University of Alberta

GENE SET REDUCTION FOR A CONTINUOUS PHENOTYPE

by

Farzana Yasmin

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Epidemiology

Department of Public Health Sciences

Spring 2014
Edmonton, Alberta

Dedicated to my parents

Abstract

Introduction: Gene set analysis (GSA) examines the association between predefined gene sets and a phenotype and is becoming a topic of growing interest in DNA microarray studies. However, when a gene set is identified to be significant, often not all the genes within the gene set are responsible for the significance. Identifying core subsets improves understanding of biological mechanisms and reduces costs from diagnosis to treatment. Few methods have been introduced to isolate core genes from significant gene sets. There are no methods for continuous phenotype that eliminate redundant genes and effectively reduce the gene sets to core subsets, explaining the observed association.

Objective: Our research objective is to reduce gene sets associated with a continuous phenotype to subsets of genes that chiefly contribute to the association.

Methods: Our method tests subsets of a differentially expressed gene set by gradually eliminating genes not associated with the phenotype. A computationally efficient method, namely Linear Combination Test (LCT), is used to test the association between each gene set and the phenotype of interest. Within the significant gene sets we used Significance Analysis of Microarrays (SAM) to order individual gene-phenotype association. Again, LCT is used to get the most differentially expressed subset of genes which is obtained by the ordered genes.

Results: We studied our proposed method using a real microarray data consisting of gene expression levels of 13,233 genes measured on 33 African-American prostate cancer patients and 1403 gene sets obtained from C2 catalog of the Molecular Signature Database (http://www.broadinstitute.org/gsea/msigdb). We showed results

of both individual gene analysis and gene-set analysis on this data using SAM and LCT, respectively. LCT identified 30 statistical significant gene sets. We used our gene reduction method to extract core subsets of genes and calculate percent reduction in each of the 30 sets. We calculated frequencies of core genes among all the significant sets.

Conclusion: This work enables us to effectively reduce the gene sets to the most important genes that contribute to disease. This approach may bring faster and more cost efficient diagnosis and treatment of chronic diseases by focusing only on differentially expressed genes in the reduced sets.

Acknowledgements

First and foremost, I want to express my sincere gratitude to my supervisor Dr. Irina Dinu for her warm encouragement and thoughtful direction on this research project. I have been extremely lucky to have a supervisor who cared so much about my research work. Her assistance and guidance helped me improving my analysis skill and writing thesis.

I also want to thank Dr. Yan Yuan and Dr. Edit Gombay for participating on my thesis committee. I want to thank Dr. Yutaka Yasui for taking the time for a wonderful discussion on this research project and answering my questions. This research is also in collaboration with Shabnam Vatanpour, a PhD student. Discussions and criticisms about each other's work improved our research ideas.

My deepest thanks to my lovely parents for their love and encouragement. I am grateful to my father, Hussain Zakir and my best friend mother, Hasina Zakir, for their inspiration and support all my life. I must express my thanks to my elder brother Hasibul Hasan Romel, for his influential guidance to follow the path of science.

Completing my MSc thesis work would have never been a success without the support of my dear husband, Ashique Mahmood Rupam. I was continually amazed by his guidance, sincere support and strong criticism throughout the past two and half years of my student life. I am grateful for his thorough proofreading and feedback on this thesis work. He has been a great friend in my life in good times and bad times.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

DNA microarray studies open a new platform to us with an opportunity to study and compare thousands of genes at the same time. Identifying differentially expressed genes helps with early and more accurate diagnosis as well as improved tailored treatment. Individual Gene Analysis (IGA) and Gene Set Analysis (GSA) target identification of differentially expressed genes, or sets of genes. Therefore, IGA and GSA quickly became popular analysis methods for data measured by DNA microarray investigations.

GSA methods group all the genes based on similarity in chromosomal location and functions, therefore making use of biological knowledge in interpreting the results becomes easy (Geoman and Buhlman, 2007). The achievement of GSA is that it enables the analysis to be interpretable. This biologically interpretable results make GSA methods more feasible than the IGA methods to the biologists. Based on these advantages, scientists often prefer GSA over IGA.

Many GSA methods have been proposed, especially for a binary phenotype. Equally, if not more, important is the ability to test the enrichment of a gene signature or pathway against a continuous phenotype which are routinely commonly observed, e.g., in clinico-pathological parameters. They include tissue features as a tumor size, staining based readouts; cellular characteristics such as the amount of lymphocytic infiltration in a tumor environment; and subject-specific measurements such as diagnostic or prognostic marker protein or metabolic concentrations. It may

not always be easy or meaningful to dichotomize continuous phenotypes into two classes, which may lead to inaccurate classification of the samples thus affecting the downstream gene-set analysis. Our proposed work builds upon recent efforts to incorporate correlation structure within gene sets and pathways into the GSA test statistic. To address the issue of continuous phenotypes directly without the need for artificial discrete classification, and thus increase the power of the test while ensuring computational efficiency and rigor, new GSA methods that can incorporate the covariance matrix estimator for a continuous phenotype present an effective approach (Dinu et. al., 2013).

All the members within a gene set are not responsible for the significance, so identifying core subsets from the gene set can improve understanding the biological mechanism. Therefore, reducing the sets to their core subsets is an important step. To the best of our knowledge, there is only one method for identifying core subsets for a binary phenotype, and no methods for continuous phenotype.

In this thesis, we address the problem of finding differentially expressed core genes for continuous phenotype. In the next sections, we discuss some aspects of DNA microarray studies, the challenges of analyzing microarray data and the contributions of this work.

1.1 Brief Overview of DNA Microarray Studies

Microarray data is a highly technical and instrumental process combining robotics, chemistry, computer science, and biology in a biology laboratory. For more than three decades, it has been an attractive platform of studying a massive amount of data in genomic studies. DNA microarrays allow the researchers to study thousands of genes or entire genome in a single assay. It enables us to explore the disease state by recognizing any difference in gene expressions comparing with healthy state. Thus the genes that change their functions during disease can be captured through their expressions with microarray technology. This technology captures the expressions of thousands of mRNA species simultaneously using transcriptional profiling,

a technological breakthrough in the analysis of biological specimens (Pusztai et al., 2003). Using microarrays, researchers study and compare the diseased tissues with healthy normal tissues and understand disease mechanisms.

DNA microarray enables researchers to improve personalized medicine, understand the complex biology of chronic diseases and to update the knowledge of the change of gene expression during a therapy. DNA microarray studies can measure gene expressions with a high degree of accuracy (Pusztai et al., 2003). Microarray studies provide very precise state of the cell because mRNA contains the current reflection of the cell condition.

1.2 Challenges in Analysis Methods for DNA Microarray Studies

Understanding disease mechanisms is one of the most rising concerns in microarray gene analysis and other computational biological studies. The main characteristic of genomic data is that it consists of much larger number of features p than the sample size N. We denote this characteristic as $p \gg N$.

High variance and overfitting are a big concern for $p \gg N$ types of data. Therefore, methods for analyzing them by reducing high variance and controlling overfitting have been introduced and still improving. Highly regularized approaches such as Lasso and Ridge regression are used to solve this problem. The problem of Lasso shrinkage is that it fails to incorporate a priori biological knowledge (Hastie et al., 2001). Inconsistency of Lasso shrinkage regression is also a big concern to the researchers.

Although Gene Set Analysis (GSA) methods are introduced to incorporate a priori biological knowledge in understanding disease mechanisms, methods for finding core genes from significant gene sets still need to improve. It is very important to obtain only those genes that are differentially expressed with the phenotype, as subset. We know that a gene set can show significance only because a subset of genes inside the set is actually differentially expressed, and the rest of the of genes

can be redundant in terms of the association of our phenotype of interest. Identifying core subsets is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis and tailored treatment.

1.3 Contributions

If a gene-set analysis identifies differential expression of a gene set in the microarray data, a natural step would be to ask: "are all members of this gene set essential, or is a subset sufficient, in considering its link with the phenotype of interest?" Our contribution here is to introduce a method for extracting a core set of genes that chiefly contribute to the statistical significance of differential expression of a given gene set by a phenotype. Our proposed method to identify core subsets of genes is a new direction of GSA methods. While this direction has been explored for a binary phenotype, there is no work for the continuous phenotype. We illustrate our gene-set reduction method on a real microarray study of prostate cancer patients with a continuous phenotype.

1.4 Thesis Organization

We organized our thesis in five chapters. In chapter 2, we discuss background of microarray studies. Here we discuss both individual gene analysis and gene set analysis. In chapter 3, we describe our analysis for identifying significant gene-sets and extracting core sub-sets. In chapter 4, we describe our microarray expression data and processing of the data to fit in our analysis. We also present all the results of our analysis and biological interpretations in chapter 4. In chapter 5, we discuss some aspects of LCT method and future directions of this research work.

Chapter 2

Background

In this chapter, we critically review the major microarray study methods that have a huge impact on bringing gene analysis studies to its current state. We describe different types of individual gene and gene set analysis methods for both binary and continuous phenotype. We thoroughly discuss the method of Significance Analysis of Microarrays (SAM) in the section of single gene analysis. We summarize some gene set analysis methods such as Over-Representation analysis and Gene Set Enrichment Analysis. We describe an extension method of Linear Combination Test (LCT) for continuous phenotype. In the last section we describe briefly about the gene-set reduction method for a binary phenotype.

2.1 Methods for Individual Gene Analysis

Many individual gene analysis methods have been proposed, for example Fold Change (DeRisi et al., 1996), Significance Analysis of Microarrays (Tusher et al., 2001), Regularized t-test (Baldi & Long, 2001) and Regression modeling (Thomas et al., 2001).

Fold change is a measure of change of a gene expression from one condition to another condition. This value compares with a pre-specified non-zero fold change value, t. If $\overline{x}_{i1}$ and $\overline{x}_{i2}$ denote the average expression levels of gene i under two different conditions of patients then a positive significant gene must satisfy $|\overline{x}_{i2}/\overline{x}_{i1}| \geq$

t and a negative significant gene must satisfy $|\overline{x}_{i1}/\overline{x}_{i2}| \leq 1/t$. Fold change method does not perform any statistical test that can identify differentially or non-differentially expressed genes Chu et al., (2002).

Perhaps among all the single gene analysis methods, SAM is the most popular method that tests the significance of the genes. We discus SAM elaborately in the next section.

