by forming a dimer upon ligand binding (Figure 1-3).

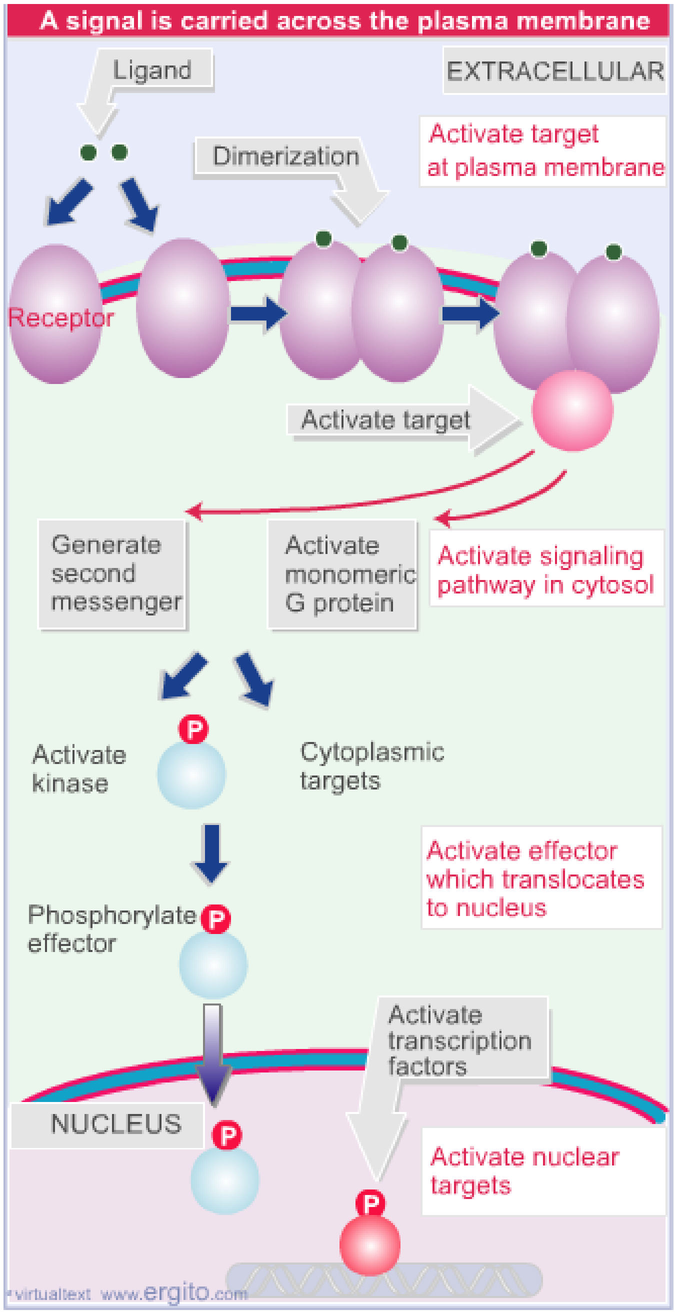


Figure 1-3: Signal transduction pathway (Gene VIII). This is a cascade of signal pathway, including receptors in the cell membrane, kinases, transcription factors, and target genes.

Activated receptors will interact and in turn activate downstream proteins in the cytosol to carry on the signals. A common means to propagate the signal pathway through the cytosol is to activate a series of protein kinases. These enzymes are able to modify other proteins by adding chemical groups, like phosphate. Ultimately a stimulus signal leads to the activation of effectors, either in cytosol or in the nucleus. The effectors that carry the signal into the nucleus have ultimate goal to activate transcription factors and thereby alter the target gene expression level. Transcription factors are a set of proteins that can bind to specific DNA sequence of certain genes, and regulate the transcription level of these genes (details in 1.4). In this way, a signal pathway is turned on. The reverse process can be fulfilled in any element of the pathway to turn off the signal pathway when signals are no longer enriched. For example, kinases will be inactivated by losing phosphate groups or pathway can be blocked by inhibitors.

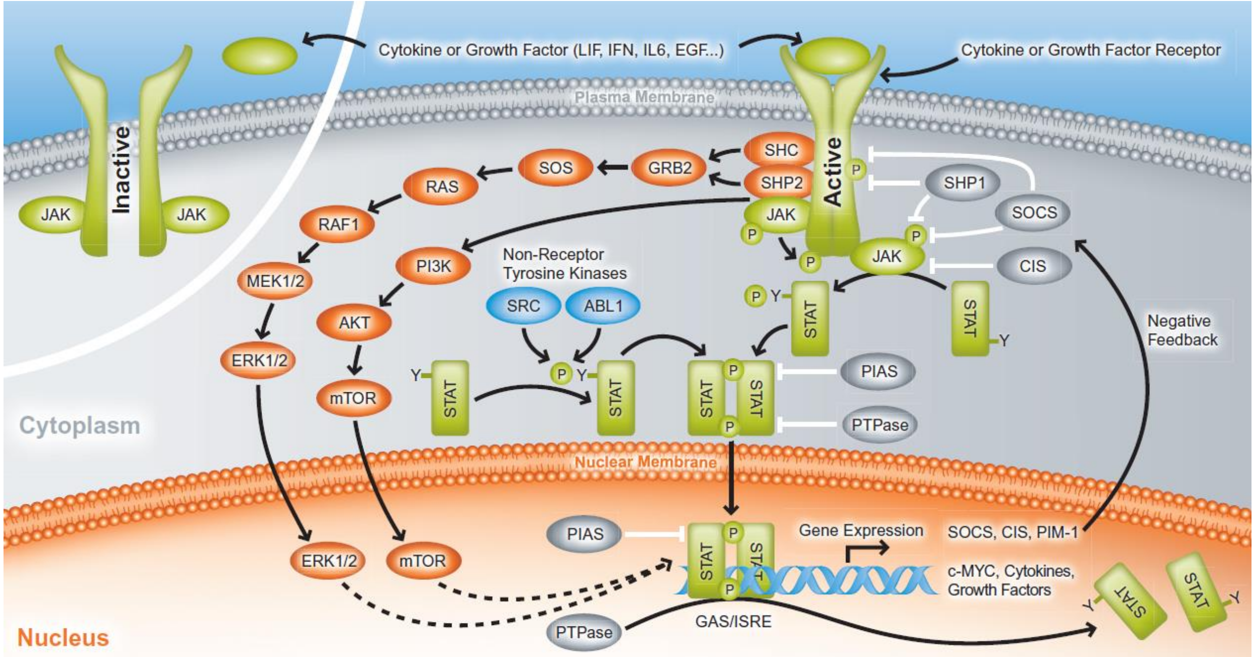


Figure 1-4: JAK/STAT pathway. The green components are in the *JAK/STAT* pathway. The orange components are in the *PI3K/AKT/mTOR* pathway. (http://docs.abcam.com/pdf/stemcells/JAK-STAT-pathway.pdf)

Signal transduction pathways are essential for organisms to react to various stimuli by sensing environments at cellular levels. One principal signaling pathway that stimulates cell proliferation, cell differentiation, cell migration and apoptosis is the JAK/STAT signaling pathway. Abnormal constitutive activation of *JAK/STAT* pathways has been implicated in various cancers and immune disorders. The core components of the pathway are receptor, JAK kinase, and STAT transcriptional factor (green components in Figure 1-4). JAK kinases bind to the receptor in the inside cell region; upon ligand binding, the receptor can be dimerized and bring JAKs close to each other. In this case, JAKs phosphorylates each other and thereby stimulates transcriptional factor STAT. The activated STAT will be transported into nucleus and activate or repress the transcription of several target genes: *SOCS, Nmi, Bcl-XL, p21, MYC, NOS2*, and etc. Although the mechanism of *JAK/STAT* signaling is relatively simple in theory, the biological consequences of pathway activation are complicated by interactions with other signaling pathways (Rawlings, 2004). For example, *JAK/STAT* is always cooperating with *RTK/Ras/MAPK* pathway and *PI3K/AKT/mTOR* pathway (orange components in Figure 1-4).

Differentially expressed pathways refer to signal pathways that are specifically activated in particular types of cells or certain diseases, but not in normal controls. Activation of these pathways is usually accompanied by increased translational level and/or protein level modification. However, these modifications may not necessarily result from dramatic increase at mRNA level. For example, Nrf2 is a transcription factor that induces the expression of genes encoding for antioxidant enzymes. Under normal or unstressed conditions, Nrf2 is tethered by another protein and degraded in cytoplasm. Nrf2 has a half-life of only 20 minutes under normal conditions (Kobayashi, et al., 2004). Consequently, although Nrf2 is transcribed all the time, its protein remains at a low level. Upon stress, Nrf2 will be built-up due to block of degradation system. In this way, the Nrf2 pathway will be activated without obvious increase of NRF2 gene expression.

Another scenario is that proteins are composed at certain levels but remain inactive. When signals are aggregated, proteins will receive modification and thereby activate the signal pathway they are involved. As described earlier, receptors can be activated by dimerization, and kinases are stimulated by phosphorylation. All these modifications occur on the pre-existing proteins, and may not induce large quantity of de novo synthesis of proteins.

Therefore, it is reasonable to separate a signal pathway into two parts. “Upstream factors” refer to the pathway factors from receptors to transcription factors, which may or may not result in significant change at transcription level of the corresponding genes. “Downstream factors” are meant for those target genes recognized and regulated by transcription factors, which should show evident alteration at mRNA level whenever activation or inhibition of transcription is initiated. For example, receptors, JAK kinases and STAT transcriptional factors would be considered as “upstream factors”, while target genes, such as SOCS, Nmi, Bcl-XL, p21, MYC, NOS2, would be considered as “downstream factors”.

1.4 Transcriptional Factors

A *transcription factor* is a protein that binds to specific DNA sequences (*promoter*), thereby controlling the transcription of genetic information from DNA to mRNA. One defining feature of transcription factors is that they contain one or more *DNA-binding domains*, which attach to the promoters of the genes they regulate (Mitchell & Tjian, 1989). These DNA-binding domains provide transcription factors with up to 10^6-fold higher affinity for their target sequences than for the remainder of the DNA strand. In fact, transcription factors are frequently classified on the basis of their DNA binding domains (Latchman, 1997).

Transcription factors are able to influence the rate of transcription of target genes in two opposite ways: activation and repression. One type of important transcription factors are general transcription factors (e.g. TFIIB, TFIID), which interact with RNA polymerase directly and form a basal transcriptional complex. This transcriptional complex is essential for transcription to occur. At the same time, many transcription factors, known as *activators*, contain specific regions which are necessary for the activation of transcription. These activators interact with the basal transcriptional complex through their specific regions, and stimulate transcription (Figure 1-5).

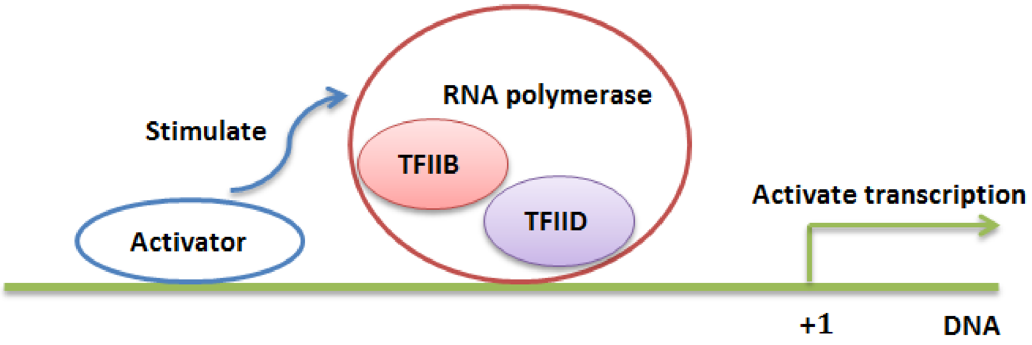


Figure 1-5: Activator stimulates basal transcriptional complex. TFIIB and TFIID are core transcription factors that form a complex with RNA polymerase to initiate transcription of target gene in the help of activators.

Although most of the transcription factors act by stimulating transcription, a variety of factors act by inhibiting the transcription of specific genes. This can be achieved, for example, by occupying certain promoter regions and preventing binding of other activators.

Hence, the balance between transcription activators and transcription repressors will determine the transcription rate of specific genes. Upon different stimuli or in various cell types, the balance changes to alter the transcription levels. The mechanism refers to regulation of transcription factor itself.

It is common in biology that important process receives multiple layers of regulation. Not only do transcription factors regulate target genes, the factors themselves receive regulation. There are two levels of regulations: regulation of synthesis and regulation of activity. Regulation of synthesis refers to transcription factors being synthesized in one particular tissue or cell type but not the others. This is an important and straightforward control mechanism. However, regulation of activity is more popular in response to a particular stimulus and is a more delicate method.

Many transcription factors are synthesized to certain levels and remain inactive in unstipulated cells. Upon stimulus, these pre-existing transcription factors can be activated via a number of different ways, including ligand binding, protein-protein interaction and post-translation modifications. In this mechanism, there may not be dramatic increase in the expression of these transcription factors since the major regulation is performed post-transnationally.

Transcription factor is always an important component in signal pathway. One transcription factor usually regulates several target genes. For example, in the JAK/STAT pathway, transcription factor STAT can recognize the promoters of many genes: SOCS, Nmi, Bcl-XL, p21, MYC, NOS2, and etc. A transcription factor may be able to regulate hundreds or thousands of target geens, however, it is not the case that transcriptions of all the target genes will be stimulated at the same time; instead, a transcription factor usually functions with other co-factors and selectively turns on certain targets based on various conditions. For example, the DNA strands are maintained in a complicated supercoiling structure so that promoters may be concealed and protected. Enzymes that remove the methyl group on the DNA and help reveal the promoter sequences are essential to activate transcription. Co-activators that form a scaffold to stabilize transcription factors are also critical in the initiation of transcription. Therefore, which target genes are to be transcribed are determined by both stimulus signals and also the cell context.

1.5 DNA Microarray

DNA microarrays, also known as DNA chips, are tools that can simultaneously measure expression levels of a large number of genes, by the identifying and quantifying of corresponding mRNA transcripts in cells (Schena, Shalon, Davis, & Brown, 1995). Most popular microarrays include cDNA microarrays and oligonucleotide microarrays. The core principle behind microarrays is the hybridization between two DNA strands (Figure 1-6).

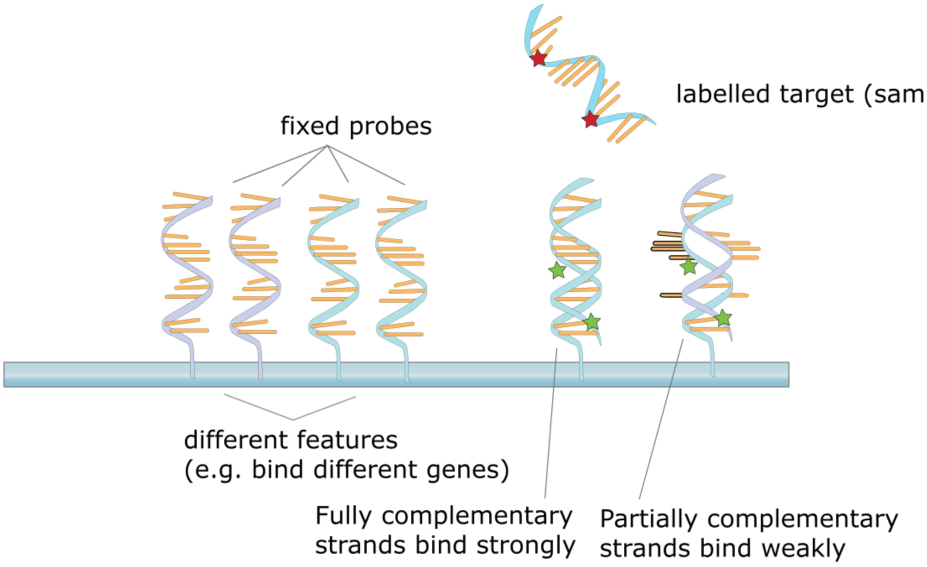


Figure 1-6: Hybridization of the targets to probes in microarray.
(http://en.wikipedia.org/wiki/DNA_microarray)

A typical microarray application is to look for differentially expressed genes (DEGs) between two different conditions (e.g. cancer cells versus normal cells). For this purpose, mRNA from two different biological samples is reversely transcribed back to complimentary DNA (cDNA), and labeled with fluorescent dyes. There are two types of cDNA microarray: one-color and two-color. In the chip, there are thousands or tens of thousands of DNA spots. Each spot contains a short section of a gene, known as *probe*, which is fixed in the surface of the chip. Each probe represents one gene, and is complementary to cDNA. Upon hybridization, cDNAs will bind to their specific probes on the same chip, by hydrogen bond between complementary nucleotide base pairs. In one-color microarray, each sample is hybridized to one chip; while in two-color microarray, cDNAs from two different conditions are hybridized to one chip and compete for probes. After washing off the non-specific bonding sequences, fluorescent signals will be collected and processed. More intense fluorescent signals typically represent higher expression level of a specific gene. Two-color microarray is measuring the ratio of expression levels between two conditions; while one-color microarray is measuring the absolute expression level. In this study, we would be focusing on single-channel microarray. Several popular single-channel systems include: Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant" (DNA_microarray).

While gene expression microarrays are powerful, variability arising the high-throughout measurement process can obscure biological signals of interest (Parmigiani, Garett, Irizarry, & Zeger, 2003). Variability may result from five different phases of data acquisition: microarray manufacturing, preparation of mRNA from biological samples, hybridization, scanning, and imaging. Therefore, a series of preprocessing are required before statisticians can perform any analysis. The pixel intensities obtained by the image scanning are thought of as the raw data. Image analysis is performed to summarize the pixel-level data, followed by quality control and normalization. For most visualization, background subtraction, logarithmic transformations, within-array normalization and across-array normalization will be applied to the data. Ultimately, the data will be stored in an expression matrix where each row represents one gene and each column represents one observation.

For DEG analysis, several statistical methods have been developed, such as t tests, Significance Analysis of Microarray (SAM) method (Tusher, Tibshirani, & Chu, 2001), empirical Bayesian method (Efron & Tibshirani, Empirical bayes methods and false discovery rates for microarrays., 2002), and etc. After obtaining a list of DEGs, further approaches are needed to group or classify samples and genes: hierarchical and K-means clustering, principal component analysis, self-organizing maps, and etc.

However, there are several limitations for these approaches:

(1) There can be thousands or tens of thousands of genes in one microarray, after correcting for multiple hypothesis testing, few genes can reach the statistical significance due to the modest change of gene expression relative to noise.

(2) Even though there is a list of statistically significant genes, it is difficult to interpret the mechanism behind the change since there is no unifying biological theme.

(3) Different statistical procedures produce lists of significant genes with little overlap.

Instead of searching for individual gene, pathway analysis using microarray data is one promising approach to increase power and became popular in the past ten years. Further details will be introduced in Chapter 2.

1.6 Database

With accumulation of biological knowledge over the past decades, many databases about physiological pathways and transcription factor target genes have been developed to facilitate the research community. Here we will review some widely used ones that can be potentially used in this research: particularly, Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), C2 databases for upstream factors; TRANSFAC and JASPAR for downstream factors.

KEGG database, the Kyoto Encyclopedia of Genes and Genomes, was initiated by the Japanese human genome program in 1995. It is a collection of online databases dealing with genomes, pathways, and biological chemicals (Kanehisa, Goto, Kawashima, Okuno, & Hattori, 2004). There are nine main databases: KEGG Pathway, KEGG Genes, KEGG Ligands, KEGG Disease, KEGG Brite, KEGG Module, KEGG Drug, KEGG Orthology, and KEGG Genome. KEGG Pathway is a collection of manually drawn pathway maps, including Metabolism Pathways, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases, and Drug Development. KEGG map includes edge and node information for signaling pathway which is an essential source for network analysis on gene pathways emerging in the previous five years.

GO database, Gene Ontology, is a bioinformatics aiming to unify the representation of gene and gene product attributes across all species (Consortium, 2008). There is no universal standard terminology in biology and related fields, and term usages may vary due to different species, research areas or even different research groups. This makes communication and sharing of data more difficult. The Gene Ontology project provides ontology of defined terms representing gene product properties.

C2 database was released by the authors of the famous GSEA paper (Subramanian, et al., 2005), and it was curated from various source of gene sets. In their 2005 publication, there were 522 gene sets in C2 database, 472 sets containing genes in metabolic and signaling pathways while 50 sets containing genes involved in response to genetic and chemical perturbations (Subramanian, et al., 2005).

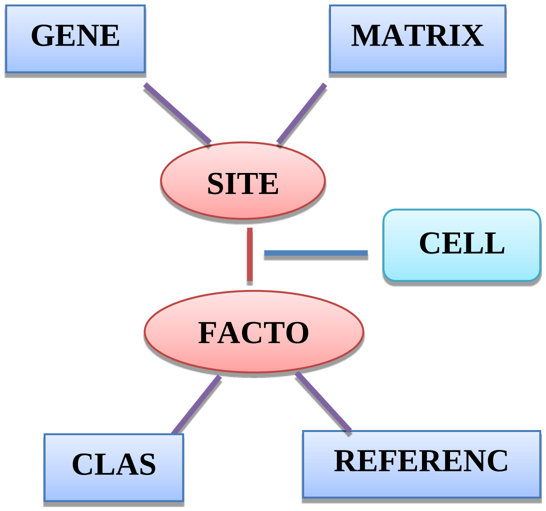
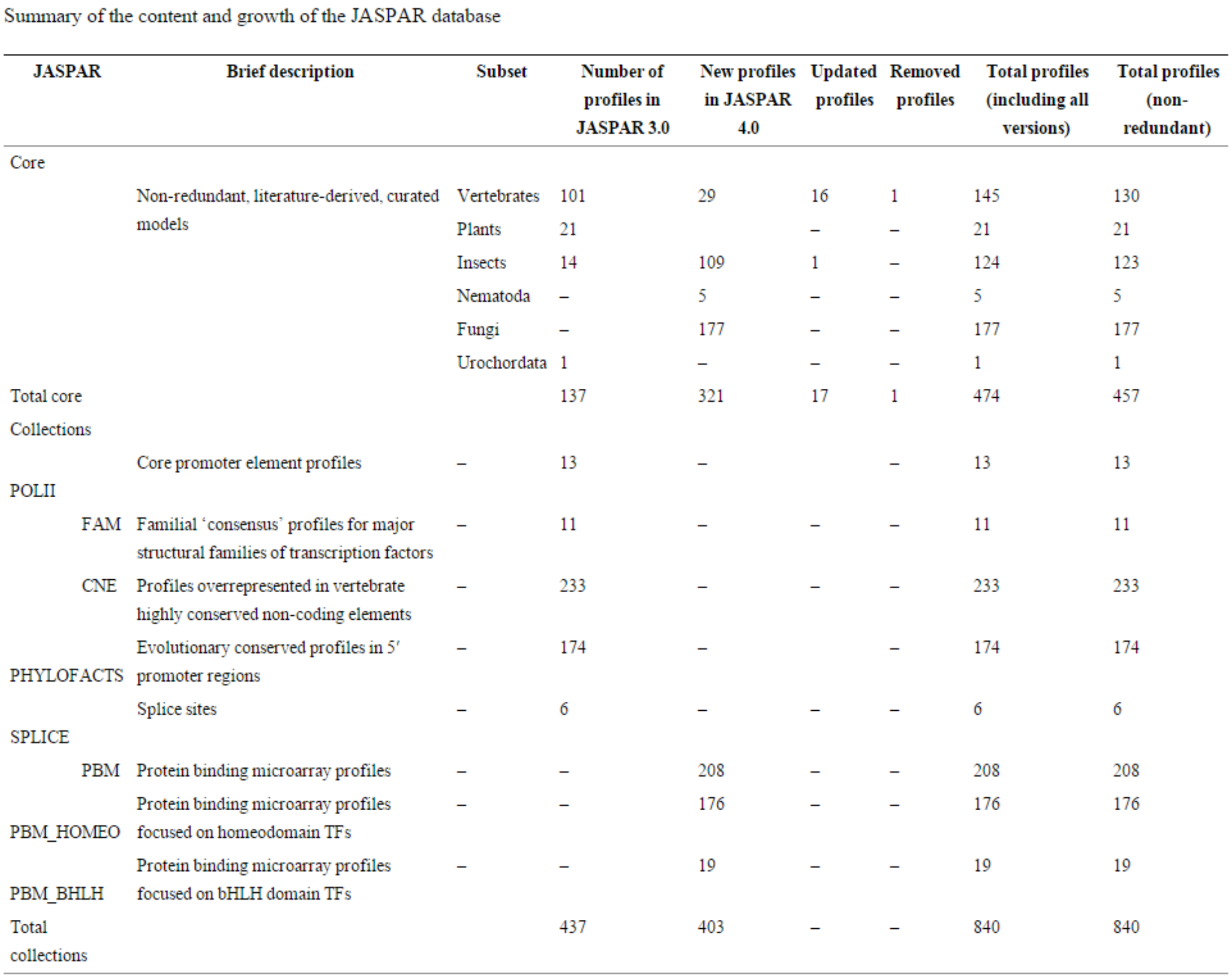


Figure 1-7: The database structure of TRANSFAC. (http://www.edgar-wingender.de/TRANSFAC.html)

TRANSFAC is a database about transcription factors and the specific DNA sequence which transcription factors bind to and regulate through in eukaryotes. This database was first compiled and released by Wingender et al (Wingender, 1988). After development over a decade, the database updated in 2012 contains information about 18,614 transcription factors (TRANSFAC Statistics). Important tables in TRANSFAC database are FACTOR, SITE, GENE, CELL, CLASS, MATRIX, and REFERECNE. FACTOR describes the transcription factors, while SITE gives information of transcription factor binding sites in eukaryotes. The central axis between FACTOR and SITE represents DNA-protein interaction (Figure 1-7). GENE gives a short explanation of the gene where a site belongs to. MATRIX gives nucleotide distribution matrices for binding sites of transcription factors. CELL gives information about the cellular source of proteins that have been shown to interact with the sites. CLASS contains some background information about the transcription factor classes. REFERENCE gives information about studies that have been done on the proteins.

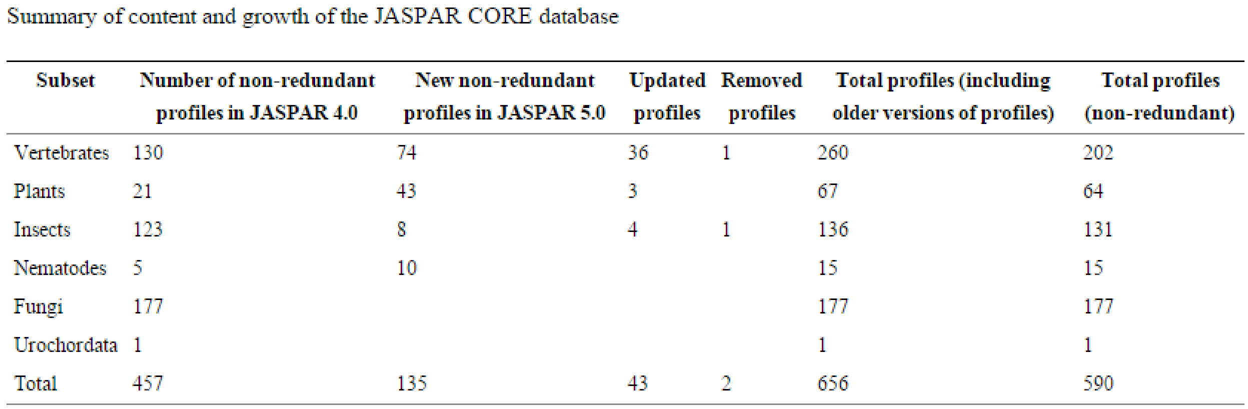


Summary of the content and growth of the JASPAR database

JASPAR	Brief description	Subset	Number of profiles in JASPAR 3.0	New profiles in JASPAR 4.0	Updated profiles	Removed profiles	Total profiles (including all versions)	Total profiles (non-redundant)
Core								
	Non-redundant, literature-derived, curated models	Vertebrates	101	29	16	1	145	130
		Plants	21		–	–	21	21
		Insects	14	109	1	–	124	123
		Nematoda	–	5	–	–	5	5
		Fungi	–	177	–	–	177	177
		Urochordata	1	–	–	–	1	1
Total core			137	321	17	1	474	457
Collections								
POLII	Core promoter element profiles	–	13	–		–	13	13
FAM	Familial 'consensus' profiles for major structural families of transcription factors	–	11	–	–	–	11	11
CNE	Profiles overrepresented in vertebrate highly conserved non-coding elements	–	233	–	–	–	233	233
PHYLOFACTS	Evolutionary conserved profiles in 5′ promoter regions	–	174	–		–	174	174
SPLICE	Splice sites	–	6	–	–	–	6	6
PBM	Protein binding microarray profiles	–	–	208	–	–	208	208
PBM_HOMEO	Protein binding microarray profiles focused on homeodomain TFs	–	–	176	–	–	176	176
PBM_BHLH	Protein binding microarray profiles focused on bHLH domain TFs	–	–	19	–	–	19	19
Total collections			437	403	–	–	840	840

Table 1-1: Summary of the content and growth of the JASPAR database (E, et al., 2010).

JASPAR is a popular open-access database for matrix models of transcription factors binding sites. In the earlier releases, the binding sites were determined by SELEX experiments (Pollock & Treisman, 1990), or by the data from experimentally confirmed binding regions of actual promoter regions (Sandelin, Alkema, Engström, Wasserman, & Lenhard, 2004). With the development of high-throughput techniques, the latest version of JASPAR database was expanded substantially in November 2013 (Mathelier, et al., 2013). Different collections and growth of JASPAR database is shown in Table 1-1, which was published in 2010. The most popular and essential module in JASPAR database is JASPAR CORE, where the matrix models of binding sites are recorded. The summary of newest version of JASPAR CORE is shown in Table 1-2.

