University of Alberta

AN APPLICATION OF GENE SET ANALYSIS FOR A COMPARISON OF TWO GROUPS

by

Ya Meng

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

Biostatistics

Department of Mathematical and Statistical Sciences

Fall, 2011

Edmonton, Alberta

Abstract

Background: Microarrays are biotechnological advancements measuring expressions of thousands of genes in a single assay. A two-group microarray study yields gene expression measurements for patients with a disease of interest and for healthy controls. Successful identification of genes differentiating between the two groups leads to new and improved treatments. While microarrays represent an exciting avenue for clinicians, the analysis of the large amount of data coming from these experiments comes with many challenges. The large number of features measured on a relatively small number of patients (the so-called $p \gg N$ problem), the small variability in some of the genes, and the correlations across genes are characteristics of microarray data and need to be addressed in analysis.

Objective: Our objective is to apply the methods of microarray data analysis to kidney transplant patients. The two groups consist of patients experiencing a more severe type of rejection, T-Cell Mediated rejection (TCMR), versus patients experiencing a borderline rejection.

Methods: We apply Significance Analysis of Microarray (SAM) to explore genes differentially expressed between the two groups. While SAM is a sound statistical method useful for exploring the data at the gene level, the output of thousands of significant genes is hard to interpret. There has been a shift of focus towards analysis at the gene set level. Biologists put together databases consisting of genes grouped by biological function, called biological pathways, or gene sets. The analysis at a gene set (pathway) level, called Gene Set Analysis (GSA), is easier to interpret, and more robust, in the sense that

significant gene sets are more likely to be replicated across studies and microarray platforms. GSA addresses the $p \gg N$ problem via permutation tests. GSA methods can be broadly classified into self-contained methods, based on group labels permutations, and competitive methods based on subject permutations. We prefer the former, as it is preserves the correlations among genes in a pathway. We present the two top self-contained methods, called Significance Analysis of Microarrays for Gene Sets (SAM-GS) and Multivariate Analysis of Variance for Gene Sets (MANOVA-GSA), as they perform best according to previous simulation studies and real applications. We also present results of the most popular GSA, which is a hybrid between self-contained and competitive methods. False Discovery Rates are calculated to address multiple hypothesis testing.

Results: Our data consists of expression measurements for 54,675 probes on 17 kidney transplant patients experiencing TCMR and 27 kidney transplant patients experiencing borderline rejection. The 54,675 expression measurements were reduced to 20,736 unique genes. For gene sets, we use the most recent version of the C2 catalogue consisting of 1892 gene sets, representing metabolic and signalling pathways from online pathway databases, gene sets from biomedical literature including 340 PubMed articles, and gene sets compiled from published mammalian microarray studies. We restricted the size of gene sets to be between 5 and 500, resulting in 1,839 gene sets used for our analysis. We found 957 significant genes with FDR values smaller than 5.71%. SAM-GS identified 58 pathways with p-value< 0.001 (FDRs $< 1.8\%$). Among these, CDK5 and Interferon-gamma are only two examples of pathways previously established as associated with kidney transplant rejection.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Brief overview of DNA microarray studies

DNA microarrays are assays for quantifying the types and amounts of messenger RNA (mRNA) transcripts, known as complementary chains of DNAs, present in a collection of cells. In a DNA microarray study, blood is collected from each subject, the RNA is extracted and then the mRNA is isolated, and placed on the microarray platform. The microarray consists of a solid surface on which strands of polynucleotides have been attached in specified positions. We refer to the polynucleotides immobilized on the solid surface as probes. The probes consist of complementary DNA (cDNA) printed on the surface of a microarray chip. Research laboratories order one such microarray chip for each subject considered in their study. One example of a company manufacturing these chips is Affymetrix. The mRNA from a subject binds with the cDNA on the chip, if they share sufficient sequence complementarities. The intensity of binding is quantified into numbers via a complex process called hybridization. These numbers represent the probe expression measurements

for all targeted probes on the microarray chip. For example, the number of probes for an Affymetrix microarray chip is approximately 40,000, mapping to around 20,000 genes. Scientists are interested in identifying those genes whose expressions are different between cases and healthy controls, or between two groups of patients.

1.2 Challenges in analysis methods for DNA microarray studies

Although there are other kinds of data storing information on gene expressions, microarray data contains the largest, most complete information of gene expression. The huge number of genes measured on a relatively small number of samples presents a difficult challenge in the analysis of DNA microarray data. This is referred as the $p \gg N$ problem, also called the high-dimensionality problem. Because of the high dimensionality problem, the classical analysis techniques are no longer applicable to DNA microarray data.

Another challenge in analysis of microarray data is the small variability in the gene expression measurements for some of the genes. Based on previous experiences with microarray studies, the signal to noise ratio decreases with decreasing gene expressions. However, even for a given level of expression, the fluctuations are gene specific. To account for gene specific fluctuations, a ratio of change in gene expression to the standard deviation in the data for that gene is employed. To compare values of those ratios across all genes, the distribution of these ratios should be independent of the levels of gene expressions. At low expression levels, variances in the ratios can be high because of small values of

the denominators. To ensure that the variances of the ratios are independent of gene expressions, a small positive constant, estimated from the data, is added to the denominator.

An important challenge in analysis of microarray data is inherent to the multiple hypothesis problem. For example, among 10,000 genes, even if we set the threshold for p-values as low as 0.01, we will still get over 100 genes significant, which is a large proportion. Various adjustments for multiple testing in microarray data have been proposed, and it is recommended that they should be part of the output.

In response to these challenging characteristic of microarray data, Significant Analysis of Microarray (SAM) proposed a moderated t-test statistic, together with a False Discovery Rate type of adjustment, calculated based on subject label permutation tests. The high dimensionality problem calls for permutation tests, which are at the basis of calculating significance of measures of association between a gene and the disease of interest. Once a test statistic is calculated for the original data, its significance is evaluated by calculating the test statistic for permuted versions of the data set. Under the null hypothesis of no association, the group labels are interchangeable. A null distribution of the test statistic is estimated and the p-value is calculated based on it, as the proportion of times the permuted test statistic is as extreme, or more extreme than the observed test statistic.

Analyzing microarray data at an individual gene level usually leads to a list of thousands of significant genes, even after multiple comparison adjustments have been made. The process of trying to interpret such a large list of genes is cumbersome. Moreover, replication of the findings in different microarray experiments, or across different platforms of thousands of genes, is another

serious challenge with gene level analysis. Starting in the past decade, molecular biologists have put together lists of genes grouped by function, also called biological pathways. Various pathway databases have been put together, for example KEGG [15], Gene Ontology [6], Biocarta [2] and Molecular Signature Data Base [3].

There has been a shift in focus from gene level analysis to pathway level, or gene set level. Many Gene Set Analysis (GSA) methods have been proposed in the past decade. The most popular one is Gene Set Enrichment Analysis (GSEA). A GSA works with the matrix of gene expressions, as well as a disease-relevant collection of biological pathways. The gene sets are a-priori determined and based on biological knowledge. GSA assigns each gene a significance value. An interpretation of this list of p-values leads to identification of biologically relevant pathways that can be used for early diagnosis, or treatment of disease. A schematic illustration of GSA is given in Figure 1.1.