The SAM Method

SAM (Significance Analysis of Microarrays), proposed by Tusher et al. (2001) and later extended and developed by Storey and Tibshirani (2003), is a popular analytical method that searches for statistically significant genes associated with a phenotype of interest in a microarray data set. SAM identifies the differentially expressed genes associated with the response variable (phenotype) by using repeated permutations of the data. This repeated permutation accounts for correlation of the genes and avoids parametric assumptions about the distribution of the genes. SAM can be used for different formats of the response types of data.

The advantage of SAM over other techniques is that we do not have to assume equal variance and independence of genes like other techniques (e.g., ANOVA and Bonferroni method). Another important advantage of SAM is that it does not require applying the same cut point to the positive and negative values of test statistic to get the differentially expressed genes. Separate cut points can be used for the two cases.

SAM can be used for different phenotypes of data, for example, continuous phenotype, binary phenotype, multi-class response, censored survival data etc. In the following we discuss the technical details of the SAM method only for continuous phenotype, since continuous phenotype case is the focus of our proposed gene-set reduction method.

Let us consider that our gene expression data is in a matrix x and the response

data is in a vector y. More precisely, x_{ij} denotes gene expression measurement for gene i and patient j, and y_j denotes phenotype measurement for patient j, where $i = 1, 2, \ldots, p$ represent the genes, and $j = 1, 2, \ldots, n$ represent the patients. The test statistic of SAM calculates relative difference of the gene expression with a phenotype. SAM uses moderated gene specific t-test. Here the test statistic d_i measures the change of the gene expression for gene i adding a constant s_0 to the denominator.

For each gene i, null hypothesis of SAM can be defined as follows:
H_0 : There is no association between the gene expressions and the phenotype. We note that this can be reformulated in the context of simple linear regression coefficient for gene i, testing a linear association between gene expressions x_{ij} and continuous phenotype y_j.

The test statistic d_i is defined as

$$d_i = \frac{r_i}{s_i + s_0}, i = 1, 2, ..., p, \tag{2.1}$$

where r_i is the linear regression coefficient of gene i on the outcome:

$$r_i = \frac{\sum_j y_j(x_{ij} - \overline{x}_i)}{\sum_j (y_j - \overline{y})^2}, \tag{2.2}$$

where

$$\overline{x}_i = \sum_j \frac{x_{ij}}{n}. \tag{2.3}$$

Here s_i is a standard error of r_i:

$$s_i = \frac{\hat{\sigma}_i}{\left[\sum_j (y_j - \overline{y})^2\right]^{1/2}}, \tag{2.4}$$

and $\hat{\sigma}_i$ is the square root of residual error:

$$\hat{\sigma}_i = \left[\frac{\sum_j (x_{ij} - \hat{x}_{ij})^2}{n-2}\right]^{1/2}. \tag{2.5}$$

where

$$\hat{x}_{ij} = \hat{\beta}_{i0} + r_i y_i \tag{2.6}$$

$$\hat{\beta}_{i0} = \overline{x}_i - r_i \overline{y}. \tag{2.7}$$

In SAM analysis, s_0 is the exchangeability factor or a constant. The calculation of this s_0 is described below.

As noted above, s_i is the standard error of r_i-th gene. Let us assume that s^α be the α-th percentile of all s_i values. Let us denote by

$$d_i^\alpha = \frac{r_i}{s_i + s^\alpha}. \tag{2.8}$$

Now, 100 quantiles of s_i values are calculated. They are ordered as, $q_1 < q_2 < \cdots < q_{100}$. We choose α among the following values: $(0, 0.05, 0.1, \ldots, 1.0)$. For each value of α, v_j of d_i^α is calculated as:

$$v_j = \mathrm{mad}\Big(d_i^\alpha | s_i \in [q_j, q_{j+1})\Big), j = 1, 2, \ldots, n; \tag{2.9}$$

where mad is the median absolute deviation from the median divided by 0.64. Now, Let $cv(\alpha)$= coefficient of variation of v_j values, which is the measure of the range of variability from the population mean. Now we can choose $\hat{\alpha}$ such that, coefficient of variation of v_j values is minimum:

$$\hat{\alpha} = \mathrm{argmin}_\alpha[\mathrm{cv}(\alpha)] \tag{2.10}$$

Then $\hat{s}_0 = s^{\hat{\alpha}}$ is calculated, and s_0 is fixed at the value of $\hat{s}_0$. Sometimes s_0 can show better performance by setting a fixed value rather than estimating automatically. The value s_0 plays an important role in estimating the moderated t-statistic. If s_0

is too large, then the non-null genes with small s_i will be lost with noise. On the other hand, if s_0 is comparatively too small, then for null genes with very small s_i, t-statistic will become very large. The value of s_0 is chosen such that the estimated coefficient of variation of d_i is minimized. For detailed information, please see the SAM users guide and technical document by Chu et al. (2002).

Steps of SAM Procedure:

1. First calculate SAM statistic d_i for each gene i in microarray study.
2. Then calculate and rank the d_i values according to their order, $d_{(1)} \leq d_{(2)} \leq d_{(3)} \leq \cdots \leq d_{(p)}$.
3. Then permute the phenotype levels to get a new data set for the same measurements for each gene i. For each permutation b, calculate SAM statistic d_i^{*b} and calculate order statistic again after the permutation, $d_{(1)}^{*b} \leq d_{(2)}^{*b} \leq d_{(3)}^{*b} \leq ... \leq d_{(p)}^{*b}$.
4. From the set of B permutations, estimate the expected order statistics by $\bar{d}_i = 1/B \sum_b d_{(i)}^{*b}$ for $i = 1, 2, \ldots, p$.
5. For each $d_{(i)}$ values, we get corresponding expected $\bar{d}_{(i)}$ values after permutation. Plot the moderated d_i values versus the expected $d_{(i)}$ values.
6. At this stage we set a pre-specified threshold value set up by researchers. For a pre-specified threshold Δ, if $d_i - \bar{d}_{(i)} > \Delta$, genes are called significant positive and if $\bar{d}_{(i)} - d_i > \Delta$, genes are called significant negative. The smallest and the largest thresholds, Δ are denoted by $cut_{low}(\Delta)$ and $cut_{up}(\Delta)$ respectively.

Multiple Hypothesis Testing

SAM tests thousands of genes associated with the phenotype of interest simultaneously, and we need to estimate an overall measure of error for this multiple hypothesis testing. A measure of error for single hypothesis testing is type I error. A variety of generalizations of the type I error for multiple hypothesis are possible. One of them is family-wise error rate (FWER). FWER is the probability of at least

one false rejection among multiple tests. If type I error is α, then FWER of the collection of test is $(1-(1-\alpha)^M)$, where M is the number of total genes. When M is very large, which is typical in microarray analysis, this FWER becomes very high (close to one).

Another simple approach for multiple testing is Bonferroni method. It controls the FWER. In order to make the FWER equal at most α, we reject H_{0i} with a type I error of α/M, for $i = 1, 2, \ldots, M$ number of genes. Bonferroni method can be useful for testing small number of genes. But for large number of genes this method is too conservative in the sense that only very few number of genes can be significant in this case.

Table 2.1: Possible outcomes from M hypothesis tests.

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M-R$	R	M

A different approach for error rate in multiple testing is False Discovery Rate (FDR). FDR focuses on the proportion of falsely significant genes. From table 2.1, type I error = V/M_0, type II error = T/M_1 and power = $1 - T/M_1$. Here, $FDR = V/R$.

SAM reports FDR values for each significant genes. Storey and Tibshirani (2003) proposed FDR by estimating the proportion of true null genes or the unaffected genes in the data set. Details about calculation of FDR value for SAM is given below.

False Discovery Rate (FDR) Calculation in SAM

1. For a grid of Δ values, calculate the total number of significant genes (from step 6 of SAM procedure). Median number and 90th percentile of falsely called genes are calculated by computing median number and 90th percentile of values from each of the B permutation sets of d_i^{*b} that fall above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$ values.

2. Calculate 25% and 75% points of the permuted d values.

3. Compute $\hat{\pi}_0$= $\{d_i \in (q^{25}, q^{75})\}/(0.5p)$, where p is the number of genes. Let $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$, That is, $\hat{\pi}_0$ is truncated at 1.

4. Finally, FDR is calculated as the ratio of number of 50th or 90th percentile of falsely called genes times $\hat{\pi}_0$ divided by the number of significant genes (Chu et al., 2002).

SAM can be downloaded for free as a user friendly Excel Add-Inn. The different formats of the response for SAM can be quantitative, two class paired and unpaired, multi class, survival data time course paired and unpaired etc.

2.2 Methods for Gene Set Analysis

Individual gene analysis result in a long list of significant genes, that is hard to interpret. Biologists have put together sets of genes that share common biological functions, called biological pathways. Interpreting results of pathway or gene-set analyses makes more biological sense than interpreting results of individual gene analysis.

Another important limitation of individual gene analyses is that they extensively depend on using cutoff threshold values. According to these methods, significance of genes is highly affected by the cutoff values which is often chosen arbitrarily by the researchers (Nam & Kim, 2008). Hence, using different threshold values can severely change the list of significant genes. A large list of significant genes is hard to interpret. Again, genes that are moderately significant can get eliminated from the significance list if any researcher decided to use a strict cutoff threshold. This approach can reduce the power of the test (Nam & Kim, 2008).