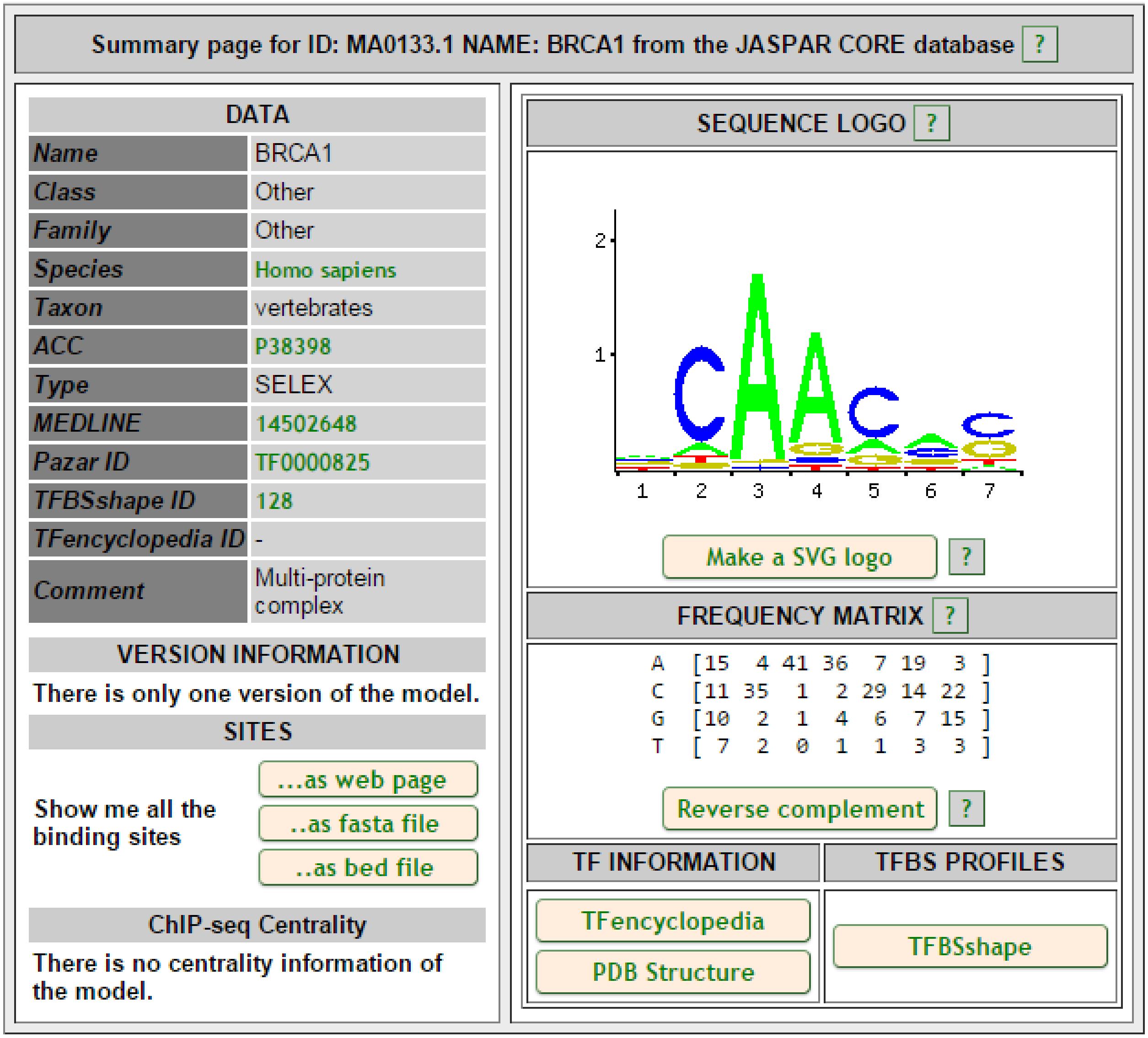


Summary of content and growth of the JASPAR CORE database

Subset	Number of non-redundant profiles in JASPAR 4.0	New non-redundant profiles in JASPAR 5.0	Updated profiles	Removed profiles	Total profiles (including older versions of profiles)	Total profiles (non-redundant)
Vertebrates	130	74	36	1	260	202
Plants	21	43	3		67	64
Insects	123	8	4	1	136	131
Nematodes	5	10			15	15
Fungi	177				177	177
Urochordata	1				1	1
Total	457	135	43	2	656	590

Table 1-2: Summary of content of JASPAR CORE database updated in 2013 (Mathelier, et al., 2013).

Based on a set of known transcription factor binding sites (TFBSs) for a given transcription factor, a position frequency matrix (PFM) or a position weighted matrix (PWM) derived from PFM is used to represent the binding preference (Sandelin, Alkema, Engström, Wasserman, & Lenhard, 2004). Example is given in Figure 1-10, where the PFM for BRCA1 binding sites is shown. A set of potential target genes could be identified by matching the promoter sequences to the PFM or PWM for each transcription factor in JASPAR CORE. Essential packages and instructions could be downloaded from Bioconductor (http://www.bioconductor.org/help/workflows/gene-regulation-tfbs/) for R environment.



Summary page for ID: MA0133.1 NAME: BRCA1 from the JASPAR CORE database ?

DATA	
Name	BRCA1
Class	Other
Family	Other
Species	Homo sapiens
Taxon	vertebrates
ACC	P38398
Type	SELEX
MEDLINE	14502648
Pazar ID	TF0000825
TFBSshape ID	128
TFencyclopedia ID	-
Comment	Multi-protein complex

VERSION INFORMATION

There is only one version of the model.

SITES

Show me all the binding sites

ChIP-seq Centrality

There is no centrality information of the model.

SEQUENCE LOGO ?

FREQUENCY MATRIX ?

```
A  [15  4 41 36  7 19  3 ]
C  [11 35  1  2 29 14 22 ]
G  [10  2  1  4  6  7 15 ]
T  [ 7  2  0  1  1  3  3 ]
```

TF INFORMATION

TFBS PROFILES

Figure 1-8: A summary of BRCA1 transcription factor in JASPAR CORE database. (http://jaspar.genereg.net/)

These databases are very important, as gene set enrichment analyses, which will be described in the next chapter, were mainly based on pre-determined gene sets from them. Due to the rapid development of biotechnology, these databases are still growing every day.

Chapter 2 Review of Previous Studies on Gene Set Analysis

Recent biomedical studies have suggested that diseases such as cancer are associated with differential expressions of multiple genes with coordinated biological functions, which are referred to “pathways”. Approaches to detect differential gene pathways have been developed and extensively investigated during the past ten years. There are 3 generations of gene set analysis methods: Over-Representation Analysis (ORA), Functional Class Scoring Approaches (FCS), and Pathway Topology (PT)-Based Approaches (Khatri, Sirota, & Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, 2012). In this chapter, we will first briefly review existing gene set analysis methods in each generation, and then we will introduce three most popular FCS methods in details: Gene Set Enrichment Analysis (Subramanian, et al., 2005), Principal Component Analysis (Ma & Kosorok, 2009), and Canonical Discriminant Analysis (Tsai & James, 2009).

2.1 Two Types of Hypotheses

The overall objective for gene set analysis is to test if the genes in a given gene set have coordinated association with the phenotype. Not like the study of single differentially expressed genes that has a clear definition of null hypothesis, 2 different null hypotheses (Q_1 and Q_2) were proposed and corresponding permutation methods were studied for gene pathway analysis, which are formulated as:

$$Q_1(competative): H_0: The\ genes\ in\ the\ gene\ set\ are\ at\ most\ as\ often\ differentially\ expressed$$
$$as\ the\ genes\ not\ in\ the\ gene\ set$$

$$Q_2(Self-contained): H_0: No\ genes\ in\ the\ gene\ set\ are\ differentially\ expressed$$

In this section, we will compare these 2 null hypotheses according to the notable review paper published by Goeman and Buhlmann in 2007 (Goeman & Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, 2007).

The difference between these two hypotheses is quite obvious: a competitive test compares the gene expression pattern in a gene set to the background genes, while a self-contained test only focuses on the genes within a given gene set and its result would not be affected by the genes outside the gene set. Each method shows advantage in different scenarios. For example, if we have a gene set with majority of genes having minor change in expression level, as long as their expression pattern is different from that of the background genes, competitive tests is sensitive enough to detect this difference. This is the most popular type of tests in early studies of gene set expression data, e.g. the notable GSEA is designed based on this hypothesis (details in 2.3). On the other hand, if there is only a small fraction of genes are differentially expressed in a gene set, self-contained tests would be more powerful.

Different hypotheses lead to different method of calculating p values. In competitive tests, the permutation test is designed based on a *gene sampling* method, which indicates that the association between samples and the phenotypes is fixed and genes in a given gene set are randomly picked up from the whole list under null hypothesis. From here, we can easily see the issues behind competitive tests. Firstly, the fixed relationship between gene expression data and phenotypes is not how microarray experiment is designed. If a new experiment needs to be repeated, only new samples were recruited but not new genes. Secondly, competitive tests assume independency between genes in a gene set, which is not realistic. Therefore, gene sampling is likely to have inflated power due to the lack of consideration of correlation structurer. Based on these, Goeman and Buhlmann strongly opposed to applying gene sampling in gene set analysis (Goeman & Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, 2007).

On the other hand, self-contained tests use subject as permutation unit: *subject sampling*. This design compares the association between gene set and phenotype with that of random phenotype, which matches the design for microarray experiment. This method also indicates that the gene membership in a gene set is fixed, which makes more sense in the study of biologically pre-defined gene set.

In the early age of gene set study, competitive tests are more popular, such as using hypergeometric test to compare the number of differentially expressed genes in each gene set. However, after a long-time development and discussion, newly proposed methods were more based on self-contained tests, including Globaltest (Goeman, van de Gee, de Kort, & van Houwelingen, 2003), GlobalAncova (Hummel, Meister, & Mansmann, 2008), Principal Component Analysis (Tomfohr, Lu, & Kepler, 2005), Hotelling's T test (Kong, Pu, & Park, 2006), MANOVA (Tsai & James, 2009), and etc.

We well accepted the points stated by Goeman and Buhlmann about competitive tests, and agreed that gene sampling is not a reasonable method for gene set analysis. However, we also have concerns about self-contained test as it tends to reject any gene set that has some differentially expressed genes even the particular gene set does not show any 'enrichment' compared to other gene sets. In practice, self-contained methods always reject more gene sets than competitive methods (Goeman & Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, 2007). The widely studied GSEA is a very special case of competitive tests. It was designed to identify the gene set with different expression patterns than the rest of the genes outside the gene set. However, the permutation test was based on subject sampling (details in 2.5). Therefore, this method would lose some power due to the ‘penalization’ for significant genes outside the gene set, and the discrepancy between the null hypothesis which method was built on and the null hypothesis the significance was accessed. But this method also shows some merit in certain conditions (details in 3.1) that will be introduced in details in 2.5.

2.2 Overview of Existing Gene Set Analysis

In this section, we would like to give a brief overview of existing gene set analysis methods. Regardless of the generation of methods, the common idea behind them is to seek a statistic that can well represent the change at expression level across the whole pathway, and can be used to compare among pathways to access significance. Figure 2-1 gives an overview of the three generations of gene set analysis by P. Khatri et al.

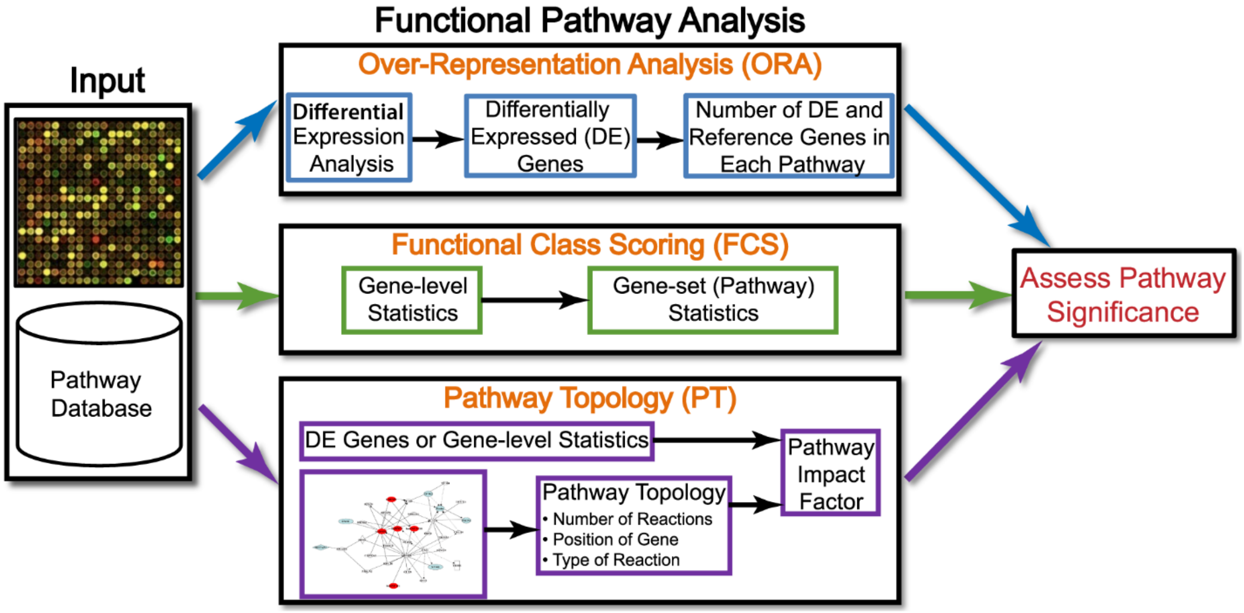

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Impact Factor
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Assess Pathway Significance

Figure 2-1: Overview of 3 generations of gene set analysis: ORA, FCS and PT. The input is genes in pre-defined gene pathway based on existing pathway database. Each method uses different gene set statistic to access pathway significance.
(Khatri, Sirota, & Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, 2012)

2.2.1 Over-Representation Analysis

Over-Representation Analysis (ORA) is the first generation of gene set analysis methods that emerged in early 21st century (Khatri, Draghici, Ostermeier, & Krawetz, 2002). This type of methods aims at identifying the pathways including such a large fraction of DEGs that can be distinguished from the pathways with randomly fallen-in DEGs. Among those published methods, common strategies were adopted: firstly, a list of DEGs are identified (either up-regulated genes or down-regulated genes or both); secondly, for each given gene pathway, the number of DEGs is counted and recorded; next, each pathway is evaluated in terms of the number of DEGs in the set by Hypergeometric Test, Chi-square Test or Binomial Distribution (Khatri, Sirota, & Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, 2012). The limitations for these methods are obvious. An arbitrary cutoff is required to define DEGs and different cutoff will lead to different results. For example, if $p\ value = 0.05$ is set as cutoff, then the genes with $p\ value = 0.051$ would be missed out in the analysis but it could contain essential information for the gene pathway as well as those with $p\ value = 0.049$. Furthermore, only the number of DEGs in the gene pathway is considered but no extent of regulation (e.g., fold-change) is made use of. Also the correlation structure between genes intra-pathway or inter-pathway is not considered at all. This loss of information would reduce the power of the test. Lastly, the biggest problem of ORA methods is that the competitive null hypothesis which they are built on is not practical in gene set study as we described in section 2.1. Therefore, this type of methods are rarely used independently nowadays.

2.2.2 Functional Class Scoring (FCS)

Methods in this class are developing enormously in the last ten years. Various assumptions and statistical methods were proposed and tested in a large amount of publications. Here we review FCS based on a paper by Marit Ackermann1 et al. in 2009 (Ackermann & Strimmer, 2009). Although this is not the most up-to-date review, the concept and framework in FCS study do not differ much from then.

An overview of the strategies for FCS is given in Figure 2-2 (Ackermann & Strimmer, 2009). The input of data is again grouped into gene sets according to gene set database. To summarize pathway level change in expression level, there are 2 types of methods: the "model-based" one that takes all the genes in a gene set into consideration at the same time, and the "non- model-based" one that is more complicated, including gene-level statistics, transformation of gene-level statistics, and gene set level statistics.

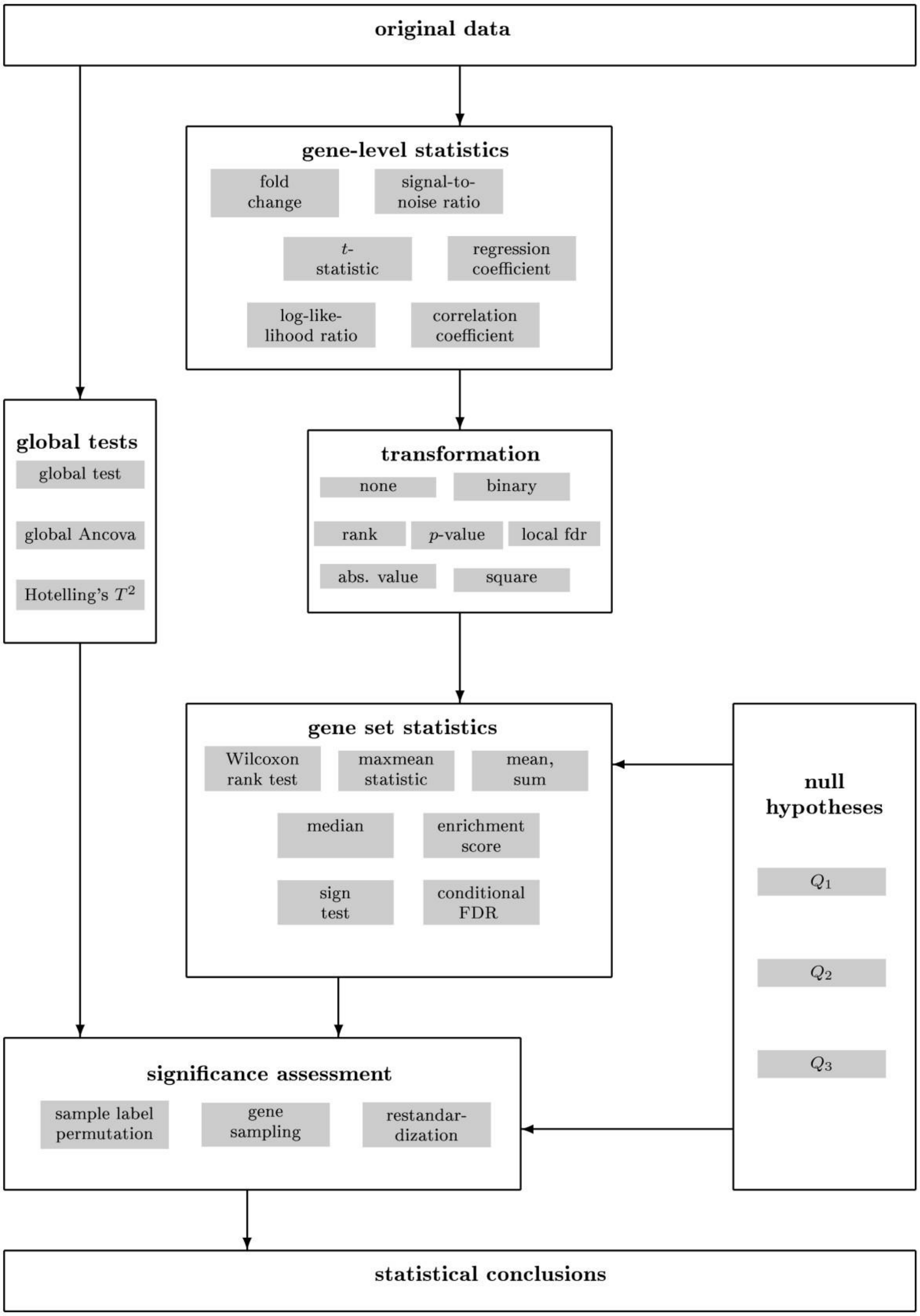


Figure 2-2: Overview of the strategies for existing FCS approaches. Input is the gene expression data of a pre-defined gene set. One type of the tests is ‘model-based’, while the other type includes 3 levels of strategies. Three different null hypotheses are considered, resulted in 3 different permutation methods.

2.2.2.1 Model-based Method

For this group, most popular ones are Globaltest (Goeman, van de Gee, de Kort, & van Houwelingen, 2003), GlobalAncova Test (Mansmann & Meister, 2005) (Hummel, Meister, & Mansmann, 2008), Principal Component Analysis (Tomfohr, Lu, & Kepler, 2005) (Ma & Kosorok, 2009), and Multivariate Analysis (Canonical Discriminant Analysis) (Kong, Pu, & Park, 2006) (Tsai & James, 2009).

Among these methods, Globaltest is constructed in a Bayesian framework. When the outcome is categorical such as disease vs. control, a logistic regression can be used to model the data in a gene set (Goeman, van de Gee, de Kort, & van Houwelingen, 2003). However, a common issue is the so-called "curse of dimensionality", as we usually have more genes than sample size in a gene set, in which the traditional logistic regression is not appropriate. Goeman et al. assumed the parameters $\beta_1, \dots, \beta_k$ are samples from some common distribution with expectation zero and variance τ^2. The null hypothesis then becomes $H_0\colon \tau^2 = 0$, and this solves the high-dimensionality problem. By using a logistic regression, Globaltest is virtually testing the predictive power of gene expression for a certain phenotype.

In 2005, Mansmann and Meister developed an ANCOVA-based method and compared it to Globaltest (Mansmann & Meister, 2005). Later on, Hummel et al. extended this method to a more general framework (Hummel, Meister, & Mansmann, 2008). They showed that linear model can also be used for gene set analysis when the role of gene expression and phenotype are switched. The advantage of GlobalAncova method is that it allows adjustment for confounders such as age, gender, and etc. It also takes correlation between genes into consideration. It may outperform Globaltest when correlations are not negligible.

Principal Component Analysis (PCA) is a popular dimension reduction method and was widely used in many fields. It was firstly introduced into gene set analysis in 2005 by Tomfohr et al. The framework for this method is that PCA is applied on the gene expression data in a given gene set, then the first principal component is picked up to represent the original data and a t-test is performed when the phenotype is binary (Tomfohr, Lu, & Kepler, 2005). This method is computational efficient and understandable, however, it has several limitations. Firstly, no phenotype information is considered when the data is transformed, and it will lead to tremendous loss of power in certain situations (details in 2.3). Secondly, only the first principal component is considered while information in other PCs is ignored. To improve this method, Ma and his colleagues extended the method to different types of phenotypes and compared the power in choosing different numbers of principal components (details in 2.3).

Multivariate Analysis (Canonical Discriminant Analysis) is another model-based method first adopted by Kong et al. in 2006. The framework they proposed is actually a combined method of principal component analysis and Hotelling's T test. Due to the high-dimensionality issue and the consequent singular correlation matrix between genes, Hotelling's T test is not applicable to gene set expression data. Therefore, Kong and his colleagues suggested firstly apply principal component analysis on the within class covariance matrix and choose the principal components corresponding to eigenvalues larger than a threshold (e.g. 10^{-4}). Then the Hotelling's T test is applied on the transformed data. Since the rank of within class covariance matrix is $\min(p, n-k)$ (p is number of genes, n is sample size, k is number of groups), the transformed data will no longer have singular within class covariance matrix. This method can successfully solve the singular matrix problem, but the power of the test is reduced at certain level due to PCA. Another multivariate model is proposed by Tsai et al. in 2009. They adopted the framework of multivariate analysis of variance (MANOVA), which is the multivariate version of ANOVA. To solve the singular matrix problem, Tsai and his colleagues suggested use a shrinkage estimate of covariance matrix (details in 2.4).

Canonical discriminant analysis (CDA) is another data-transformed method similar to PCA but it takes the class information into consideration. When the outcome is binary, CDA is essentially equivalent to MANOVA (details in 2.4).

All the above methods use the ‘self-contained’ hypothesis, which means they are testing if there is any differentially expressed gene in the gene set. Therefore, this set of tests is more appropriate to detect the gene sets with at least a small fraction of genes of large change in expression level. In both simulations and real data study, Tsai showed that MANOVA was outperforming Globaltest, GlobalAncova, PCA combined with Hotelling’s T test, and some other tests we will introduce in the next section. We suppose the authors did not compare PCA method and MANOVA method because they were published in the same volume of the same journal in 2009. Therefore, we studied the PCA and MANOVA (or CDA) in more details and chose them to be candidate methods for our proposed combined method.

2.2.2.2 Non-model-based Method

More methods have been published as ‘non-model-based’ methods because of the flexibility of choosing different statistics at each level of the study. The framework for ‘non-model-based’ methods includes three steps. Firstly, we need to access the gene expression change at gene level. The possible statistic includes fold change, signal-to-noise ratio, t statistic, correlation coefficient, regression coefficient, log-likelihood ratio and any other statistics that can measure the association between single gene expression level and phenotype. Secondly, the gene level statistics need to be transformed. Just as choosing an appropriate link function for generalized linear regression models, transformation is performed according to the property of methods to be applied on gene set level. The easiest way is to use the identity of gene level statistic, or we can use quadratic transformation, absolute value transformation, ranking, p value and etc. The last step is to summarize the gene set level statistic by various tests: sum, mean, median, Kolmogorov-Smirnov test, maxmean statistic, Wilcoxon rank test, and etc.

Ackermann and Strimmer compared 261 combinations of the methods as well as gene sampling method vs. subject sampling method for significance assessment. Since we are concerned about the validity of gene sampling method, we only focus on the results from subject sampling method. The most popular methods published are GSEA (Subramanian, et al., 2005), PAGE (Parametric Analysis of Gene Set Enrichment) (Kim & Volsky, 2005), Maxmean (Efron & Tibshirani, On testing the significance of sets of genes, 2007), SAM-GS (Dinu, et al., 2007), and etc.

All these methods differ at one of three 3 levels. For example, GSEA is using signal-to-noise statistic and a ranking transformation at gene level, and it adopts a weighted Kolmogorov-Smirnov Test at gene set level. PAGE was developed on the basis of GSEA. It proceeds by averaging over fold change or other gene level statistics, which is compared to a standard normal distribution. Maxmean statistic deals with positive scores and negative scores separately. The advantage of this method is that it facilitates the detection of gene sets with both up-regulated genes and down-regulated genes. Efron and Tibshirani also proposed a ‘restandardization’ process that combines both gene sampling and subject sampling. Dinu and his colleagues pointed out several problems of GSEA method and proposed an improved method call ‘SAM-GS’, which is a sum up of quadratic transformed SAM statistic proposed by Tusher et al in 2001 (Tusher, Tibshirani, & Chu, 2001).

All the above methods were developed based on self-contained hypothesis except for GSEA. Several groups have compared existing methods by both simulation and real data, and consistently found that Hotelling's T test (or MANOVA) outperformed other methods (including GSEA, Maxmean, SAM-GS, Globaltest, GlobalAncova, and etc. but not including PCA) in most cases (Tsai & James, 2009) (Ackermann & Strimmer, 2009). Therefore, we pick out PCA and MANOVA to represent the self-contained tests. At the same time, GSEA is the most widely used competitive method, although there are many concerns about this type of method (details in 2.1 and 3.1), it still outperforms the self-contained methods in some scenarios. We will keep it as a candidate for our proposed combined method.

2.2.3 Pathway Topology-based Approaches

The third generation of gene set analysis methods is Pathway Topology (PT)-based approaches, which tries to take into consideration of each gene according to their position in the gene set to design of the tests. For example, if we have a pathway as in Figure 2-3, the activation of protein A is weighted highest, the activation of protein B, E, and F is weighted next highest, and the activation of C and D is weighted least. The idea behind this design is that gene A has the ability to influence the activity of all the rest of the genes, which indicates a higher possibility of differentially expression of the whole pathway if gene A is differentially expressed. On the other hand, if genes C or gene D are differentially expressed, it is not guaranteed the pathway is indeed turned on because these two genes could be activated by the components from other real activated pathways. In the most popular PT-based method SPIA, this weight of position is expressed by a gene perturbation factor, which is defined as an aggregation of the perturbation effects of all genes in the gene set (Tarca, et al., 2009), and only those genes declared as DEG by certain pre-set cutoff contribute to the perturbation. For example, if the gene A in Figure 2-3 is DEG, its effect would be aggregated into B, E and F. If gene B is also DEG, then the effect of A would further pass on to gene C and D. However, if gene B is not DEG, the effect of perturbation will not be aggregated to gene C and D. The total perturbation score is the sum of the individual perturbation score of all genes.

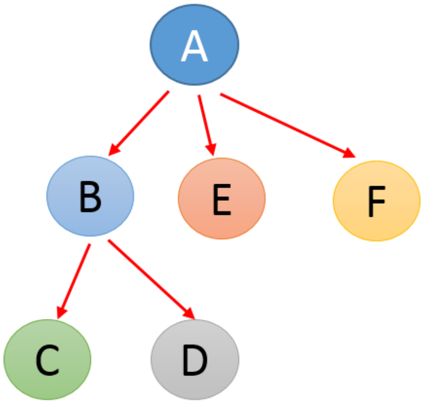


Figure 2-3: A pathway with topology. Protein A activates protein B, E, and F. Protein B further activates protein C and D.

Although all gene set analysis depends on existing database, PT-based method requires more information of the topology of the gene set. KEGG database provides topology data and is a good source for SPIA. Limited source on database is still an issue for this type of methods. Another problem we are concerned for SPIA is that it pre-selects genes before calculating perturbation score that different selection of cutoff would result in inconsistent results.

2.3 Permutation Test

Traditional statistical test requires an explicit distribution function that can be used to compute p values. With the help of current computer technology, a permutation test is usually adopted to access significance which does not require any assumption of distribution. The idea of permutation test is to estimate an empirical null distribution for computing the p values. For example, in self-contained tests, the null hypothesis is: *No gene in the gene set is differentially expressed*. If this is true, we can permute the phenotype label to break the association between genes and phenotypes, therefore, we can estimate the null distribution of self-contained test by subject sampling (Figure 2-4).

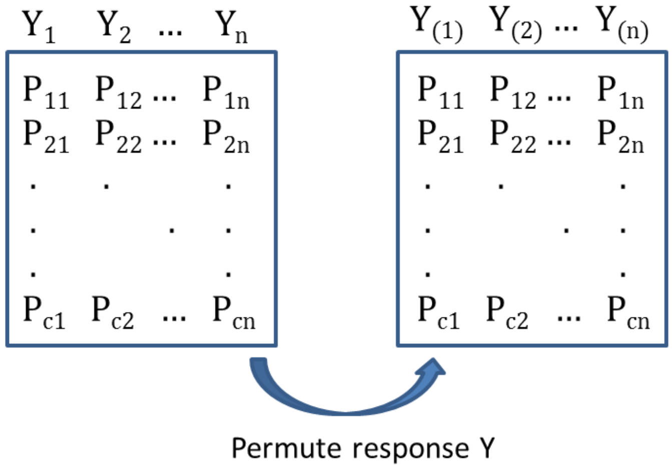

Y_1 Y_2 ... Y_n
P_{11} P_{12} ... P_{1n}
P_{21} P_{22} ... P_{2n}
P_{c1} P_{c2} ... P_{cn}
$Y_{(1)}$ $Y_{(2)}$ ... $Y_{(n)}$
P_{11} P_{12} ... P_{1n}
P_{21} P_{22} ... P_{2n}
P_{c1} P_{c2} ... P_{cn}
Permute response Y

Figure 2-4: A schematic diagram for permutation test based on subject sampling.

Suppose the statistic calculated based on the observed data is T*. Then we randomly permute the response Y, refit the model and calculate the corresponding statistic T. If we repeat permutation for B times, we could have statistic $T_1, T_2, \dots, T_B$, which can be used to estimate the empirical distribution of statistic T under null distribution. The P-value is given by

$$P - value = \frac{number\ of\ T \geq T^*}{B}$$

On the other hand, if we want to calculate p values for competitive tests where the association between gene and phenotype is fixed but the gene membership in a given gene set if random under null hypothesis, we need to resample genes instead of subject label.

2.4 False Discovery Rate

P value is a widely adopted method to summarize statistical significance. However, when there are multiple tests at the same time, the decision based on each $p - value < \alpha$ cannot guarantee to control the "family-wise error rate (FWER)" of α, which is the probability of rejecting at least one H_0 given H_0 is true for the whole family. In fact, to retain the FWER within α, the error rate for each test must be more stringent than α. Many researchers proposed various adjustment for p value when multiple comparisons exist. For example, the notable Bonferroni correction is to use α/k as the decision criterion instead of α. By Boole's inequality, it is easy to prove that Bonferroni correction can well-control FWER; however, when the large amount of comparisons are correlated , this method could be too conservative to identify any significance. This limitation is especially severe in microarray data where there are usually more than $10{,}000$ of genes, or more than 100 of gene sets. Therefore, the study of false discovery rate (FDR) became popular since 1995 (Multiple comparisons).

Benjamini and Hochberg pointed out that not only the question of whether there is an error is important, but the number of erroneous rejections should be taken into consideration in multiple comparisons. They proposed the definition of false discovery rate to be the expected proportion of errors among the rejections (Benjamini & Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995).

	Rejection	Not rejection	Total

True null hypothesis	F	$m_0 - F$	m_0
True alternative hypothesis	T	$m_1 - T$	m_1
	S	$m - S$	m

Table 2-1: Number of errors committed when testing m hypothesis. m is the total number of hypotheses tested; m_0 is the number of true null hypotheses; m_1 is the number of true alternative hypotheses; F is the number of false positives (Type I error); T is the number of true positives; m_0-F is the number of true negatives; m_1-T is the number of false negatives (Type II error); S is the number of rejected null hypotheses; m-S is the number of not rejected null hypothesis.