1.3 Thesis Organization

This thesis is organized in four chapters. In this chapter, we have given a brief overview of DNA microarray technology and describe challenges associated with analyzing data measured by microarray studies. In Chapter 2, we discuss statistical analysis methods for microarray studies, starting with analysis at individual gene level by the most popular method called Significance Analysis of Microarrays (SAM). We present the framework for a more biologically meaningful type of analysis, called gene set analysis. We discuss one of the first gene set analysis methods called Over Representation Analysis (ORA). We present the most popular gene set analysis method, called

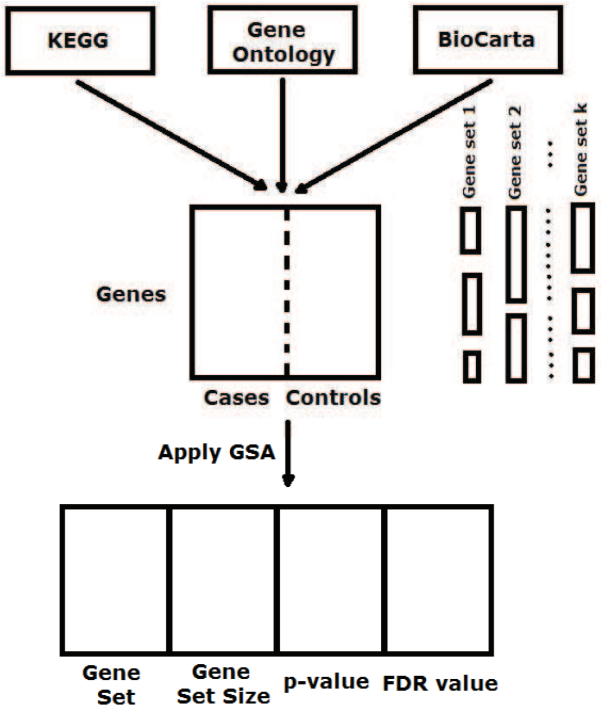

Figure 1.1: Schematic illustration of Gene Set Analysis(GSA)

Gene Set Enrichment Analysis (GSEA), and two more recent analysis methods called Significance Analysis of Microarrays for Gene Sets (SAM-GS), and Multivariate Analysis of Variance for Gene Sets Analysis (MANOVA-GSA). The third section mainly discusses the issues of multiple hypothesis testing in microarray studies, and the calculations of Family-wise Error Rate (FWER) and False Discovery Rate (FDR) values. Chapter 3 compares results of SAM, GSEA, SAM-GS and MANOVA on a microarray study comparing two groups: patients experiencing a more severe type of rejection after kidney transplant (T-cell mediated rejection), versus patients experiencing a borderline rejection. In Chapter 4, we discuss future directions of gene set analysis.

Chapter 2

Methods

2.1 Single Gene Analysis: SAM Method

Significance Analysis of Microarrays (SAM) [25] is the most popular single gene analysis method specifically designed for data measured by microarray studies. Microarray data is an important gene expression data set and usually contains a large amount of information. It represents expressions of thousands of genes for different biological states, as measured by a phenotype of interest. SAM uses permutations of the phenotype labels to determine the significance of the genes used to interrogate the differences between two groups under the phenotype.

Suppose we have a microarray data set on two groups of patients (see Table 2.1 for example). For each gene i, we denote by x_{ij}^1 the gene expression measurements corresponding to patient j in group 1, and by x_{ik}^2 the gene expression measurements corresponding to patient k in group 2.

SAM tests the null hypothesis of no difference between the two states of

	Group 1	$\cdots$	Group 1	Group 2	$\cdots$	Group 2
Gene 1	4.980	$\cdots$	5.000	5.307	$\cdots$	5.165
Gene 2	7.590	$\cdots$	8.600	8.251	$\cdots$	8.619
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gene k	6.931	$\cdots$	7.289	7.926	$\cdots$	6.944

Table 2.1: Example of a gene expression data set

the phenotype, or the two groups, for each gene i:

$$H_0 : \quad \mu_{1i} = \mu_{2i}.$$

SAM uses a moderated t-test, denoted as d_i, to measure the change of expression of gene i between the two groups by adding a constant s_0 to the denominator:

$$d_i = \frac{x^1_{i\cdot} - x^2_{i\cdot}}{s_i + s_0},$$

where $x^1_{i\cdot}$ and $x^2_{i\cdot}$ are the average expressions of gene i in each of the two groups and s_i is the pooled standard deviation across the two groups of patients of gene i:

$$s_i = \sqrt{a(\sum_{j=1}^{n_1}[x^1_{ij} - x^1_{i\cdot}]^2 + \sum_{k=1}^{n_2}[x^2_{ik} - x^2_{i\cdot}]^2)}, \tag{2.1}$$

where

$$a = \frac{(\frac{1}{n_1} + \frac{1}{n_2})}{n_1 + n_2 - 2},$$

n_1 and n_2 are the number of measurements in the two groups. This is equivalent to the pooled standard deviation of two sample unequal variance t-test:

$$s_i = s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$\text{where} \quad s' = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}}$$

$$sd_1 = \sqrt{\frac{1}{n_1-1}\sum_{j=1}^{n_1}[x_{ij}^1 - x_{i\cdot}^1]^2}$$

$$sd_2 = \sqrt{\frac{1}{n_2-2}\sum_{k=1}^{n_2}[x_{ik}^2 - x_{i\cdot}^2]^2}$$

Details about calculating s_0 are as the following:

1. Let s^α be the α percentile of the s_i values. Let $d_i^\alpha = \frac{r_i}{s_i + s^\alpha}$.

2. Compute the 100 quantiles of the s_i values, denoted by $q_1 < q_2 < \ldots < q_{100}$.

3. For $\alpha \in (0, 0.05, 0.1, \ldots, 1)$, compute $v_j = mad(d_i^\alpha | s_i \in [q_j, q_{j+1}))$, $j = 1, 2, \ldots, n$, where mad is the median absolute deviation from the median, divided by 0.64.

4. Compute $cv(\alpha) =$ coefficient of variation of the v_j values.

5. Choose $\hat{\alpha} = argmin[cv(\alpha)]$. Finally compute $s_0 = s^{\hat{\alpha}}$.

Moderating the test statistic using the small positive constant s_0 is necessary to make the distribution of d_i independent of the levels of gene expressions, the values of measurements. More precisely, to compare values of d_i across all genes, the distribution of d_i should be independent of the levels of gene expressions. At low expression levels, variance in d_i can be high because of small values of s_is. To ensure that the variance of d_i is independent of gene expressions, we added a small positive constant s_0 to the denominator of the modified test statistic. Although increasing the threshold would also result in

larger p-values, the cut-off would have to be decided in a very ad-hoc manner, varying across data sets studied. On the other hand, choosing s_0 minimizes the coefficient of variation of d_i as a function of s_i in moving windows across the data. It also helps with the calculation of FDR values as well as an automatic way of analyzing microarray data sets.

Based on the microarray data, we rank all the d_i values of the original data and get an order statistic, denoted by $d_{(i)}$. Then $d_{(1)}$ is the largest relative difference among all the genes. To find the potentially significant changes in expression, we permute the group labels and get a new data set with the same measurements for each gene i and different measurements for the two groups. Usually 1,000 random permutations of the group labels are performed. For each permutation p, $p = 1, 2, \ldots, 1000$, the statistic for each gene i under the p-th permutation is calculated, and denoted by d_i^p. We also get the order statistics $d_{(1)}^p \geq d_{(2)}^p \geq \ldots \geq d_N^p$ under each permutation p, where N is the total number of genes. Based on all $1,000$-permutations, we calculate the expected value of the moderated t-tests as:

$$\bar{d}_{(i)} = \frac{\sum_p d_{(i)}^p}{p}.$$

Then we compare the observed moderated t-test $d_{(i)}$ vs. the expected value $\bar{d}_{(i)}$. Gene i will be considered as not associated to the phenotype if $d_{(i)} \cong \bar{d}_{(i)}$. Typically, most of the genes are not associated with the phenotype. Otherwise, gene i exhibits a larger distance between $d_{(i)}$ and $\bar{d}_{(i)}$. Genes exhibiting differences larger than a pre-specified threshold Δ are labelled as associated with the phenotype. All genes with positive relative differences and satisfying $d_{(i)} - \bar{d}_{(i)} > \Delta$ are called "positive significant". Similarly, all genes with

negative relative differences and satisfying $\bar{d}_{(i)} - d_{(i)} > \Delta$ are called "negative significant". The smallest and largest cutpoints are denoted by $d^p_{(i_1)}$ and $d^p_{(i_2)}$ respectively.