Gene-set analysis (GSA) represent a cut-off free approach to analysis of microarray studies. GSA methods use pre-defined gene sets, grouped based on biological knowledge. Finally, GSA identifies significant gene sets for even a slight change of the coordinated similar functioning genes. The benefit of using coordinated genes

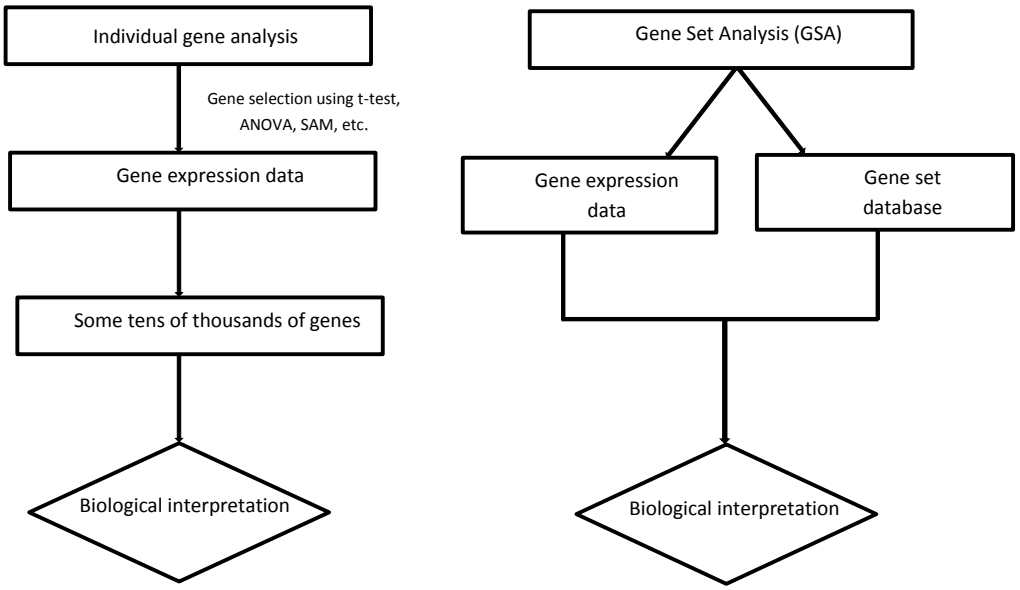


Figure 2.1: Individual gene analysis vs. gene set analysis.

enables to understand the process of the biological pathways and gives a better understanding behind disease mechanism. Some individual gene analysis methods are developed assuming the genes are independent. This leads to large number of false positive rates. Actually, genes within a gene set can be highly correlated as they share the common biological function, chromosomal location or regulations. So, GSA gives more consistent results than the result of single gene analysis among microarray studies. Hence, the biological interpretations from GSA methods are more consistent and convenient than individual gene analyses.

GSA methods are developed based on different hypothesis testing and methodologies (Nam & Kim, 2008). According to Tian et al. (2005) GSA methods are classified with two different tests of hypotheses. To test the association between gene sets and the phenotype of interest, one hypothesis is Q1 or competitive methods and the another one is Q2 or self-contained methods. These terms are given in and discussed by Geoman and Buhlman (2007).

Competitive or Q1 hypothesizes that the level of association of a gene set with a given phenotype is same as the complement of the gene set. The second type, Q2 or self-contained method considers only the genes within the gene set and hypothe-

sizes that there is no gene in the gene set associated with the phenotype. Predictions in self-contained (Q2) methods are strong, because it uses the information within a gene set. On the other hand, competitive method (Q1) is based on the assumption that the genes are independent. In fact, the genes can be highly correlated within a gene set (Dinu et al., 2008). The key methodological difference between the two approaches is that the competitive approach uses genes as the sampling units, whereas self contained method uses subjects as sampling units. Geoman and Buhlmann (2007) discussed the problem of testing gene sampling (competitive) method assuming independence of genes. They compared the performance of three methods; self-contained, competitive and GSEA by simulating data under Q1 or competitive hypothesis, and highly discouraged using competitive method mentioning its wrong assumption of independence across genes. Q3 is a another null hypothesis of GSEA, that correlations of the genes in the set with the phenotype are clustered. Q3 is a hybrid between Q1 and Q2. GSEA is hybrid between self-contained and competitive method. An extensive review on methodological issues of gene set analysis has been performed by Nam and Kim (2008) and Dinu et al., (2008).

2.2.1 Over-Representation Analysis

Over-representation method was introduced by Draghichi et al. (2003) to demonstrate the validity and utility among the results of two different microarray studies. The authors developed Onto-Express tool to be able to automatically translate the list of differentially expressed genes according to their functional impact. This analysis is one of the early suggested gene set analysis methods by Draghichi et al. (2003). The main idea of this method is to analyze microarray data at the pathway level, rather than individual gene level. This enables researchers to understand the biological function and its process.

Over-representation method first determines a list of significant genes. Based on these significant genes a measure of over representation is calculated for each gene set. This measure of overrepresentation is measured by tagging each gene of the

gene set F or NF considering whether they are significant or not. With the researcher's own selection procedure, the genes are found F (significant genes) or NF (not significant genes). From the observed F (significant genes) the probability of getting significant genes are calculated to check whether they are selected either by chance or not. The p-value for over represented categories is calculated with hyper-geometric distribution or alternatively χ^2 test for equality of proportions and Fishers exact test for small samples arranging in a 2×2 table to obtain over-representation of differentially expressed genes from the gene sets as shown in table 2.2.

Table 2.2: A 2×2 table for calculating over-representation.

	Diff. expressed gene	Non-diff. expressed gene	Total
Within gene set	x	$M - x$	M
Not within gene set	$K - x$	$N - K - M + x$	$N - M$
Total	K	$N - K$	N

Let us consider that the total number of genes is N, where M of them are within the gene set. If K is the total number of differential expressed genes where x of these genes are coded as F, then the probability of being F can be calculated. The probability of occurring x differentially regulated genes is modeled as hyper-geometric distribution using sampling without replacement:

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}. \tag{2.11}$$

Then the p-value of having x genes or fewer in F can be calculated by summing the probabilities of a random list of K genes having $1, 2, \ldots, x$ genes of category F:

$$p = \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}. \tag{2.12}$$

This p-value is one sided test which finds the probability of underrepresented categories of genes. The p-value of overrepresented categories of genes is calculated as

$$p = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}, \tag{2.13}$$

when the sum is larger than 0.5 For large number of sample N, hypergeometric distribution tends to be binomial distribution.

Although over-representation method was developed to explain the biological pathways by combining similar functional genes in a gene set, it has some limitations. Firstly, in spite of combining genes in a gene set according to their functions, they do not consider correlations among the genes inside a gene set. The assumption of Fishers exact test considers all genes to be independent. This violates the most important characteristics of pathways that genes within a set are highly correlated with each other. Another limitation of this method is that the test statistic does not include the correlation of the phenotype with the significant genes. An important aspect of this overrepresentation is that the input of identifying statistically significant gene ontology is the list of already defined differentially expressed genes for a certain phenotype. The major problem of this approach is that those genes which are not differentially expressed are not included in the list. As a result, marginally significant genes are discarded from the output, and it is well known that marginally significant genes may work together and achieve significance when considered as a set, or pathway.

2.2.2 Gene Set Enrichment Analysis (GSEA)

Gene set enrichment analysis (GSEA) was first proposed by Mootha et al. (2003) and later improved by Subramanianan et al. (2005). Among Gene set analysis methods, GSEA has the most comprehensive and accessible output. In spite of being heavily criticized, it is still a gold standard among gene-set analysis users.

Method of GSEA:

In GSEA method, genes are ordered in a list L according to their differential expression levels with the binary phenotype labeled as 1 or 2. For a priori defined gene set S, the goal of GSEA is to find whether the genes in S are either primarily located at the top or bottom of the list L or they are found by chance. An enrichment score (ES) is used to rank the genes in a list L.

In the calculation of enrichment score (ES), a running sum statistic is calculated by picking genes from the L list to be found in a gene set S gradually using a for loop procedure. If the gene from list L is found in the gene set S, the running sum statistic increases and when the gene is not found in the gene set the running sum statistic decreases. The correlation of the gene with the phenotype decides the magnitude increment of the running sum statistic. Now, the enrichment score for a gene set is calculated by getting the maximum deviation from zero encountered in the random walk which is similar to the *weighted Kolmogorov-Smirnov statistic*.

Let us define r_j as the correlation coefficient between the expression measurements in the list of total number of genes, $L = \{g_1, ...g_N\}$. Genes are listed in the list according to their correlation $r_{(g_j)} = r_j$ order with gene expression profiles C. For all the genes N, estimate fraction of genes in S ("hits") weighted by their correlation and fraction of genes not in S ("misses") present up to a given position i in L.

$$P_{hit}(S,i) = \sum_{g_j \in S \& j \leq i} \frac{|r_j|^p}{N_R}, \tag{2.14}$$

$$P_{miss}(S,i) = \sum_{g_j \notin S \& j \leq i} \frac{1}{N - N_H}. \tag{2.15}$$

where $N_R = \sum_{g_j \in S} |r_j|^p$ and $N - N_H$ is the number of genes in the gene set S and p is the exponent component to control the weight of the "hit" genes. By putting $p = 0$ and $p = 1$ we get a list of paired scores P_{hit} and P_{miss}. Enrichment score (ES) is the maximum deviation from the zero $P_{hit} - P_{miss}$. For $p = 0$, by walking

down the L list a reduced standardize Kolmogorov-Smirnov statistic is obtained and when $p = 1$ genes are weighted in S by their correlation with C normalized by the sum of the correlations over all of the genes in S. Significance of ES is calculated by phenotype label permutation. For randomly distributed S, $E[ES]$ is relatively small and for non-randomly distributed S, $E[ES]$ is relatively high. That is, when the genes are concentrated to the top or to the bottom, $E[ES]$ is differentially expressed.

For estimating the significance of a gene set, first the authors computed ES_{NULL} with randomly assigned phenotypes. ES is compared with the set of scores ES_{NULL}. Then performing this for 1 to 1000 permutations, a histogram of the corresponding ES_{NULL} is created. After getting the significant gene sets, multiple hypothesis testing is adjusted.

1. First estimate the $ES(S)$ for each gene set from the data base.
2. For each S and 1000 fixed permutations π of the phenotype labels, reorder the genes in L and determine $ES(S, \pi)$.
3. Adjust the variation in gene set size and normalize $ES(S, \pi)$ the observed $ES(S)$. Dividing by the mean of $ES(S, \pi)$, positive and negative scores are separately rescaled. This way normalized scores for both $NES(S, \pi)$ and $NES(S)$ are obtained.
4. Then FDR values are calculated by gaining a fixed level of significance for both (positive and negative) $NES(S)$ and $NES(S, \pi)$ to control the ratio of FDR to the total number of gene sets.
5. A histogram of $NES(S, \pi)$ values are achieved and FDR q values are computed by using null distribution values. FDR is the ratio of the percentage of all (S, π) values when $NES(S, \pi) \geq 0$.