Table 2-1 gives the number of errors committed when testing m hypotheses. It defines some random variables related to multiple hypotheses testing.

From the definition of FDR, we can formulate it as follows:

$$FDR = E(F/S)$$

A simple Benjamini-Hochberg Procedure (Benjamini & Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995) has been described to control FDR: let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p values. Define

$$k = \max\{i: p_{(i)} \leq \frac{i}{m} q\}$$

and reject $H^0_{(1)}, H^0_{(2)}, \ldots, H^0_{(k)}$. Benjamini and Hochberg showed that when the test statistics are independent, the above procedure controls the FDR at level $q * \frac{m_0}{m} \leq q$, where q is the pre-set false discovery rate.

In 2001, Benjamini proposed another procedure that allows controlling FDR under dependency (Benjamini & Yekutieli, The control of the false discovery rate in multiple testing under dependency, 2001). The procedure became: let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, and set up an expected FDR value q. Then we will search for r that satisfies

$$r = \max\{i: p_{(i)} \leq \frac{i}{N} \times \frac{q}{C(N)}\}, where\ C(N) = \sum_{i=1}^{N} \frac{1}{i}$$

The tests corresponding to $p_{(1)} \ldots p_{(r)}$ are rejected.

This FDR method is based on the assumption that p-values from N tests are dependent, which is appropriate in pathway analysis since different pathways are likely to share common genes. However, the Benjamini-Hochberg Process requires pre-set criterion, and will need repeated calculation if the criterion changes. If we compare it to the significance assessment in univariate test, this criterion is just like the critical value.

Storey and Tibshirani proposed a method to compute $q\ value$, which is analogue to $p\ value$ in significance assessment. The difference between these two probabilities is that p value is the measure of significance in terms of false positive rate while q value is measuring false discovery rate (Storey & Tibshirani, 2003). In Table 2-1, we list all related variables where the number of rejection S can be observed but the number of null distribution m_0 and the number of false positives F are not observable. If we choose a t value where any $p - value \leq t$ indicates rejection, then the corresponding FDR is defined as

$$FDR(t) = E(\frac{F(t)}{S(t)})$$

$$\approx \frac{E\big(F(t)\big)}{E\big(S(t)\big)}$$

where $E\big(S(t)\big)$ can be simply estimated by $S(t) = \#\{p_i \leq t; i = 1,2, \dots, m\}$, which is the observed number of rejections. Given that p values follow a uniform distribution between 0 and 1 under null distribution, $E\big(F(t)\big)$ can also be estimated by $F(t) = m_0 * t$. However, m_0 is an unknown parameter that is not observable and needs to be estimated. Equivalently we can estimate the proportion of null distribution among all the tests: $\pi_0 = \frac{m_0}{m}$, where m is the total number of tests and it is already known.

In their publication, Storey and Tibshirani formulated the estimate of π_0 as

$$\widehat{\pi_0}(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)}$$

where λ is a tuning parameter with range of $[0,1)$ to accommodate a position from which the distribution of p values is approximately uniform (Storey & Tibshirani, 2003).

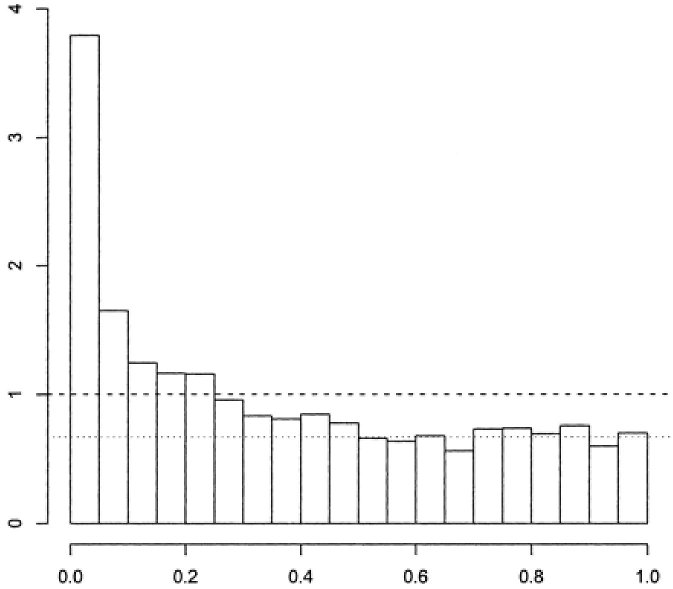


Figure 2-5: A density histogram of 3170 p values from Hedenfalk et al. data. The p values larger than 0.5 are approximately uniformly distributed. If $\lambda = 0.5$ is chosen, about 67% of the tests are under null distribution. (Storey & Tibshirani, 2003)

One example is given in Figure 2-5, which is the density histogram of 3,170 p values from a data obtained from *BRCA1*- and *BRCA2*-mutation-positive tumors (Hedenfalk, Ringner, Trent, & Borg, 2001). We can see that some p values are cumulated near 0 which indicates the existence of differentially expressed genes. The histogram becomes approximately flat when p values larger than 0.5. Therefore, it is reasonable to estimate the proportion of null distribution using $\lambda = 0.5$, which gives an estimate $\widehat{\pi_0}(0.5) \approx 0.67$. Storey and Tibshirani introduced an automated method to estimate $\widehat{\pi_0}$ by using natural cubic spline to compute $limit_{\lambda \to 1}\widehat{\pi_0}(\lambda)$ (Storey & Tibshirani, 2003).

Finally, we have

$$\widehat{FDR}(t) = \frac{\widehat{\pi_0} m * t}{\#\{p_i \leq t\}}$$

If we order the p values, then the q value calculated correspondingly will be

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\mathrm{FDR}}(\mathrm{t})$$

This process computes q value based on positive FDR $(\mathrm{pFDR}) = \mathrm{E}(\frac{\mathrm{F}}{\mathrm{S}} | \mathrm{S} > 0)$, which makes perfect sense because we only care about false discovery rate when there is a discovery. Furthermore, q value guarantees that a smaller p value shows the evidence of significance at least as strong as it is based on minimum possible pFDR. In certain sense, q value is very similar to p value as it measures the significance of the feature and helps making decision based on a chosen critical value.

Chapter 3 Existing Statistical Methods on Gene Set Analysis

In the previous chapter, we gave a brief review of the existing gene set analysis methods (Khatri, Sirota, & Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, 2012) (Ackermann & Strimmer, 2009). In this chapter, we will introduce three gene set analysis methods in details: Gene set enrichment analysis (GSEA), Principal Component Analysis (PCA), and Canonical Discriminant Analysis (CDA) (or equivalently MANOVA). We chose these three methods as candidates for our proposed combined method (Chapter 4), because these three are the representatives of non-model-based and model-based methods described in Chapter 2.

3.1 Gene Set Enrichment Analysis

3.1.1 Introduction

GSEA is a method that evaluates microarray data at the level of gene sets. This method was first proposed by Mootha et al. (Mootha, et al., 2003), and was further developed by Subramanian et al. (Subramanian, et al., 2005). Instead of identifying individual differential genes, GSEA considers a set of functionally co-regulated genes. This method takes advantage of prior biological knowledge, and typically considers experiments with genomewide expression profiles from samples belonging to two classes (e.g. cancer versus normal). Given a *priori* defined set of genes S (e.g. genes encoding products in a signal pathway), and a ranked list L with all genes ordered according to the strength of their association to a phenotype, we expect that the members of S would be found at the top or bottom of L, instead of randomly distributed throughout L, if the gene set S is correlated with the phenotypic class. The statistic used in GSEA is called enrichment score (ES) that reflects how well the set is concentrated at the extremes.

GSEA considers all genes in an experiment, not just focusing on those beyond an arbitrary cutoff. Since it uses a non-parametric statistic, there is less restriction for applying the method. Furthermore, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes.

3.1.2 Mathematical Description of GSEA

Suppose we have an expression data set D with m genes and n samples, and all these samples are from two phenotypes. Gene set S is independently derived from existing database (e.g. GO, KEGG, etc.), and contains s genes. The goal is to investigate if genes in the gene set S are differentially expressed between two groups of samples, e.g. disease and control. The general procedures of GSEA are described as follows.

(1) Ranked list: First of all, we compute the gene level statistics which measures the correlation between gene expression profiles and phenotypes (e.g. t statistics, signal-to-noise ratio, Pearson's correlation coefficient); then we transform these m statistics for all genes into an ordered list $L = \{g_{(1)}, \dots, g_{(m)}\}$.

(2) Enrichment score (ES):

ES is a gene set level statistic for the gene set S, which is designed as a weighted Kolmogorov-Smirnov-like statistic:

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, where\ N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{m-s}$$

$$ES(S) = \max_{1 \le i \le m} \{P_{hit}(S,i) - P_{miss}(S,i)\}$$

$P_{hit}(S,i)$ evaluates the cumulative probability of genes in set S, and $P_{miss}(S,i)$ evaluates the cumulative probability of genes not present in S, up to a given location i in L. Enrichment score is the maximum deviation from zero of $P_{hit} - P_{miss}$(Figure 3-1). In the formula, r_j is the statistic for gene j, which is present in S. When the power of $r_j, p = 0$, $ES(S)$ reduces to the standard Kolmogorov-Smirnov statistic; When $p = 1$, the genes in S were weighted by their gene level statistics, normalized by the sum of statistics over all the genes in S. Genes ranked before the maximum of running sum are called 'leading-edge subset', which are the members contributing most to $ES(S)$.

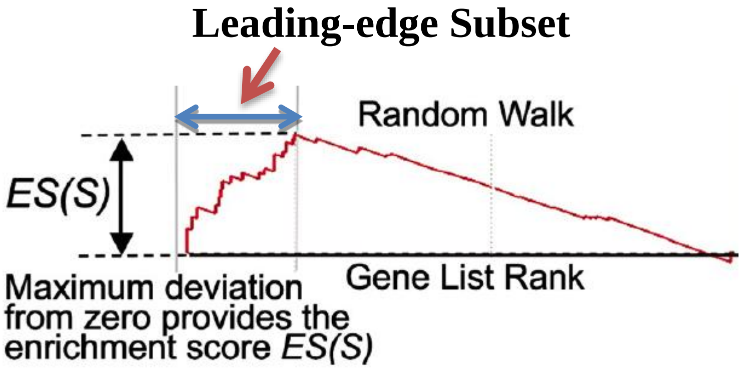


Figure 3-1: Enrichment score (Subramanian, et al., 2005). Genes in Leading-edge subset appear before the running sum reaches its maximum deviation from zero.

Remark: ES statistic is a comparison of the distribution of genes in S to that of the genes not in S. Since genes not in S are expected to be uniformly distributed, $ES(S)$ in fact tests the hypothesis of whether genes in S are randomly distributed throughout the ranked list.

The use of weighted steps can cause asymmetry in the distribution of observed ES scores. Therefore, when estimating the significance levels, we need to consider separately the positive and negative ES scores. Furthermore, as the size of a gene set does affect the enrichment score (intuitively, larger gene set size indicates bigger chance of containing highly ranked genes), a normalization by dividing the mean permutated ES is performed (Figure 3-2).

(3) Estimating significance:

Estimation of the significance level will be fulfilled by permutation tests. The original ES score is recorded as $ES(S)*$. Random permutation of the phenotype labels, reordering genes and recalculation of $ES(S)$ for each permutation will be performed. For example, this procedure may be repeated for 1000 times, and p value will be given by counting how many permutations give $ES(S) \geq ES(S)*$ if $ES(S)*$ is a positive value; or how many $ES(S) \leq ES(S)*$ if $ES(S)*$ is negative, divided by total number of permutations.

Remark: By permuting the phenotype labels, the correlation structure between genes is reserved. Therefore, GSEA is more like a combination of self-contained t and competitive tests.

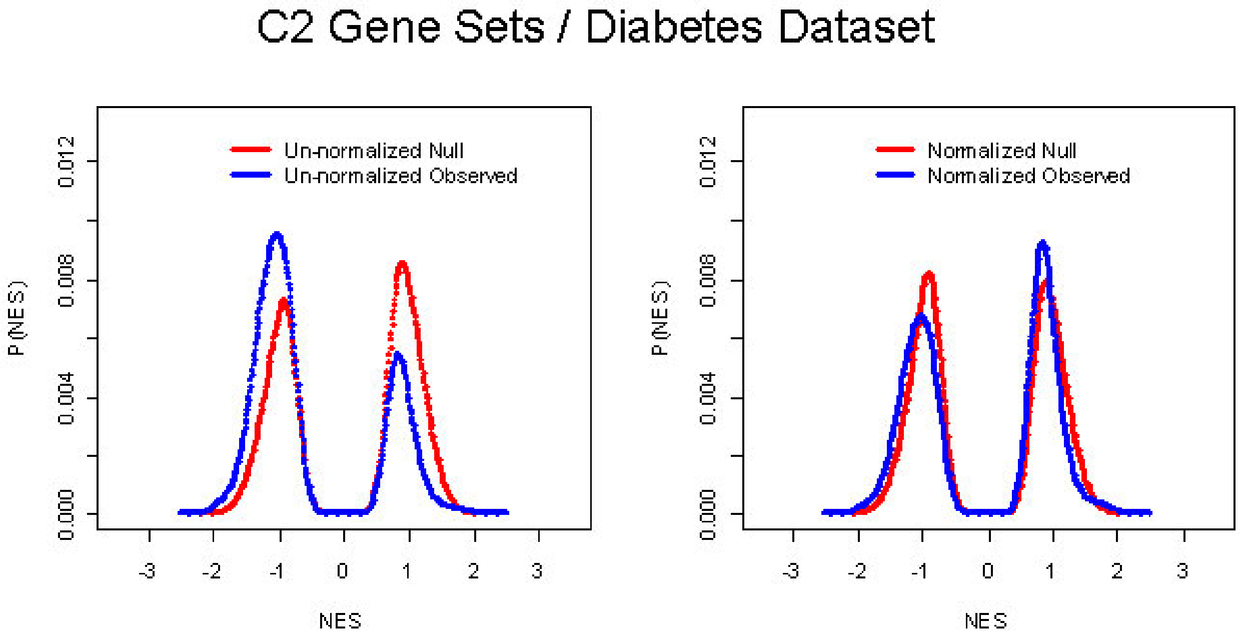


Figure 3-2: Asymmetry of GSEA results due to unbalanced global phenotype expression and gene set collection bias. (Subramanian, et al., 2005)

(4) Multiple hypothesis testing:

In practice we may be interested in several gene sets instead of one. In this case, we need to control the FDR as described in Chapter 2. For each gene set S and a fixed permutation π, the corresponding $ES(S,\pi)$ can be computed. Each $ES(S,\pi)$ is normalized accounting for size of . Since the distribution is bimodal, a standarization is performed by rescaling the positive and negative scores separately with their mean values, yielding $NES(S,\pi)$ (Figure 3-2). A global null distribution can be obtained by pooling all the $NES(S,\pi)$ over S and π. Meanwhile, we denote the normalized score of the original observed data set as $NES(S)$. Then for a given $NES^* \geq 0$, the FDR q value is calculated by

$$q = \frac{(\# \, of \; NES(S,\pi) \geq NES^*)/(\# \, of \; NES(S,\pi) \geq 0)}{(\# \, of \; NES(S) \geq NES^*)/(\# \, of \; NES(S) \geq 0)}$$

Remark: The numerator represents the proportion of gene sets exceeding NES^* among all permutated gene sets with positive NES, and the denominator represents the proportion of gene sets exceeding NES^* among all gene sets with positive NES in the observed data. Similar calculation can be done for $NES^* < 0$. Significant gene sets can be identified according to these q values.

3.1.3 Limitation

Some limitations of GSEA are given below.

(1) GSEA is a method based on competitive test, however, the permutation test to access significance is based on subject sampling. This discrepancy leads to a reduced power of this method.

(2) GSEA is sensitive to the gene sets of genes with expression level change in consistent direction. If both up-regulated and down-regulated genes exist in a large number, GSEA perform poorly to detect this gene set. A quadratic transformation could be adopted to improve the

performance. However, it sacrifices the power when gene expression levels are changing in the same direction.

(3) GSEA is designed to identify the gene sets with most consistent expression patterns that a gene set with a certain percent of differentially expressed genes may not reach significance because of the comparison to other gene sets. In other words, GSEA requires more extreme evidence to declare significance compared to self-contained tests.

(4) The performance of GSEA is affected tremendously by covariance structure of the genes in a given gene set.

All these limitations would be further demonstrated in our simulation studies in Chapter 4.

3.2 Gene Pathway Analysis by Principal Component Analysis

3.2.1 Introduction

Testing for differential expression of many genes with small samples is problematic (Yang H, 2007). A rigorous approach to gene expression analysis must explore the characteristic structure of the data. In this case, principal component analysis (PCA) can be a valuable tool in obtaining such a characterization.

PCA is a dimension reduction method to simplify complex data sets, identify patterns in data, and re-express the data such that the similarities and differences of the variables are highlighted (Pearson, 1901). This mathematical procedure is implemented by orthogonal transformation. A set of observations of possibly correlated variables are converted into a set of linearly uncorrelated variables, which are called principal components.

Since patterns in gene expression data set can be hard to find due to high dimension and weak signals, dimension reduction is usually needed to extract a small number of representative features for the effects of all genes, that is, to identify the most meaningful basis to re-express a data set. PCA can be done by eigenvalue decomposition of covariance or correlation matrix of a data or by singular value decomposition (SVD) of the data matrix.

3.2.2 Eigenvalue Decomposition (Johnson & Wichern, 2007)

Let $\sum$ be the covariance matrix associated with the random vector $\boldsymbol{X}' = \left(X_1, X_2, \ldots, X_p\right)$. Let $\sum$ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e_1}), (\lambda_2, \mathbf{e_2}), \ldots, \left(\lambda_p, \mathbf{e_p}\right)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, $\mathbf{e}_{\boldsymbol{i}}'\mathbf{e}_{\boldsymbol{i}} = 1$ and $\mathbf{e}_{\boldsymbol{i}}'\mathbf{e}_{\boldsymbol{j}} = 0$ for $i \neq j$. Then the i^{th} *principal component* is given by

$$Y_i = \mathbf{e}_{\boldsymbol{i}}'\boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1,2, \ldots, p$$

With these choices,

$$Var(Y_i) = \mathbf{e}_{\boldsymbol{i}}'\sum\mathbf{e}_{\boldsymbol{i}} = \mathbf{e}_{\boldsymbol{i}}'\lambda_i\mathbf{e}_{\boldsymbol{i}} = \lambda_i \qquad i = 1,2, \ldots, p$$

$$Cov\left(Y_i, Y_j\right) = \mathbf{e}_{\boldsymbol{i}}'\sum\mathbf{e}_{\boldsymbol{j}} = \mathbf{e}_{\boldsymbol{i}}'\lambda_j\mathbf{e}_{\boldsymbol{j}} = 0 \qquad i \neq j; i, j = 1,2, \ldots, p$$

If some λ_i are equal, the choices of the corresponding coefficient vectors, $\mathbf{e}_{\boldsymbol{i}}$**,** and hence Y_i, are not unique.

Proof:

First step:

Goal: To find the first principal component, we want to maximize $Var(Y_1) = \mathbf{a_1'}\sum\mathbf{a_1}$ subject to $\mathbf{a_1'}\mathbf{a_1} = 1$.

Method: By using Lagrange multipliers, we maximize the function

$$\mathbf{a_1'}\sum\mathbf{a_1} - \gamma_1(\mathbf{a_1'}\mathbf{a_1} - 1)$$

With respect to $\mathbf{a_1}$ by differentiating with respect to $\mathbf{a_1}$.

Result:

$$\frac{d}{d\mathbf{a_1}}\left(\mathbf{a_1'}\sum\mathbf{a_1} - \gamma_1(\mathbf{a_1'}\mathbf{a_1} - 1)\right) = \sum\mathbf{a_1} - \gamma_1\mathbf{a_1} = 0$$

$$\sum \mathbf{a_1} = \gamma_1 \mathbf{a_1}$$

This should be recognized that $\mathbf{a_1}$ is one of the eigenvectors of $\sum$ corresponding to eigenvalue γ_1. But which eigenvector should be chosen?

Let's take a look at the variance of the first principal component:

$$Var(Y_1) = \mathbf{a_1'} \sum \mathbf{a_1} = \mathbf{a_1'} \gamma_1 \mathbf{a_1} = \gamma_1 \mathbf{a_1'} \mathbf{a_1} = \gamma_1$$

This gives the answer to the question above. Since we want the variance of the first principal component to be as large as possible, the vector chosen to compute the first principal component is the eigenvector of covariance matrix $\sum$ corresponding to the maximum eigenvalue λ_1.

Second step:

<u>*Goal*</u>: To find the second principal component, which means we want to maximize $Var(Y_2) = \mathbf{a_2'} \sum \mathbf{a_2}$ subject to $\mathbf{a_2'} \mathbf{a_2} = 1$ and $\mathbf{a_1'} \mathbf{a_2} = 0$

<u>*Method*</u>: By using Lagrange multipliers, we maximize the function

$$\mathbf{a_2'} \sum \mathbf{a_2} - \gamma_2 (\mathbf{a_2'} \mathbf{a_2} - 1) - \phi \mathbf{a_2'} \mathbf{a_1}$$

With respect to $\mathbf{a_2}$ by differentiating with respect to $\mathbf{a_2}$.

<u>*Result*</u>:

$$\frac{d}{d\mathbf{a_2}} (\mathbf{a_2'} \sum \mathbf{a_2} - \gamma_2 (\mathbf{a_2'} \mathbf{a_2} - 1) - \phi \mathbf{a_2'} \mathbf{a_1}) = \sum \mathbf{a_2} - \gamma_2 \mathbf{a_2} - \phi \mathbf{a_1} = 0$$

If we multiply $\mathbf{a_1'}$ to the both sides of the equation, we would have

$$\mathbf{a_1'} \sum \mathbf{a_2} - \mathbf{a_1'} \gamma_2 \mathbf{a_2} - \mathbf{a_1'} \phi \mathbf{a_1} = 0 - 0 - \phi = 0$$

Then we need to solve

$$\sum \mathbf{a_2} - \gamma_2 \mathbf{a_2} = 0$$

Therefore, $\mathbf{a_2}$ is also eigenvector of $\sum$ which is associated with the second largest eigenvalue λ_2.

Third step:

The process will be repeated for $k = 1,2, \dots, m$ and yields m principal components with variance corresponding to the ordered eigenvalues of $\sum$.

We can see that principal components are weighted averages of the original variables. They are constructed to be uncorrelated with each other and capture as much of the original variability as possible. The first principal component represents the largest proportion of the total variance; the second PC represents the second largest, and so on. Therefore, we may select the most representative components based on the fraction of variability to reduce the dimension of variables and gain more power in the analysis.

In practice, the microarray dataset is a $m \times n$ matrix with each row representing the transcriptional response of a gene and each column representing expression profile of an assay. Empirical covariance matrix S of the m genes can be used to perform eigenvalue decomposition and obtain the eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_{\mathbf{1}}), \dots, (\hat{\lambda}_m, \hat{\mathbf{e}}_{\boldsymbol{m}})$. Therefore, the *sample principal components* are

$$\hat{Y}_k = \hat{\boldsymbol{e}}'_{\boldsymbol{k}} \boldsymbol{x}_{\boldsymbol{l}}, \qquad k = 1,2, \dots, m; l = 1,2, \dots, n$$

where $\boldsymbol{x}_{\boldsymbol{l}}$ is the l^{th} observation on *gene 1, gene 2,…, gene m*. Elements in vector $\hat{\boldsymbol{e}}_k$ are called *loadings*, giving weights to be applied to each gene’s expression. A loading close to 0 implies that a gene does not have much variation across arrays. Large loadings (positive or negative) imply considerable variation of a gene across arrays. Numbers of principal components depend on the rank of covariance matrix S; if $rank = r \leq m$ then there would be r principal components. In a typical microarray dataset, $rank\ r$ is much smaller than the number of genes, therefore, principal component analysis can effectively reduce the number of variables.

3.2.3 Singular Value Decomposition (Berrar, Dubitzk, & Granzow, 2002)

Let $\boldsymbol{X}$ denote an $m \times n$ matrix of real-valued data and rank r**.** In the case of microarray data, X_{ij} is the expression level of the *i*th gene in the *j*th observation. Since a typical microarray data always consists of much more genes than observations, we will focus on the situation $m \geq n$, and therefore $r \leq n$.

Before applying SVD to the data, we would like to center each row of the data by subtracting row means: $\overline{\boldsymbol{x}} = n^{-\mathbf{1}}\left(\boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)} + \cdots + \boldsymbol{x}^{(n)}\right)$**, where** $\boldsymbol{x}^{(k)} = (x_{1k}, x_{2k}, \ldots, x_{mk})',\ \ for\ k = 1,2,\ldots,n.$ We replace the original data by the centered data: $\boldsymbol{X} = (\boldsymbol{x_1} - \overline{\boldsymbol{x}}, \boldsymbol{x_2} - \overline{\boldsymbol{x}}, \ldots, \boldsymbol{x_n} - \overline{\boldsymbol{x}})$

The equation for singular value decomposition of $\boldsymbol{X}$ is the following:

$$\boldsymbol{X} = \boldsymbol{UDV'}$$

where $\boldsymbol{U}$ is an $m \times n$ matrix with orthonormal columns ($\boldsymbol{u_i'u_i} = 1, \boldsymbol{u_i'u_j} = 0\ for\ i \neq j$), called the *left singular vectors,* while $\boldsymbol{V}$ is an $n \times n$ orthonormal matrix ($\boldsymbol{V'V} = \boldsymbol{I}$) with columns called the *right singular vectors*, and $\boldsymbol{D}$ is an $n \times n$ diagonal matrix with positive or zero elements, called the *singular values*. Thus,$\boldsymbol{D} = diag(d_1, \ldots, d_n)$. Furthermore, $d_k > 0$ for1 $\leq k \leq r$, and $d_k = 0$ for $(r + 1) \leq k \leq n$.

From $\boldsymbol{X}$ we can construct two positive-definite symmetric matrices, $\boldsymbol{XX'}$ and $\boldsymbol{X'X}$, each of which we can decompose

$$\boldsymbol{XX'} = \boldsymbol{UD^2U'}$$

$$\boldsymbol{X'X} = \boldsymbol{VD^2V'}$$

Remember $m \geq n$, thus we can see that

(1) The left singular vectors of $\boldsymbol{X}$ are eigenvectors corresponding to the non-zero eigenvalues of $\boldsymbol{XX'}$;

(2) The right singular vectors of $\boldsymbol{X}$ are eigenvectors of $\boldsymbol{X'X}$**;**

(3) The non-zero singular values of $\boldsymbol{X}$ are the square roots of the non-zero eigenvalues of both $\boldsymbol{XX'}$ and $\boldsymbol{X'X}$.

We can transform the data as

$$\boldsymbol{Y} = \boldsymbol{U'X} = \boldsymbol{DV'}$$

where $\boldsymbol{Y}$ has a diagonal empirical covariance matrix:

$$\boldsymbol{C} = \boldsymbol{n^{-1}}(\boldsymbol{YY'}) = \boldsymbol{n^{-1}}(\boldsymbol{DV'VD}) = \boldsymbol{n^{-1}D^2}$$

Each row of the transformed data $\boldsymbol{Y}$ represents n observations of sample principal components, and each column vector in $\boldsymbol{U}$ gives loadings for the corresponding principal component.

It is obvious that SVD and eigenvalue decomposition are essentially equivalent if we use centered data for SVD and covariance matrix for eigenvalue decomposition. The two methods will give the same set of loadings for principal components; furthermore, it is obvious that d_k^2 is proportional to the variances of principal components.

3.2.4 Application in Gene Pathway Analysis

A gene pathway may contain a large number of genes (more than the number of observations); therefore, a straightforward regression fitting may result in saturated models (Ma & Kosorok, 2009). Variable selection methods can be used when there are a small number of genes with considerable variability. However, in gene pathway, a more common situation is that there exist a large number of genes with moderate changes. In this case, dimension reduction with PCA may perform. Tomfohr and his colleagues proposed to use PCA on gene set analysis (Tomfohr, Lu, & Kepler, 2005), but they only used the first PC and applied a t test on it. Later, Ma et al. extended the PCA method to a more general model that can include more PCs on gene set analysis (Ma & Kosorok, 2009). They also investigated the second-order non-linear effects and claimed that non-linear effects may identify a small number of key pathways that could not be found by models merely including linear effects, and need to be considered in practice (Ma & Kosorok, 2009).

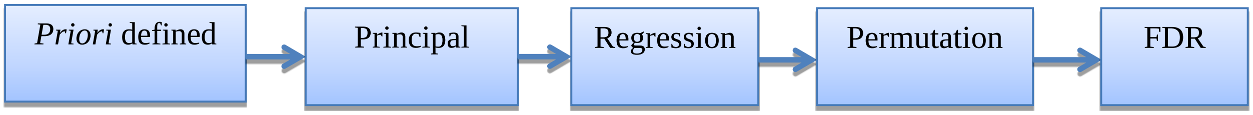


Figure 3-3: Flowchart of gene pathway analysis by PCA.

The following are details of gene pathway analysis with PCA proposed by Ma et al:

Step 1: We first retrieve pathway information from priori biological knowledge, and group data into gene sets. The validity of the following analysis depends on the accuracy of the pathway information.

Step 2: For each gene pathway, PCA is performed to compute a set of PCs. Ma et al. showed that including a few top PCs did improve the power. However, in practice, the number of PC to be included in the regression model is limited by sample size. For example, if there are only 40 samples (20 for case, 20 for control), it is recommended not to incorporate more than 2 PCs into the regression model.

Step 3: Regression models and statistics are selected based on the type of outcomes, that is:

(1) Continuous outcomes → Linear regression model and mean squared error;

(2) Categorical outcomes → Logisitc regression model and deviance;

(3) Survival clinical outcomes → Cox proportional hazards model and the statistic of the score test.

Step 4: The p value of the regression model is given by permutation test.

Step 5: Repeat step 2 to step 4 for each pathway in the microarray data.

Suppose there are N pathways, then there will be N p-values in total. Since all these pathways are test simultaneously, we also need to control false discovery rate (FDR).

3.2.5 Limitation

Although PCA can be easily incorporated and is a widely used method, there were two obvious limitations of applying it on gene pathway analysis.

Firstly, PCA is an unsupervised method, which only depends on the data matrix. The first several principal components capture the majority of the data variation, however, no outcome information is taken into consideration during the data transformation. There is no guarantee that the first principal component should be the factor most correlated with the outcome, and the second PC be the second most correlated factor, and etc. Therefore, the PCA based transformation may completely fail for the goal of identifying factors that are significantly associated with outcome,.

Secondly, standardization of the data matrix is usually performed before applying PCA. That is,

$$z_{ij} = \frac{x_{ij} - \overline{x_\iota}}{Std(x_i)}$$

where x_{ij} is the observation for gene i in sample j, $\overline{x_\iota}$ is the sample mean of gene i, $Std(x_i)$ is the sample standard deviation of gene i.

The empirical covariance matrix for the standardized data is identical with the empirical correlation matrix for the data before standardization. Therefore, the first principal component is essentially capturing the variables that are in large correlation clusters.