Adjusting for multiple hypothesis testing is essential in microarray data analysis methods: with an alpha level of 0.05 applied to 10,000 tests we expect approximately 500 false positive findings. SAM uses False Discovery Rate (FDR) to address the multiple hypothesis testing problem. FDR is calculated as the ratio of estimated falsely significant genes over the number of significant genes declared. The number of falsely significant genes is estimated by the average of number of significant genes from all permutations.

Details about calculating FDR value for SAM is given below:

1. For a grid of δ values, compute the total number of significant genes. Also, compute the median number (for i) of values among the d^p_is that falls between $d^p_{i_1}$ and $d^p_{i_2}$ under each permutation p, which are declared as number of falsely significant genes.

2. Compute the 25%, $q75\%$ quartiles, $q25$, $q75$, of all permuted d values (in total $p \times N$ values).

3. Compute $\pi_0 = d_i \in (q25, q75)/(0.5N)$.

4. Let $\pi_0 = min(\pi_0, 1)$ (i.e. truncate at 1). The median percentile of the number of median falsely significant genes is multiplied by π_0 and the result is the final number of median falsely significant genes.

5. The FDR is calculated by the rate of median falsely significant genes over the number of genes called significant.

For a given threshold Δ, the number of significant genes in each permutation is calculated as the number of genes with $d^p_{(i)} > d^p_{(i_1)}$ or $d^p_{(i)} < d^p_{(i_2)}$, where, $d^p_{(i_1)}$ and $d^p_{(i_2)}$ are the cutpoints discussed before. Generally speaking, when the threshold Δ increases, the number of significant genes will decrease and so does FDR. Alternative way of choosing the quartiles of FDR is using 90th percentiles in the above steps. Values of FDR (25%) are considered acceptable in analysis of microarray data. We note that published bioinformatics methods are based on ad-hoc choice of parameters, and many are not tested for type I and type II error via simulations. For example, no theoretical justification has been given to choice of 25% and 75% quartiles in estimating SAM-FDR.

SAM has been extended to cover a wide range of phenotypes such as multi-class independent (i.e. more than two independent groups), multi-class correlated (i.e. paired data or repeated measurements over time), continuous response, and even censored survival time data.

2.2 Gene Set Analysis

Extracting clear and coherent information from microarray studies has always been challenging. Many sound methods have focused on the development of techniques for accurate identification of genes associated with a disease and with evaluation of their statistical significance in a variety of experimental designs [21]. However, an analysis at the individual-gene level may yield a list of thousands of significant genes, with only a small number of them being truly associated, making it very difficult for the scientist to interpret it. Therefore, recent efforts have focused on the discovery of significant biological pathways, or sets of genes, rather than individual genes. Usually those gene sets are

"priori" defined by previous biological researches. Another important aspect motivating this shift of focus in the analysis of microarray studies is that the gene set of interest may involve moderate effects of individual members that are not captured by analysis at the individual gene level, but will be captured by gene set analysis.

The null hypothesis for individual gene analysis can be expressed as:

$$H_0 : \quad \mu_1(i) = \mu_2(i),$$

where $\mu_1(i)$ and $\mu_2(i)$ represent the population means of the i-th gene expression measurement for each of the two groups, respectively. In the gene-set analysis context, this becomes a **multivariate** null hypothesis

$$H_0 : \quad \mu_1 = \mu_2,$$

where μ_1 and μ_2 are the mean expression vectors over the gene set of interest in each of the two groups. Gene set analysis comes with many challenges, due to important characteristics of the data:

1. The number of genes is far larger than the number of observations (large p, small n problem).

2. Gene expression measurements, especially within each gene set, can be highly correlated.

3. As new pathways are discovered, efficient gene set analysis methods are needed to deal with the computational burden of testing thousands of sets.

We note that the traditional Hotelling's T^2 test for testing the multivariate null hypothesis is no longer feasible in the gene set analysis setting, because of the large p, small n problem. Permutation based tests are necessary to respond to this problem. An extensive review of gene set analysis methods has been performed by Nam and Kim [19]. The importance of distinguishing between gene-versus subject-sampling methods was first raised by Tian et al. [23]. The term "sampling" here refers to permutation tests employed by gene set analysis methods. Goeman and Buhlmann [14] discussed methodological ideas underlying gene set analysis and established the distinction between testing "self-contained null hypotheses" and testing "competitive null hypothesis". "Self-contained null hypotheses" are defined as testing that no gene in the gene set we are interested in is differentially expressed. A self-contained null hypothesis is defined as The mean vectors of gene expressions corresponding to the genes in a set are equal between the two groups. The corresponding test statistic is calculated based on subject permutations. More precisely, under the null hypothesis of equality between the two groups, the group labels are interchangeable, and a null distribution can be estimated based on permuting the labels of the two groups. On the other hand, a competitive null hypothesis is defined as The proportion of genes being different between the two groups is the same as the proportion of genes being different among the genes outside the gene set. The corresponding test statistic is calculated based on gene permutations. Goeman and Buhlmann [14] strongly recommended against the testing of competitive null hypotheses with the use of gene sampling methods, on the grounds of the untenable statistical independence assumption across genes. Delongchamp et al. [9] also commented on how ignoring the correlations within the sets can overstate significance, and proposed meta-analysis

methods for combining p-values with a modification to adjust for correlation. In their review paper, Nam and Kim [19] revisited the two different null hypotheses, self-contained and competitive, for testing the association of a gene set with a phenotype, as introduced by Tian et al. [23]. The first type of hypothesis, called $Q1$, is "competitive" and tests whether the level of association of a gene set with the phenotype is equal to those of the other gene sets. The second type of hypothesis, called $Q2$, is "self-contained" and tests whether gene expressions of a gene set differ by the phenotype.

Next we present Over-representation analysis (ORA), an example of the gene set analysis method, testing a competitive null hypothesis, then Gene Set Enrichment Analysis (GSEA) which is a hybrid between competitive and self-contained, and also the most popular gene set analysis method with over two hundred citations on PubMed. We also present two self-contained methods (Significance Analysis of Microarrays, or SAM-GS; and Multivariate Analysis of Variance for Gene Sets, or MANOVA-GSA) which provide the best power results in an extensive simulation study among all seven self-contained methods.

2.2.1 Over-Representation Analysis

One of the early gene set analysis methods is Over-representation analysis (ORA) [11]. First, ORA determines a list of genes differentially expressed/statistically significant among all genes in the data set and then based on that, a measure of over-representation is calculated for each gene set. The measure of "over-representation" is calculated as follows: each gene in the list is assigned a value of 1 if it was found significant, and 0 otherwise. The membership of

each gene to the gene set list is also coded as 1 if the gene is in the set, and 0 otherwise. The measure of over-representation and corresponding p-value is calculated using Fisher's Exact test statistic for two-by-two table, Table 2.2, of significance status versus gene set membership.

	diff.expr.gene	non-diff.expr.gene	total
in gene set	l	$N-l$	N
not in gene set	$K-l$	$M-K-N+l$	$M-N$
total	K	$M-K$	M

Table 2.2: A 2×2 table for assessing over-representation

Fisher's exact test assumes that all genes are independent. The p-value based on the test statistics is given by:

$$p\text{-value} = \sum_{l=x}^{K} \frac{\begin{pmatrix} N \\ l \end{pmatrix} \begin{pmatrix} M-N \\ K-l \end{pmatrix}}{\begin{pmatrix} M \\ K \end{pmatrix}}$$

where M is the number of all genes, N is the number of genes in the gene set, K is the number of genes expressed differentially among all genes, and x is the number of genes differently expressed and belonging to the gene set.