2.3 Gene Set Analysis for Continuous Phenotype

The urge of improving gene set analysis (GSA) method for continuous phenotype has risen based on the fact that the gene expression variables sometimes are taken in continuous measurements. Such variables can be tumor size or measurements of the marker proteins. It is inadvisable to the researchers to set the continuous measurements into binary or categorical variables by giving a range. It is inappropriate in the sense that some specific ranges may fail to express the biological functioning capacity for each patient. Different specialists may want to use different ranges according to the patient's health condition. So, directly analyzing the continuous variables may result in an improved GSA method in genomics study.

In the following we discuss a GSA method for continuous phenotype known as Linear Combination Test (LCT).

Linear Combination Test Analysis

Let us consider, our gene expression data consists of n total subjects. If a gene set has pre-defined genes $\{X_1, X_2, ..., X_p\}$, then we test the hypothesis that the pre-defined gene set is not associated with the phenotype. This multivariate hypothesis can be written in a univariate way, such as, H_0 : no linear combination of $X_1, X_2, ..., X_p$ is associated with the phenotype of interest. For $X_1, X_2, ..., X_p$ genes, the linear combination can be written as $Z(\beta) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$. For a given vector β of combination coefficients, whether the combination $Z(\beta)$ is associated with the phenotype or not can be tested in the following univariate model: $Y_i = \alpha_0 + \alpha_1 Z_i(\beta) + e_i$, where α_0 and α_1 are the intercept and slope respectively, $e_i \sim N(0, \sigma^2)$, and i denotes subjects $1, ..., n$. This is a classical simple linear regression problem. For testing H_0, we can consider the most-significant linear combination of $\{X_1, X_2, ..., X_p\}$, that is, the linear combination with the maximum correlation with the phenotype among all possible linear combinations. We have,

$$\beta^* = \text{argmax}_\beta \, \rho^2_{Y,Z(\beta)}, \tag{2.16}$$

where correlation between Y and $Z(\beta)$ is

$$\rho^2_{Y,Z(\beta)} = \frac{\text{Cov}(Y, Z(\beta))^2}{\sigma^2_Y \sigma^2_{Z(\beta)}}. \tag{2.17}$$

If we ignore the σ_Y^{-2} in the square of correlation, then we can write

$$\rho^2_{Y,Z(\beta)} = \frac{\text{Cov}(Y, Z(\beta))^2}{\sigma^2_{Z(\beta)}}. \tag{2.18}$$

Now, we can write,

$$\rho^2_{Y,Z(\beta)} = \frac{\text{Cov}(Y, Z(\beta))^2}{\beta^T \hat{\Omega} \beta}. \tag{2.19}$$

where,

$$\begin{aligned}
\sigma^2_{Z(\beta)} &= \text{E}[Z(\beta) - \text{E}[Z(\beta)]^2] && (2.20)\\
&= \text{E}\left[(\beta^\top X - E[\beta^\top X])^2\right] && (2.21)\\
&= \text{E}\left[\{\beta^\top (X - E[X])\}^2\right] && (2.22)\\
&= \text{E}\left[\beta^\top (X - E[X]) \beta^\top (X - \text{E}[X])\right] && (2.23)\\
&= \beta^\top \hat{\Omega} \beta. && (2.24)
\end{aligned}$$

Here $\hat{\Omega}$ is the gene expression covariance matrix, where hh'-th entry can be written as

$$\omega_{hh'} = \frac{1}{n-1} \sum_{l=1}^{n} (x_{hl} - \overline{x}_h)(x_{h'l} - \overline{x}_{h'}) \tag{2.25}$$

From the numerator of equation 2.19 we can write,

$$
\begin{aligned}
\mathrm{Cov}(Y, Z(\beta))^2 &= \mathrm{Cov}(Y, Z(\beta))\mathrm{Cov}(Y, Z(\beta)) && (2.26)\\
&= \mathrm{E}^2[(Y - \mathrm{E}[Y])(Z(\beta) - \mathrm{E}[Z(\beta)])] && (2.27)\\
&= \mathrm{E}^2[(Y - \mathrm{E}[Y])(\beta^\top X - \mathrm{E}[\beta^\top X])] && (2.28)\\
&= (\beta^\top \mathrm{E}[(Y - \mathrm{E}[Y])(X - \mathrm{E}[X])])^2 && (2.29)\\
&= \beta^\top \mathrm{E}\big[(Y - \mathrm{E}[Y])(X - \mathrm{E}[X])\big] \\
&\quad \times \mathrm{E}[(Y - \mathrm{E}[Y])(X - \mathrm{E}[X])]^\top \beta && (2.30)\\
&= \beta^\top \mathrm{Cov}_{Y,X}\mathrm{Cov}_{Y,X}^\top \beta && (2.31)
\end{aligned}
$$

Hence equation 2.19 can be written as,

$$
\rho^2_{Y,Z(\beta)} = \frac{\beta^\top \mathrm{Cov}_{Y,X}\mathrm{Cov}_{Y,X}^\top \beta}{\beta^\top \hat{\Omega} \beta}, \tag{2.32}
$$

where $\mathrm{Cov}_{Y,X} = (\mathrm{Cov}(Y, X_1), \ldots, \mathrm{Cov}(Y, X_p))^T$. Now the optimization problem can be written as,

$$
\rho^2_{Y,Z(\beta)} = \frac{\beta^\top A \beta}{\beta^\top B \beta}, \tag{2.33}
$$

where $A = \mathrm{Cov}_{Y,X}\mathrm{Cov}_{Y,X}^\top$ and $B = \hat{\Omega}$. The solution to this optimization problem is the maximal eigen-vector of AB^{-1} and $\rho^2_{Y,Z(\beta^*)}$ is the corresponding eigenvalue (Johnson, 2002).

When the size of the gene set is larger than the sample size in genomic data sets, that is $p > N$, matrix B from equation (2.33) is singular. Similar to the adjustment implemented in MANOVA-GSA (Tsai and Chen, 2009), a possible solution for the singularity is to incorporate a shrinkage covariance matrix as proposed previously by Schafer and Strimmer (2005). Thus the singular covariance matrix $\hat{\Omega}$ can be replaced with shrinkage covariance matrix $\hat{\Omega}^*$ given by $\omega^*_{hh'} = \rho^*_{hh'}\sqrt{\omega_{hh}\omega_{h'h'}}$ with shrinkage co-efficients $\rho^*_{hh'} = 1$, if $h = h'$ and $\rho^*_{hh'} = \rho_{hh'}\min\{1, \max(0, 1 - \hat{\lambda}^*)\}$,

if $h \neq h'$, where $\rho_{hh'}$ is the sample correlation between h-th and h'-th genes. The optimal shrinkage intensity $\hat{\lambda}^*$ is estimated as, $\hat{\lambda}^* = \sum_{h \neq h'} \text{Var}(\rho_{hh'}) / \sum_{h \neq h'} \rho^2_{hh'}$.

Incorporating the covariance matrix estimator into the test statistic reflects high computational cost which introduces lack of computational efficiency. Thus after identifying this computational efficiency problem for continuous phenotype Dinu et al. (2013) proposed a strategy. The strategy is to use an orthogonal transformation of the original gene expression measurements. To get the orthogonal basis vectors, Dinu et al. performed an eigenvalue decomposition of the shrinkage covariance matrix, $\hat{\Omega}^* = UDU^\top$. Then obtain $(V_1, \ldots, V_p) = (X_1, \ldots, X_p)UD^{-1/2}$ as orthogonal basis vectors. Now the square of the correlation is,

$$\rho^2(\gamma) = \frac{\gamma^\top \text{Cov}_{Y,V} \text{Cov}_{Y,V}^\top \gamma}{\gamma^\top \gamma} \tag{2.34}$$

where, $\gamma = D^{1/2}U^\top \beta$ and $\text{Cov}_{Y,V} = (\text{Cov}(Y, V_1), \ldots, \text{Cov}(Y, V_p))^\top$. According to a calculation of matrix algebra (Schafer & Strimmer, 2005), the coefficients of the most significant combinations are given by $\gamma^* \propto \text{Cov}_{Y,V}$. The Linear Combination Test statistic is proportional to the sum over the gene set of the square covariance between the phenotype and gene expression measurements. After taking orthogonal transformation,

$$\rho^2(\gamma^*) = c \sum_{j=1}^{p} \text{Cov}(Y, V_j)^2, \tag{2.35}$$

where c is a constant. This c can be ignored in the permutation test. We use permutation test to evaluate the statistical significance against the null hypothesis by permuting the phenotype labels. This approach is computationally advantageous as $\hat{\Omega}^* = UDU^\top$ is evaluated only once for the original data; we do not need to evaluate this covariance matrix for each permuted data set.

2.4 Gene Set Reduction for Binary Phenotype

We discuss here the concept and analysis procedure of gene set reduction proposed by Dinu et al. (2008). For that we first discuss some of the concepts of two different hypothesis of GSA methods described by Dinu et al., (2008). Some GSA methods are based on a priori defined pathway databases, such as Gene Ontology, KEGG and BioCarta. There are also some other applications which are not well defined a priori. In those cases researchers are more interested to know the information provided by the microarray data sets. In any of these analyses, it is interesting to find out whether all the genes or some of the genes are actually contributing to the disease. If a gene set contains a large number of genes, it may be possible to achieve the statistical significance easily, but all the members of that gene set may not be differentially expressed for the phenotype. For this, Dinu et al., (2008) have explored a new direction of finding a core subset of genes from a gene set. Initially authors have performed a study to get differentially expressed gene sets illustrating the difference of Q1 and Q2 hypothesis that we discussed in section 2.2, and showed supportive argument for self- contained methods over competitive methods.