For example, if we have three variables x_1, x_2, x_3 and their empirical correlation matrix is like this

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_2 \\ 0 & \rho_2 & 1 \end{bmatrix}, where\ \rho_1 > \rho_2$$

The ordered eigenvalues for this matrix are: $\lambda_1 = 1 + \sqrt{\rho_1^2 + \rho_2^2}, \lambda_2 = 1, \lambda_3 = 1 - \sqrt{\rho_1^2 + \rho_2^2}$. The corresponding eigenvector to the largest eigenvalue is $(\frac{1}{\sqrt{2}}, \frac{\rho_1}{\sqrt{2(\rho_1^2+\rho_2^2)}}, \frac{\rho_2}{\sqrt{2(\rho_1^2+\rho_2^2)}})'$. Since $\frac{1}{\sqrt{2}} > \frac{\rho_1}{\sqrt{2(\rho_1^2+\rho_2^2)}} > \frac{\rho_2}{\sqrt{2(\rho_1^2+\rho_2^2)}}$, we can easily see that in the first principal component, the largest weight is given to x_1, which is involved in two correlation clusters, the second largest weight is given to x_2, which is in the cluster with larger correlation, and the smallest weight is given to x_3. We can make this example more extreme by setting ρ_2 to 0. In this case, the eigenvector for the first principal component becomes $(\frac{1}{\sqrt{2}}, \frac{\rho_1}{\sqrt{2(\rho_1^2+\rho_2^2)}}, 0)'$. This means x_3 does not contribute to the first principal component at all.

Therefore, the first principal component is determined completely by the correlation clusters of the data matrix, which could be quite biased in gene pathway analysis.

More details will be shown in our simulation study in Chapter 4.

3.3 Canonical Discriminant Analysis/MANOVA

3.3.1 Introduction

Canonical Discriminant Analysis (CDA) is another PCA-related dimension-reduction technique. However, unlike PCA that only summarizes the total variation of data without including the outcome information, canonical discriminant analysis takes account the group information. Given a categorical variable (outcome) and a set of continuous variables (predictors), CDA derives a set of linear combinations of the predictors (*canonical variables*), which maximize the ratio of between-group variation to within-group variation. The vector that generates the linear combination is *canonical coordinate* and the elements in the vector, which are the coefficients of the linear combination, are called *canonical coefficient*.

The first canonical variable has the highest possible multiple correlation with groups. The second canonical variable is obtained by finding the linear combination uncorrelated to the first canonical variable, which has the second highest possible multiple correlation with groups. And the process goes on until the number of canonical variables equals the number of classes minus one.

3.3.2 Mathematical Implementation

In the following, we briefly show how to perform the general CDA. Let $\boldsymbol{X}$ be the $m \times n$ data matrix with each row representing observations of a continuous variable from n samples. Let $\boldsymbol{a}$ be a $m \times 1$ vector, and therefore, $\boldsymbol{z} = \boldsymbol{a}'\boldsymbol{X}$ is also a $m \times 1$ vector.

Suppose the data were from k groups, and each group includes $n_r, r = 1,2,\dots,k$ samples. The ratio of between group variations to within group variation would be

$$ratio = \frac{\sum_{r=1}^{k}\sum_{i=1}^{n_r}(\bar{z}_{.r} - \bar{z}_{..})^2}{\sum_{r=1}^{k}\sum_{i=1}^{n_r}(z_{ir} - \bar{z}_{.r})^2}$$

Where $\bar{z}_{.r} = \frac{\sum_{i=1}^{n_r} z_{ir}}{n_r}$ is the group mean for the r^{th} group, $\bar{z}_{.r} = \frac{\sum_{r=1}^{k}\sum_{i=1}^{n_r} z_{ir}}{\sum_{r=1}^{k} n_r}$ is the overall mean across k groups.

To find the canonical coordinate that maximizes the ratio, the ratio can be written in a matrix form:

$$ratio = \frac{\boldsymbol{a'X}(\boldsymbol{I_k} \otimes \boldsymbol{J} - \frac{\boldsymbol{J_n}}{\boldsymbol{n}})\boldsymbol{X'a}}{\boldsymbol{a'X}(\boldsymbol{I_n} - \boldsymbol{I_k} \otimes \boldsymbol{J})\boldsymbol{X'a}} = \frac{\boldsymbol{a'Ba}}{\boldsymbol{a'Wa}}$$

$$\boldsymbol{J} = \left(\frac{\boldsymbol{J_{n_1}}}{n_1}, \frac{\boldsymbol{J_{n_2}}}{n_2}, \dots, \frac{\boldsymbol{J_{n_k}}}{n_k}\right)', \boldsymbol{B} = \boldsymbol{X}\left(\boldsymbol{I_k} \otimes \boldsymbol{J} - \frac{\boldsymbol{J_n}}{\boldsymbol{n}}\right)\boldsymbol{X}', \boldsymbol{W} = \boldsymbol{X}(\boldsymbol{I_n} - \boldsymbol{I_k} \otimes \boldsymbol{J})\boldsymbol{X}'$$

where $\boldsymbol{J}_{\boldsymbol{n}_r}$ is a $n_r \times n_r$ matrix with all elements equal to 1.

By Cholesky Decomposition,

$$\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{U}$$

where $\boldsymbol{U}$ is an upper triangular matrix with dimension of $m \times m$. Therefore, the ratio can be expressed as

$$ratio = \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{a}} = \frac{\boldsymbol{b}'(\boldsymbol{U}^{-1})'\boldsymbol{B}\boldsymbol{U}^{-1}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{b}} = \frac{\boldsymbol{b}'\boldsymbol{D}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{b}}$$

where $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{a}, \boldsymbol{a} = \boldsymbol{U}^{-1}\boldsymbol{b}, \boldsymbol{D} = (\boldsymbol{U}^{-1})'\boldsymbol{B}\boldsymbol{U}^{-1}$.

We want to maximize $ratio = \frac{\boldsymbol{b}'\boldsymbol{D}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{b}}$ with subject to $\boldsymbol{b}'\boldsymbol{b} = 1$. Then the question becomes maximization of $\boldsymbol{b}'\boldsymbol{D}\boldsymbol{b}$ with subject to $\boldsymbol{b}'\boldsymbol{b} = 1$. Now the question became similar to compute principal components. The first canonical variable is given by the eigenvector of matrix $\boldsymbol{D}$ associated with the largest eigenvalue of $\boldsymbol{D}$, the second canonical variable is computed by the eigenvector that is corresponding the second largest eigenvalue, and etc.

3.3.3 Comparison with MANOVA/Hotelling's T Test

After we apply CDA on the original gene expression data and obtain the transformed data, an appropriate statistic and the corresponding test would be required to access the significance for each pathway. Since the CDA algorithm is designed to separate the original data by groups, the maximum possible ratio of between group variability to within group variability is an intuitive statistic to measure how well the data are apart. In other words, we choose the eigenvalue of matrix $\boldsymbol{D}$ described above to represent how correlated the data are to the outcome. The larger of the eigenvalue is, the stronger correlation exists between gene expression levels and the outcome. This method is virtually equivalent to multivariate analysis of variance (MANOVA) as shown below.

There are four different MANOVA tests: *Pillai's trace*, *Hotelling-Lawley's trace*, *Wilk's lambda* and *Roy's largest root* (Mardia & J.T. Kent, 1979). These methods are all based on the matrix $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$, and they give identical results when there are only two groups. The tests are defined as follows:

Pillai's trace $= trace(\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{W})^{-\mathbf{1}}) = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$

Hotelling-Lawley's trace$=$ $trace(\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}) = \sum_{i=1}^{m}\lambda_i$

Wilk's lambda$=$ $\frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|} = \prod_{i=1}^{m}\frac{1}{1+\lambda_i}$

Roy's largest root $= \max(\lambda_i)\,, i = 1,2,\dots,m$

where λ_i is the ordered eigenvalue of $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$. In certain circumstances, these four tests give an exact F ratio and in other situations F ratio is approximated. Some researchers consider the Pillai's trace to be the most powerful and robust method while Wilk’s lambda was the first derived MANOVA test and widely used. When there are only two groups, the rank of $\boldsymbol{B}$ would be one and only one nonzero eigenvalue λ existing for $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$.

Pillai's trace $= \frac{\lambda}{1+\lambda}$

Wilk’s lambda $= \frac{1}{1+\lambda}$ = 1- Pillai's trace

Hotelling-Lawley's trace = Roy’s largest root $= \lambda$

Therefore, these four tests lead to identical result.

Now let’s take Wilk’s lambda as an example to show that using the maximum ratio of between group variability to within group variability as statistic after application of CDA is essentially equivalent to MANOVA test in the case of two classes. In other words, this is to show the eigenvalue of matrix $\boldsymbol{D}$ is equal to eigenvalue of $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$.

Proof:

$$\boldsymbol{D} = (\boldsymbol{U}^{-\mathbf{1}})'\boldsymbol{B}\boldsymbol{U}^{-\mathbf{1}}$$

$$\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B} = (\boldsymbol{U}'\boldsymbol{U})^{-\mathbf{1}}\boldsymbol{B} = (\boldsymbol{U}^{-\mathbf{1}})(\boldsymbol{U}^{-\mathbf{1}})'\boldsymbol{B}$$

Therefore,

$$\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B} = \boldsymbol{U}^{-\mathbf{1}}\boldsymbol{D}\boldsymbol{U}$$

Say λ is the eigenvalue for $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$ and $\boldsymbol{x}$ is the corresponding eigenvector, then

$$(\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B} - \lambda\boldsymbol{I})\boldsymbol{x} = \mathbf{0}$$

That is

$$(\boldsymbol{U}^{-\mathbf{1}}\boldsymbol{D}\boldsymbol{U} - \lambda\boldsymbol{I})\boldsymbol{x} = \mathbf{0}$$

$$(\boldsymbol{D} - \lambda\boldsymbol{I}\boldsymbol{U}\boldsymbol{U}^{-\mathbf{1}})\boldsymbol{x} = \mathbf{0}$$

$$(\boldsymbol{D} - \lambda\boldsymbol{I})\boldsymbol{x} = \mathbf{0}$$

Therefore, we have proved that $\boldsymbol{W}^{-\mathbf{1}}\boldsymbol{B}$ and $\boldsymbol{D}$ share the same eigenvalue.

In practical, either CDA algorithm or one of the four MANOVA tests could be adopted for gene pathway analysis in the case of binary outcome.

3.3.4 Singular Matrix Problem

In the algorithm described above for CDA, matrix $\boldsymbol{W}$ represents the within-group variability with the dimension of $m \times m$. However, the number of genes in a pathway vary widely so that it is possible to have more variables than samples ($m \gg n$), and this will lead to singularity of the $\boldsymbol{W}$ matrix. Consequently, the $\boldsymbol{U}$ matrix obtained by Cholesky Decomposition of $\boldsymbol{W}$ is not full-ranked so it is not invertible, and matrix $\boldsymbol{D}$ is not computable. When the dimension of variables is smaller than sample size ($m < n - 1$), CDA can be applied in a straightforward way. When dimension of variables is larger than sample size ($m \geq n - 1$), the algorithm proposed above is not direcdtly applicable.

Several different approaches have been proposed to solve this problem. For example, Heydebreck et al. ignored the correlation between genes and made a simple modification to set the off-diagonal elements in $\boldsymbol{W}$ to be zero (Heydebreck, Huber, Poustka, & Vingron, 2001). Kong and his group proposed to apply PCA on $\boldsymbol{W}$ matrix, and pick up the first several principal components based on a threshold (e.g. 10^{-4}) to form a new nonsingular $\boldsymbol{W}$ matrix, and then Hotelling's t test was adopted for significance test (Kong, Pu, & Park, 2006). For more sophisticated modification of $\boldsymbol{W}$ matrix, different penalized methods were proposed, such as Regularized Discriminant Analysis (Friedman, 1989), Penalized Discriminant Analysis (Hastie, 1995), and the method to keep the diagonal of $\boldsymbol{W}$ but shrink centroid for each group (Tibshirani, Hastie, Narasimhan, & Chu, 2002). Most recently, Tsai and Chen chose the shrinkage covariance matrix estimator proposed by Schafer and Strimmer (Schafer & Strimmer, 2005) to make the $\boldsymbol{W}$ matrix well-conditioned (Tsai & James, 2009). The estimator is given as follows:

$$S_{ij}^{*} = \begin{cases} S_{ii} & if\ i = j \\ r_{ij}^{*}\sqrt{s_{ii}s_{jj}} & if\ i \neq j \end{cases}$$

and

$$r_{ij}^{*} = r_{ij}\min\{1, \max(0,1 - \hat{\lambda}^{*})\}$$

where s_{ii} and r_{ij} denote the empirical sample variance and sample correlation, respectively. $\hat{\lambda}^{*}$ is called optimal shrinkage intensity and is estimated by

$$\hat{\lambda}^{*} = \frac{\sum_{i\neq j}\widehat{Var}(r_{ij})}{\sum_{i\neq j} r_{ij}^{2}}$$

The function to estimate the shrinkage covariance matrix is embedded in an R package "corpor". For the eigenvalue calculated based on the modified $\boldsymbol{W}$ matrix, its distribution does not have a closed form. Therefore, the p value could be computed by permutation test as most of the gene set analyses do.

3.4 Intersection-Union Test and Union-Intersection Test

In the previous sections, we introduced three popular gene set analysis methods in detailed. But our study did not focus on an individual method, but a combination of methods. Therefore, we would introduce two types of tests that can combine the results from various methods in this section.

For a study including more than one test, a question naturally raised would be how to combine the results from different tests. It is intuitive to see that the answer depends on the purpose of the study. For example, if we have several microarray data sets and we are interested in finding consistently differentially expressed genes among all data to make the results more reliable and robust, we would like to identify genes with H0 rejected in tests for all data sets. This type of “reject all” test is called Intersection- Union Test (IUT). The complementary test to IUT is Union-Intersection Test (UIT) (Casella & Berger), which would reject the global null hypotheses as long as there is at least one null hypothesis is rejected. One possible scenario is that when we are testing the hazard of a new drug using different index, we would say the drug to be hazardous as long as one test shows hazardous. Definitions and details of IUT and UIT would be described in the following section.

3.4.1 Definition

Union-Intersection Test (UIT) is applicable when the null hypothesis can be conveniently expressed as an intersection of a set of null hypotheses:

$$H_0: \bigcap_{i=1}^{k} H_{0i}$$

While the alternative hypothesis is a union of a family of alternative hypotheses:

$$H_A: \bigcup_{i=1}^{k} H_{Ai}$$

If we suppose a suitable test for each hypothesis $H_{0i}: \theta \in \Theta_i$ versus $H_{Ai}: \theta \in \Theta_i^c$, we can rewrite the global hypothesis as

$$H_0: \theta \in \bigcap_{i=1}^{k} \Theta_i$$

Therefore, the reject region is

$$H_A: \theta \in \bigcup_{i=1}^{k} \Theta_i^c$$

From this definition, we can see that if one of the null hypothesis is rejected, the global null hypothesis would also be rejected.

On the other hand, Intersection- Union Test (IUT) is defined as follows:

Suppose we wish to test the null hypothesis which is expressed as a union:

$$H_0: \bigcup_{i=1}^{k} H_{0i}$$

and the alternative hypothesis is:

$$H_A: \bigcap_{i=1}^{k} H_{Ai}$$

Say the reject region for the test H_{0i} is $\{x: T_i(x) \in R_i\}$ (Hasler, 2007), the reject region for intersection-union test now become:

$$\bigcap_{i=1}^{k} \{x: T_i(x) \in R_i\}$$

The global null hypothesis would be rejected if and only if all local null hypotheses are rejected.

3.4.2 Method

There are various methods developed for both union-intersection test and intersection-union test, which are called *Meta-analysis*. In simple words, meta-analysis is to identify a common statistical measure that is shared by each study. For example, we can use *effect size* as the common measure if we expect to reject the global null hypothesis when $\{x: T_i(x) < c\}$, then the most straightforward statistic for UIT would be:

$$T(x) = \min_{i=1,\ldots,k} T_i(x)$$

This indicates as long as there is one local null hypothesis rejected, the global null hypothesis gets rejected.

On the other hand, the most intuitive statistic for IUT would be:

$$T(x) = \max_{i=1,\ldots,k} T_i(x)$$

which means only when all the local null hypotheses get rejected, we have the confidence to reject the global null hypothesis.

The above example was based on assumption that all the tests share the same rejection region. However, it is not always applicable in real data analysis where different endpoints could come from totally different measures that they do not share identical cutoff. Or we may apply different statistical tests on different endpoints which make the assumption even more unstandable.

To overcome the problem raised by inconsistence of multiple tests, p value is another popular statistic used to combine results. P value is defined as the probability of obtaining a test statistic result at least as extreme as the actually observed one. No matter what specific test is adopted, p value is measuring how strong the evidence is against the null hypothesis, and follows a uniform distribution between zero and one when the null hypothesis is true. With a known distribution under null hypothesis and consistent explanation among tests, p value combination was the basis of many published methods for meta-analysis. Here we introduce several most popular ones.

3.4.2.1 Fisher's Method

Let $p_1, p_2, \dots, p_k$ be the p values obtained by k independent tests, the global null hypothesis is $H_0 \colon \bigcap_{i=1}^{k} H_{0i}$ versus $H_A \colon \bigcup_{i=1}^{k} H_{Ai}$. Ronald Fisher (Fisher, 1925) proposed a method to sum up log-transformed p values:

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \sim \chi_{2k}^2$$

Under the global null hypothesis, the statistic X^2 is following a chi-squared distribution with $2k$ degree of freedom.

<u>*Proof:*</u>

Under local null hypothesis, P_i follows distribution of $Uniform(0,1)$. The negative natural logarithm of a uniformly distributed variable follows an exponential distribution with parameter of one. Multiplying a scale of two to the exponential distribution described above yields a chi-squared distribution with 2 degree of freedom. Since all the tests are independent and the global hypothesis is an intersection of all local null hypotheses, X^2 is now a sum of k independent chi-squared distribution; therefore, it also follows a chi-squared distribution and the degree of freedom is $2k$ under global null hypothesis.

Fisher's method essentially computes the distribution of $\prod_{i=1}^{k} P_i$ where P_i is a random variable of p value. Since we know the distribution for X^2, for any observed $\prod_{i=1}^{k} p_i$ we can calculate the probability of observing a product of $p_i's$ at least as extreme as the current one, in other words, the p value based on X^2 test. Fisher's method has been used for a long time because of its explicit distribution and easy calculation. However, there are two assumptions we need to pay attention to. Firstly, this method requires independency between tests, which could be easily violated in biological studies because genes or gene pathways always function interactively. This limitation can be overcome by Satterwhite's approximation which will be shown in the following section. Secondly, this method is appropriate for UIT only. Fisher's method is essentially a multiple of p values and would encounter some problem when some of the p values are extremely small (Whitlock, 2005). More details will be shown in 3.4.4.

3.4.3 Satterwhite's Approximation for Dependent Hypotheses

Satterwhite's approximation is an extension for Fisher's method when dependency between tests exists. Again, let $p_1, p_2, \dots, p_k$ be the p values obtained by k individual tests, and these tests are not necessarily independent. The global null hypothesis is $H_0: \bigcap_{i=1}^{k} H_{0i}$ versus $H_A: \bigcup_{i=1}^{k} H_{Ai}$. We define $z_i = -2\ln(p_i)\,, i = 1,2, \dots, k$, as we proved in the previous section, each z_i is following a chi-squared distribution with 2 degree of freedom. In this case, since z_i's are not necessarily independent the distribution of sum of all z_i's is then unknown.

To solve this problem, we can assume the sum of dependent chi-squared distribution follows a scaled chi-squared distribution under the global null hypothesis (Li S. , 2011):

$$T = \sum_{i=1}^{k} z_i \ \dot{\sim}\ a\chi_g^2$$

The scale a and the degree of freedom g can be estimated by equating the first and second moments of $\sum_{i=1}^{k} z_i$ to the first and second moments of $a\chi_g^2$, respectively.

$$E(T) = E\left(\sum_{i=1}^{k} z_i\right) = 2k$$

$$Var(T) = Var\left(\sum_{i=1}^{k} z_i\right) = \sum_{i=1}^{k} Var(z_i) + 2\sum_{i<j} Cov(z_i, z_j) = 4k + 8\sum_{i<j} \rho_{ij}$$

where ρ_{ij} is the correlation between z_i and z_j.

On the other hand,

$$E\left(a\chi_g^2\right) = ag$$

$$Var\left(a\chi_g^2\right) = 2a^2g$$

Therefore, we have

$$2k = ag;\ \ 4k + 8\sum_{i<j} \rho_{ij} = 2a^2g$$

By some simple algebra, we can solve a and g as follows (Li S. , 2011):

(1) $\hat{a} = \frac{4k+8\sum_{i<j}\rho_{ij}}{4k} = 1 + \frac{2\sum_{i<j}\rho_{ij}}{k}$;

(2) $\hat{g} = \frac{2k}{\hat{a}} = \frac{2k^2}{k+2\sum_{i<j}\rho_{ij}}$.

When tests are independent, $\rho_{ij} = 0\ \forall i \neq j$. Then $\hat{a} = 1, \hat{g} = 2k$, this is exactly identical to the resul obtained by Fisher's method. When there exists dependency between tests, the question is how to estimate the correlation ρ_{ij}. Li et al. again suggested a general method that is based on permutations (Li S. , 2011).

The idea is to generate an empirical distribution of z (natural logarithm of p value) under null hypothesis so that we can estimate the correlation by calculating sample correlation. For each permutation, we would have a vector of p values: $\boldsymbol{p}^b = \left(p_1^b, p_2^b, \dots, p_k^b\right), b = 1,2, \dots, B$; consequently, we have a vector of z as $\boldsymbol{z}^b = (z_1^b, z_2^b, \dots, z_k^b)$. After finishing all permutations, we could calculate the correlation for any pair of z under global null hypothesis by the sample correlation of the permuted random sample.

Since Satterwhite's approximation is just an extension of Fisher's method, again one limitation of this method is that it is only suitable for UIT.

3.4.4 Z-transform Method

Another widely used p value combination method is called Z-transform method or Stouffer's method (Stouffer, 1949). The Z-transform method takes advantage of the one-to-one mapping of the cumulative distribution function of standard normal distribution to the p values under null hypothesis (Whitlock, 2005). The assumptions for Z-transform method are those used in Fisher's method: Independency between tests, intersection of local null hypotheses.

If we use $\Phi(x)$ to denote the cumulative distribution function of standard normal distribution with mean of 0 and standard deviation of 1, then for each p value the transformed Z score is

$$Z_i = \Phi^{-1}(1 - P_i) \sim Normal(0,1)$$

The statistic $Z_s = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$ is also following a standard normal distribution, therefore, a p value corresponding to Z_s can be easily calculated given any set of observed p value from individual tests.

The largest advantage of Z-transform method is the symmetric distribution of statistic Z_s compared to Fisher's method. It is not hard to show that X^2 in Fisher's method is skewed to the right, which indicates that Fisher's method is asymmetrically sensitive to small p values compared to large p values (Whitlock, 2005). The problem can be elaborated with the following example. Suppose we want to combine the results of 2 tests: in the first scenario, $p_1 = 0.0001$ and $p_2 = 0.3$, and in the second scenario, $p_1 = 0.01$ and $p_2 = 0.01$. If we just review these numbers by eyes, we probably would think the results in the second scenario is more significant because both local tests show significance. However, after applying Fisher’s method, we end up with $X_1^2 = 20.82$ in scenario 1, which is more extreme than the statistic $X_2^2 = 18.42$ obtained in scenario 2. This asymmetry results in bias in combining multiple results, although the bias may not be as great as in the example described above (Whitlock, 2005). On the other hand, Z-transform method is virtually a sum of variables with standard normal distribution under the global null hypothesis, which guarantees the symmetry of the combined statistic.

Another advantage of Z-transform method is that weights could be introduced to the statistic. It is called weighted Z-method:

$$Z_w = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

The statistic Z_w still follows a standard normal distribution. This is a very practical design, for some tests may provide more meaningful explanation or show more importance when comparing to others. A large number of criteria were proposed on how to choose weights. For example, if we want to combine results from different microarray data, we may use the sample size as weight. Or if all the studies are using t-tests, weight can be defined as the degree of freedom. More generally, weights could be the inverse of the squared standard error of the effect size (Whitlock, 2005). More methods to choose weights in genomic study were studied in the work of Li et al. (Li & Ghosh, 2012)

3.4.5 Minimum P Value and Maximum P Value

Here we are going to introduce 2 more traditional meta-analysis methods: minimum p value (Tippett, 1931) and maximum p value (Wilkinson, 1951).

$$minP = \min_{i=1,\dots,k} P_i$$

$$maxP = \max_{i=1,\ldots,k} P_i$$

When we are considering union-intersection hypothesis, we want the reject region to be a union of local reject regions:

$$\bigcup_{i=1}^{k} \{x: T_i(x) \in R_i\}$$

If p value is used to decide the rejection region, now the above formula becomes

$$\bigcup_{i=1}^{k} \{p_i < \alpha\} = \min_{i=1,\ldots,k} P_i < \alpha$$

where α is the cutoff for rejection region.

On the other hand, if what we care is the consistency of the local tests and will reject the global null hypothesis if and only if all local tests are rejected, the reject region is

$$\bigcap_{i=1}^{k} \{x: T_i(x) \in R_i\} = \bigcap_{i=1}^{k} \{p_i < \alpha\} = \max_{i=1,\ldots,k} P_i < \alpha$$

Therefore, minimum p value is suitable for UIT while maximum p value is appropriate for IUT.

Now let’s take a look at the distribution of minimum p value under the global null hypothesis. If we assume the local tests are independent, it is straightforward to derive the distribution of minimum p value as follows

$$P\left(minP \leq \alpha | \bigcap_{i=1}^{k} H_{0i}\right) = 1 - P\left(minP > \alpha | \bigcap_{i=1}^{k} H_{0i}\right)$$

$$= 1 - P(P_1 > \alpha, P_2 > \alpha, \ldots, P_k > \alpha | \bigcap_{i=1}^{k} H_{0i})$$

$$= 1 - P(P_1 > \alpha | H_{01}) P(P_2 > \alpha | H_{02}) \ldots P(P_k > \alpha | H_{0k})$$

$$= 1 - (1 - \alpha)^k$$

Therefore, we can easily calculate the p value for each observed minimum p value according to the distribution function described above. If the dependency between tests is not negligible, a permutation test can be applied to incorporate this concern. For each permutation, we can obtain a list of p values and the corresponding minimal p value: $minP^b$. Say we perform B permutations, then in the end we will have B permutated minimum p value $(minP^1, minP^2, \dots, minP^B)$, which forms an empirical distribution under global null hypothesis for $minP$. The p value for the observed $minP$ can be calculated by computing the ratio of permuted minimum p value smaller than the observed minimum p value.

The above methods are designed for UITs, however the study of IUT is usually more challenging. As we described in previous section that using maximum p value for IUT is very intuitive and understandable, however, the distribution of maximum p value under global null hypothesis is not derivable.

$$P\left(maxP < \alpha \middle| \cup_{i=1}^{k} H_{0i}\right) = P(P_1 < \alpha, P_2 < \alpha, \dots, P_k < \alpha | \bigcup_{i=1}^{k} H_{0i})$$

Even if all local tests are independent, there is no explicit formula for the distribution of maximum p value because the global null hypothesis is a union of local null hypothesis, which means that local alternative hypothesis is also allowed to be true for some of the tests..

3.4.5.1 Methods to Compute Maximum P Value

In this section, three methods will be introduced which are normally employed to compute maximum p value: Berger’s method, Relaxed IUT (RIUT), and meta-analysis method.

Roger. L. Berger first proved that if every local test has a size α, the IUT is also a level α test (Berger, 1982). In other words, if we use the observed maximum p value to represent the global p value of the combined test, the size is well-controlled at the level of α, which is the size for each local test.

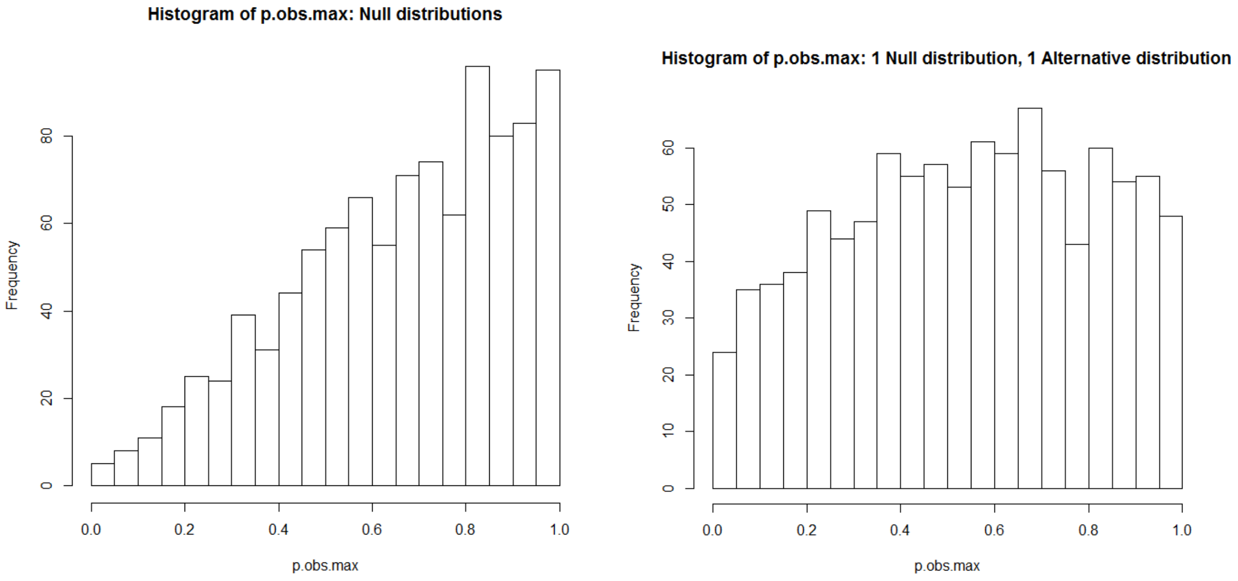
Proof:

$$P\left(maxP \leq \alpha \middle| \cup_{i=1}^{k} H_{0i}\right) = P(\bigcap_{i=1}^{k} R_i \mid \bigcup_{i=1}^{k} H_{0i})$$

$$\leq P(R_i \mid \bigcup_{i=1}^{k} H_{0i})$$

$$\leq \alpha$$

This theorem built up a basis for using maximum p value as the global p value. When applying this method to data, we can just pick up the maximum of the observed p value and access significance from here. However, this method is way too conservative in practice. For example, from a simulation composed of 2 local null distribution, the empirical distribution of maximum p value based on 1000 simulations does not follow a uniform distribution but more concentrated on the large p value region as shown in Figure 3-4A. If we choose $\alpha = 0.05$ then the size of combined test is 0.005 based on the simulation result. Another example is if one of the local distributions is null (p value is p_{H_0}), another is alternative (p_{H_A}) as shown in Figure 3-4B. The distribution of maximum p value depends on how large the effect size is under the local alternative distribution. If the effect size is extremely large and all the p values under alternative distribution is close to 0 ($p_{H_A} = 0$), the maximum p values would be equal to the p values from local null distribution ($maxp = p_{H_A}$). In this case, the distribution of maximum p value would be approximately uniform. Otherwise, the distribution of the maximum p value still skews to the large value region. The size of the test in Figure 3-4B is 0.024 at a significance level of 0.05.



a b

Figure 3-4: Histogram of maximum p value based on 1000 simulations. a. In each simulation, there are 2 local tests and both are under null distribution. Maximum p value is too conservative as concentrated to the large p value region. b. In each simulation, one local test is under null distribution and another one is under alternative distribution. This is one of the situations for global null distribution of intersection-union test. The distribution of maximum p value is again cumulated more to the larger values.