Three important limitations are associated with ORA. First, only genes differentially expressed are considered to determine significance of gene sets. Those genes non-differentially expressed are excluded. Second, the independence assumption of Fisher's exact test does not take into account the correlation between genes in a gene set. This is an important limitation, as genes in a biological pathway are usually highly correlated. Third, the p-value calculation only includes the information on number of genes differentially expressed,

not their correlations with the phenotype.

2.2.2 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) [22] was designed to overcome limitations of individual gene analysis methods, by grouping genes based on their biological function into gene sets, or pathways. Searching for pathways associated with disease is more meaningful than searching for individual genes for various reasons: biologists are more interested in pathways rather than single genes; interpreting a list of significant pathways based on a priori knowledge of the biological functions makes more sense than interpreting individual genes; biological signals may be modest relative to the noise inherent to microarray data and pathway analysis is more powerful than individual gene analysis especially in detecting modest signals; significant pathways are more robust to replication across different studies and platforms.

Based on a phenotype, all genes in the microarray data set are ranked in a list based on the correlations between gene expressions and the phenotype. Then we get a gene set enrichment score for each gene set which indicates the distribution of genes in the gene set locating in the ranked gene list. For each gene set, an enrichment score based on the correlations of the genes in the set with the phenotype is calculated. Permutations tests are used to evaluate the significance of the enrichment score for each gene set. An FDR value is calculated for a collection of sets of genes. GSEA can be summarized as follows:

Step 1 Calculation of an Enrichment Score for a gene set

1. For each gene i, the correlation coefficient r_i between the expression

measurements and the phenotype is calculated. Then a ranked list L of all genes is obtained based on all the correlations. For each gene set S and each gene j in the ranked list L, we calculate:

$$P_{hit}(S,j) = \sum_{gene\ i \in S\ and\ i \leq j} \frac{|r_i|^t}{r_S}, \quad where \quad r_S = \sum_{gene\ i \in S} |r_i|^t,$$

$$P_{miss}(S,j) = \sum_{gene\ i \bar{\in} S\ and\ i \leq j} \frac{1}{N - N_S},$$

where N is the number of all the genes in the data set, N_S is the number of genes in the gene set S, j is a given position in the ordered gene list L and t is a defined exponent to control the weight of the measure of "hits" genes. P_{hit} is the fraction of genes in S ("hits") weighted by their correlation and P_{miss} is the fraction of genes not in S ("misses") present up to a given position i in L. When $t = 0$, the above expressions are reduced to standardize Kolmogorov-Smirnov statistic:

$$P_{hit}(S,j) = \sum_{gene\ i \in S\ and\ i \leq j} \frac{1}{N_S},$$

$$P_{miss}(S,j) = \sum_{gene\ i \bar{\in} S\ and\ i \leq j} \frac{1}{N - N_S},$$

When $t = 1$, the function uses the correlations as the weights of all genes in the gene set. Usually, $t = 1$ is preferred.

2. Based on the above calculation, we can get a list of paired scores of P_{hit} and P_{miss} by walking down the list L for each position j. The score for the gene set S, $ES(S)$ is the maximum deviation from zero of $P_{hit} - P_{miss}$. A small $ES(S)$ indicates that the genes in the

gene set S are randomly distributed in the ranked list L. A large $ES(S)$ indicates that genes in the gene set are concentrated in the extreme positions (at the top or bottom) of the list.

Step 2 Estimation of Significance Level of ES

1. Permute the phenotype labels and repeat Step 1 for a large number of permutations. Usually a minimum of 1,000 permutations are recommended. Statistical significance of the gene set S is evaluated by comparing the ES from Step 1 to the 1,000 permuted ES values from Step 2.

Step 3 Adjustment for Multiple Hypothesis Testing

1. For multiple gene sets S_k, $k \geq 1$, calculate the observed $ES(S_k)$ for each gene set S_k in the original data set. For each permutation, denoted as π, and each gene set S_k, calculate the $ES(S_k, \pi)$.

2. Since different gene set may have various sizes, we need to adjust for the gene set size and get a new score in order to compare analysis results across gene sets. A normalized enrichment score (NES), such as $NES(S_k, \pi)$ or $NES(S_k)$, is calculated by dividing the corresponding observed enrichment score, $ES(S_k, \pi)$ or $ES(S_k)$, by the average of $ES(S_k, \pi)$ over all permutations.

3. The FDR for each gene set is calculated as the expected proportion of false positives of the gene sets declared significant. For a given positive $NES(S^*)$ score from a specific gene set S^*, the numerator of FDR is the percentage of all positive $NES(S_k, \pi)$ for all k and

π, whose

$$NES(S_k, \pi) \geq NES(S^*).$$

The denominator is the percentage of those observed positive $NES(S_k)$ for all k whose

$$NES(S_k) \geq NES(S^*).$$

The FDR is the ratio of these two percentages, the numerator containing NES from all the permutations of all gene sets and the denominator containing NES from only significant gene sets. Calculation is similar if a given $NES(S^*)$ is negative.

Step 2 provides the p-value of an individual hypothesis test but Step 3 gives the FDR which takes into account the multiple hypothesis testing of all gene sets.

2.2.3 SAM-GS

Although GSEA is the most popular method focusing on gene sets, previous work identified two important limitations associated with it:

1. According to the way the enrichment score is calculated, GSEA identifies those gene sets as significant whose members are clustered along the correlation axis, no matter if the clustering occurs in the weak, moderate or strong correlation region. For example, both simulated and real pathways whose members are weakly correlated with the phenotype (i.e. correlations ranging from -0.1 to 0.1) have been identified as significant by GSEA. These pathways are truly null hypothesis pathways, and the fact that they are declared significant illustrates a poor control of GSEA

over the Type I error.

2. GSEA fails to identify as significant gene sets whose members exhibit both positive and negative associations with the phenotype. This kind of behaviour is not uncommon for a pathway, and it is scientifically referred to as feedback loops. Both simulated and real pathways whose members are moderately to strongly correlated with the phenotype (i.e. absolute values of correlations larger than 0.6) are not identified as significant by GSEA. These pathways are truly associated with the phenotype, and the fact that they are not called significant illustrates a poor control of GSEA over the Type II error.

To correct for these limitations, Significance Analysis of Microarrays (SAM-GS) was proposed [10]. This method extends SAM from individual gene level type of analysis to sets of genes.

For a given gene set S, SAM-GS statistic is calculated as the squared L_2 norm of the SAM test statistics for each gene in the set:

$$\text{SAM-GS} = \sum_{i=1}^{S} d_i^2.$$

An alternative way to define the SAM-GS statistic is the L_1 norm:

$$\text{SAM-GS} = \sum_{i=1}^{S} |d_i|.$$

The squared L_2 norm based test statistic puts more weight on the genes exhibiting stronger signals. We note that both the L_1 norm and squared L_2 norm are useful in addressing the second limitation of GSEA regarding pathways exhibiting feedback loops. Similar to other gene set analysis methods, SAM-GS

employs permutation based tests to evaluate significance. When a collection of gene sets is tested, various FDR adjustments for multiple hypothesis testing are available [7, 12]

2.2.4 MANOVA-GSA

Analysis of Covariance (ANCOVA) method [17, 18] is based on a global test score by modelling gene expressions as random effects in a logistic regression model. Most of the gene set analysis methods are based on a binary phenotype. ANCOVA test can be directly used to comparisons of two groups, as well as adjusting for other covariates, such as demographic or clinical variables.