Dinu et al. (2008) compared p-values obtained from Q1, Q2 and Q3 hypotheses with simulated data and point out the difference of these analytical procedure. They have generated 4000 gene expression profiles with two groups each containing 20 samples. One group with 2000 genes divided into 100 gene sets followed multivariate normal distribution, data generated based on Q2 hypothesis where mean vector is zero and constant off diagonal entry is generated from Uniform (0.5, 0.9). Correlated pair of genes are generated within a set uniformly from 0.5 to 0.9. Another group with 2000 genes were sampled following standard normal distribution considering gene expressions are not correlated. Dinu et al. (2008) computed average t-statistic in a gene set and compared the three hypotheses. Their result showed that Q2 hypothesis does not identified any differentially expressed gene sets and Q1 hypothesis incorrectly recognized 27 out of 100 gene sets which are differentially expressed with 0.05 cutoff level. Authors have also mentioned that by increasing the genes in gene sets with constant correlation (0.9) across genes showed more

misleading result. On the other hand, GSEA method identified 64 gene sets as differentially expressed. Dinu et al., (2008) nullify the idea proposed by Nam and Kim (2008) that GSEA is a mixed approach that works in between competitive and self-contained. The reason behind nullifying the idea is that GSEA method gives inaccurate result as a consequences of testing the genes using their correlation order. This approach can be problematic when it declares genes clustered in low correlation region as highly associated with the phenotype (Dinu et al., 2007), more specifically GSEA tests the clustering of the genes within a gene sets along the correlation with the phenotype axis. When the clustering occurs in the high correlation region, the gene-set is correctly identified as associated with the phenotype. But GSEA incorrectly identifies as significant genes clustered in the low to very low correlation region. Another limitation of GSEA is that it fails to identify those significant gene sets that are not clustered. For example, if a gene set consists of a mix of moderate to highly positive and negative correlations, GSEA will fail to identify that set is significant.

One aspect of simulation study is that, data generated based on competitive or self-contained method would naturally support either of the two methods. So the results of simulation study should be considered carefully. A limitation of self-contained method is that a gene-set may be identified as significant even if only a small number of genes are associated with the phenotype. Therefore, Dinu et al., (2008) proposed a method of finding core subsets of genes.

To get the significant gene sets Dinu et.al. (2008) used SAM-GS, which combines t-like statistics of single genes into a measure of association of a gene set with the phenotype. For a given gene set S, SAM-GS is the L_2 norm of the t-like statistics,

$$\text{SAM-GS} = \sum_{i=1}^{|s|} d_i^2, \tag{2.36}$$

where, $d_i = (\bar{x}_1(i) - \bar{x}_2(i))/(s(i)+s_0)$ is estimated for each gene i, $\bar{x}_1(i)$ and $\bar{x}_2(i)$ are the sample average of the two groups of phenotype, $s(i)$ is a pooled standard deviation over the two groups and s_0 is a small positive constant. Permutation test

is used to get the significance of gene set S. The principle for reducing the gene sets using SAM-GS is: for a pair of genes in S of gene i and gene j, $|d_i| > |d_j|$ suggests j belongs to a subset only if i belongs to that same subset. SAM-GSR is their proposed gene reduction method that gradually partitions the whole gene set into two subsets of genes and evaluates their associations with the phenotype.

SAM-GSR proceeds as follows. For each gene set select k genes with largest statistic $|d|$ gradually. For a reduced set, the stopping rule for the analysis is that the p-value of SAM-GS reaches to a certain threshold value. Following this procedure, core subsets of genes are obtained. Threshold values can be arbitrarily chosen by the researchers according to the biological association with the genes and the phenotype.

Chapter 3

Proposed Method

In this chapter, we describe our proposed method for obtaining core subsets of genes from significant gene sets. First, we describe the method for identifying gene sets. Then we discuss our algorithm for obtaining core genes to concentrate on specific genes that are chiefly contributing to the association of the set with the phenotype.

3.1 Identifying Significant Gene Sets for Continuous Phenotype

First, We need to identify the significant gene sets. Genes within a gene set are correlated as a consequence of sharing similar biological functions and same chromosomal locations. We use LCT, a gene-set analysis method described in section 2.3, which incorporates covariance matrix estimator into the test statistic. However, the covariance matrix is singular because of the number of genes in the sets exceeding the sample size. Incorporating the shrinkage covariance matrix estimator can be beneficial in this situation. To reduce the high computational cost of incorporating a shrinkage covariance matrix estimator, we perform an eigenvalue decomposition of the shrinkage covariance matrix which needs to be calculated only once for the original data. If the covariance estimator is, $\hat{\Omega}^* = UDU^\top$, then the orthogonal basis vectors are $(V_1, ..., V_p) = (X_1, ..., X_p)UD^{-1/2}$. The correlation among the genes in a set can be calculated as

$$\rho^2(\gamma) = \frac{\gamma^\top \text{Cov}_{Y,V} \text{Cov}_{Y,V}^\top \gamma}{\gamma^\top \gamma}. \tag{3.1}$$

Permutation test is used to evaluate the statistical significance against the null hypothesis by permuting phenotype labels. It is computationally advantageous because the shrinkage covariance matrix $\hat{\Omega}^* = UDU^\top$ for the orthogonal basis vector is computed only once in our original data. For this, we do not need to compute this covariance matrix for each permuted data. Details about this method are described in the section 2.3.

To get the p-values of the gene sets we permute the phenotype level 1000 times. We set up the significance level at 0.05.

3.2 Selecting Core Genes for Continuous Phenotype

After we obtain significant gene sets using LCT, we use a gene-set reduction method to obtain core genes. To the best of our knowledge, there are no current methods for reducing gene-sets to their core members. In this section we discuss our proposed algorithm for gene set reduction for a continuous phenotype. We state the steps of the reduction method. We also propose the criteria for choosing the cut off value that separates the core genes from the rest within a gene set.

Following are the steps of our proposed algorithm for gene set reduction for a continuous phenotype:

step 1: Obtain the list of significant gene sets applying Linear Combination Test (LCT) for continuous phenotype.

step 2: For each significant gene set, repeat the following. For all genes in a significant gene set apply Significant Analysis of Microarrays (SAM). For each gene, a SAM statistic d_i is obtained.

step 3: If S is the total number of genes in a gene set, then for $k = 1, \ldots, |S| - 1$, select the first k genes with largest statistic $|d_i|$ to form a reduced set R_k. Let

$\bar{R}_k$ be the complement gene set of R_k in S, and p_k be the corresponding LCT p-value of the complement gene set.

step 4: The final reduced set R_k is chosen such that p_k is larger than a certain threshold c, chosen by the analyst.

To accomplish the first step, we use Linear Combination Test, described in section 2.3.

In step 2, we obtain the SAM statistic for all the genes for continuous response type using R package (SAMr). The SAMr statistic for continuous response is defined as

$$d_i = \frac{r_i}{s_i + s_0}, i = 1, 2, ..., p \tag{3.2}$$

where r_i is the linear regression coefficient of expression measurements for gene i on the outcome, s_i is a pooled standard error of r_i and s_0 is the exchangeability factor or a small positive constant that adjusts for the variability in the microarray measurement. The calculations of r_i, s_i and s_0 are described in section 2.1.

In step 3, we order the genes inside a gene set according to SAM statistic with their decreasing order. That is, the first gene of a gene set is found to be the most differentially expressed, the second gene is the second most differentially expressed and so on. In this stage we actually want to check whether all the genes in this set are differentially expressed associated with the phenotype or a small number of the subset among the most differentially expressed genes are causing the whole gene set to be significant. To get the core genes, first we remove the gene that has the largest statistic and check whether the complement gene set is still differentially expressed by conducting the LCT method and obtaining p-values for the complement set.

In step 4, we choose a cut-off value as a stopping rule for taking the genes gradually and test the complement set with the LCT method. The procedure stops after the LCT p-value of the complement subset reaches a specified threshold. Actually researchers can arbitrarily choose any cut-off point based on the biological importance of the genes associated with the phenotype. Selecting this cut-off point for

getting core genes can be made more flexible by using different cut-offs for different gene sets. We used 0.1 cut-off as previously used by Dinu et al. (2008) for gene set reduction with binary phenotype. We used this large threshold so that we do not overlook the genes in a significant set that may not be differentially expressed by themselves but connecting with other genes as a teamwork, many have a biological impact associating with the phenotype. When the LCT p-value of a complementary set is reached at the threshold point, we can remove those genes of the complement set from the original gene set and obtain the core genes.

False Discovery Rate (FDR) is calculated as described by Storey 2002. FDR values can be used to adjust multiple comparison for testing multiple gene sets.

Chapter 4

Data Description and Results

In this chapter, we described the real microarray gene expression data that we used for our study, LEPTIN phenotype with which gene expression association is analyzed, and the processing of data to fit in our analysis. In the result section, we reported significant gene-sets with their p-values, described gene-set reduction with an example and reported core genes obtained from reducing the significant gene sets.

4.1 Data Description

We apply our method for obtaining significant gene sets and their reductions on a real Affymetrics microarray dataset. This dataset consist of genome-wide transcriptomic measurements of prostate tumor samples from African-American prostate cancer patients (Wallace et al., 2008) against the continuous phenotype of human LEPTIN gene expression values. Using surgical procedure primary prostate tumors were removed and the gene expression profiles were collected from these men. These patients did not receive any therapy prior to prostatectomy. These tumor samples were obtained from the National Cancer Institute-supported Cooperative Prostate Cancer Tissue Resource (CPCTR) and the Department of Pathology at the University of Maryland. According to Wallace et al. (2008), the macro dissected CPCTR tumor specimens were reviewed by a CPCTR-associated pathologist and

confirmed the presence of tumor in the specimens. The tissues were collected between 2002 and 2004 at four different sites.

We downloaded the expression data from Gene Expression Omnibus (Edgar et al. 2002). Our accession ID is GSE 6956. RNAs were labeled and hybridized using Affymetrix standard protocols. Detail RNA extraction, labeling and hybridization are discussed in the paper by Edgar et al. (2002). The gene expressions were centered and scaled across samples before applying a GSA method.

The incidence and mortality rate of prostate cancer actually varies in different regions and ethnic groups. In particular, African-American men have the highest risk of developing prostate cancer (Edgar et al. 2002). Edgar et al. showed in their paper that tumor immunobiology is different for African-American men and European-American men and explained the factors that make the differences. The authors suggested that the presence of genetic factors that increase the risk of prostate cancer may be higher when African-American men experience such disease. Therefore, we only use the gene expression levels of 33 African American men, which is the part of a larger microarray study into immunobiological differences in prostate cancer tumors between African-American and European-American men. The LEPTIN expression levels may be different between these two groups. So, for our analysis, we focus on the data consisting of 13,233 gene expressions measured in 33 African-American prostate cancer patients.

For the data we considered, LEPTIN levels may also be influenced by patient specific covariates, such as BMI, age and smoking status. Smoking status did not show a significant association with LEPTIN (p-value=0.36). BMI and age were not available for our analysis. The median age of protatectomy was 61 and the median prostate-specific antigen (PSA) at diagnosis was 6.1 ng/ml.

In table 4.1, we show an example of our gene expression data which represents genes on the rows and samples on the columns. Each cell represents the continuous measurement of gene expression level.