Deng et al. noticed this problem and proposed an adjusted version of maximum p value, which is called Relaxed IUT (RIUT) (Deng, Xu, & Wang, 2007). The key idea is to compute the true family wise error rate of the combined test and adjust the observed maximum p value to reduce the conservative behavior. They formulated the adjusted maximum p value expression under the condition with only two local tests and independency between these two tests were assumed (Deng, Xu, & Wang, 2007).

$$FWER = Pr\left(Reject\ H_{01}\ and\ H_{02} \middle| at\ least\ one\ H_{0i}\ is\ true\right)$$

$$= Pr\left(max\left(P_1, P_2\ \right) \leq \alpha \middle| at\ least\ one\ H_{0i}\ is\ true\right)$$

$$= Pr\left(P_1 \leq \alpha,\ P_2 \leq \alpha\ \middle| at\ least\ one\ H_{0i}\ \ is\ true\right)$$

$$= \frac{\Pr\left(P_1 \le \alpha,\ P_2 \le \alpha,\ \left(H_{01}\ or\ H_{02}\right)\right)}{\Pr\left(H_{01}\ or\ H_{02}\right)}$$

$$= \left. \begin{matrix} \Pr\left(P_1 \le \alpha,\ P_2 \le \alpha, H_{01}, H_{02}\right) \\ +\Pr\left(P_1 \le \alpha,\ P_2 \le \alpha, H_{01}, H_{A2}\right) \\ +\Pr\left(P_1 \le \alpha,\ P_2 \le \alpha, H_{A1}, H_{02}\right) \end{matrix} \right/ (1 - \Pr(H_{A1})\Pr(H_{A2}))$$

$$= \left. \begin{matrix} \Pr(P_1 \le \alpha|\mathrm{H}_{01})\Pr(P_2 \le \alpha|\mathrm{H}_{02})\Pr(H_{01})\Pr(H_{02}) \\ +\Pr(P_1 \le \alpha|\mathrm{H}_{01})\Pr(P_2 \le \alpha|\mathrm{H}_{A2})\Pr(H_{01})\Pr(H_{A2}) \\ +\Pr(P_1 \le \alpha|\mathrm{H}_{A1})\Pr(P_2 \le \alpha|\mathrm{H}_{02})\Pr(H_{A1})\Pr(H_{02}) \end{matrix} \right/ (1 - \Pr(H_{A1})\Pr(H_{A2}))$$

$$= \frac{\alpha^2\pi_1\pi_2 + \alpha(1-\beta_2)\pi_1(1-\pi_2) + (1-\beta_1)\alpha(1-\pi_1)\pi_2}{1-(1-\pi_1)(1-\pi_2)}$$

where α is the size for each local test, π_1 is the probability of null distribution in the first test, π_2 is the probability of null distribution in the second test, β_1 is type II error rate for the first test, β_2 is type II error rate for the second test. This formula is to calculate the true family wise error rate based on the global null hypothesis in IUT. For example, if $\alpha = 0.05$, the probability of local null distribution for test 1 and test 2 are both equal to 0.5, type II error rate for test 1 and test 2 are both 0 (must be a very sensitive test and extremely large effect size), then the true FWER is 0.034. If we plug in the observed maximum p value from local tests into the formula, what we are estimating is the true family wise error rate defined based on using maximum p value as the global p value, which we call “adjusted maximum p value”.

$$Adjusted\ maxp = \frac{maxp_{obs}^2\pi_1\pi_2 + maxp_{obs}(1-\beta_2)\pi_1(1-\pi_2) + (1-\beta_1)maxp_{obs}(1-\pi_1)\pi_2}{1-(1-\pi_1)(1-\pi_2)}$$

Now we have 4 parameters to estimate: $\pi_1, \pi_2, \beta_1, \beta_2$. Deng et al. suggested to adopt a method in adjusting multiple comparisons to estimate the probability of null hypothesis, which was proposed by Storey and Tibshirani in 2003 (Storey & Tibshirani, 2003). The formula to estimate π_i is

$$\widehat{\pi_i} = \frac{\#\{\, p_i(j) > \lambda\}}{m(1-\lambda)}, i = 1,2, \dots, k; j = 1,2, \dots, m$$

Type II error rate, the probability of not rejecting the test when the alternative hypothesis is true, is very difficult to estimate without any information of the effect size in alternative distribution. Therefore, Deng et al. simply set $\beta_1 = 0$ and $\beta_2 = 0$ for both simulation study and real data study in their paper.

Relaxed IUT (RIUT) is an effective method to drag those large p values to the smaller region and make maximum p value less conservative. More details will be shown in the simulation study in Chapter 4.

The third method to deal with intersection-union test was proposed by Kui Shen and his colleagues in 2010 (Shen & Tseng, 2010). Their goal is to combine results of gene set analysis across different data sets from different tissues. Only those consistent results are reliable, therefore, they are dealing with IUT problem. Instead of adopting the maximum p value as a global p value, they treated it as a statistic and performed meta-analysis on it. However, the global hypothesis given in the paper is not the same as IUT:

$$HS_A\colon \{H_0\colon \theta_{1g} = \cdots = \theta_{Kg} = 0 \; versus \; H_A\colon \theta_{kg} \neq 0, \forall 1 \leq k \leq K\}$$

The global alternative hypothesis is an intersection of local alternative hypotheses. However, the global null hypothesis is not a union of local null hypotheses but instead is an intersection again. This hypothesis does not consist of the overall parameter space, and the test based on it cannot well control the type I error rate. For example, if we have two tests, and only one of them is significant, the p value of meta-analysis calculated based on the hypothesis described above would be very small and suggests rejection of the global null hypothesis. At the same time, we want to declare significance only if both tests give evidence of rejection, therefore, the example given above becomes a false positive.

In our study, we would like to define the gene pathways that show enough evidence of differentially expressed in both upstream factors and downstream factors. Therefore, IUT would be an appropriate test, and the Relaxed IUT proposed by Deng et al. can well control the type I error and give reasonably good results. We will adopt it for our combined methods in the next chapter.

Chapter 4 The Combined Method

4.1 Main Goal and Significance of the Study

Various methods have been developed for gene set analysis in the past decade and we have reviewed three most popular and representative methods in Chapter 2 and Chapter 3. However, for both self-contained tests or competitive tests, these methods were designed based on existing gene set database including only upstream factors which are defined in Chapter 1. *Upstream factors* usually include receptors, enzymes, and transcriptional factors, which are very likely to have protein level modification instead of enormous transcription level alteration upon activation of the pathway. However, the target genes of this pathway, we call *downstream factors*, are the components expected to have dramatic gene expression level change in different phenotype if the pathway is really differentially expressed. Therefore, it is sound to add this critical information from downstream factors into gene set analysis.

To include downstream factors, there are two possible choices:

(1) Combine upstream factors and downstream factors into a whole gene set, and apply a gene set analysis method on this new defined larger gene set;

(2) Apply gene set analysis methods on upstream factors and downstream factors separately; then combine the results from these two tests.

We studied both, but found the second combined method works better. The main reasons may be as follows:

(1) The data structures differ significantly between upstream factors and downstream factors.. For example, genes in the upstream gene sets are those functioning interactively, while target genes in the downstream gene sets are those sharing same transcription factor binding

sites. In a certain cellular context, one pathway may only regulate a limited number of target genes, even though it has a large number of potential target genes across the whole genome. Hence, in a differentially expressed gene pathway, upstream factors are supposed to have similar expression pattern while only a small percent of the downstream factors are really activated or inhibited. If we pool all the genes from both upstream and downstream together to form a new gene set, it would be a mixture of two types of data, which may lead to loss of power in statistical analysis. Instead, if we analyze these two subgroups individually, appropriate methods could be chosen according to divergent distributions, resulting in power improvement.

(2) Gene pathways are very complicated and interconnected. For example, in Figure 4-1, both upstream factors 1 and upstream factors 2 can regulate one given downstream factors. In one situation, upstream factors 1 are activated while in another cell type, upstream factors 2 are turned on. In this case, if both upstream and downstream factors are grouped together to form a new gene set,if would be difficult to distinguish these two pathways. Therefore, we will have more confidence to declare that a gene pathway is differentially expressed only if we gather enough evidence of significance from both upstream factors and downstream factors.

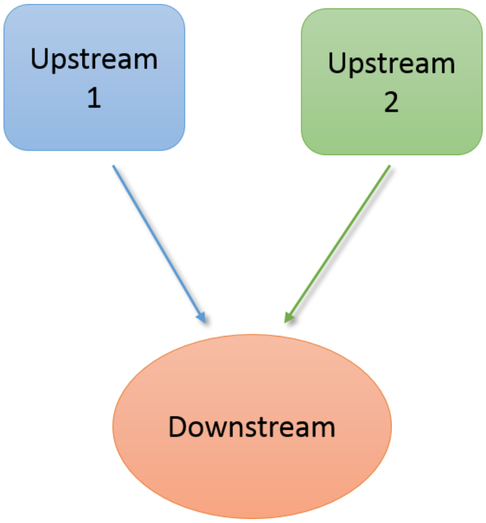


Figure 4-1: Example of complex gene pathway. Both upstream factors 1 and upstream factors 2 are regulating the given downstream factors but in different physiological environment.

4.2 Method

Our proposed gene set analysis method is a combined method. Given the different features in upstream factors and downstream factors, we propose applying GSEA on upstream factors, CDA on downstream factors, and using Relax IUT (RIUT) to combine the p values from these two tests to obtain a global p value for significance assessment (Figure 4-2).

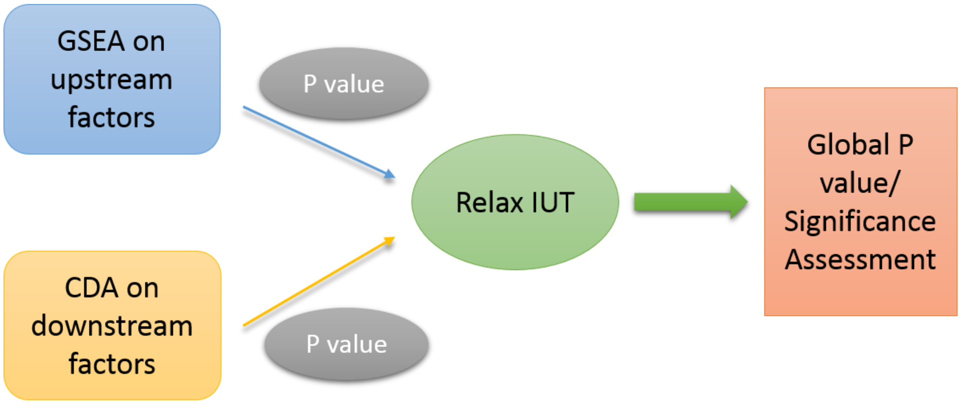

GSEA on upstream factors
P value
CDA on downstream factors
P value
Relax IUT
Global P value/ Significance Assessment

Figure 4-2: Flowchart of our proposed combined method of gene set analysis.

We chose GSEA for upstream factors because this method is more designed to detect the concordant pattern of expression change in a given gene set. We would declare significance if more components from receptors to transcriptional factors show evidence of differentially expression. On the other hand, we assume a pathway might be turned on as long as a few target genes are differentially expressed. Therefore, a self-contained method is suitable for downstream factors, and CDA is the most powerful one in a variety of scenarios.

In the following sections, we will first compare GSEA, PCA and CDA under different simulated data structures. Then we will compare the combining methods by using different combinations of existing methods. We will also show how our proposed combining method outperforms the overall method which treats the upstream factors and downstream factors as a big gene set.

4.3 Simulations

4.3.1 Simulation for PCA

We simulated two scenarios with all parameters set to be identical except for correlation structure (Table 4-1). In simulation 1, we set all the differentially expressed genes into a correlation cluster with correlation equal 0.5, while the correlation between differentially expressed genes and non-differentially expressed genes, and the correlation among non-differentially expressed genes were set to be zero. In simulation 2, we set all the non-differentially expressed genes into a correlation cluster, as opposite to the design of simulation 1. Each scenario was simulated 1000 times..

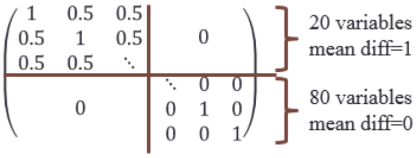
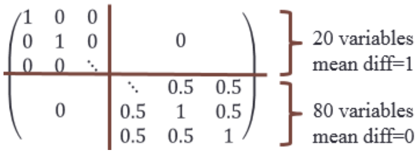
Parameters	Simulation 1	Simulation 2
Sample size	20	20
Number of pathway	1	1
Number of genes in the pathway	100	100
Number of DEG in the pathway	20	20
Difference in group mean	1	1
Variance	uniform (0.1, 10)	uniform (0.1, 10)
Correlation	 $\begin{pmatrix} \begin{matrix}1&0.5&0.5\\0.5&1&0.5\\0.5&0.5&\ddots\end{matrix} & 0 \\ 0 & \begin{matrix}\ddots&0&0\\0&1&0\\0&0&1\end{matrix}\end{pmatrix}$	 $\begin{pmatrix} \begin{matrix}1&0&0\\0&1&0\\0&0&\ddots\end{matrix} & 0 \\ 0 & \begin{matrix}\ddots&0.5&0.5\\0.5&1&0.5\\0.5&0.5&1\end{matrix}\end{pmatrix}$

Table 4-1: Parameters for simulations of illustrating the limitation of PCA.

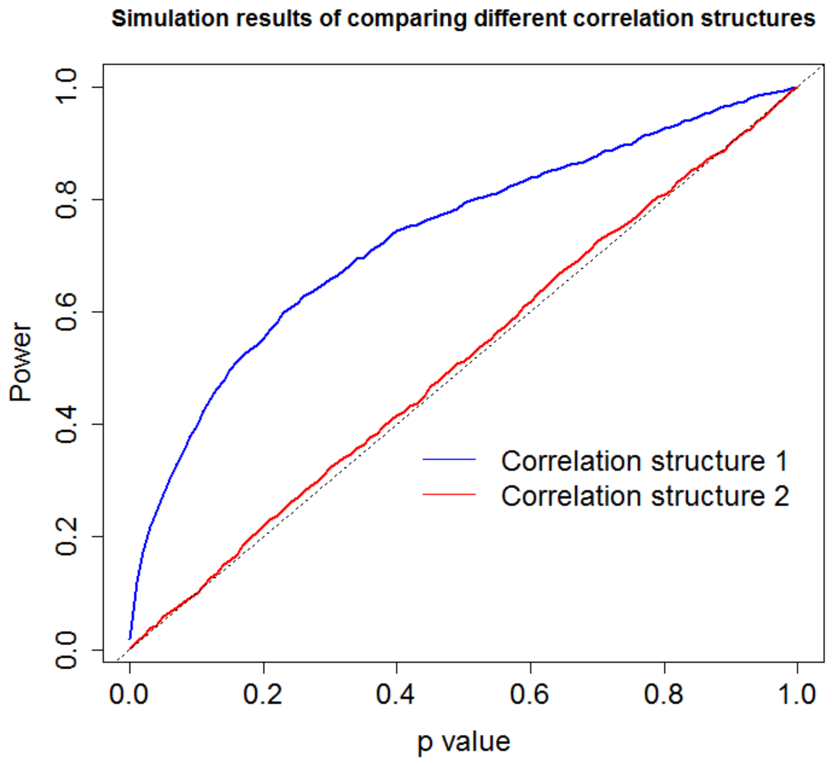


Figure 4-3: Result of simulations comparing PCA under two different correlation structures. The blue line corresponds to simulation 1, where all differentially expressed genes are in a correlation cluster with correlation equal 0.5.The red line is from simulation 2, where all non-differentially expressed genes are in the correlation cluster. X axis is the cut-off p value to declare significance, while Y axis shows the corresponding power based on 1000 simulations.

The result is shown in Figure 4-3, where the blue line corresponds to simulation 1 and the red line is from simulation 2. When all differentially expressed genes are in the correlation cluster with correlation equal to 0.5, PCA can identify the differentially expressed pathway with a power of 0.275 if we control the size to be 0.05. The power is not impressing because we only adopted the first PC in the regression. What is worse is that, if only the non-differentially expressed genes are in the correlation cluster, PCA has a power of 0.064 to identify the differentially expressed pathway at a level of 0.05, which is not much better than a random decision when there is no information of the gene expression level provided. Although the scenarios are somehow extreme, we can still have a sense of how PCA could fail under certain circumstance. Therefore, we do not suggest using PCA in gene set analysis unless one can ascertain that there is no special correlation structure that may affect the result tremendously. This simulation also confirmed the two limitations of PCA we discussed in Chapter 2: no consideration of outcome and weights of each gene decided merely by correlation structure

4.3.2 Simulation for GSEA

The following settings were designed for both GSEA and CDA. We followed the designs from references (Tsai & James, 2009) (Ackermann & Strimmer, 2009), and added more combinations to study GSEA and CDA more thoroughly.

Simulation settings:

Each generated gene set consists of $p = 100$ genes and $n = 40$ samples (20 in control group, 20 in case group). Genes in control group were simulated with random means $\mu_0 \sim uniform(0,10)$, and random variance $\sigma_0^2 \sim uniform(1,10)$. Genes in case group were simulated with mean $\mu_1 = \mu_0 + \Delta\mu$, with a shift from μ_0, and variance σ_1^2. We considered various scenarios detailed as follows.

- Background 1 (genes not in the gene set): 100 genes were simulated with $\Delta\mu = 0$ and $\sigma_1^2 = \sigma_0^2$, which indicates no differentially expressed gene exists.
- Background 2 (genes not in the gene set): 10000 genes were simulated with $\Delta\mu = 0$ and $\sigma_1^2 = \sigma_0^2$, which indicates no differentially expressed gene exists.
- Background 3 (genes not in the gene set): 10000 genes were simulated, the first 2500 genes with $\Delta\mu \sim Normal(0,1)$, the rest of the genes with $\Delta\mu = 0$, and $\sigma_1^2 = \sigma_0^2$. This means 25% of the backgrounds genes were differentially expressed.

Set	Gene set size	Percent of DEGs	Number of DEGs	Mean difference	Correlation	Identical Covariance

						Matrix
S1A	100	10%	10	0	0	Yes
S2A	100	10%	10	0.2	0	Yes
S3A	100	10%	10	0.5	0	Yes
S4A	100	10%	10	1	0	Yes
S5A	100	10%	10	0	0.25	Yes
S6A	100	10%	10	0.2	0.25	Yes
S7A	100	10%	10	0.5	0.25	Yes
S8A	100	10%	10	1	0.25	Yes
S9A	100	10%	10	0	Block 0.25*	Yes
S10A	100	10%	10	0.2	Block 0.25	Yes
S11A	100	10%	10	0.5	Block 0.25	Yes
S12A	100	10%	10	1	Block 0.25	Yes
S13A	100	10%	10	0	0.5	Yes
S14A	100	10%	10	0.2	0.5	Yes
S15A	100	10%	10	0.5	0.5	Yes
S16A	100	10%	10	1	0.5	Yes
S17A	100	10%	10	0	Block 0.5*	Yes
S18A	100	10%	10	0.2	Block 0.5	Yes
S19A	100	10%	10	0.5	Block 0.5	Yes
S20A	100	10%	10	1	Block 0.5	Yes
S21A	100	50%	50	0	0	Yes
S22A	100	50%	50	0.2	0	Yes
S23A	100	50%	50	0.5	0	Yes
S24A	100	50%	50	1	0	Yes
S25A	100	50%	50	0	0.25	Yes
S26A	100	50%	50	0.2	0.25	Yes
S27A	100	50%	50	0.5	0.25	Yes
S28A	100	50%	50	1	0.25	Yes
S29A	100	50%	50	0	Block 0.25	Yes
S30A	100	50%	50	0.2	Block 0.25	Yes
S31A	100	50%	50	0.5	Block 0.25	Yes
S32A	100	50%	50	1	Block 0.25	Yes
S33A	100	50%	50	0	0.5	Yes
S34A	100	50%	50	0.2	0.5	Yes
S35A	100	50%	50	0.5	0.5	Yes
S36A	100	50%	50	1	0.5	Yes
S37A	100	50%	50	0	Block 0.5	Yes
S38A	100	50%	50	0.2	Block 0.5	Yes
S39A	100	50%	50	0.5	Block 0.5	Yes
S40A	100	50%	50	1	Block 0.5	Yes
S41	100	25%, 25%	25, 25	0	0	Yes
S42	100	25%, 25%	25, 25	0.2, -0.2	0	Yes
S43	100	25%, 25%	25, 25	0.5, -0.5	0	Yes

S44	100	25%, 25%	25, 25	0.75, -0.75	0	Yes
S45	100	25%, 25%	25, 25	1, -1	0	Yes
S46	100	35%, 15%	35, 15	0	0	Yes
S47	100	35%, 15%	35, 15	0.2, -0.2	0	Yes
S48	100	35%, 15%	35, 15	0.5, -0.5	0	Yes
S49	100	35%, 15%	35, 15	0.75, -0.75	0	Yes
S50	100	35%, 15%	35, 15	1, -1	0	Yes
S51	500	5%, 5%	25, 25	0	0	Yes
S52	500	5%, 5%	25, 25	0.2, -0.2	0	Yes
S53	500	5%, 5%	25, 25	0.5, -0.5	0	Yes
S54	500	5%, 5%	25, 25	1, -1	0	Yes
S55	500	5%, 5%	25, 25	2, -2	0	Yes
S56	500	5%, 5%	25, 25	3, -3	0	Yes
S57	500	5%, 5%	25, 25	5, -5	0	Yes

Table 4-2: Parameters of simulations for GSEA and CDA. Block correlation is defined as that correlation $\rho = 0.25\ or\ 0.5$ only exists between DEGs but not any other two genes.

Set S1B ~ S40B shared the same parameters as their corresponding A sets except that the variance in control group was not identical to that in case group. These B sets were designed to study the effect of unequal covariance matrix on GSEA and CDA.

As described in previous chapter, GSEA is designed based on competitive assumption but significance assessment is instead based on subject sampling, which makes it more conservative in detecting significance. In the following simulation study, we showed how the power of GSEA would be affected by fold change, percent of DEG in gene set, mixture of both up-regulated and down-regulated genes, correlation between genes, and percent of DEG in background genes.

<u>*Results:*</u>

(1) Correlation between genes in a given gene set reduces the power of GSEA, especially when the correlation between differentially expressed genes and non-differentially expressed genes is not negligible (Figure 4-4). All gene sets in the plot include 10% of DEGs and the mean difference increases from 0 to 1. Size was set to be 0.05. GSEA showed largest power when

there was no pairwise correlation between genes in the gene set (red solid line). Power decreased dramatically when correlation among all genes exist (blue and black dotted lines). This pattern is consistent when the percentage of DEG increases to 50% and background genes increases to 10000 (Figure 4-5).

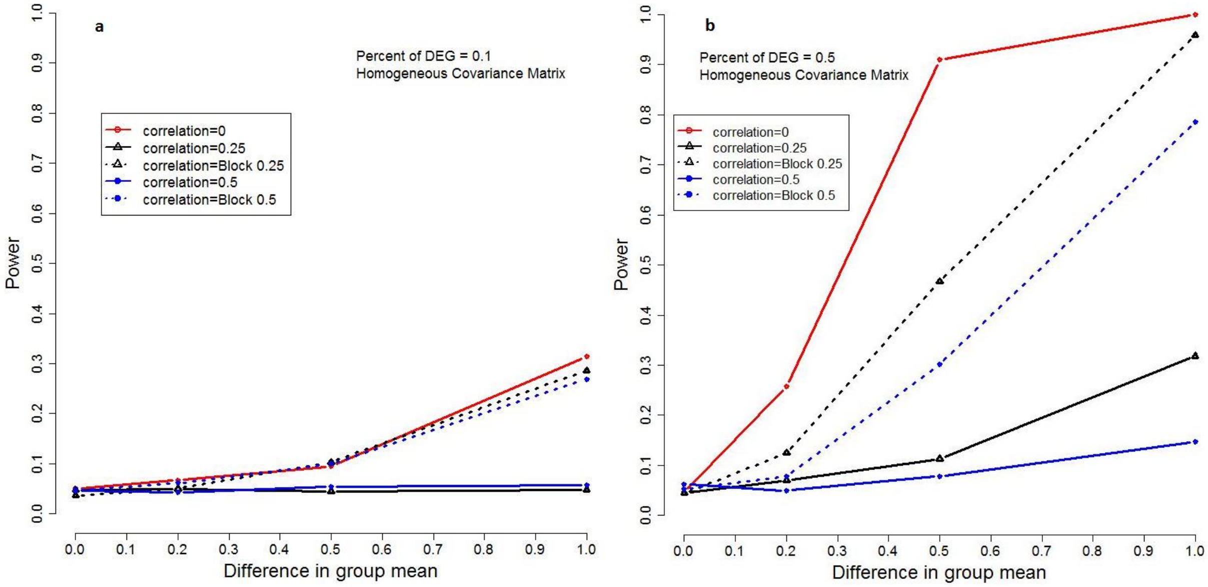


Figure 4-4: Simulation results of GSEA on gene set with 10% of DEGs (a) and 50% of DEGs (b). Red solid line corresponds to gene set with no pairwise correlation; black solid line correponds to gene set with pairwise correlation of 0.25; black dotted line corresponds to gene set with pairwise correlation of 0.25 between DEGs but no correlation between any other 2 genes; blue solid line represents gene set with $\rho = 0.5$ while blue dotted line corresponds to gene set with block correlation structure of $\rho = 0.5$. 100 background genes (background 1) were considered in the simulation. Results were based on 1000 simulations. $\alpha = 0.05$.

(2) Higher percentage of differentially expressed genes corresponds to higher power of GSEA (Figure 4-4 and Figure 4-5). This is reasonable because more DEGs ranking higher in the rank list lead to stronger evidence of differentially expression of the gene set.

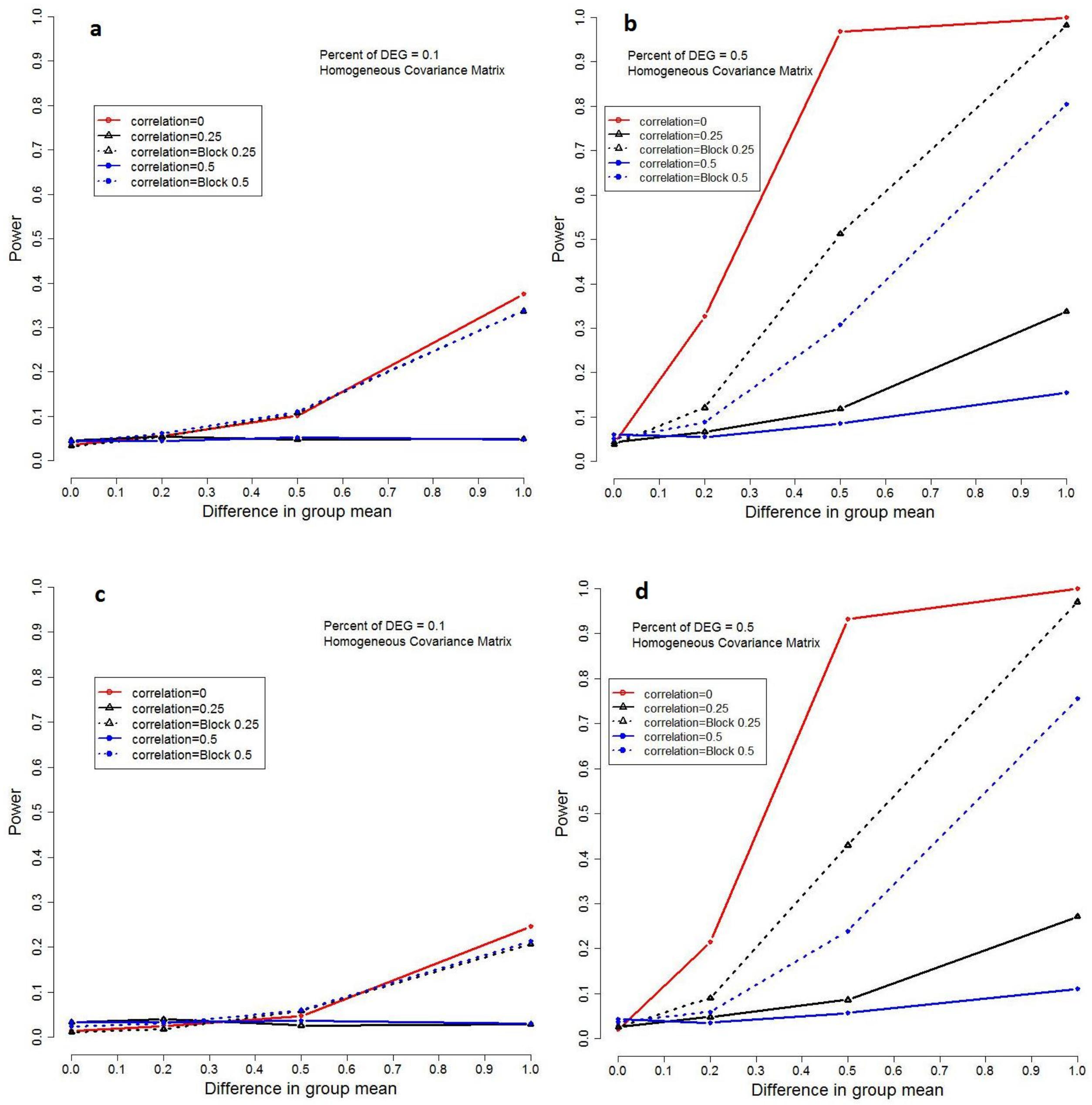


Figure 4-5: Simulation results of GSEA on gene set with 10% of DEGs (a, c) and 50% of DEGs (b, d). Red solid line corresponds to gene set with no pairwise correlation; black solid line correponds to gene set with pairwise correlation of 0.25; black dotted line corresponds to gene set with pairwise correlation of 0.25 between DEGs but no correlation between any other 2 genes; blue solid line represents gene set with $\rho = 0.5$ while blue dotted line corresponds to gene set with block correlation structure of $\rho = 0.5$. 10000 background genes with no DEGs (background 2) were considered in a and b; while 10000 background genes with 2500 DEGs (background 3) were considered in c and d. Results were based on 1000 simulations. $\alpha = 0.05$.

(3) Larger number of background genes improves the power of GSEA slightly. In Figure 4-

4, both results were based on competition to 100 non-differentially expressed background genes, while in Figure 4-5 a and b, results were from a rank list of 10000 non-differentially expressed background genes. When the background list increased from 100 to 10000, power increased slightly from 0.257 to 0.326 in the gene set with 50% DEGs, $\Delta\mu = 0.2$ and $\rho = 0$. This improvement became negligible when $\Delta\mu$ increased to 1. We also tested the effect of DEGs in the background list to the given gene set. Figure 4-5 c and d gave results from simulations based on background genes with 25% of DEGs. If we compare results in c and d to those in a and b, respectively, we can see that power decreased when DEGs exist in background list. It is also as expected since GSEA compares the genes in the given gene set to those genes not in the gene set. If more DEGs exist in the background list, it requires more evidence of differentially expression for the genes in given gene set to be identified as significant by GSEA. 25% of DEGs in the background list is quite realistic as we will see in Chapter 5.

(4) Mixture of up-regulated and down-regulated genes dramatically reduced the power of GSEA (Figure 4-6). This is an obvious limitation for GSEA because genes in a given gene set are not guaranteed to be regulated in a consistent direction in the real data. A quadratic transformation of the gene level scores (e.g. squared of t score) seems to be a solution to improve GSEA (Ackermann & Strimmer, 2009). We compared the quadratic version and original version of GSEA under 3 scenarios: 1) 100 percent of the DEGs were up-regulated (Set 21A~24A), 2) 70 percent of the DEGs were up-regulated and 30 percent were down-regulated (Set 46~50), 3) 50 percent of the DEGs are up-regulated and the rest 50 percent were down-regulated (Set 41~45). Original GSEA was most powerful when all the DEGs changed in the same direction but failed to identify the tested gene set when mixture of up- and down-regulated genes existed. Quadratic GSEA did save some power in the extreme case where half of the DEGs

were up-regulated and half were down-regulated. However, this method resulted in much lower true positive rates than original GSEA when majority of the genes were in concordance (red and blue lines). In general, the quadratic transformed GSEA was not a promising alternative to improve GSEA and we did not recommend it as an alternative in gene set analysis.