Later a modified multivariate analysis of variance (MANOVA) [24] test is proposed to model comparisons of two or more groups. When there are only two states, MANOVA is equivalent to Hotelling's T^2 test. A significant strength of MANOVA is that it incorporates the correlations among gene expression measurements for genes in the same pathway via a shrinkage covariance matrix [20]. When the number of genes is larger than the number of samples, i.e. $p \gg N$, the sample covariance matrix is singular and therefore a shrinkage version is needed to evaluate the test statistic. A naive use of a generalized inverse covariance matrix would result in a less efficient method and loss of power. A more efficient estimator of the covariance matrix has been proposed by Schäffer and Strimmer (2005) [20].

The details of MANOVA follow. We denote by $n_1, n_2, \ldots, n_c$ the sample sizes for each of the c groups of the multi-class phenotype. As MANOVA is a self-contained method, it is safe to only consider the genes in the set S, and ignore the rest of the genes measured by the microarray study. Let m denote

the number of all genes in the set S. The MANOVA model can be expressed as:

$$y_{ij} = \mu_i + e_{ij},$$

where y_{ij} is the vector of gene expressions in the i-th classes and for j-th sample, $i = 1, 2, \ldots, c$ and $j = 1, 2, \ldots, n_i$, μ_i is a vector of length m, representing the means of gene expressions in state i, e_{ij} is the error vector in the model and $Var(e_{ij}) = \Sigma_{m \times m}$.

The null hypothesis is:

$$\mu_1 = \mu_2 = \cdots = \mu_c,$$

that there is no difference among gene expression of all c classes. The alternative hypothesis is that gene expressions measurements means are different for at least two classes.

There are four different test statistics for MANOVA and all of them are equivalent to Hotelling's T^2 test for a binary phenotype. MANOVA-GSA uses "Wilk's Λ":

$$\Lambda = \prod_{k=1}^{K} \frac{1}{(1 + \lambda_k)},$$

where λ_k is the k-th eigenvalue of the matrix $S_{m \times m} = E^{-1}H$. Here E is the within class covariance/sample covariance matrix and H is the between class covariance matrix. The number of eigenvalues of S is K which is equal to the minimum of m and $c - 1$. The null distribution of the test statistic can be approximated by the F-distribution and p-values can be evaluated. The problem is that, as noticed before, when the number of genes is greater than the number of samples, E becomes a singular matrix. Then a modified "Wilk's

Λ" test statistic using the shrinkage covariance matrix estimator is proposed to be used in MANOVA-GSA:

$$s^*_{hh'} = \begin{cases} s_{hh} & \text{if } h = h' \\ r^*_{hh'}\sqrt{s_{hh}s_{h'h'}} & \text{if } h \neq h' \end{cases}$$

and

$$r^*_{hh'} = r_{hh'}\min\{1, \max(0, 1 - \hat{\lambda}^*)\},$$

$$\hat{\lambda}^* = \frac{\sum_{h \neq h'} \hat{Var}(r_{hh'})}{\sum_{h \neq h'} r^2_{hh'}},$$

where s_{hh} denotes sample variance of gene h and $r_{hh'}$ denotes sample correlation between gene h and h'. As usual, permutation based tests are used to calculate p-values.

2.3 Multiple Hypotheses Testing in Microarray Studies

Adjustments for multiple hypothesis testing need to be made in the analysis of microarray data, as thousands of genes are being tested. Multiple hypothesis adjustments are also needed for gene set analysis, as a large number of gene sets are being tested.

Consider the problem of testing simultaneously m null hypotheses H_j, $j = 1, 2, \ldots, m$, and denote by R the number of rejected hypotheses. This situation can be summarized in the following table:

The specific m hypotheses are assumed to be known in advance, the numbers m_0 and $m_1 = m - m_0$ of true and false null hypotheses, respectively,

Number of	Number not rejected	Number rejected	
True null hypothesis	U	V	m_0
Non-true null hypothesis	T	S	m_1
	$m-R$	R	m

are unknown parameters, R is an observable random variable and S, T, U and V are unobservable random variables. A variety of generalizations of the Type I error are possible. The Family-wise error rate (FWER) is defined as the probability of at least one Type I error, i.e., FWER=Pr($V \geq 1$). The false discovery rate (FDR) of Benjamini and Hochberg [7] is the expected proportion of Type I errors among the rejected hypothesis, i.e., FDR = $E(Q)$, where

$$Q = \begin{cases} V/R, \ if \ R > 0 \\ 0, \quad if \ R = 0 \end{cases} \tag{2.2}$$

A multiple testing procedure is said to control a particular Type I error rate at level α, if this error rate is less than or equal to α when the given procedure is applied to produce a list of R rejected hypotheses.

We describe below the following four multiple hypothesis adjustments used in microarray analysis:

1. Bonferroni
2. Bonferroni Step-down (Holm)
3. Westfall and Young Permutation
4. Benjamini and Hochberg False Discovery Rate (B&H FDR)

These methods are listed from the most stringent (Bonferroni [4]) to the least stringent (B&H FDR). The more stringent a multiple testing correction, the

fewer false positive genes are allowed. The most stringent one, Bonferroni method causes more false negative genes (genes that are called non-significant when they are) whereas the B&H FDR [7] causes more false positive genes. In our analysis, we try to control the false positives and B&H FDR is used.

Details about the four methods are discussed here.

A. Bonferroni Correction

The α value of all hypothesis tests together will be set as 0.05, the cut-off value that is usually used. Bonferroni method [4] set the cutoff value for each gene individually satisfying:

$$\text{Corrected } p - \text{value} = p - \text{value} \times N(\text{number of genes in test}) < 0.05$$

As a result, the cutoff for individual p-values is equal to the total accepted error rate 5% divided by the number of genes tested in total that the total error rate for multiple hypothesis testing is controlled. If testing 1,000 genes at a time, the highest accepted individual p-value is 0.00005, making the correction very stringent.

B. Bonferroni Step-down (Holm) Correction

This Bonferroni step-down (Holm) method [16] is quite similar to Bonferroni. The steps are as follows:

- Get the ordered p-values from the smallest to the largest for each of the genes that are tested:

$$p_1 < p_2 < p_3 < \ldots < p_N$$

- The cutoff value for the first (smallest) p-value, p_1, is equal to the total accepted error rate 0.05 devided by the number of genes tested in total. The gene will be declared significant if:

$$\text{Corrected } p - \text{value} = p_1 \times N < 0.05$$

- The cutoff value for the second smallest, p_2, is equal to the total accepted error rate 0.05 devided by the number of genes tested in total minus 1. Then the gene will be declared significant if:

$$\text{Corrected } p - \text{value} = p_2 \times (N - 1) < 0.05$$

- Then the gene with the third smallest p-value, p_3, will be declared significant if:

$$\text{Corrected } p - \text{value} = p_3 \times (N - 2) < 0.05$$

- It follows that sequence until no gene is found to be significant.

The step-down correction is a little less corrective as the p-value increases, this correction is less conservative.

C. Westfall and Young Permutation

Both Bonferroni and the Holm method are also called single-step procedure, because they correct each of the p-values independently. The Westfall and Young permutation [26] still follows step-down procedure similar to the Holm's correction but takes into consideration the dependencies between gene expression measurements. Here are the steps

summarizing Westfall and Young procedure:

1. The first step is similar to Holm's correction. The p-values for genes are calculated and ranked from the smallest to the largest.
2. Usually the data set has been separated to two groups, group 1 (treatment) and group 2 (control). Permutations are done over the group labels (phenotype-permutations) and result in pseudo-data sets.
3. Under the permutations, p-values for all genes are re-calculated for each of the pseudo-data sets.
4. The successive minima of the new p-values are retained and compared to the original ones.
5. This process is repeated a large number of times, and the proportion of re-sampled data sets where the minimum pseudo-p-value is less than the original p-value represents the adjusted p-value.

This test gives a more powerful test than the Bonferroni or Holm procedure. However, the permutation process takes much longer to calculate.