Table 4.1: An example of microarray gene expression data set.

Gene name	Sample 1	Sample 2	Sample 3	...	Sample N
Gene 1	88.00161	70.52682	76.43981	...	213.1661
Gene 2	71.51296	42.41818	37.55037	...	104.91981
⋮	⋮	⋮	⋮	⋱	⋮
Gene p	13579.23568	15149.50266	13089.98246	...	40574.49724

LEPTIN Phenotype

LEPTIN is a widely known marker protein for human adiposity, where excessive levels of adiposity damage health and lead to various chronic diseases. Circulating levels of *LEPTIN* in the blood are directly proportional to the total amount of body fat. *LEPTIN* is also associated with various metabolic and inflammatory conditions. Researchers found that increased plasma or serum *LEPTIN* levels are associated with the development of prostate cancer (Chang et al., 2001, Saglam et al., 2003 and Singh et al., 2010).

In our analysis, we screened sets of genes for the association with *LEPTIN* gene expression measurements. Here, we used *LEPTIN* gene expression measurements of the patients' as a surrogate measure of serum *LEPTIN* in blood. We used *LEPTIN* gene as a response variable structured in a vector of 33 expressions of the samples.

4.2 C2 Curated Gene Sets

In order to perform gene-set analysis, we need a data set of pre-defined gene sets. We downloaded C2 catalog from Broad Institute of MIT and Harvard, (`http://www.broad.mit.edu/gsea`) for a priori defined gene sets. This C2 catalog consists of 1,892 pre-defined gene sets. They are collected from on-line databases biomedical literatures including 340 PubMed articles, gene sets from published mammalian studies, and knowledge of domain experts. Sources of the gene sets are provided with gene set files in the C2 catalog. Gene set sizes in C2 catalog were restricted between 15 to 500 following Subramanian et al. (2005), so we used 1403

gene sets from C2 catalog that satisfied this restriction. Each gene set was screened for the association with LEPTIN gene expression, a well studied marker of adiposity, and various metabolic and inflammatory conditions. Currently the developers are working to create more automated method of curating gene sets from published literatures. In C2 catalog, rows represent gene-sets and columns represent genes. Each gene set contains pre-defined number of genes. An example of C2 catalog is given in table 4.2.

Table 4.2: An example of C2 curated gene set.

Gene set name	Gene name 1	Gene name 2	...	Gene name p
Gene set 1	Gene 1.1	Gene 1.2	...	Gene 1.p
Gene set 2	Gene 2.1	Gene 2.2	...	Gene 2.p
⋮	⋮	⋮	⋱	⋮
Gene set 1403	Gene 1403.1	Gene 1403.2	...	Gene 1403.p

4.3 Processing Data for Permutation (0/1 Matrix)

Conventional statistical tests such as t-test strictly assumes the data follows normal distribution. But using a permutation test we can avoid assuming the data is normal, as the permutation test does not use such an assumption. Even when the data is normal and has same variance for two groups, permutation test gives close to the result using equal variance for sufficiently large sample size (Ewens et al., 2006). Using a permutation test we can obtain an empirical distribution that do not assume normality and get p-values for statistical significance.

Based on the list of genes in the gene expression data and the gene sets in the C2 catalog, we process a new data set and refer to it as the 0/1 matrix. It enables us to check whether a gene from the gene expression data exists in the C2 catalog. The rows of the 0/1 matrix represents genes and each columns represents gene sets. The elements of this data file are 0 or 1. Let us denote the 0/1 matrix as M. Then M_{ij} denotes the cell entry for the i-th row and j-th column. If the entry is 1, it means

that the i-th gene from the gene expression data exists in the j-th gene set from C2 catalog. If the entry is 0, then it means that the i-th gene from the gene expression data does not exist in the j-th gene set from C2 catalog. This matrix is used as an input to the Linear Combination Test. This matrix helps during permutation by searching genes of gene expression levels in the gene sets defined in C2 catalog. An example of the 0/1 matrix is given in table 4.3.

Table 4.3: An example of mapping data: 0/1 Matrix.

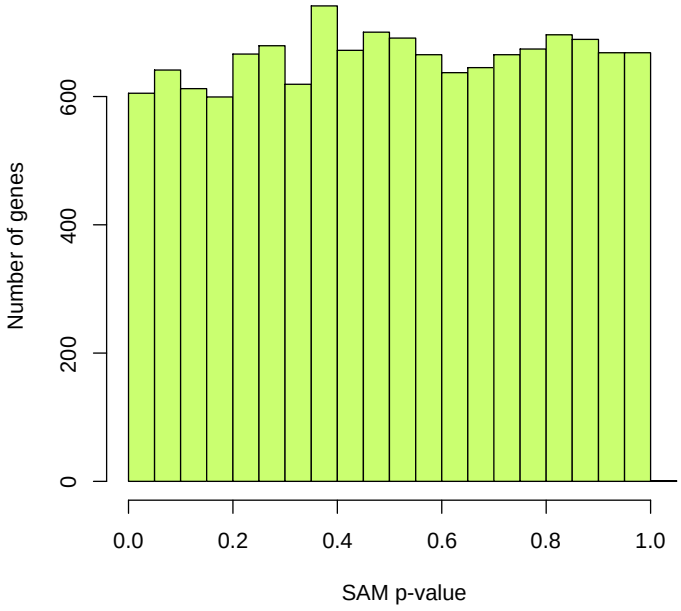
Gene name	Gene set 1	Gene set 2	Gene set 3	...	Gene set 1403
Gene 1	0	0	1	...	0
Gene 2	0	1	0	...	0
⋮	⋮	⋮	⋮	⋱	⋮
Gene 13233	1	0	0	...	0

4.4 Results

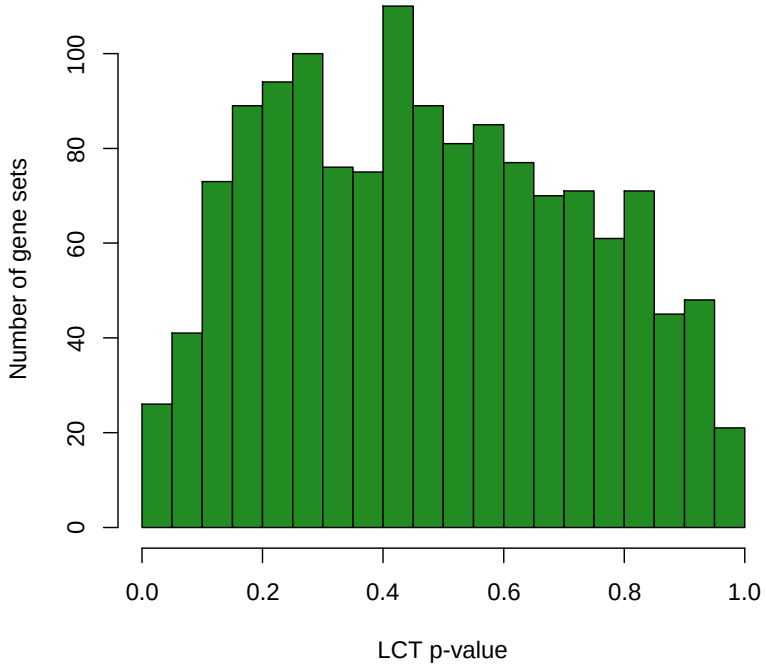
In this section, we first report results of individual-gene analysis obtained using SAM and gene-set analysis obtained using LCT. Then we describe gene-set reduction with an example and report core genes obtained from reducing the significant gene-sets.

4.4.1 Results Using SAM and LCT

First, we present results of individual gene analysis as an explanatory step before running gene set analysis. A histogram distribution of p-value from SAM are presented in Figure 4.1 (a). Here, Y-axis is representing number of genes and X-axis is denoting p-values of the SAM analysis. From this individual gene analysis, there are about 600 genes which have p-values between 0 to 0.05. Although a large number of genes are showing significance in SAM analysis but before running any analysis SAM can play an important tool as an initial step to identify differentially expressed genes in microarray studies.



(a)



(b)

Figure 4.1: Distribution of p-values using SAM and LCT. (a) Histogram of SAM p-values for individual gene analysis. (b) Histogram of LCT p-values for gene set analysis.

In figure 4.1 (b), we present a histogram of LCT p-values for the 1403 gene sets. Here, Y-axis is representing number of gene sets and X-axis is denoting the p-values of LCT analysis. LCT yields 66 gene sets when p-value is < 0.1. We report 30 significant gene sets from Dinu et al's (2013) paper. This same data set was used by Dinu et al. (2013) and they reported gene sets that are found significant by at least one of the four different GSA methods: LCT, LCT_2, SAM-GS and Global test with 5% significance level. We obtained gene sets from this list that has p-value< 0.1 from LCT analysis and use them for our gene set reduction analysis.

We report the 30 significant gene sets in table 4.4. The biological interpretation of the association of these 30 gene sets with *LEPTIN* phenotype is described in the section 4.4.3.

Table 4.4: LCT analysis: 30 gene sets associated with LEPTIN phenotype.