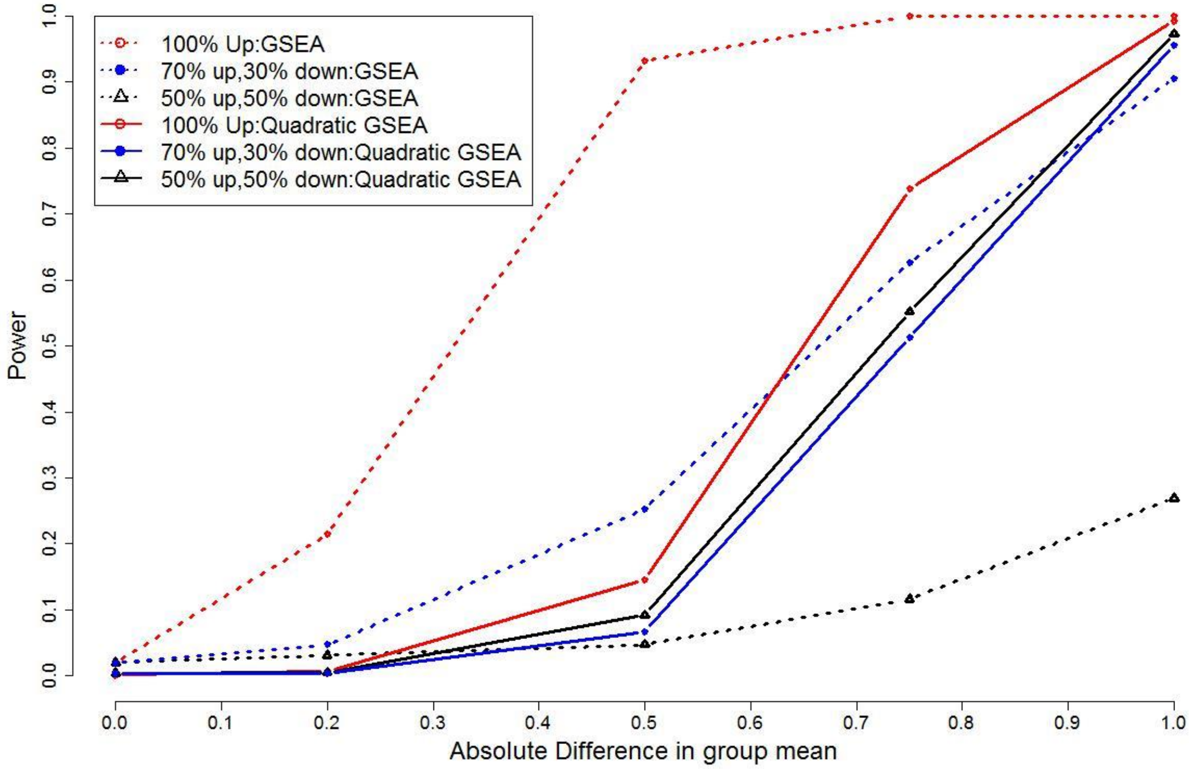


Figure 4-6: Simulation results of GSEA and Quadratic GSEA on gene sets of both up- and down-regulated genes. Dotted lines represented GSEA; solid lines corresponded to Quadratic transformed GSEA. Red was for gene set with up-regulated genes only; blue was for gene set with 70% up-regulated and 30% down regulated genes; black was for gene set with 50% up-regulated and 50% down-regulated genes.

In summary, GSEA is most powerful when

(1) Large percent of genes in the gene set are differentially expressed;

(2) Fold changes of the DEGs are at least moderate and in consistent direction;

(4) Pairwise correlations between genes are small;

(5) No or small percent of DEGs exist in the background list.

4.3.3 Study of CDA

CDA was designed based on self-contained assumption, and the subject-sampling permutation suits the assumption perfectly, which promises high power of CDA in a variety of scenarios (Tsai & James, 2009). One key assumption for CDA, which is easily ignored by users, is the homogeneous within-group covariance matrices. Violation of this assumption may result in large bias if we adopt a parametric method to assess significance. An unequal covariance can be easily incorporated by permutation test in practice. In the following series of simulation studies, we showed how the unequal covariance affects the power of CDA (Figure 4-7).

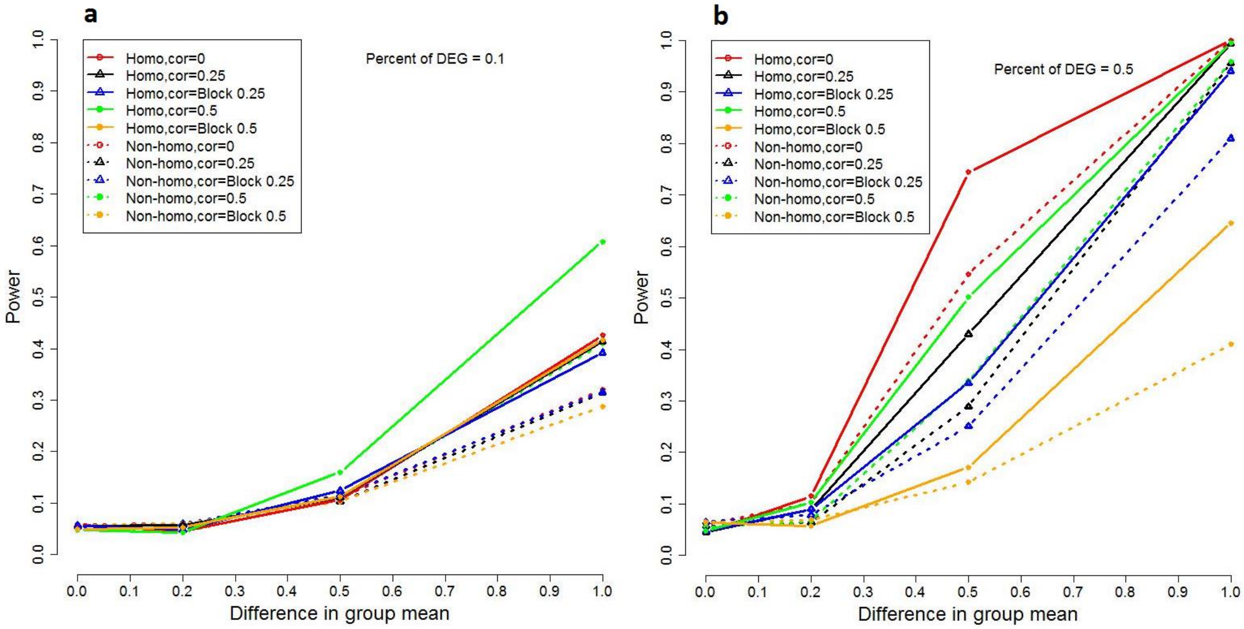


Figure 4-7: Simulation results of CDA on gene set with 10% of DEGs (a) and 50% of DEGs (b). Solid lines represented the gene sets with homogeneous covariance among groups; Dotted lines corresponded to gene sets with unequal covariance among groups. Different colors corresponded to different pairwise correlation. Red: No correlation; Black: $\rho = 0.25$; Blue: $\rho = block\ 0.25$; Green: $\rho = 0.5$; Orange: $\rho = block\ 0.5$. Results were based on 1000 simulations. $\alpha = 0.05$.

We noticed several interesting observations from these simulation results:

(1) CDA yielded larger power in gene sets with homogeneous within-group covariance,

which is as expected since the within-group covariance matrix is pooled across all groups. Two groups with more homogeneous covariance structure are better distinguished by CDA.

(2) In general, stronger correlation reduced the power of CDA, if the calculation was calculated based on empirical sample covariance matrix. However, when the percentage of DEGs is small (10% in Figure 4-7 a,), CDA detected most true positives in the gene sets with pairwise correlation of 0.5. This is due to the shrinkage covariance estimator (Schafer & Strimmer, 2005) we chose for computing CDA. The block correlation structure reduced the power tremendously, especially when more DEGs existing in the gene set. In real data shown in previous section, pairwise correlations between genes are too weak to severely affect the power of CDA.

4.3.4 Comparison of PCA, GSEA and CDA

In this section, we will show the simulation results of comparisons among CDA, PCA and GSEA in different scenarios where we changed the mean difference, percentage of DEGs in gene set, and correlation structure between genes. In Figure 4-8 and Figure 4-9, gene sets contained up-regulated genes only while in Figure 4-10, both up-regulated and down-regulated genes exist. The background genes for GSEA were from the scenario of “background 2” (10000 genes, 2500 of them are DEGs). We only showed results when within-group covariance matrices were homogeneous, but the pattern was similar if the within-group covariance were not equal (results not shown).

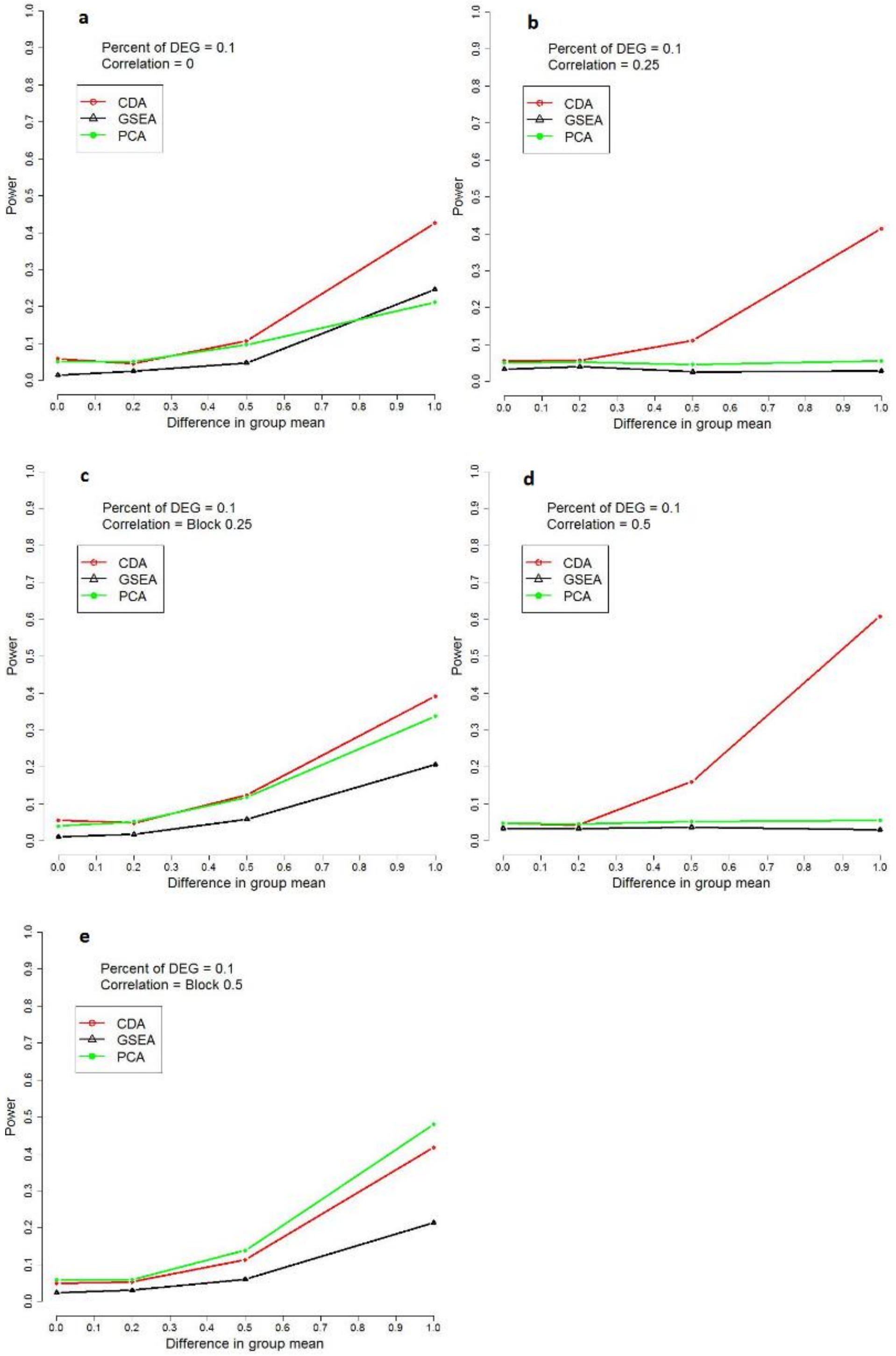


Figure 4-8: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with 10% DEGs. a. No pairwise correlation; b. $\rho = 0.25$; c. block correlation $\rho = 0.25$; d. $\rho = 0.5$; e. block correlation $\rho = 0.5$. All DEGs were up-regulated. Red line represented CDA; Black line corresponded to GSEA; Green line was from PCA. Results were based on 1000 simulations. $\alpha = 0.05$.

When the percentage of DEGs was small (10% in Figure 4-8), CDA generally outperformed the other two methods, except for the scenario where moderate correlations only existed between DEGs (Figure 4-8 e), in which PCA gave slightly larger power than CDA. This is consistent with our illustration about PCA that it gives weights according to correlation structures. We expect that PCA would be more powerful when the correlation between DEGs increases. However, this is not realistic as we will see in real data analysis.

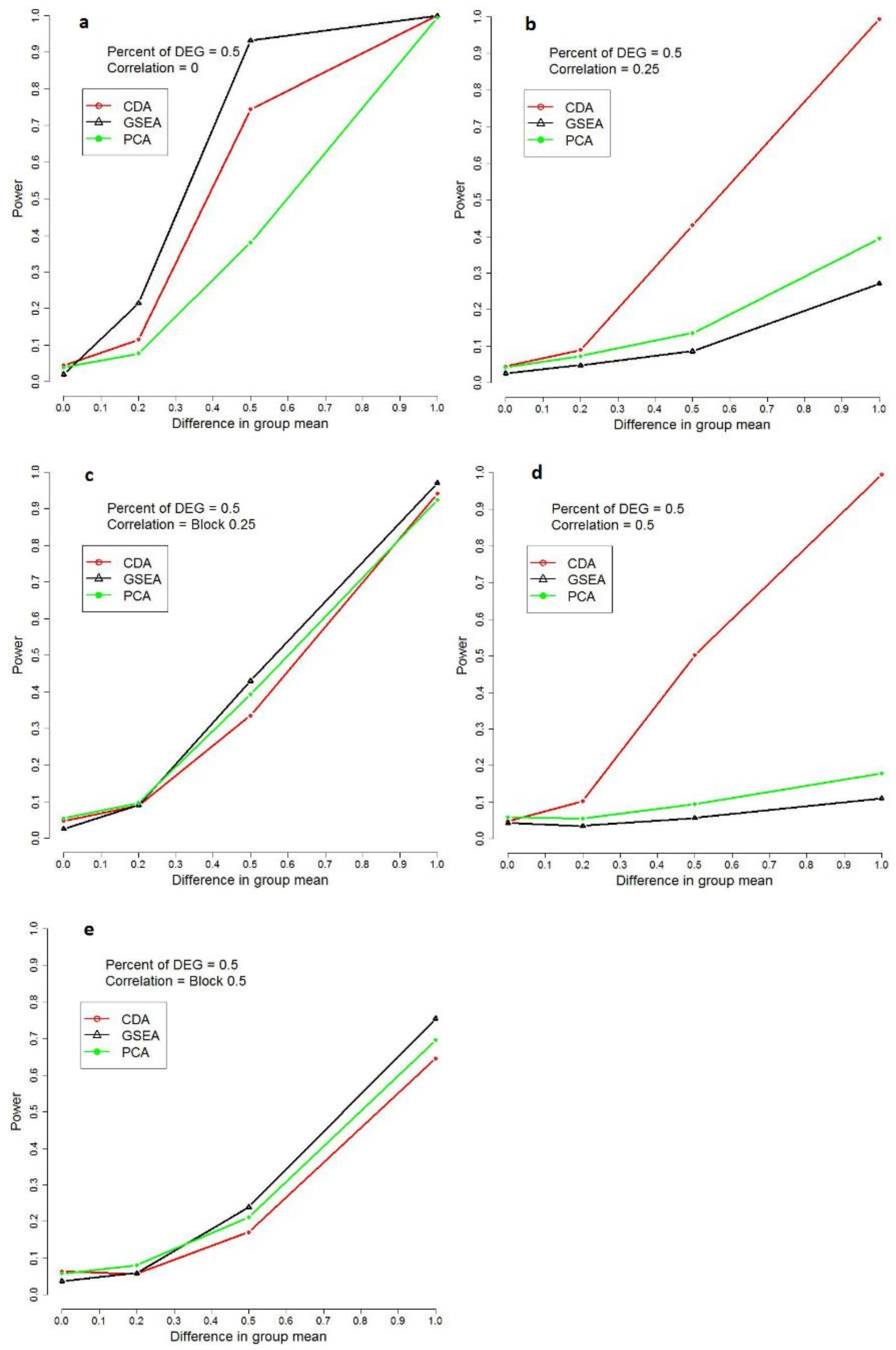


Figure 4-9: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with 50% DEGs. a. No pairwise correlation; b. $\rho = 0.25$; c. block correlation $\rho = 0.25$; d. $\rho = 0.5$; e. block correlation $\rho = 0.5$. All DEGs were up-regulated. Red line represented CDA; Black line corresponded to GSEA; Green line was from PCA. Results were based on 1000 simulations. $\alpha = 0.05$.

When the percentage of DEGs was at least moderate (50% in Figure 4-9), GSEA outperforms the other two if there was minor pairwise correlation or only correlation between DEGs (Figure 4-9 a, c, e). However, if moderate pairwise correlation existed among all genes, GSEA lost power as we showed in the previous section. At the same time, CDA was least affected by correlation structure when half of the genes were DEGs.

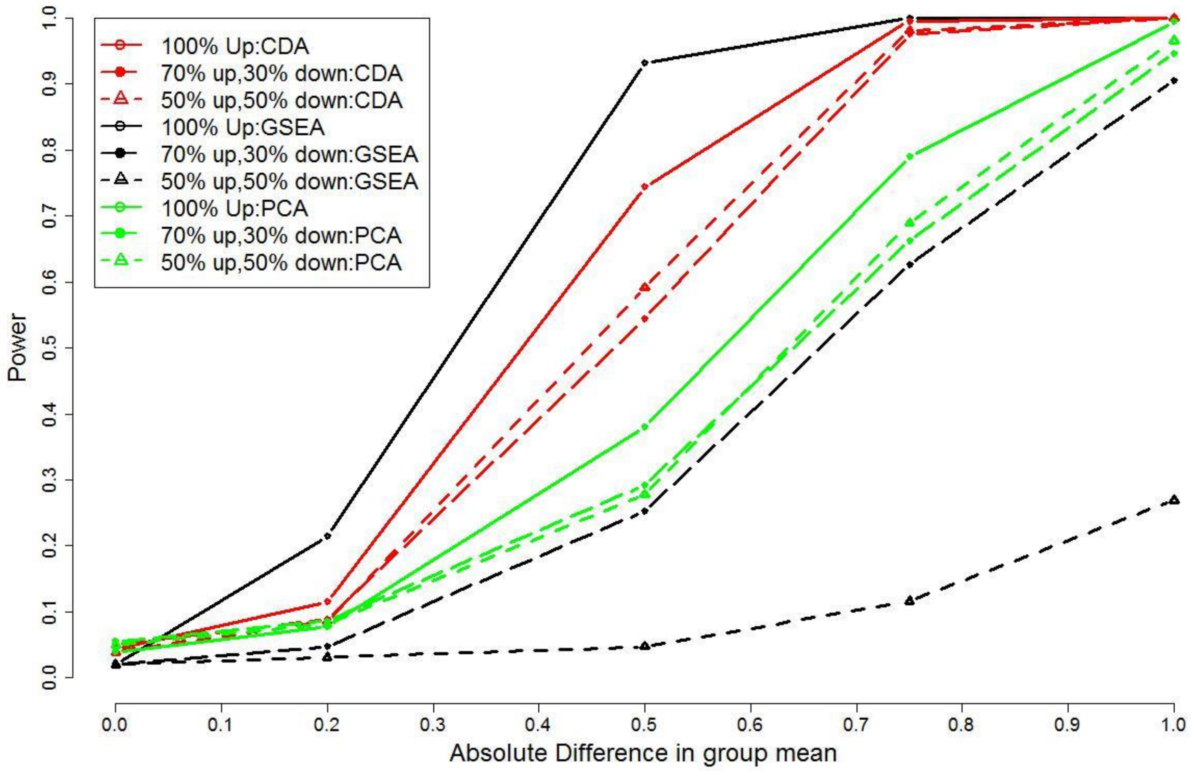


Figure 4-10: Simulation results of comparisons among CDA, GSEA, and PCA in gene sets with mixture of up-regulated and down-regulated DEGs. Solid lines were from gene sets with up-regulated DEGs only; long-dash lines were from gene sets with 70% up-regulated DEGs and 30% down-regulated DEGs; Dotted lines were from gene sets with 50% up-regulated DEGs and 30% down-regulated DEGs. Red lines represented CDA; Black lines corresponded to GSEA; Green lines were from PCA. Results were based on 1000 simulations. $\alpha = 0.05$.

Finally we compared CDA, GSEA and PCA when mixture of up- and down-regulated genes existed (Figure 4-10). We saw the poor performance of GSEA in previous section when the

DEGs were divided into two equal portions with opposite gene expression change directions. Now we can compare it to CDA and PCA in similar situations. Although the inconsistency of gene regulation direction reduced the power of both CDA and PCA slightly, these two methods still identified the true positives in a comparatively high rate. Out of these two methods, CDA gave larger power than PCA in all three scenarios.

Generally speaking, CDA is more sensitive in majority of the situations due to the concordance of its assumption and the subject sampling method to assess significance. PCA is less trustable since it tends to capture the correlated gene clusters instead of the difference in group mean. GSEA gives weakest power because the discrepancy between the “competitive” assumption and the “self-contained” permutation test.

4.3.5 Combining Upstream and Downstream Factors

In this section, we will show the simulation results of combining upstream factors and downstream factors by Relax IUT (RIUT) (Deng, Xu, & Wang, 2007). We tested different combinations of simulation settings. In general, upstream factors were chosen from gene sets 1A~40A, and downstream factors were from gene sets 51~57.

Firstly, we showed the comparison between RIUT and the overall test, which treats upstream factors and downstream factors as a whole gene set. We tried different combinations of GSEA, CDA and PCA using Set 21A and 23A as upstream set and Set 51 and 55 as downstream set. In the scenarios with either $\mu_{up} = 0$ or $\mu_{down} = 0$, the whole pathway should not be treated as differentially expressed. RIUT correctly controlled the false positive rate while overall test failed to control type I error (Table 4-3). This is as expected since overall test treated upstream factors and downstream factors as a whole set, and it would declare significance as long as there was evidence of differential expression in either part. For example, both CDA and PCA gave power of 1 when $\mu_{up} = 0$ and $absolute\ \mu_{down} = 2$ because strong DEGs existed in downstream and were capture by CDA and PCA, ignoring the fact that upstream factors were not differentially expressed at all.

μ_{up}	μ_{down}	RIUT									Overall		
		GSEA/ CDA	GSEA/ PCA	GSEA/ GSEA	PCA/ CDA	PCA/ GSEA	PCA/ PCA	CDA/ PCA	CDA/ GSEA	CDA/ CDA	GSEA	CDA	PCA
0	0	0.024	0.026	0.006	0.043	0.009	0.043	0.05	0.006	0.051	0.007	0.044	0.046
0	2 -2	0.02	0.02	0.008	0.039	0.018	0.039	0.044	0.023	0.044	0.117	1	1
0.5	0	0.056	0.052	0.005	0.042	0.007	0.034	0.038	0.004	0.042	0.116	0.242	0.119
0.5	2 -2	0.932	0.932	0.183	0.38	0.108	0.38	0.75	0.155	0.932	0.555	1	1

Table 4-3: Simulation results of power in RIUT and overall test. $\alpha = 0.05$. Each estimate was based on 1000 simulations.

Secondly, we showed the power of different combinations of methods on various upstream and downstream matches: the upstream factors of 100 genes with 10% DEGs or 50% DEGs and no pairwise correlation, the downstream factors of 500 genes with 10% DEGs, half of which were up-regulated while the other half were down-regulated. When the percent of DEGs was small (10%) in upstream set, the combination using CDA for upstream factors and CDA for downstream factors (CDA/CDA) gave slightly larger power than all other combinations (Table 4-4). However, when the percent of DEGs increased to 50%, GSEA/CDA combination outperformed the other methods, especially when mean difference was moderate in upstream ($\mu_{up} = 0.5$) and moderate to large in downstream ($\mu_{up} = 1$) (Table 4-5).

μ_{up}	μ_{down}	RIUT								
		GSEA/ CDA	GSEA/ PCA	GSEA/ GSEA	PCA/ CDA	PCA/ GSEA	PCA/ PCA	CDA/ PCA	CDA/ GSEA	CDA/ CDA
0	0	0.019	0.023	0.004	0.039	0.011	0.048	0.041	0.009	0.04
0	0.2 -0.2	0.023	0.023	0.008	0.04	0.01	0.053	0.047	0.008	0.05

0	0.5 -0.5	0.008	0.016	0.005	0.027	0.012	0.048	0.052	0.016	0.03
0	1 -1	0.013	0.013	0.012	0.049	0.034	0.041	0.046	0.025	0.055
0	2 -2	0.014	0.014	0.009	0.051	0.024	0.051	0.058	0.027	0.058
0.2	0	0.028	0.029	0.005	0.056	0.011	0.055	0.059	0.01	0.046
0.2	0.2 -0.2	0.028	0.032	0.01	0.036	0.012	0.042	0.045	0.01	0.04
0.2	0.5 -0.5	0.017	0.032	0.01	0.04	0.013	0.039	0.041	0.014	0.034
0.2	1 -1	0.024	0.022	0.023	0.048	0.034	0.05	0.034	0.04	0.041
0.2	2 -2	0.025	0.025	0.011	0.05	0.02	0.05	0.045	0.022	0.045
0.5	0	0.037	0.049	0.01	0.044	0.008	0.047	0.048	0.012	0.04
0.5	0.2 -0.2	0.035	0.049	0.012	0.041	0.008	0.045	0.043	0.006	0.048
0.5	0.5 -0.5	0.03	0.045	0.016	0.053	0.007	0.051	0.048	0.01	0.053
0.5	1 -1	0.04	0.037	0.036	0.089	0.028	0.066	0.078	0.031	0.096
0.5	2 -2	0.047	0.047	0.019	0.096	0.042	0.096	0.106	0.045	0.106
1	0	0.044	0.044	0.008	0.045	0.006	0.054	0.035	0.005	0.032
1	0.2 -0.2	0.047	0.046	0.006	0.055	0.011	0.04	0.036	0.009	0.041
1	0.5 -0.5	0.076	0.061	0.008	0.072	0.007	0.067	0.077	0.007	0.114
1	1 -1	0.226	0.157	0.028	0.194	0.032	0.13	0.224	0.02	0.398
1	2 -2	0.246	0.246	0.086	0.211	0.067	0.211	0.426	0.111	0.426

Table 4-4: Simulation results of power in RIUT in upstream factors with 10% DEGs. $\alpha = 0.05$. Each estimate was based on 1000 simulations.

μ_{up}	μ_{down}	RIUT GSEA/ CDA	 GSEA/ PCA	 GSEA/ GSEA	 PCA/ CDA	 PCA/ GSEA	 PCA/ PCA	 CDA/ PCA	 CDA/ GSEA	 CDA/ CDA
0	0	0.024	0.026	0.006	0.043	0.009	0.043	0.05	0.006	0.051
0	0.2 -0.2	0.021	0.02	0.004	0.037	0.011	0.038	0.051	0.013	0.021
0	0.5 -0.5	0.018	0.019	0.011	0.035	0.009	0.046	0.046	0.018	0.019
0	1 -1	0.02	0.017	0.02	0.038	0.032	0.037	0.037	0.036	0.02

0	2 -2	0.02	0.02	0.008	0.039	0.018	0.039	0.044	0.023	0.044
0.2	0	0.043	0.04	0.005	0.035	0.008	0.049	0.044	0.008	0.042
0.2	0.2 -0.2	0.042	0.047	0.01	0.048	0.018	0.051	0.045	0.009	0.042
0.2	0.5 -0.5	0.083	0.055	0.012	0.05	0.013	0.046	0.047	0.011	0.065
0.2	1 -1	0.206	0.158	0.028	0.073	0.026	0.058	0.083	0.027	0.11
0.2	2 -2	0.215	0.215	0.072	0.077	0.035	0.077	0.115	0.056	0.115
0.5	0	0.056	0.052	0.005	0.042	0.007	0.034	0.038	0.004	0.042
0.5	0.2 -0.2	0.063	0.053	0.009	0.043	0.005	0.043	0.047	0.009	0.057
0.5	0.5 -0.5	0.175	0.107	0.008	0.113	0.009	0.062	0.102	0.007	0.157
0.5	1 -1	0.849	0.432	0.035	0.347	0.029	0.211	0.356	0.027	0.682
0.5	2 -2	0.932	0.932	0.183	0.38	0.108	0.38	0.75	0.155	0.932
1	0	0.056	0.052	0.006	0.057	0.006	0.054	0.052	0.006	0.056
1	0.2 -0.2	0.066	0.057	0.01	0.068	0.01	0.058	0.057	0.01	0.066
1	0.5 -0.5	0.184	0.111	0.008	0.188	0.008	0.114	0.111	0.008	0.184
1	1 -1	0.911	0.457	0.035	0.906	0.035	0.458	0.457	0.035	0.911
1	2 -2	1	1	0.2	0.995	0.196	0.995	1	0.2	1

Table 4-5: Simulation results of power in RIUT in upstream factors with 50% DEGs. $\alpha = 0.05$. Each estimate was based on 1000 simulations.

4.3.6 Summary of the Simulation Studies

In conclusion, PCA gave weights to genes mainly according to their correlation structure instead of the mean difference between groups, therefore, it lost power to detect the real differentially expressed pathway in a large variety of scenarios, and we do not recommend it for gene set analysis. CDA showed highest power in majority of the simulation scenarios, and it is least affected by correlation between genes. The hypothesis of CDA is that there is no DEG in the gene set, hence CDA would declare significance even there is only one gene showing significant change in expression level. Based on our understanding of the gene pathways, CDA is more appropriate for testing downstream factors. GSEA seemed to have weakest power compared to the other two methods, but it could still outperform CDA and PCA when moderate or large percent of the genes in the gene set showed at least moderate change in expression level in the same direction. This is consistent with our expectation for the upstream factors. Therefore, we suggest GSEA for upstream factors and CDA for downstream factors.

Relax IUT (RIUT) is essentially an adjusted version of maximum p value, which could well control the size in a less conservative range and therefore improved power than the original maximum p value. Compared to the overall test which treated upstream factors and downstream factors as a whole big gene set, RIUT could correctly control type I error when either upstream or downstream was differentially expressed. Therefore, we suggest apply RIUT for combining p values from upstream factors and downstream factors, and assess significance for each combined gene pathway.

Chapter 5 Real Data Analysis

In this section, we will apply our proposed combining method on two distinct data: the p53 and Essential thrombocythaemia (ET) data set. We will compare their results with those from the overall test treating upstream and downstream as one big gene set.