D. Benjamini and Hochberg False Discovery Rate

The B&H FDR correction [7] is the least stringent among the four methods:

1. The p-values for each of the genes are ranked from the smallest to the largest.
2. The largest p-value remains as it is and is compared with cut-off of 0.05.

3. The second largest p-value, p_{N-1}, is adjusted as:

$$\text{Corrected } p - \text{value} = p_{N-1} \times \frac{N}{N-1} < 0.05$$

Then the corrected p-value is compared with 0.05.

4. The third largest p-value, p_{N-2}, is adjusted as:

$$\text{Corrected } p - \text{value} = p_{N-2} \times \frac{N}{N-2} < 0.05$$

5. The adjustments are made for the entire list of genes and the smallest p-value, p_1, is adjusted as:

$$\text{Corrected } p - \text{value} = p_1 \times N < 0.05$$

The first three methods control the Family-wise error rate. The Bonferroni correction is the most stringent test of all, and it offers the most conservative approach to control false positives. The Westfall and Young Permutation is the only correction accounting for dependency of genes. B&H method controls the false discovery rate. If the error rate is 0.05, 5% of the genes declared as significant are truly null genes. We report B&H FDR correction in our analysis.

Chapter 3

Results

3.1 Data Pre-Processing

Our data consist of microarray expression measurements corresponding to 54,675 probes on 17 kidney transplanted patients who had been experiencing T-cell-mediated rejection, a more severe form of kidney transplant rejection, and 27 kidney transplant patients experiencing borderline rejection. RNA extraction, dsDNA and cRNA synthesis, hybridization to HG_U133_Plus_2.0 (GeneChip, Affymetrix®), washing and staining were carried out according to [1]. The data were normalized using RMA, according to GeneSpring™ software (Version 7.2, Silicon Genetics, CA, USA).

The microarray data were obtained by hybridizing mRNA to Affymetrix HG_U133_Plus_2.0 microarrays. These arrays contain 54,675 probe sets whose expressions were reduced from the probe level to the gene level of 20,736 unique genes by a method described in the GSEA website [3], by taking the maximum probe set expression of each gene in each sample.

3.2 Gene Set catalogue

For gene sets/pathways, we used the most recent version of the $C2$ catalogue downloaded from the Gene Set Enrichment Analysis webpage [3], and consisting of 1892 gene sets, representing metabolic and signalling pathways from online pathway databases, gene sets from biomedical literatures including 340 PubMed articles, and gene sets compiled from published mammalian microarray studies. Following Subramanian et al. (2005) [22], we restricted the size of the gene sets to be between 5 and 500, resulting in 1,839 gene sets used for our analysis.

3.3 Individual gene level analysis

First, we explore our dataset by running an analysis at individual gene level. We applied the most popular individual gene level analysis method, Significance Analysis of Microarrays (SAM), described in Chapter 2. SAM is implemented as Excel Add-On (`R` software is required), and is freely available for download and public use on the SAM website: *http://www-stat.stanford.edu/ tibs/SAM/*.

We used 100 permutations, and a cutoff for Δ of 0.73, corresponding to an FDR of approximately 5%. SAM output presents the observed score over the expected score of 100 permutations, Figure 3.1. The total number of significant genes is 957, including both positive and negative expressed genes. The plot has three lines, one centre line crossing the origin, and the other two lines defined by the Δ cutoff, more precisely going below and above the centre line by Δ. Genes with observed SAM scores above the upper line are positive significant genes. Genes with observed SAM scores below the lower line are

negative significant genes. Based on an FDR of 5.71%, we get approximately 55 false positive genes among the 957 significant genes.

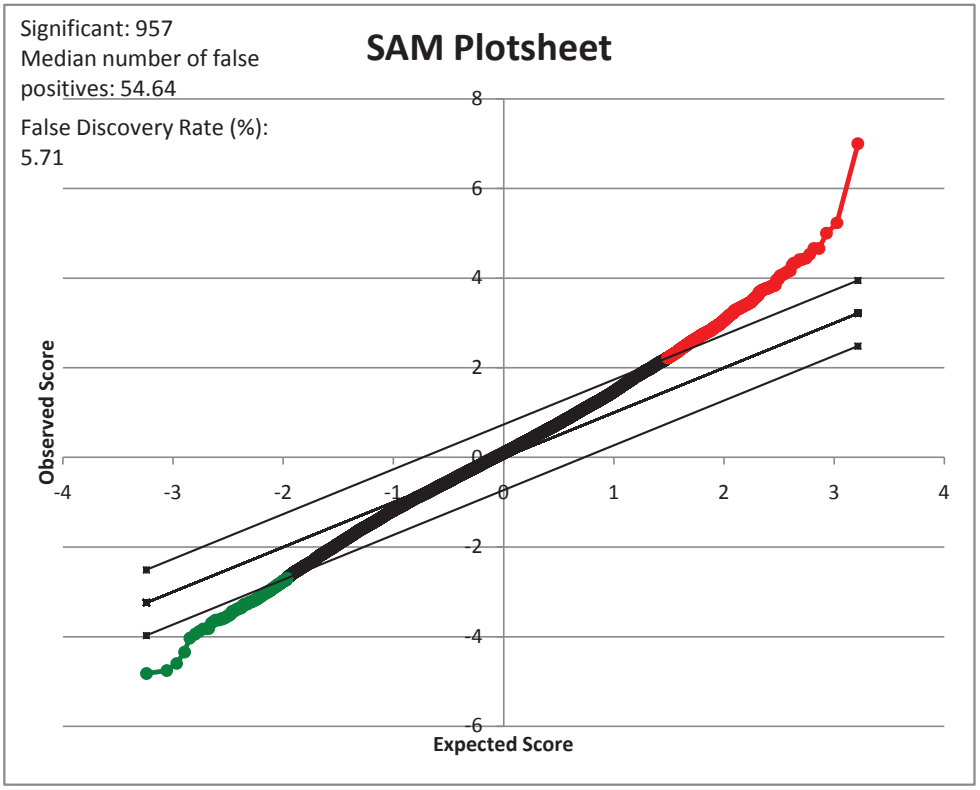

Figure 3.1: SAM plot of kidney cancer microarray data set. 957 significant genes with False Discovery Rate $< 5.71\%$. The plot is the observed score versus expected score for each of the genes. Black part are dots of insignificant genes within the area of two critical lines. Red dots and Green dot represent positive and negative significant genes, respectively.

SAM output gives a list of all 20,736 genes in the dataset, their corresponding SAM test statistic, i.e. d_i, and FDR values. We summarized these results in Table 3.1 by calculating the frequencies of positive and negative genes, as well as percentage of genes by various FDR ranges.

Based on the table, we get 425 significant genes (2%) within a small range of FDR of (0, 5%). For the FDR in the range of (0,25%), we have (sum-up everything up to 25) in total 1630 significant genes.

FDR range [a,b)(%)	Frequency of Genes		Percentage of
	Positive genes	Negative genes	genes
[0,5)	345	80	2.05%
[5,10)	211	71	1.36%
[10,15)	198	62	1.25
[15,20)	241	86	1.58%
[20,25)	233	103	1.62%
[25,50)	1329	696	9.77%
[50,100)	8512	8569	82.37%
Total	11069	9667	100%

Table 3.1: Percentages of significant genes and frequency of positive and negative genes respectively in each range of FDRs.

3.4 Gene set analysis using SAM-GS

We used SAM-GS to run gene set analysis for the kidney transplant rejection patients data set, using the C2 catalogue of pathways/gene sets. We used a total of 1,000 permutations to calculate p-values. A p-value was calculated for each of the 1,839 gene sets. We calculated FDR values according to Benjamini and Hochberg (1995) [7]. Table 3.2 displays the gene sets with p-values $<$ 0.001 and FDR values $<$1.8%.