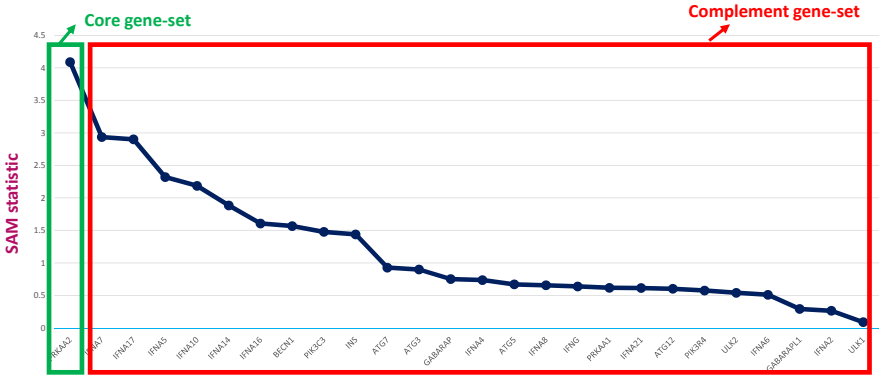
Gene Set	Size	P-value
NADLER_OBESITY_UP	46	0
HSA04920_ADIPOCYTOKINE_SIGNALING_PATHWAY	68	0.003
HSA04140_REGULATION_OF_AUTOPHAGY	26	0.004
HIF1_TARGETS	32	0.006
DORSEY_DOXYCYCLINE_UP	29	0.011
SHIPP_DLBCL_CURED_UP	28	0.013
JNK_UP	24	0.015
PROSTAGLANDIN_SYNTHESIS_REGULATION	28	0.016
CARDIACEGFPATHWAY	16	0.019
CITED1_KO_HET_UP	23	0.022
XU_CBP_DN	32	0.022
CHREBPPATHWAY	16	0.027
OXSTRESS_BREASTCA_UP	24	0.027
AGUIRRE_PANCREAS_CHR17	61	0.029
ST_GAQ_PATHWAY	27	0.031
HSA04340_HEDGEHOG_SIGNALING_PATHWAY	46	0.032
NFATPATHWAY	47	0.034
HYPOXIA_REVIEW	75	0.035
HSA04614_RENIN_ANGIOTENSIN_SYSTEM	16	0.04
CPR_NULL_LIVER_DN	16	0.041
HSA00380_TRYPTOPHAN_METABOLISM	49	0.043
HSA04630_JAK_STAT_SIGNALING_PATHWAY	135	0.045
DIAB_NEPH_UP	58	0.046
TRYPTOPHAN_METABOLISM	57	0.049
INSULIN_SIGNALING	93	0.049
PASSERINI_GROWTH	32	0.049
TNFA_NFKB_DEP_UP	18	0.05
FRUCTOSE_AND_MANNOSE_METABOLISM	24	0.055
ANDROGEN_AND_ESTROGEN_METABOLISM	21	0.058
POMEROY_DESMOPLASIC_VS_CLASSIC_MD_DN	38	0.091

4.4.2 Gene-Set Reduction for Continuous Phenotype with LCT

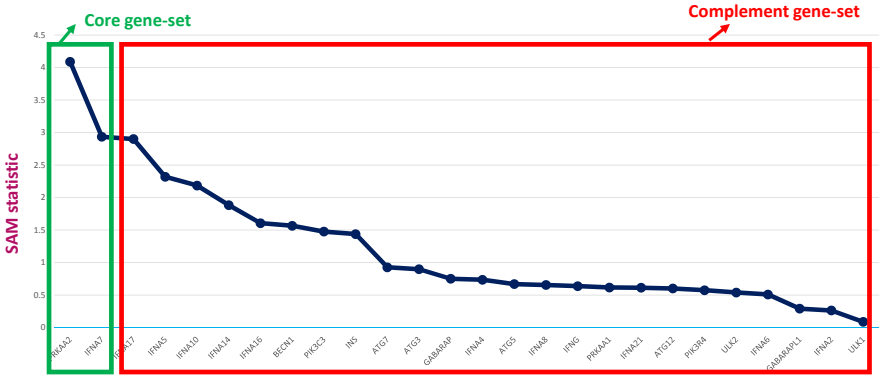
We used the SAM statistic to rank the differentially expressed genes inside a gene set in a decreasing order, so that we can gradually discover the core genes associated with *LEPTIN* phenotype. SAM is an analytical tool for microarray analysis at individual gene level. It can be used as an exploratory data analysis step before running gene-set analysis, and we presented a histogram with p-values in the previous section. Here we use SAM to rank the genes in a set, which is a step in our gene-set reduction algorithm. SAM statistics is calculated using the SAM-R package for our analysis. The SAM analysis gives result in both FDR values and p-values. But assuming that some of the FDR values can be similar for several genes, ordering the genes according to their significance would be a problem. On the other hand, each statistic for each gene is different than the others. Hence, we used the SAM statistic for continuous phenotype instead of p-values or FDR values.

We begin by presenting the reduction process for one of the sets, called REGULATION OF AUTOPHAGY. Autophay can act as a tumor-suppression mechanism. Again, defective autophagy provides oncogenic stimulus, causing malignant transformation and self-generated tumor. So, understanding the module of autophagy pathway may provide new approaches to cancer therapy and prevention (Dalby et al., 2010).

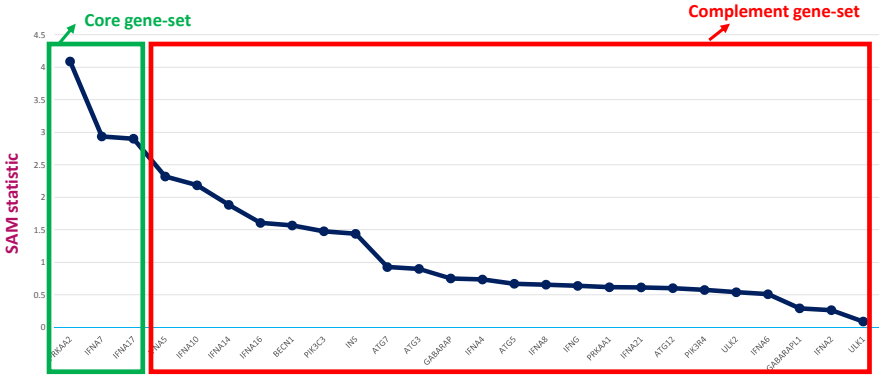
To illustrate our proposed gene-set reduction method, we present the following example using the real microarray expression data described in chapter 4. Figure 4.2 contains the plot for REGULATION OF AUTOPHAGY gene set which consists of 26 genes defined in the C2 catalog. We plot these 26 genes according to the decreasing order of the absolute value of their SAM statistic. First we select the gene with the largest SAM statistic PRKAA2 in the core set and the rest of the gene set forms the complement set. We obtain the LCT p-value of the complement set and check whether it reached our pre-specified cut off value 0.1. While the LCT p-value is still < 0.1 threshold, we gradually select the gene with the second largest SAM statistic IFNA7 and the third largest IFNA17. We found that the p-value of



(a)



(b)



(c)

Figure 4.2: Gene set reduction by example. This example shows graphically how our proposed method performs gene-set reduction. We used HSA04140_REGULATION_OF_AUTOPHAGY gene set, found significant by LCT, for this example. Each plot shows the SAM statistic (magnitude) of constituent genes of this gene set in decreasing order. Three plots corresponds to three consecutive iterations of the gene-set reduction method.

the complement set is > 0.1 after taking out PRKAA2, IFNA7 and IFNA17 genes. Genes in their complement set together are not differentially expressed with the phenotype. So the complement set represents the redundant gene set. Therefore, PRKAA2, IFNA7 and IFNA17 represent the core sub-set, associated with LEPTIN phenotype.

In table 4.5, we report the gene set sizes, core pathway sizes, percent reduction across each pathway and the core members of each pathway. By core pathway size we mean the number of core genes that we obtained from the reduction method from each significant gene sets. We calculated percent reduction by number of genes eliminated divided by the total number of genes in a set multiplied by 100.

We observed from the table that 20 of the sets are reduced to a single gene. This speaks to the fact that, a large number of genes in a gene set are not differentially expressed as we hypothesized before. The reduced subsets of these significant gene sets contain 31 unique core genes in total.

Table 4.5: Core subsets of genes associated with *LEPTIN* phenotype of 33 African American patients.

Gene set	Set size	Core pathway size	Percent reduction	Core pathway members		
				Gene1	Gene2	Gene3
NADLER _OBESITY_UP	46	1	97.83	LEP		
HSA04920 _ADIPOCYTOKINE _SIGNALING _PATHWAY	68	2	97.06	LEP	PRKAA2	
HSA04140 _REGULATION _OF_AUTOPHAGY	26	3	88.46	PRKAA2	IFNA7	IFNA17
HIF1 _TARGETS	32	1	96.88	EDN1		

DORSEY _DOXYCYCLINE _UP	29	1	96.55	REN		
SHIPP _DLBCL _CURED_UP	28	3	89.29	CYP4A11	DRP2	CLCNKB
JNK_UP	24	2	91.67	EFNB2	XRCC5	
PROSTAGLANDIN _SYNTHESIS _REGULATION	28	1	96.43	EDN1		
CARDIACEGF-PATHWAY	16	1	93.75	EDN1		
CITED1_KO _HET_UP	23	1	95.65	DDX6		
XU_CBP_DN	32	3	90.63	ERCC2	RAB14	EPHA4
CHREBP PATHWAY	16	1	93.75	PRKAA2		
OXSTRESS _BREASTCA_UP	24	1	95.83	EDN1		
AGUIRRE _PANCREAS _CHR17	61	2	96.72	SPANXA1 SPANXA2 SPA	GRP	
ST_GAQ _PATHWAY	27	3	88.89	NFKB2	PDK1	DAG1
HSA04340 _HEDGEHOG _SIGNALING _PATHWAY	46	2	95.65	WNT10B	PRKACG	
NFATPATHWAY	47	1	97.87	EDN1		
HYPOXIA _REVIEW	75	1	98.67	EDN1		
HSA04614_RENIN _ANGIOTENSIN _SYSTEM	16	1	93.75	REN		
CPR_NULL _LIVER_DN	16	1	93.75	CYP7B1		

HSA00380 _TRYPTOPHAN _METABOLISM	49	1	97.96	KMO		
HSA04630 _JAK_STAT _SIGNALING _PATHWAY	135	1	99.26	LEP		
DIAB_NEPH_UP	58	1	98.28	PTPRB		
TRYPTOPHAN _METABOLISM	57	1	98.25	TDRD12		
INSULIN _SIGNALING	93	1	98.92	LEP		
PASSERINI _GROWTH	32	1	96.88	EDN1		
TNFA_NFKB _DEP_UP	18	2	88.89	CXCL3	NFKB2	
FRUCTOSE _AND_MANNOSE _METABOLISM	24	2	91.67	PFKFB4	FBP2	
ANDROGEN_AND _ESTROGEN _METABOLISM	21	1	95.24	CYP11B2		
POMEROY _DESMOPLASIC _VS_CLASSIC _MD_DN	37	1	97.30	CITED1		

In table 4.6, we report the frequencies of 31 core genes obtained from the reduction method and their p-values from SAM analysis. Core gene *EDN1* or *Endothelin-1* is the most frequently appeared gene, appeared in 7 significant gene sets. *LEP* or *LEPTIN* appeared 4 times, *PRKAA2* or *Protein kinase, AMP-activated, alpha 2 catalytic subunit* 3 times, *NFKB2* or *Nuclear factor NF-kappa-B p100 subunit* 2 times, *REN* or *Renin* 2 times and rest of the 26 core genes appeared once in the significant gene sets.