5.1 Materials

The p53 dataset and ET dataset used in our study were referred to the Oliver study (Olivier, et al., 2002) and Bahou study (Gnatenko, et al., 2005), respectively. The p53 dataset was based on NCI-60 collection of cancer cell lines, including samples from 17 wild-type cells with normal p53 status and 33 mutated cells with p53 mutants. The tumor suppressor gene p53 is a transcription factor involved in almost all human cancers. The data was from HGU95Av2 chip, which includes 12,625 probe sets. For each sample, the expression value of a given gene is represented by taking the maximum value of all probe sets for that gene. After probe reduction, expression values of 10100 genes were obtained. P53 dataset can be downloaded from the Developmental Therapeutics Program web site (http://dtp.nci.nih.gov/mtargets/download.html). The goal of study is to identify functional gene sets correlated to p53 mutation.

ET dataset from the Bahou study could be downloaded from GEO website (GSE2006): http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2006. ET is a rare chronic blood disorder that is recognized by the overproduction of platelets. This dataset used the HGU133A chip, which include 22277 probe sets. After probe to gene matching, 12495 genes are left. In the ET dataset, 6 platelet samples were from ET patients and 5 platelet samples were from normal people. The goal of this study is to identify gene pathways that are associated with ET.

Dataset	Sample size	Number of probes	Number of genes	Platform	Year of submission	GSE ID	Link

p53	33 mutants, 17 wild type	12625	10100	Affymetrix Human Genome U95Av2	2002	NA	http://dtp.nci.nih.gov/mtargets/download.html
Essential thrombocythaemia	6 ET, 8 normal	22277	12495	Affymetrix Human Genome U133A Array	2005	GSE2006	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2006

Table 5-1: Summary of p53 dataset and ET dataset.

5.2 Database

In our analyses, for the upstream factors, we would like to adopt the C2 database (Mootha, et al., 2003) (Subramanian, et al., 2005). There were 522 gene sets with an average gene set size of 33 (Table 5-2), 472 of which contain gene sets involved in specific metabolic and signaling pathways, while another 50 of which are involved in response to genetic and chemical perturbations. For the downstream factors, we chose the publicly available JASPAR database. By scanning promoter sequences of candidate target genes, we identified 125 downstream gene sets with mean size of 4627. Since these gene sets were predicted by sequencing matching, we expect that a certain proportion of genes in the gene set may not be true targets of the given transcription factor.

Database	Number gene sets	Minimum gene set size	Median gene set size	Mean gene set size	Maximum gene set size
C2	522	2	19	33	447
JASPAR	125	1	383	782	6181

Table 5-2: Summary of gene set databases: C2 and JASPAR.

5.3 Upstream Factors and Downstream Factors

Next, we carefully studied the data structures of upstream factors defined by C2 database and downstream factors defined by JASPAR database. In order to compare with previous studies by Subramaniana in 2005 (Subramanian, et al., 2005), we set a minimum gene set size of 15 and maximum gene set size of 500 for upstream factors. According to biological meaning and computation efficiency, we set a minimum of 1 and maximum of 1000 for downstream factors. We applied two sample t test on each gene in the data set, and identify DEGs based on an arbitrary criterion of $p < 0.1$. This criterion was not strict since we only intended to explore the data structure and needed a cut-off to define genes with somewhat statistical significance. These DEGs were the ones affecting the power of GSEA or CDA, therefore, we were interested in their distribution. We studied each gene set in terms of gene set size, mean pairwise correlation, mean absolute pairwise correlation, percentage of DEGs in the gene set, percentage of up-regulated genes out of DEGs, mean absolute difference of DEGs, maximum absolute difference of DEGs, mean absolute t score of DEGs and maximum absolute t score of DEGs (Table 5-3, Figure 5-1 to Figure 5-10).

In p53 data set, about 12.7% of genes had significant difference in expression level between mutant and wild-type cells at cut-off of 0.1. ET data set had about 28.9% DEGs.

Dataset	Percent of DEGs in dataset ($p <$ 0.1)	Database	Number gene sets	Minimum gene set size	Median gene set size	Mean gene set size	Maximum gene set size
P53	12.7%	C2	308	15	25	43	358
		JASPAR	115	1	114	208	795
		C2/JASPAR	472	20	256	343	1043
ET	28.9%	C2	297	15	25	42	373
		JASPAR	104	1	140	262	994
		C2/JASPAR	348	17	201	299	1142

Table 5-3: Summary of upstream and downstream gene sets in p53 and ET.

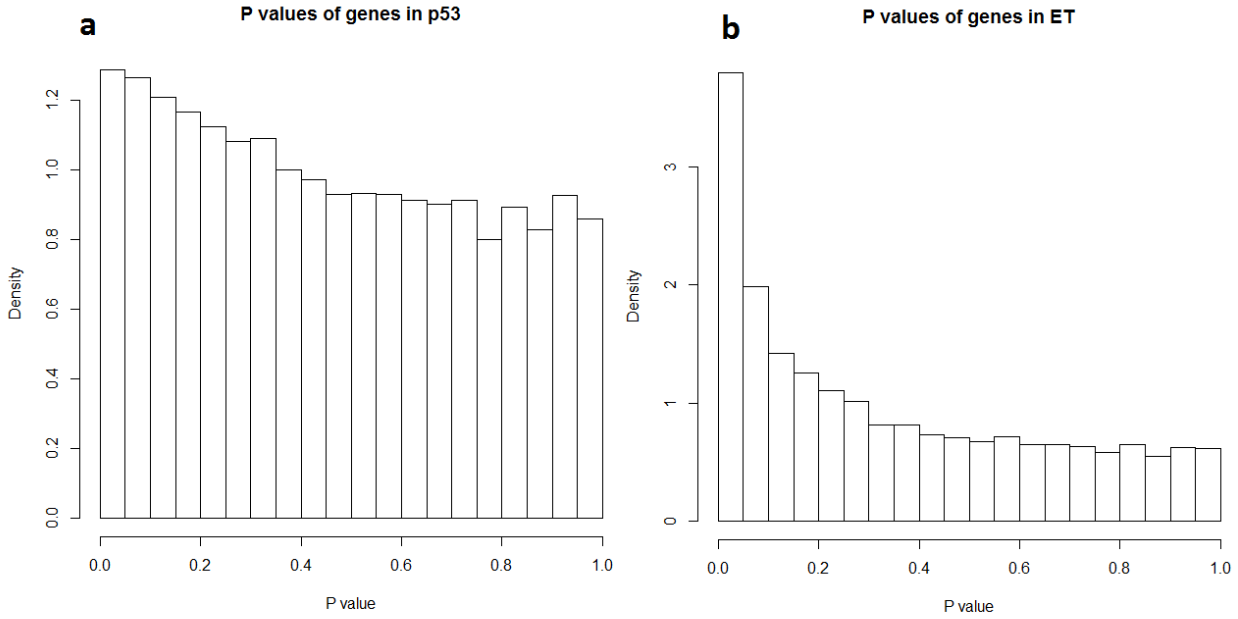


Figure 5-1: A density histogram of the p values from p53 dataset (a) and ET dataset (b).

From the comparisons between upstream factors and downstream factors, we noticed that

(1) Pairwise correlation between genes in was weak in both upstream and downstream factors (Figure 5-3 and Figure 5-4).

(2) The percentage of DEGs, defined as $\frac{\#\{p<0.1\ in\ given\ gene\ set\}}{gene\ set\ size}$, did not differ significantly between upstream gene set and downstream gene set (Figure 5-5). Majority of the upstream gene sets in p53 data set had 5% ~ 20% DEGs while the majority of the downstream gene sets in p53 data had 10% ~ 20% DEGs. Percentage of DEGs in ET data set ranged from 0% ~ 60% in upstream factors and 20% ~ 40% in downstream factors.

(3) We studied the percentage of up-regulated genes in DEGs of each gene set. This percentage would affect the performance of GSEA as shown in previous chapter. More gene sets in downstream had 50% ~ 80% of up-regulated DEGs while upstream gene sets had a percent spreading out from 0% to 100% (Figure 5-6). GSEA will lose a lot of power due to the mixture

of up-regulated and down-regulated genes in downstream gene sets, therefore, we will not apply it to downstream factors.

(4) The mean absolute group difference and mean absolute t score for each gene set did not differ too much between upstream factors and downstream factors (Figure 5-7 and Figure 5-9). However, when we considered the maximum absolute group difference and maximum absolute t score, we noticed that there are more large absolute mean difference or t score in downstream factors than in upstream factors (Figure 5-8 and Figure 5-10). This difference was more obvious in ET data set. If we recall the feature of CDA to capture at least one DEG, we can imagine that this large mean difference will benefit CDA trememdously.

Real data is very complicated, and large variation exists among different data sets. But in general, upstream factors are more in concordance with relatively moderate mean difference in DEGs. The target genes in downstream sets do not necessarily have specific relationship between each other, and the mean difference tends to be large in DEGs.

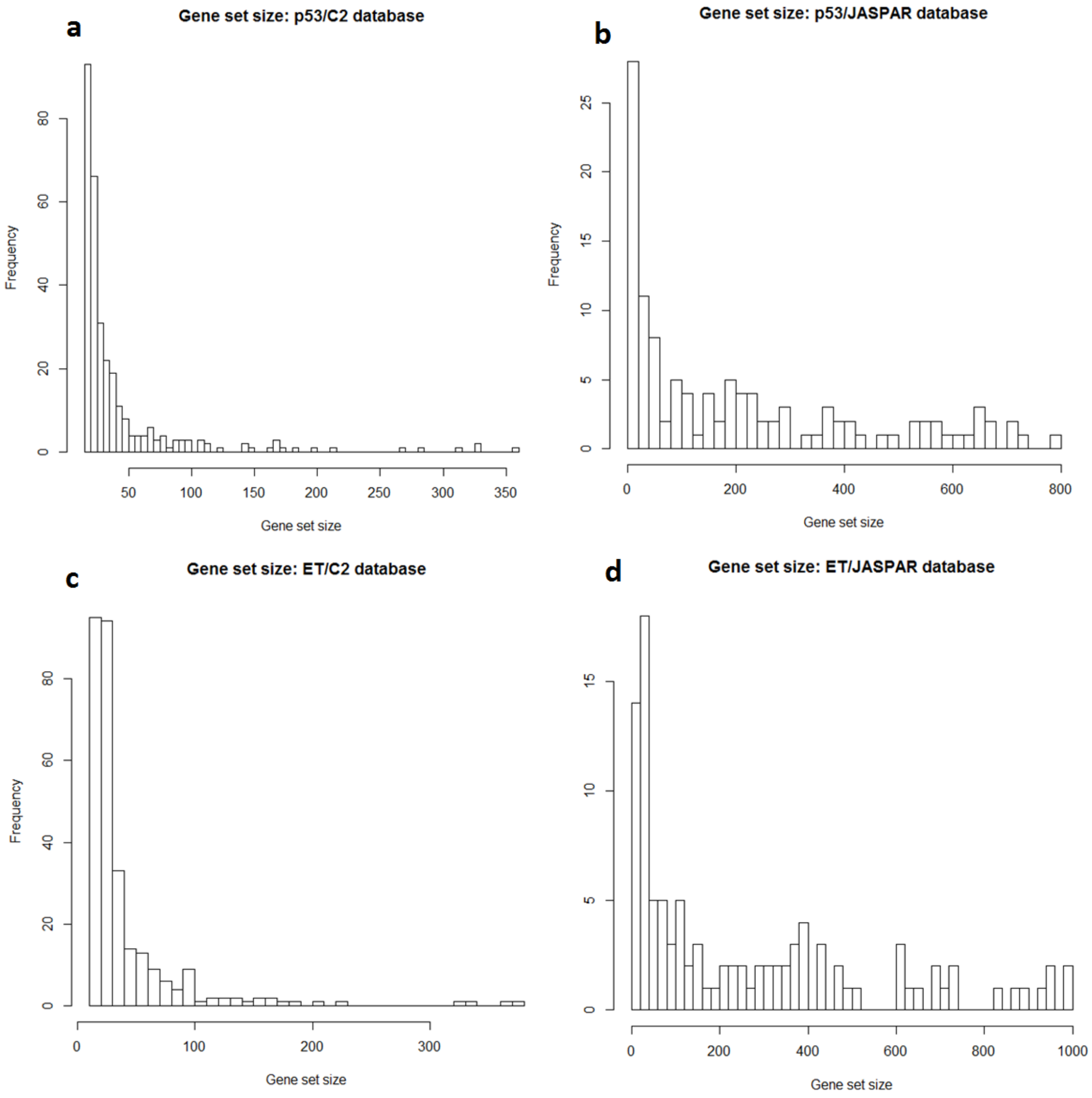


Figure 5-2: Gene set size of p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

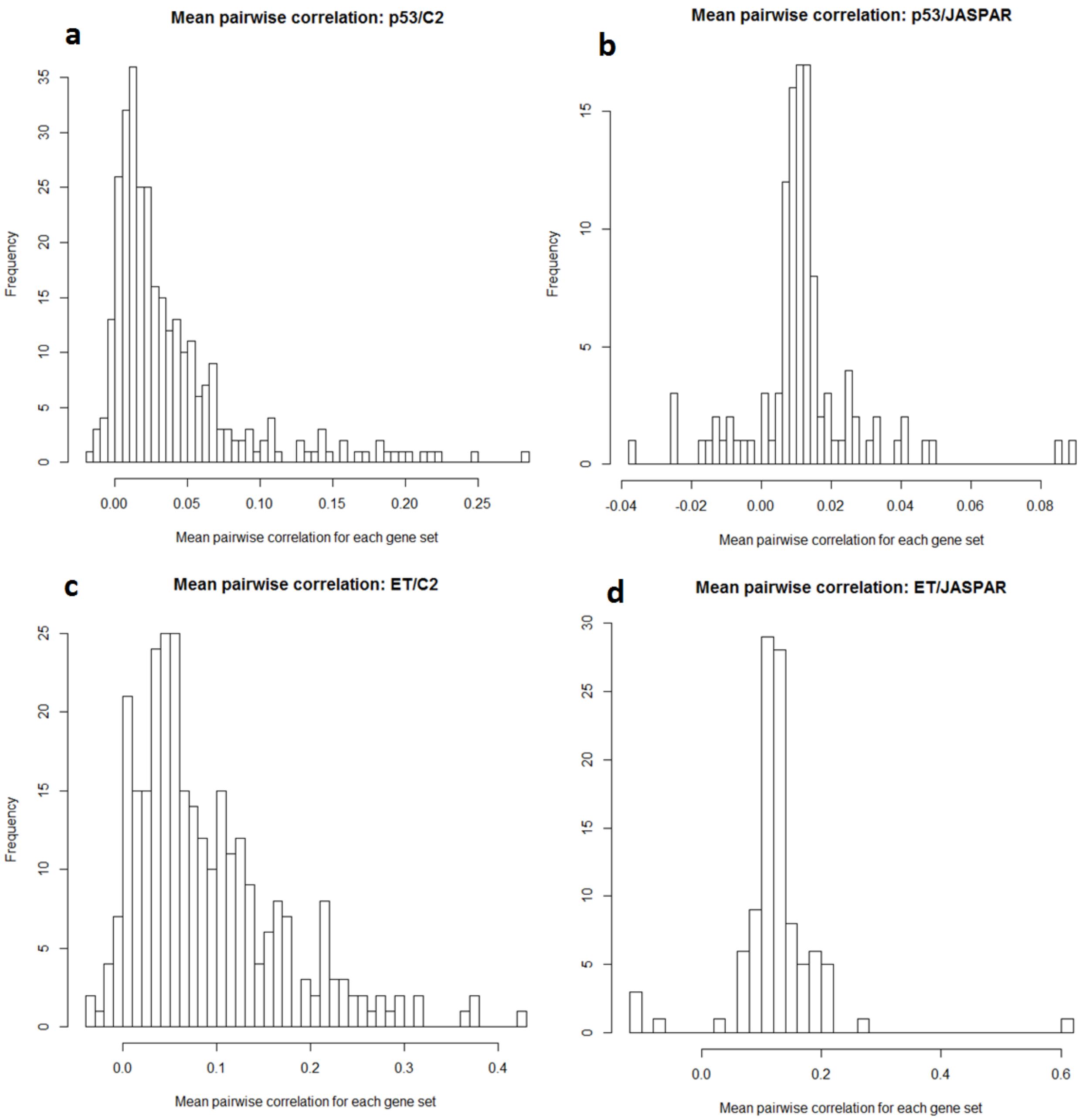


Figure 5-3: Mean pairwise correlation for each gene set in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

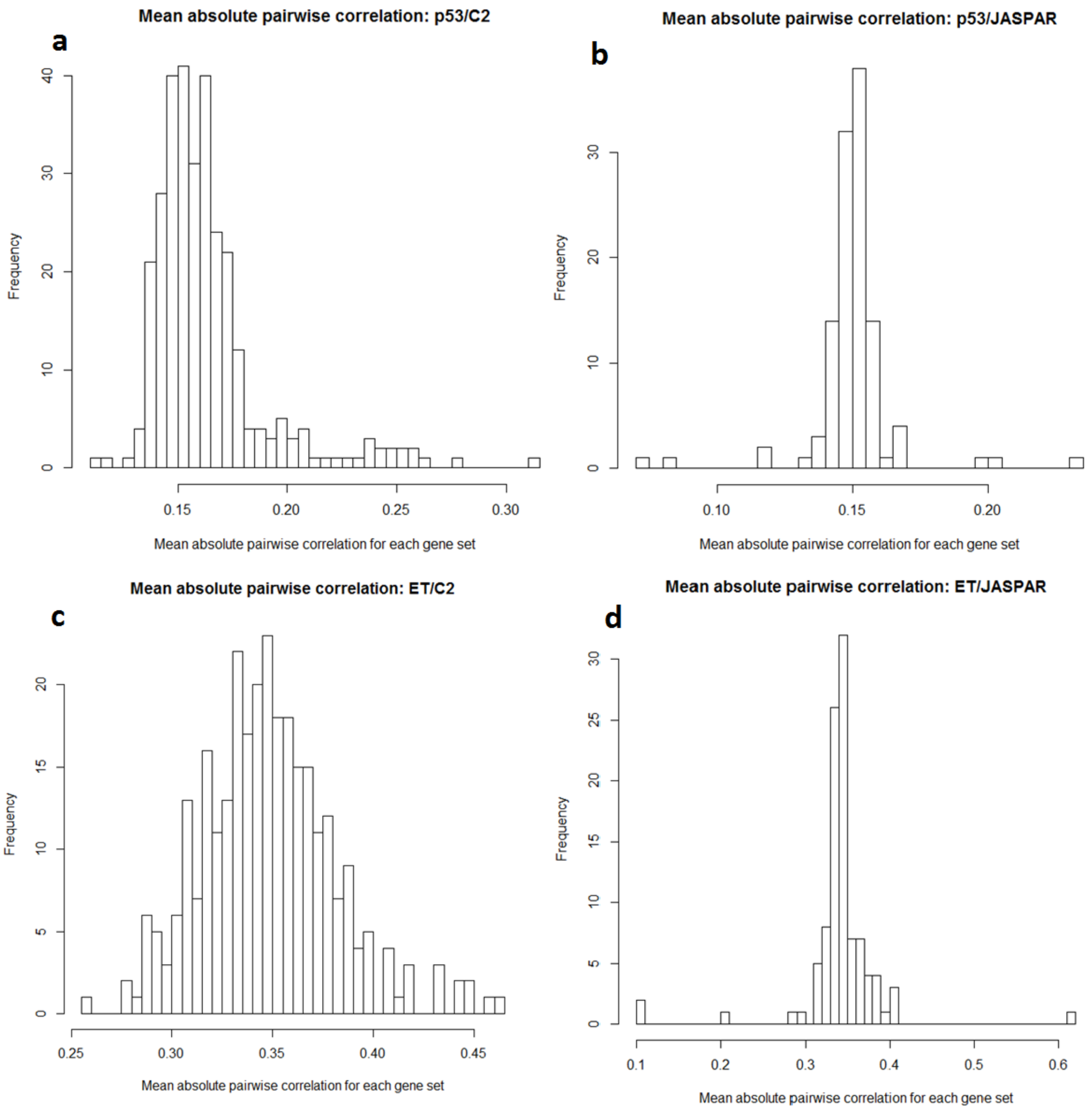


Figure 5-4: Mean absolute pairwise correlation for each gene set in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

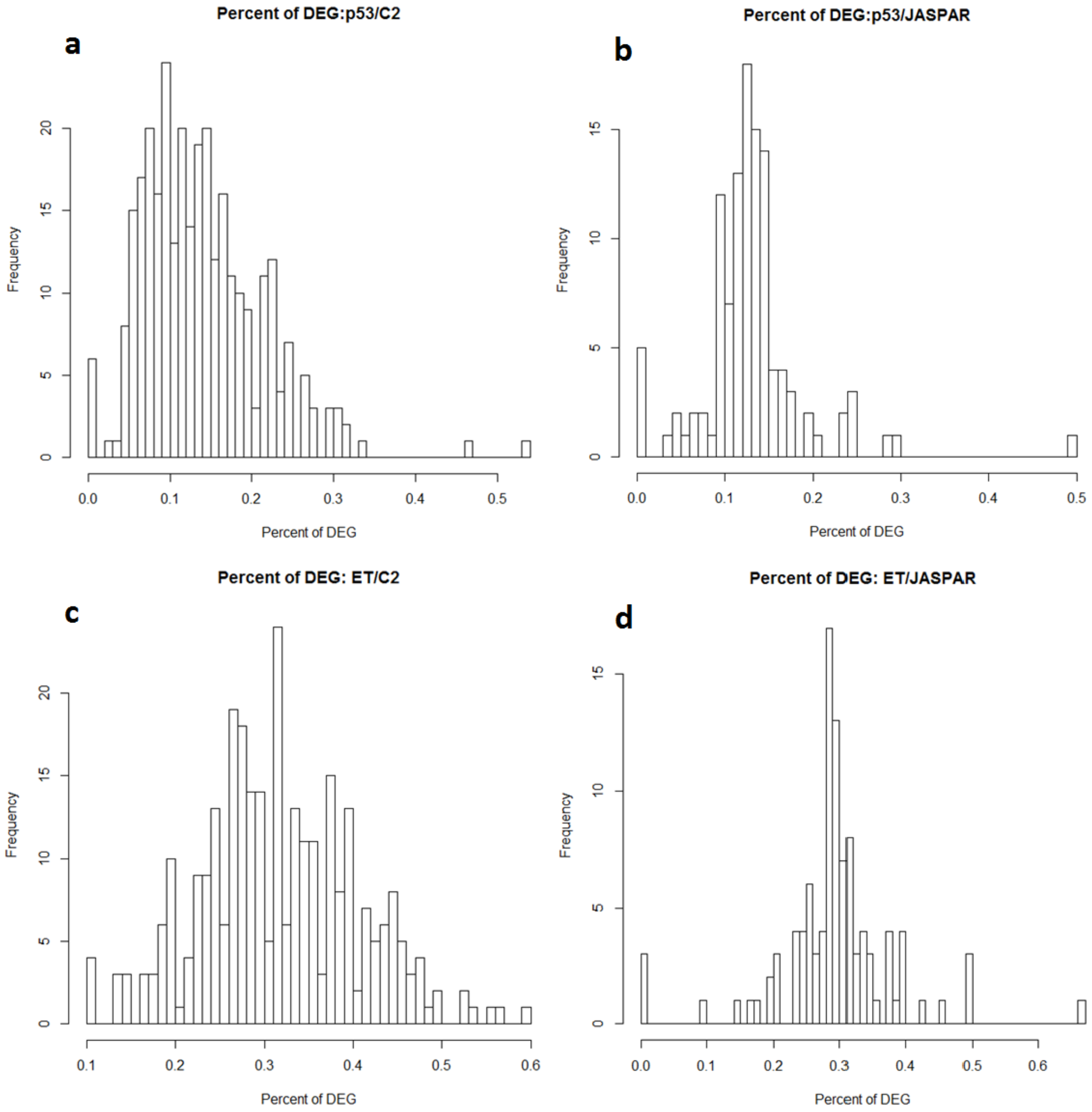


Figure 5-5: Percent of DEGs ($p < 0.1$) in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

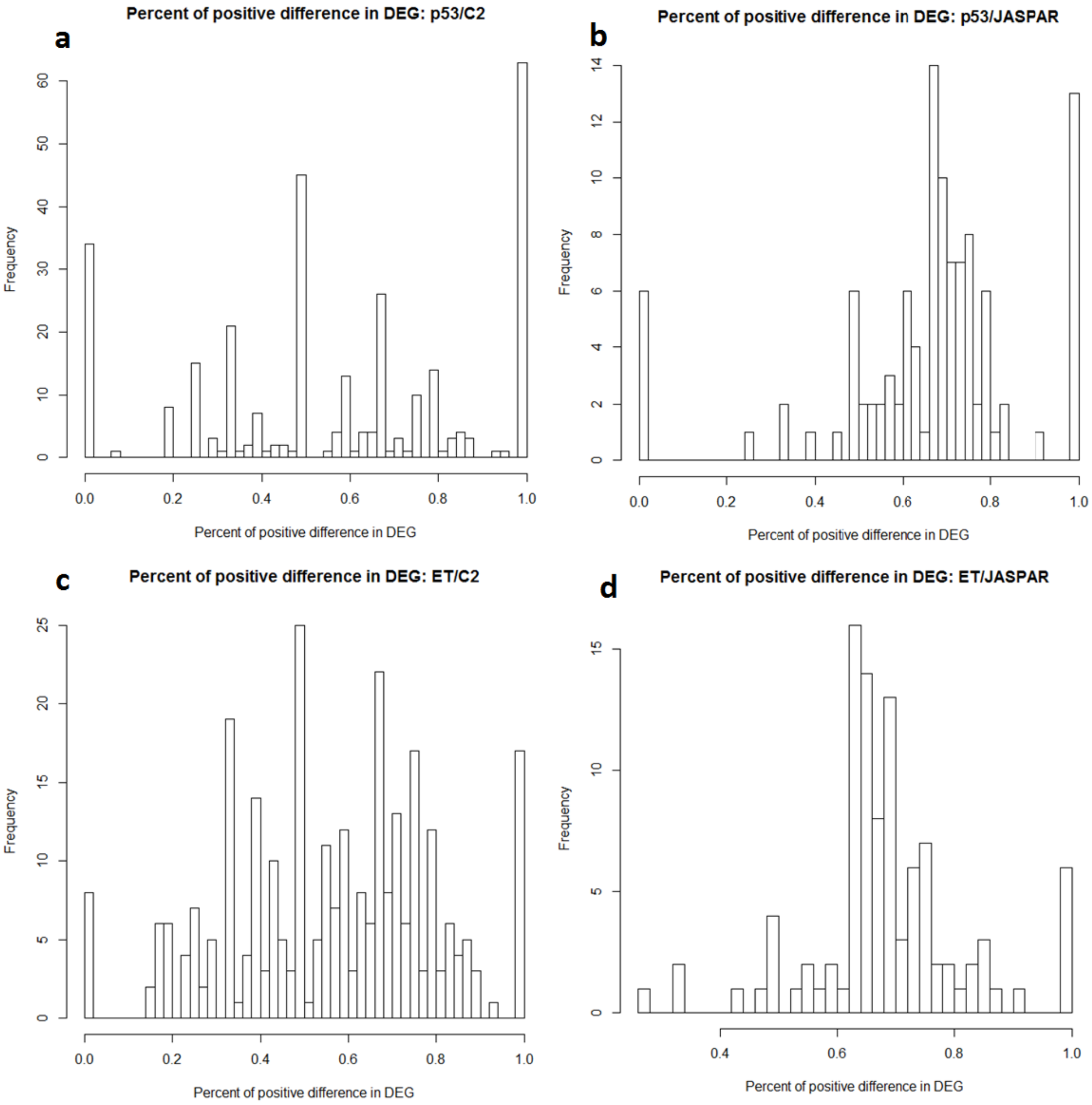


Figure 5-6: Percent of up-regulated genes out of DEGs in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

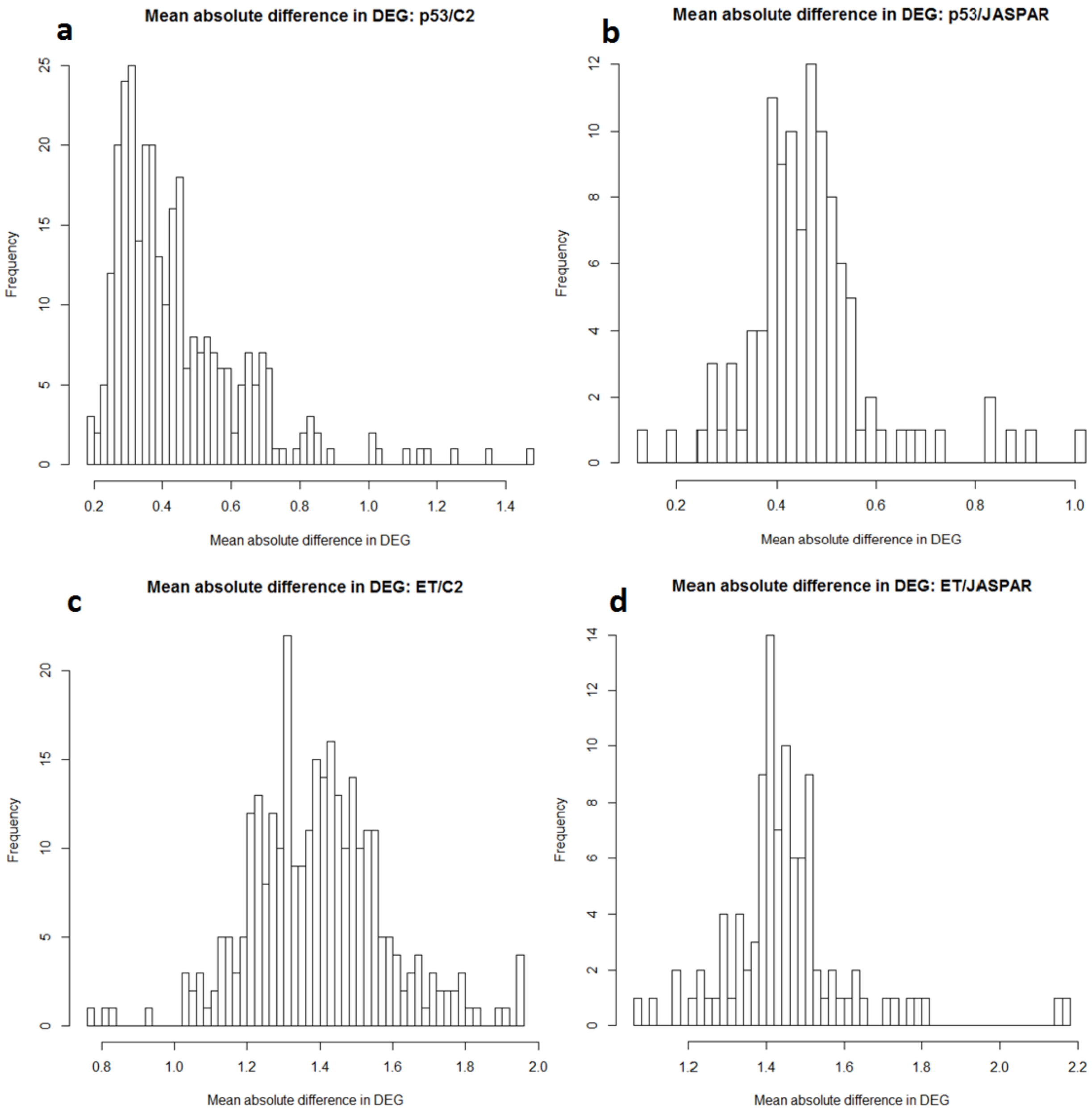


Figure 5-7: Mean absolute difference of DEGs in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