Table 3.2: Comparison of p-values by two methods

Gene Set Name	Gene Set Size	SAM-GS p-value	GSEA p-value
AGED-MOUSE-CORTEX-DN	42	<0.001	0.171
AGED-MOUSE-HIPPOCAMPUS-MULTI-UP	19	<0.001	0.335
AMIPATHWAY	23	<0.001	0.008
ARAPPATHWAY	12	<0.001	0.008
ASBCELLPATHWAY	11	<0.001	0.014
AT1RPATHWAY	32	<0.001	0.248
BBCELLPATHWAY	7	<0.001	No records

Gene Set Name	Gene Set Size	SAM-GS p-value	GSEA p-value
BIOPEPTIDESPATHWAY	37	<0.001	0.387
BLYMPHOCYTEPATHWAY	13	<0.001	0.028
CCR3PATHWAY	21	<0.001	0.561
CDK5PATHWAY	12	<0.001	0.395
CIRCADIANPATHWAY	6	<0.001	0.057
CIRCADIAN-EXERCISE	42	<0.001	0.748
CISPLATIN-PROBCELL-UP	17	<0.001	0.239
CSKPATHWAY	23	<0.001	0.008
CTLA4PATHWAY	20	<0.001	0
DCPATHWAY	21	<0.001	0.014
ECMPATHWAY	21	<0.001	0.175
EOSINOPHILSPATHWAY	11	<0.001	0.023
GABAPATHWAY	12	<0.001	0.125
GLEEVECPATHWAY	22	<0.001	0.542
H2O2-CSBRESCUED-C2-UP	8	<0.001	0.418
HBXPATHWAY	8	<0.001	0.133
HSA00130-UBIQUINONE-BIOSYNTHESIS	10	<0.001	0.153
HSA04710-CIRCADIAN-RHYTHM	10	<0.001	0.346
HSA05020-PARKINSONS-DISEASE	16	<0.001	0.595
HUMAN-TISSUE-THYMUS	15	<0.001	0.108
IFNALPHA-RESIST-DN	16	<0.001	0.081
IFNGPATHWAY	6	<0.001	0.029
IGF1RPATHWAY	15	<0.001	0.084
IL5PATHWAY	13	<0.001	0.034
INSULIN-ADIP-SENS-DN	16	<0.001	0.743
INTEGRINPATHWAY	34	<0.001	0.283
LONGEVITYPATHWAY	13	<0.001	0.291
MCALPAINPATHWAY	22	<0.001	0.715

Gene Set Name	Gene Set Size	SAM-GS p-value	GSEA p-value
MMS-HUMAN-LYMPH-HIGH-24HRS-DN	12	<0.001	0.782
MPRPATHWAY	22	<0.001	0.200
NADLER-OBESITY-HYPERGLYCEMIA	42	<0.001	0.204
NAKAJIMA-MCSMBP-EOS	27	<0.001	0.030
PARKINPATHWAY	10	<0.001	0.626
PYK2PATHWAY	28	<0.001	0.287
RASPATHWAY	21	<0.001	0.391
SPRYPATHWAY	18	<0.001	0.023
ST-TYPE-I-INTERFERON-PATHWAY	8	<0.001	0.066
TH1TH2PATHWAY	20	<0.001	0.022
TNFALPHA-ADIP-UP	9	<0.001	0.010
TNFALPHA-TGZ-ADIP-UP	13	<0.001	0.082
UBIQUITIN-MEDIATED-PROTEOLYSIS	21	<0.001	0.506
UV-ESR-OLD-UNREG	19	<0.001	0.545
VEGFPATHWAY	25	<0.001	0.193

We also ran GSEA, using 1,000 permutations. Table 3.2 consists of gene names, gene set sizes, p-values according to SAM-GS and GSEA, for gene sets with p-values according to SAM-GS <0.001, corresponding to FDR values < 1.8%. Tsai and Chen [24] demonstrated that MANOVA-GSA and SAM-GS perform very close in simulations and analysis of real data. Therefore we are not presenting MANOVA results here.

GSEA output is fairly extensive. Table 3.3 gives detailed information with respect to metric values corresponding to CDK5 genes. Their ranks are cal-

culated by running ES and form a whole list of more than 20,000 genes. The ES is displayed in Figure 3.2 and represents part of the output of GSEA run on our microarray kidney transplant data.

To interpret the results, we first pick up two of them for example, CTLA4 pathway and IFNG pathway. These two pathways are declared as statistically significance both by GSEA and SAM-GS, which indicates the consistent results between the two methods. This result is also consistent with the biological results as they were already considered as biological significance due to previous literatures.

Then we focus on Cyclin-dependent kinase-5, or CDK5 pathway. Figure 3.2 shows the enrichment plot of CDK5 pathway. The top part is the plot of enrichment profile representing the changes of scores when walking down the list of genes. The enrichment score is the maximum absolute distance from zero, among all scores when running down the ranked list. The genes are ranked according to a t-test:

$$\frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{SD_A^2}{n_A} + \frac{SD_B^2}{n_B}}},$$

where $\bar{X}$s, SDs and n denotes the estimated means, standard deviations and the number of samples in each of the two groups respectively.

CDK5 has a p-value< 0.001 according to SAM-GS, and a p-value of 0.395 according to GSEA. Table 3.3 and Figure 3.2, illustrate how GSEA fails to identify as significant a pathway whose members exhibit moderate to weak associations with kidney transplant rejection. The reason GSEA fails to identify CDK5 as associated with kidney transplant rejection can be explained by the fact that the genes in CDK5 are randomly distributed in the ranked list.

PROBE	DESCRIPTION (from dataset)	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING $ES3$
EGR1	EGR1	1017	0.715	0.093
KLK2	KLK2	3585	0.307	0.030
MAPK1	MAPK1	8038	0.070	-0.171
DPM2	DPM2	12893	-0.130	-0.379
MAPK3	MAPK3	13866	-0.174	-0.392
MAP2K1	MAP2K1	13870	-0.174	-0.357
CDK5R1	CDK5R1	15321	-0.253	-0.377
RAF1	RAF1	16806	-0.348	-0.380
CDK5	CDK5	17567	-0.411	-0.335
MAP2K2	MAP2K2	19436	-0.652	-0.295
NGFR	NGFR	19908	-0.792	-0.161
HRAS	HRAS	20347	-1.011	0.019

Table 3.3: Rank, Rank metric score and ES of 12 genes in the CDK5 gene set.

Our finding is in line with previous simulation studies and real data analysis results indicating that GSEA is testing for gene sets that are clustered along the metric axis, and not for differentially expressed sets. Another deficiency of GSEA that is illustrated here is that positive and negative effects are cancelled out. This set is found significant by SAM-GS, and it illustrates how moderate effects, no matter if positive or negative, work together in a pathway contributing to its significance. CDK5 Pathway has been previously linked to kidney transplant rejection [8].