4.4.3 Biological Interpretation of Our Findings

Biological interpretation and validation of statistically significant results is an essential step in gene-set analysis. We found that *Adipocytokine* signaling Pathway is significantly associated with *LEPTIN* phenotype with p-value 0.003 (FDR 0.7015). *Adipocytokines* including *LEPTIN* are a group of adipose tissue-derived hormones. *Adipocytokines* play an important role in the regulation of angiogenesis and tumor growth connected to obesity and diabetes (Housa et al., 2006). *Regulation of autophagy* is another significant gene set found associated with *LEPTIN* expression measurement (p-value 0.004 and FDR 0.894802). Autophagy is a fundamental process in tumorigenesis and treatment response. It acts as a tumor-supression mechanism (White et al., 2009). On the other hand, defective autophagy provides oncogenic stimulus, causing malignant transformation and self-generated growing tumor. Some studies have showed that inhibiting autophagy in cancer cells may be therapeutically beneficial in some circumstances.

EDN1 or Endothelin-1 produced by prostate epithelia is a core gene of *HIF1 TARGET* pathway. It plays an important role in the progression of prostate cancer (Nelson et al., 2005). *Endothelin-1* prohibits apoptosis in prostate cancer meaning that it hinders the normal process of Programmed Cell Death (PCD). PCD is a natural fundamental functioning for both plants and multicellular organisms. Defective process of apoptotic by *Endothelin-1* stops the processing of PCD, thus developing of cancer occurs in the body by cell proliferation. Prostate cancer is found in association with this increased amount of *Endothelin-1* gene.

Table 4.6: Frequencies of the genes selected in the core gene sets

Genes	Frequency	P-value
EDN1	7	0.00032
LEP	4	0.000076
PRKAA2	3	0.001831
NFKB2	2	0.030352
REN	2	0.001613
CITED1	1	0.006569
CLCNKB	1	0.01356
CXCL3	1	0.01181
CYP11B2	1	0.01073
CYP4A11	1	0.00363
CYP7B1	1	0.005802
DAG1	1	0.04713
DDX6	1	0.002168
DRP2	1	0.003859
EFNB2	1	0.002347
EPHA4	1	0.036321
ERCC2	1	0.009041
FBP2	1	0.029190
GRP	1	0.001197
IFNA17	1	0.010825
IFNA7	1	0.010224
KMO	1	0.000113
PDK1	1	0.030424
PFKFB4	1	0.008381
PRKACG	1	0.003645
PTPRB	1	0.00142
RAB14	1	0.025018
SPANXA1 /// SPANXA2 /// SPA	1	0.000561
TDRD12	1	0.00173
WNT10B	1	0.002774
XRCC5	1	0.003382

Chapter 5

Conclusion

Our gene-set reduction method is an extension of GSA self-contained method from binary to continuous phenotype. Many of the GSA self-contained methods are generalized for binary or categorical outcomes. But for continuous phenotype, extension of self-contained GSA methods have rarely been reported. As an extension of GSA self contained method, we obtained significant gene-sets and successfully reduced subsets to its core genes.

We use LCT for obtaining significant gene sets for continuous phenotype. There are several benefits of using LCT method for continuous phenotype. First, the extension of the enrichment test for continuous phenotype is rigorous and computationally efficient. Because there is scope of inaccurately categorizing the response variable, the variable may not still be informative about the disease progression after categorizing. Second, while most of the traditional GSA methods are unsuccessful to accommodate the correlation characteristic in the test statistic, our method incorporates correlation among the similar genes in a set. The incorporation of covariance matrix into the test statistic gives better power using permutation method (Dinu et al., 2013). But this covariance matrix gets ill-conditioned when genes in a set are larger than the sample size. Calculation of a shrinkage covariance matrix estimation can solve this problem. But the computational cost of this solution is high. The computational cost efficiency problem is overcome by taking orthogonal transformation of the gene expressions. Therefore, eigenvalue decomposition of

the shrinkage covariance matrix is performed only once for the real gene expression data and it does not need to be estimated for each permuted version of data.

Our method adds to the gene-set analysis of a continuous phenotype literature by providing the community with a tool for reducing the sets to their members. Our method incorporates some improvements into existing GSA methods. We hope this method can be used as an advantageous tool for testing the association between different molecular pathways and gene signatures. Again, a significant reduced gene set helps understand the biological mechanism underlying the gene-set associated with a phenotype of interest. Targeted therapies and intervention strategies (Ein-Dor et al., 2006) may improve by concentrating only on the reduced gene sets. On the other hand, examining the expression levels of the genes that are not differentially expressed with the phenotype can increase the unnecessary cost and fails to improve in clinical decision makings.

Future Studies

An Explanatory data analysis of our gene expression data can be done before running a formal inference. But a small number of sample size can be a limitation to check thoroughly for non-linearity. Furthermore, LCT method can be extended for non-linear data using a large number of sample size,.

The results of our analysis can play an important role as a source of information to improve personalized medicine and intervention therapy by interpreting the biological association with our obtained core genes. We attempt explaining the biological interpretation by reviewing literature in section 4.4.3. We suggest to build the biological understanding underlying a disease mechanism through an extensive literature review.

Software Packages

Linear Combination Test (LCT) has been analyzed with free R software, version 2.15.3. Free R codes for performing LCT for continuous and binary phenotype is available at http://www.ualberta.ca/∼yyasui/homepage.html. Our operating system used for running the codes is Windows 8. SAS 9.3 is used to process our mapping 0/1 matrix data using gene expression data and C2 catalog.

References

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17, 509–519.

Chang, S., Hursting, S. D., Contois, J. H., Strom, S. S., Yamamura, Y., Babaian, R. J., Troncoso, P., Scardino, P. T., Wheeler, T. M., Amos, C. I. and Spitz, M. R. (2001). Leptin and prostate cancer. *Prostate.* 46: 62–67.

Chu, G., Narasimham, B., Tibshirani, R., and Tusher, V. (2002). SAM "Significance Analysis of Microarrays", Users guide and technical document. Stanford University.

Dalby, K. N., Tekedereli, I., Lopez-Berestein, G., Ozpolat, B. (2010). Targeting the prodeath and prosurvival functions of autophagy as novel therapeutic strategies in cancer. *Autophagy*, 6(3):322–329.

Dinu, I., Potter, J., Mueller, T., Liu, Q., Adewale, A., Jhangri, G., Einecke, G., Famulski, K., Halloran, P., Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8(1): 242.

Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulsky, K.S., Halloran, P. F., Yasui, Y. (2008). Gene Set Analysis and Reduction. *Briefings in Bioinformatics*, 10(1): 24–34.

Dinu, I., Wang, X., Kelemen, E. L., Vatanpour, S. and Pyne, S. (2013). Linear combination test for gene set analysis of a continuous phenotype. *BMC Bioinformatics*, 14:212.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su Y.A., Trent, J. M. (1996). Use of a cDNA microarray to analyse gene

expression patterns in human cancer. *Nature genetics*, 14:457–460, PubMed.

Draghici, S., Khatri, P., Martins, P. R., Ostermeier, C. G., and Krawetz, A. S. (2003). Global functional profiling of gene expression. *Genomics* 81.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30, 207–210.

Ein-Dor, L., Zuk, O., Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci* USA. 103(15):5923–5928.

Ewens, J. W., and Grant, R. G. (2006). *Statistical Methods in Bioinformatics: An Introduction (Statistics for Biology and Health)*, 2nd edition.

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

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Housa, D., Housova, J., Vernerova, Z., and Haluzik, M. (2006). Adipocytokines and Cancer. *Physiol. Res.* 55:233–244.

Johnson, R. A. and Wichern, D. W., (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall.

Mootha, V. K., Lindgren, C. M., Eriksson, K.F. (2003). PGC- 1alpha-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. *Nature Genetics*, 34:267–273.

Nam, D. and Kim, S.Y. (2008). Gene-set approach for expression pattern analysis. *Brief Bioinformatics*, 9:189–197.

Nelson, J. B, Udan, M. S., Guruli, G., Pflug, B. R., (2005). Endothelin-1 inhibits apoptosis in prostate cancer. *Neoplasia*.7:631–637.

Pusztai, L., Ayers M, Stec J., Hortobagyi, G. N. (2003). Clinical application of cDNA microarrays in oncology. *Oncologist*, 8:252–258.

Saglam, K., Aydur E, Yilmaz M, Goktas S. (2003). Leptin influences cellular differentiation and progression in prostate cancer. *J Urol* 169:1308–1311.

Schafer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.* 4(32).

Singh, S.K., Grifson J.J., Mavuduru R.S., Agarwal M.M., Mandal A.K., Jha V. (2010). Serum leptin: A marker of prostate cancer irrespective of obesity. *Cancer Biomarkers*, 7(1):11–15.

Storey, J.D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* USA 100, 9440–9445.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:479–498.

Subramanianan, A., Tamayoa, P., Moothaa, K. V., Mukherjeed, S., Eberta, L. B., Gillettea A. Michael, Amanda Paulovichg, Scott L. Pomeroy, Todd R. Golub, Eric S. Landera and Jill P. Mesirova. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci.* USA, 15545–15550.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, 11:1227–1236.

Tian, L., Greenberg, S., Kong, S., Altschuler, J., Kohane, I. and Park, P. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci.* 102:13544–13545.

Tsai, C. and Chen J.J. (2009) . Multivariate analysis of variance test for gene set analysis. *Bioinformatics* 25(7):897–903.

Tusher, G. V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* USA 98: 5116–5121.

Wang, X, Dinu I., Liu W., Yasui Y. (2011). Linear Combination Test for Hierarchical Gene Set Analysis. *Statistical Applications in Genetics and Molecular Biology*, 10(1): Article 13.

Wallace, T.A., Prueitt R.L., Yi M.H., Howe T.M., Gillespie J.W., Yfantis H.G.,

Stephens R.M., Caporaso N.E., Loffredo C.A., Ambs S. (2008). Tumor Immunobiological Differences in Prostate Cancer between African-American and European-American Men. *Cancer Research,* 68:927–936.

White, E., DiPaola, R.S. (2009). The double-edged sword of autophagy modulation in cancer. *Clin. Cancer Res.* 15:5308–5316.