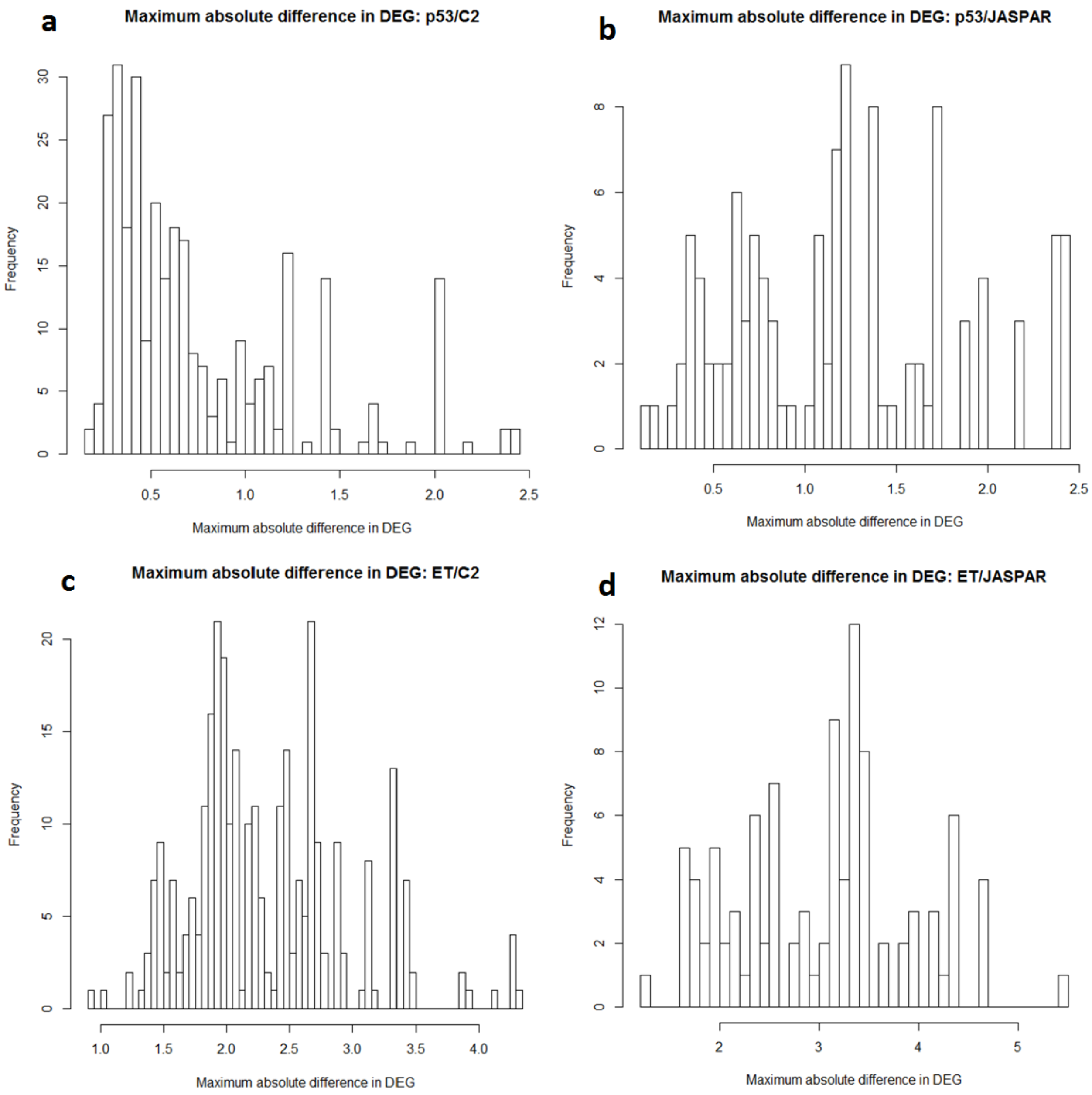


Figure 5-8: Maximum absolute difference of DEGs in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

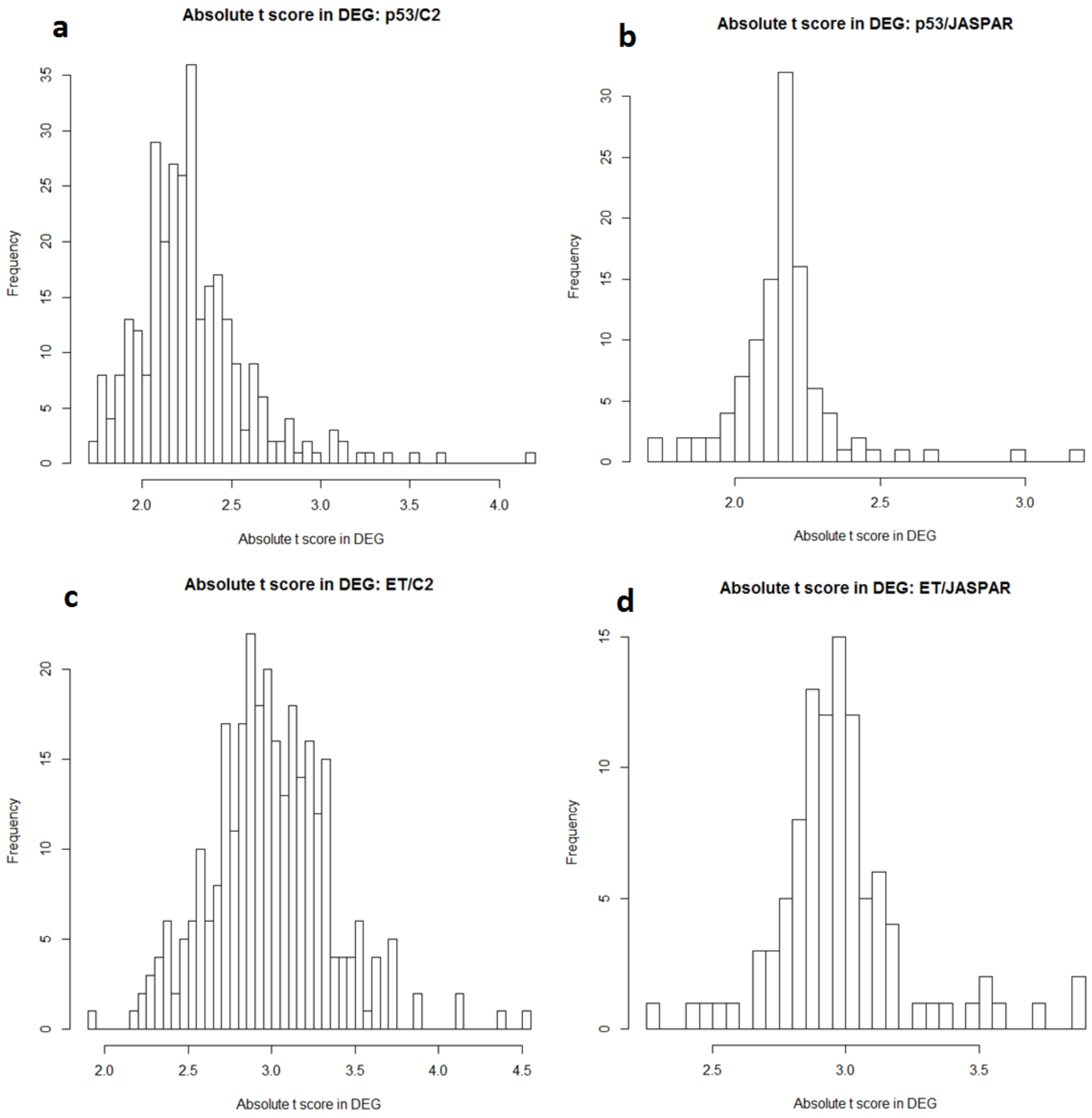


Figure 5-9: Mean absolute t score of DEGs in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

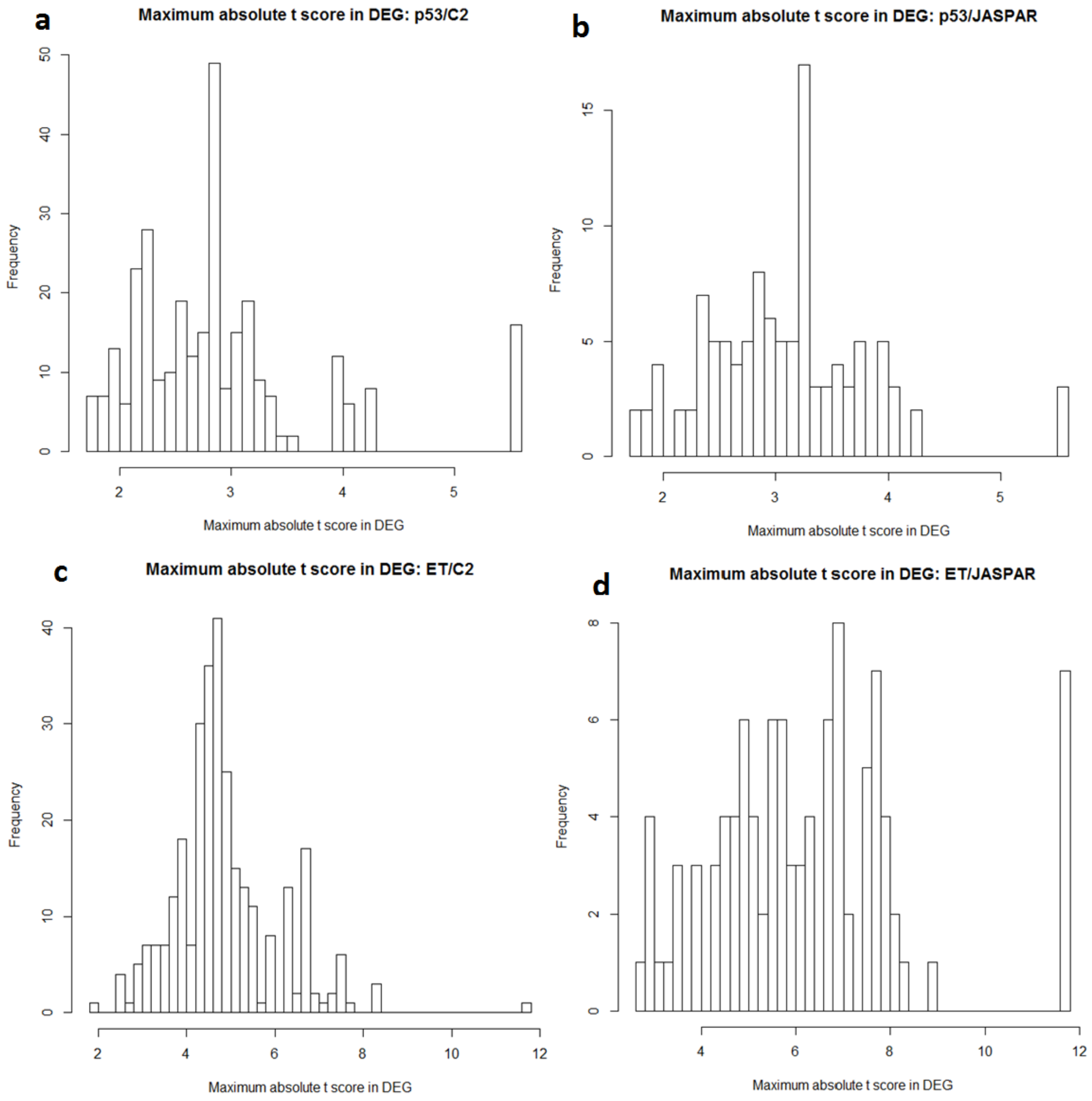


Figure 5-10: Maximum absolute t score of DEGs in p53 dataset and ET dataset. a. Upstream factors in p53 dataset based on C2 database. b. Downstream factors in p53 dataset based on JASPAR database. c. Upstream factors in ET dataset based on C2 database. d. Downstream factors in ET dataset based on JASPAR database.

5.4 Real Data Analysis

5.4.1 P53 Data Set

Table 5-4 showed the number of gene sets with q value<0.25 in p53 data set from combining test GSEA/CDA, combining test CDA/CDA, overall test GSEA and overall test CDA. This q-value criterion was considered in several publications (Subramanian, et al., 2005) (Tsai & James, 2009). We also considered the criterion p value<0.01 and 0.05 since the combining method including GSEA was too conservative to identify any gene set with q value<0.25. The overall test using CDA gave the most significant gene sets, however, this method could not control the type I error as we saw in the simulation studies in Chapter 4. The combining method CDA/CDA identified 14 significant gene sets compared to 0 identified by GSEA/CDA.

Combinations (472)	p value < 0.01	p value < 0.05	q value < 0.25
Combining test GSEA/CDA	3	8	0
Combining test CDA/CDA	15	32	14
Overall test GSEA	6	18	3
Overall test CDA	30	65	43

Table 5-4: The number of gene sets with p value <0.01, 0.05 and q value<0.25 in p53 data set.

Top ten ranked gene sets based on GSEA/CDA method were given in Table 5-5, where the corresponding results from CDA/CDA, overall GSEA and overall CDA were also shown. Since GSEA/CDA was more conservative than the other three methods, the top ranked gene sets were also ranked high in results from all the other three methods. For example, although the top three gene sets: p53hypoxiaPathway, p53Pathway and radiation sensitivity genes were not significant in GSEA/CDA method, they were identified as differentially expressed pathways by CDA/CDA, overall GSEA and overall CDA. This finding provided more evidence of differentially expression. Furthermore, these three gene sets were also reported in previous studies (Subramanian, et al., 2005) (Tsai & James, 2009). The downstream factors for all three gene sets were related to TP53 transcription factor, which again supports the finding since the target genes for TP53 were expected to have dramatic change in expression level between wild-type cells and TP53 mutant cells.

On the other hand, 43 gene sets were declared as significant by overall CDA test and the top 10 gene sets were shown in Table 5-6. However, false positive rates could be very high due to the different assumption of the overall method. For example, p53Signalling pathway with ESR1 transcription factor was identified by overall CDA test (q=0.0472). But if we considered the p values from upstream factors and downstream factors, respectively, we would noticed that there was no strong evidence of differential expression for downstream factors (p value for upstream<0.001, p value for downstream=0.277). This gene set was not identified by any of the other three methods. Therefore, the overall test required extra precaution for false positives.

Up	Down	GSEA/CDA				CDA/CDA				GSEA OVERALL				CDA OVERALL			
		Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr
p53hypoxiaPathway	TP53	0.001	0.004	0.002	0.372	0	0.004	0.003	0.110	0.001	0.015	0.000	0.000	0	0.004	0.000	0.000
p53Pathway	TP53	0.003	0.004	0.002	0.372	0	0.004	0.003	0.110	0.003	0.015	0.000	0.000	0	0.004	0.000	0.000
radiation_sensitivity	TP53	0.001	0.004	0.002	0.372	0	0.004	0.003	0.110	0.001	0.015	0.000	0.000	0	0.004	0.000	0.000
chemicalPathway	TP53	0.029	0.004	0.017	0.495	0.013	0.004	0.008	0.263	0.029	0.015	0.006	0.514	0.013	0.004	0.002	0.059
p53_signalling	TP53	0.038	0.004	0.023	0.495	0	0.004	0.003	0.110	0.038	0.015	0.034	0.514	0	0.004	0.004	0.099
p53_signalling	E2F1	0.038	0.046	0.028	0.495	0	0.046	0.030	0.435	0.038	0.374	0.393	0.514	0	0.046	0.015	0.186
p53Pathway	E2F1	0.003	0.046	0.028	0.495	0	0.046	0.030	0.435	0.003	0.374	0.364	0.514	0	0.046	0.013	0.170
HTERT_UP	NFIL3	0.06	0.079	0.049	0.495	0.175	0.079	0.123	0.435	0.06	0.134	0.059	0.514	0.175	0.079	0.135	0.312
ca_nf_at_signalling	SP1	0.08	0.081	0.050	0.495	0.055	0.081	0.054	0.435	0.08	0.246	0.118	0.514	0.055	0.081	0.071	0.310
p53_signalling	SP1	0.038	0.081	0.050	0.495	0	0.081	0.054	0.435	0.038	0.246	0.234	0.514	0	0.081	0.026	0.267

Table 5-5: Top ten ranked gene sets by GSEA/CDA method in p53 data set.

Upstream	Downstream	Pup	Pdown	OVERALL P value	q value
Cell_Cycle	E2F4	0.001	0.206	0	0
DNA_DAMAGE_SIGNALLING	TP53	0.001	0.004	0	0

p53hypoxiaPathway	TP53	0	0.004	0	0
p53Pathway	TP53	0	0.004	0	0
radiation_sensitivity	TP53	0	0.004	0	0
atmPathway	TP53	0.001	0.004	0.001	0.0472
atrbrcaPathway	TP53	0.027	0.004	0.001	0.0472
Cell_Cycle	TP53	0.001	0.004	0.001	0.0472
g2Pathway	TP53	0.001	0.004	0.001	0.0472
p53_signalling	ESR1	0	0.277	0.001	0.0472
cell_cycle_checkpoint	TP53	0.047	0.004	0.002	0.059
chemicalPathway	TP53	0.013	0.004	0.002	0.059
CR_DEATH	FOSL2	0.003	0.412	0.002	0.059
CR_DEATH	TP53	0.003	0.004	0.002	0.059
drug_resistance_and_metabolism	FOS	0	0.427	0.002	0.059
drug_resistance_and_metabolism	ESR1	0	0.277	0.002	0.059
Cell_Cycle	E2F6	0.001	0.209	0.003	0.078666667
RAP_UP	TP53	0.006	0.004	0.003	0.078666667
p53_signalling	TP53	0	0.004	0.004	0.099368421
drug_resistance_and_metabolism	TP53	0	0.004	0.005	0.102608696

Table 5-6: Top twenty ranked gene sets from overall CDA test in p53 data set.

5.4.2 ET Data Set

There were no reported studies on gene set analysis using ET data set yet. As we showed in section 5.3., larger percent of DEGs existed in ET data set than in p53 data set, which resulted in a more conservative performance for GSEA, that is, no significant pathway was found by combining method GSEA/CDA or overall GSEA test. Meanwhile, either combining test CDA/CDA or overall test CDA identified significance for almost all the gene sets. This performance was also understandable since CDA was very sensitive as long as there was at least one differentially expressed gene existing in the given gene set. To gain some insight into biology, we would consider the top ranked gene sets by GSEA/CDA test, although they were not significant due to the specificity of GSEA method.

The highest ranked gene set by GSEA/CDA test was il6Pathway with transcription factor ELK1. Il6Pathway, activated by a cytokine Interleukin-6, was reported to provoke a broad range of cellular and physiological responses, including immune response, inflammation, hematopoiesis and oncogenesis (QIAGEN). ET is a chronic blood disorder recognized by overproduction of platelets by megakaryocytes in the bone marrow. Although no publications discussed about the correlation between il6Pathway and ET, our finding gave a reasonable clue about it.

Methods including CDA gave significance to almost all the gene sets, and it did not shred a light into the puzzle by knowing majority of the gene sets were differentially expressed. Again, if we took a look at the overall CDA, some false positives would be noticed.

Combinations (348)	p value < 0.01	p value < 0.05	q value < 0.25
Combining test GSEA/CDA	0	41	0
Combining test CDA/CDA	7	169	321
Overall test GSEA	1	18	0
Overall test CDA	113	308	346

Table 5-7: The number of gene sets with p value <0.01, 0.05 and q value<0.25 in ET data set.

Up	Down	GSEA/CDA				CDA/CDA				GSEA OVERALL				CDA OVERALL			
		Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr	Pup	Pdown	Pcomb	fdr
il6Pathway	ELK1	0.012	0.004	0.011	0.332	0.02	0.004	0.019	0.073	0.012	0.082	0.204	0.476	0.02	0.004	0.002	0.019
erkPathway	ELK1	0.018	0.004	0.017	0.332	0.012	0.004	0.012	0.073	0.018	0.082	0.094	0.476	0.012	0.004	0.004	0.019
il6Pathway	JUN	0.012	0.02	0.019	0.332	0.02	0.02	0.019	0.073	0.012	0.372	0.438	0.476	0.02	0.02	0.012	0.019

FETAL_LIVER_HS_ENRICHED_TF_JP	NR2C2	0.024	0.009	0.022	0.332	0.01	0.009	0.010	0.073	0.024	0.002	0.010	0.476	0.01	0.009	0.002	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	ELF1	0.024	0.013	0.022	0.332	0.01	0.013	0.013	0.073	0.024	0.244	0.160	0.476	0.01	0.013	0.006	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	ETS1	0.024	0.003	0.022	0.332	0.01	0.003	0.010	0.073	0.024	0.092	0.014	0.476	0.01	0.003	0.004	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	NRF1	0.024	0.019	0.022	0.332	0.01	0.019	0.018	0.073	0.024	0.447	0.092	0.476	0.01	0.019	0.008	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	ELK4	0.024	0.011	0.022	0.332	0.01	0.011	0.011	0.073	0.024	0.417	0.366	0.476	0.01	0.011	0.006	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	GATA2	0.024	0.014	0.022	0.332	0.01	0.014	0.014	0.073	0.024	0.456	0.374	0.476	0.01	0.014	0.012	0.019
FETAL_LIVER_HS_ENRICHED_TF_JP	HLF	0.024	0.011	0.022	0.332	0.01	0.011	0.011	0.073	0.024	0.268	0.408	0.476	0.01	0.011	0.008	0.019

Table 5-8: Top ten ranked gene sets by GSEA/CDA method in ET data set.

Upstream	Downstream	Pup	Pdown	OVERALL P value	q value
CR_TRANSCRIPTION_FACTORS	TCF7L2	0.014	0.005	0	0
gata3Pathway	JUNB	0.008	0.373	0	0
HUMAN_CD34_ENRICHED_TF_JP	TCF7L2	0.01	0.005	0	0
HUMAN_CD34_ENRICHED_TF_JP	SREBF1	0.01	0.011	0	0
HUMAN_CD34_ENRICHED_TF_JP	ELK1	0.01	0.004	0	0
HUMAN_CD34_ENRICHED_TF_JP	REST	0.01	0.077	0	0
biopeptidesPathway	STAT1	0.034	0.006	0.002	0.019
CR_PROTEIN_MOD	ELK1	0.016	0.004	0.002	0.019
CR_SIGNALLING	STAT1	0.022	0.006	0.002	0.019
CR_TRANSCRIPTION_FACTORS	EGR1	0.014	0.037	0.002	0.019

Table 5-9: Top ten ranked gene sets from overall CDA test in ET data set.

5.5 Conclusion

In this chapter, we applied our proposed combining method on p53 data set and ET data set. Real data structure is very complicated and difficult to summarize. Consequently, it is not likely to prioritize one method over all the other methods universally.

From our findings, we noticed that methods including GSEA tended to be more conservative, which was consistent with our simulation results and was due to the competitive design of GSEA method. Combining method GSEA/CDA did not identify any significance in either data set under the criterion of q value < 0.25, however, the top ranked gene sets based on this method were illustratable and meaningful in practical sense. At the same time, methods including CDA had larger power than those including GSEA, which resulted from the nature of CDA that it can capture the change in gene expression distribution even though there existed only one differentially expressed gene. In the ET data set, the combining method CDA/CDA identified more than 90% of the gene sets as significant which is too sensitive to be helpful in real data explanation.

At the same time, the overall method was to group upstream factors and downstream factors into a big gene set and apply an existing gene set analysis method on it. The real data results confirmed with the simulation studies on the incorrect size of the test. If there existed strong evidence in differential expression in either upstream factors or downstream factors, the overall method would capture the change and declare significance to the whole set.

Therefore, the combining method GSEA/CDA is still on top of the method list because it showed better specificity and gave more reliable and explainable results to leave a clue into the mysterious truth.

Chapter 6 Discussion and Future Work

Gene set analysis emerged dramatically in the last decade, and shed new light into the complex biological study of disease. There are many gene set analysis methods built on different assumptions or using a variety of statistics. However, all the existing studies were more from mathematical or statistical direction. They focused on the gene sets including receptors to transcription factors but did not draw any attention to the target genes of the transcription factors. Our proposed combining method was developed based on practical biological insight that we incorporated the important information from the target genes into gene set analysis.

In the current work, we mainly adopted the gene set analysis methods in the category of functional class scoring (FCS), which made use of all the genes' expression data in a given gene set and the covariance structure among genes. We performed a series of simulations on the comparison of two representative methods of competitive null hypothesis and self-contained null hypothesis: GSEA and CDA, and chose GSEA for upstream factors and CDA for downstream factors. This combining method showed some promising results and gave insightful view into biology studies. But one limitation of the FCS methods is that they ignore the topology of the gene set, in other words, they give each gene the same weight no matter it is a receptor or it is a transcription factor. One direction of the future gene set analysis is topology-based method, which is more meaningful for upstream factors that have network structure between genes. One of our future studies is to incorporate topology-based method into combining method.

In this study, we used the Relax IUT method to combine p values from upstream and downstream. Although RIUT has shown improved power than the original version of maximum p value (Deng, Xu, & Wang, 2007), it has limited power due to the loss of information from the more significant test, the one with smaller p value. We would like to improve the power by developing a new combining method that takes the smaller p value into consideration. The difficulty lies in the estimate of type II error when no clue of the alternative distribution is provided.

Reference

(n.d.). Retrieved from QIAGEN: http://www.sabiosciences.com/pathway.php?sn=IL-6_Pathway

Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics, 10*, 47.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*, 1165-1188.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics, 24*(4), 295-300.

Berrar, D. P., Dubitzk, W., & Granzow, M. (2002). *A Practical Approach to Microarray Data Analysis.* Springer.

Casella, G., & Berger, R. L. (n.d.). *Statistical Inference.*

Consortium, G. O. (2008). The Gene Ontology project in 2008. *Nucleic Acids Res., 36*, D440–4.

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*, 561-563.

Deng, X., Xu, J., & Wang, C. (2007). Improving the power for detecting overlapping genes from multiple DNA microarray-derived gene lists. *BMC Bioinformatics, 9*, S14.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., . . . Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics, 8*, 242.

DNA_microarray. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/DNA_microarray

E, P.-C., Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., . . . Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research, 38*, D105–D110.

Efron , B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics, 1*(1), 107-129.

Efron, B., & Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet.Epidemiol., 23*(1), 70–86.

Essential thrombocythaemia. (n.d.). Retrieved from Wikepedia: http://en.wikipedia.org/wiki/Essential_thrombocythaemia

Fisher, R. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84*, 165.

Gnatenko, D. V., Cupit, L. D., Huang, E. C., Dhundale, A., Perrotta, P. L., & Bahou, W. F. (2005). Platelets express steroidogenic 17beta-hydroxysteroid dehydrogenases. Distinct profiles predict the essential thrombocythemic phenotype. *Thromb Haemost, 94*(2), 412-21.

Goeman, J. J., & Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987.

Goeman, J. J., van de Gee, S. A., de Kort, F., & van Houwelingen, H. C. (2003). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20*(1), 93-99.

Hasler, M. (2007). *IUT for multiple endpoints.* Leibniz University of Hannover.

Hastie, T. (1995). Pseudosplines. *Journal of the Royal Statistical Society, 58*(2), 379-396.

Hedenfalk, I. A., Ringner, M., Trent, J. M., & Borg, A. (2001). Gene-Expression Profiles in Hereditary Breast Cancer. *The New England Journal of Medicine, 344*(8), 539-548.

Heydebreck, A. v., Huber, W., Poustka, A., & Vingron, M. (2001). Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics, 17*, S107-114.

Hummel, M., Meister, R., & Mansmann, U. (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics, 24*(1), 78-85.

Johnson, R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis (6th Edition).* Pearson Prentice Hall.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res., 32*(1), D277-280.

Khatri, P., Draghici, S., Ostermeier, G., & Krawetz, S. A. (2002). Profiling gene expression using Onto-Express. *Genomics, 79*(2), 266–270.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*.

Kim, S.-Y., & Volsky, D. J. (2005). PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics, 6*, 144.

Kobayashi, A., Kang, M.-I., Okawa, H., Ohtsuji, M., Zenke, Y., Chiba, T., . . . Yamamoto, M. (2004). Oxidative Stress Sensor Keap1 Functions as an Adaptor for Cul3-Based E3 Ligase To Regulate Proteasomal Degradation of Nrf2. *Mol. Cell. Biol., 24*(16), 7130–7139.

Kong, S., Pu, W. T., & Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics, 22*(19), 2373-2380.

Latchman, D. S. (1997). Transcription factors: an overview. *Int. J. Biochem. Cell Biol., 29*(12), 1305–1312.

Lewin, B. (2004). *Gene VIII.* Pearson Prentice Hall.

Li, S. (2011). A combined p-value approach to infer pathway regulations in eQTL mapping. *Statistics and Its Interface, 4*, 389–401.

Li, Y., & Ghosh, D. (2012). Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *Bioinformatics Advance Access*.

Ma, S., & Kosorok, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics, 25*(7), 882–889.

Mansmann, U., & Meister, R. (2005). Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods of Information in Medicine, 44*(3), 449-453.

Mardia, K., & J.T. Kent, J. B. (1979). *Multivariate Analysis.* Academic Press.

Mathelier, A., Zhao, X., Zhang, W. A., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., . . . Wasserman, W. W. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 1-6.

Mitchell, P., & Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science, 245*(4916), 371–378.

Mootha, V. K., Lindgren, C., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., . . . Groop, L. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *34*, 267-273.

Multiple comparisons. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Multiple_comparisons_problem

Olivier, M., Eeles, R., Hollstein, M., Khan, M., Harris, C., & Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum. Mutat., 19*(6), 607-14.

Parmigiani, G., Garett, E. S., Irizarry, R. A., & Zeger, S. L. (2003). *The analysis of gene expression data.* Springer.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space (PDF). *Philosophical Magazine, 2*(6), 559–572.

Pollock, R., & Treisman, R. (1990). A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Research, 18*(21), 6197-6204.

Rawlings, J. S. (2004). The JAK/STAT signaling pathway. *J Cell Sci, 117*, 1281-1283.

Roy. (1953).

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research, 32*(1), D91-D94.

Schafer, J., & Strimmer, K. (2005). A Shrinkage Approach to Large-Scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology, 4*(1), 1175-1189.

Schena, M., Shalon, D., Davis, R., & Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467-470.

Shen, K., & Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics, 26*(10), 1316–1323.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS, 100*(16), 9440–9445.

Stouffer, S. (1949). *The American Soldier, Vol. 1: Adjustment during Army Life.* Princeton University Press.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS, 102*(43), 15545–15550.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., . . . Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics, 25*(1), 75-82.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *99*(10), 6567-6572.

Tippett, L. (1931). *The Methods in Statistics.* London: Williams and Norgate, Ltd,.

Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics, 6*, 225.

TRANSFAC Statistics. (n.d.). Retrieved from http://www.biobase-international.com/wp-content/uploads/2012/04/statistics_transfac.pdf

Tsai, C.-A., & James, C. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics, 25*(7), 897-903.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS, 98*(9), 5116-5121.

Virginia Goss Tusher, R. T. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 5116-21.

Whitlock, M. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *Journal of Evolutionary Biology, 18*(5), 1368-73.

Wikipedia. (n.d.). Retrieved from http://en.wikipedia.org/wiki/Meta-analysis

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin, 48*(3), 156-158.

Wingender, E. (1988). Compilation of transcription regulating proteins. . *Nucleic Acids Research, 16*(5), 1879–1902.

Yang H, C. G. (2007). Estimating p-values in small microarray experiments. *Bioinformatics, 23*(1), 38-43.