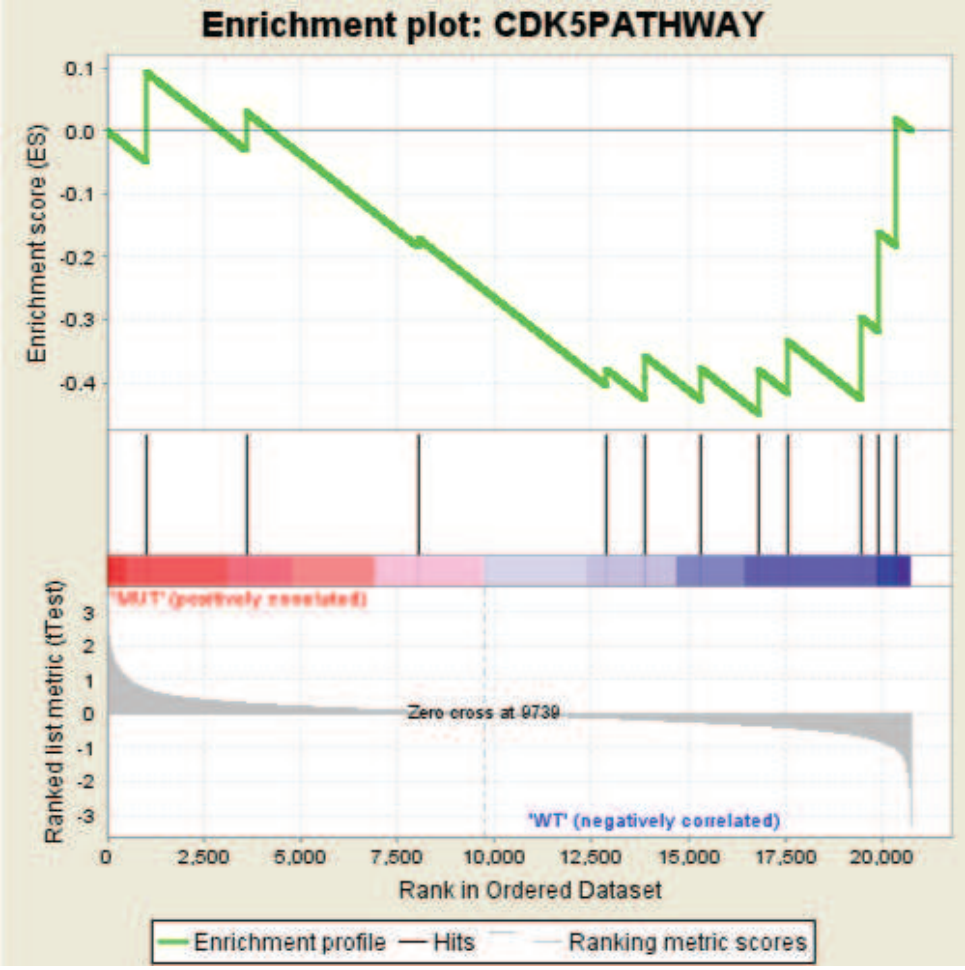

Figure 3.2: The enrichment plot of Gene CDK5 as part of GSEA output; Profile of the running ES score and positions of gene set members on the rank ordered list.

Chapter 4

Discussion

4.1 Remarks on Methods for DNA Microarray Studies

4.1.1 Availability and Usage of Methods for DNA microarray studies

DNA microarray studies are valuable biotechnological advancements, measuring tens of thousands of genes in a single assay. While the laboratories are prepared to run the procedures for generating data, the scientists are less prepared to run the analyses of such large datasets. Software can most of the times represent a "black box" for a scientist not aware of the numerical problems associated with microarray data. One example is the small variability associated with microarray data. This particular characteristic has not been addressed by most of the Gene Set Analysis Methods. With respect to single gene analysis, many scientists are using a rather naive form of identifying top genes, via a ratio of the means, called a "fold ratio". "Fold ratio" is the ratio

of the measured gene expression value for an experimental sample (group 1) to the expression value for the control sample (group 2). Only genes with a fold ratio larger than two are considered "top genes", while the rest of the genes are ignored. Another example of a characteristic of microarray data not addressed by many gene set analysis methods is the correlation among gene expression measurements within a pathway. A "black box" use of gene set analysis software leaves the scientist unaware of limitations in the results of analysis obtained.

4.1.2 Simulation Methods for DNA microarray Studies

A large amount of the published microarray methods are not necessarily statistically sound methods, leading to a large number of false positives. Many methods are published without being tested for Type `I` and `II` errors using simulation studies, and their validity is justified based on biological significance via a couple of real microarray datasets, most of the times just one dataset. Without intensive comparative simulations studies tailored specifically for data generated by microarray experiments, an objective evaluation of current and new proposals is very hard to perform. Real life applications should not be used alone in checking methods performance.

4.1.3 Applications and Excel Add-Ons versus R Software

There is a significant literature on methods for microarray studies, however many of the method implementations are not readily available to scientists. Many of the methods are made available via R functions. Scientists prefer

desktop applications or Excel Add-On implementations of microarray data analysis methods to R functions. Fortunately, SAM is available via an Excel Add-On. GSEA is experiencing increasing popularity also due to its desktop implementation. SAM-GS is available via Excel Add-On. MANOVA-GSA is available only as an R function.

4.2 Gene Set Analysis Extensions

We only presented here gene set analysis methods for comparing two groups. The Global method (ANCOVA) is based on regression models in which the distribution of the response variable is modelled as a function of the covariates. The type of regression model depends on the response. Currently implemented models are linear regression models dealing with continuous responses, logistic regression models dealing with binary responses, multinominal logistic regression dealing with multi-class responses, the Poisson regression models dealing with count responses and the Cox proportional hazards models dealing with survival responses. GSEA method which is the most commonly used method can be applied to both categorical and continuous phenotypes.

SAM-GS has been extended from comparing two groups to analysis of continuous, multi-class, survival type of response, as well as incorporating covariates [5].

Bibliography

[1] Affymetrix Technical Manual, `http://www.affymetrix.com`.

[2] BioCarta, `http://www.biocarta.com/`.

[3] Gene Set Enrichment Analysis, `http://www.broad.mit.edu/gsea`.

[4] Abdi, H., and Salkind, N.J. (ed.) (2007), "Bonferroni and *Šidák* corrections for multiple comparisons", *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA: Sage.

[5] Adewale, A.J., Dinu, I., Potter, J.D., Liu, Q., and Yasui, Y. (2008), "Pathway analysis of microarray data via regression," *Journal of Computational Biology*, 15(3), 269-277.

[6] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock,.G. (2000), "Gene ontology: tool for the unification of biology, The Gene Ontology Consortium," *Nature Genetics*, 25(1), 25-29.

[7] Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, B 57, 289-300.

[8] Chkhotua, A.B., Altimari, A., Gabusi, E., D'Errico, A., Stefoni, S., Chieco, P., Yakubovich, M., Vienken, J., Yussim, A., and Grigioni, W.F. (2003), "Increased expression of p21 (WAF1/CIP1) cyclin-dependent kinase (CDK) inhibitor gene in chronic allograft nephropathy correlates with the number of acute rejection episodes," *Transplant International*, 16(8), 600-604.

[9] Delongchamp, R., Lee, T., and Velasco, C. (2006), "A method for computing the overall statistical significance of a treatment effect among a group of genes," *BMC Bioinformatics*, 7 (Suppl 2), S11.

[10] Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P., and Yasui, Y. (2007) "Improving Gene Set Analysis of Microarray Data by SAM-GS", *BMC Bioinformatics*, 2007, 8, 242.

[11] Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., and Krawetz, S.A. (2003), "Global functional profiling of gene expression," *Genomics*, 81, 98-104.

[12] Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. (2002), "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, 12, 111-139.

[13] Dudoit, S., Shaffer, J.P., and Boldrick, J.C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18(1), 71-103.

[14] Goeman, J.J., and Bühlmann, P. (2007), "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, 23, 980-987.

[15] Goto, S., and Kanehisa, M. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27-30.

[16] Holm, S. (1979), "A Simple Sequentially Rejective Bonferroni Test Procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

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

[18] Mansmann, U., and Meister, R. (2005), "Testing differential gene expression in functional groups: Goemans global test versus an ANCOVA approach," *Methods of Information in Medicine*, 44, 449-453.

[19] Nam, D., and Kim, S.Y. (2008), "Gene set approach for expression pattern analysis," *Briefings in Bioinformatics*, 9, 189-197.

[20] Schäffer, J., and 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, 32.

[21] Speed, T. (2003), *Statistical Analysis of Gene Expression Microarray Data (1st Ed.)*, Chapman and Hall/CRC, ISBN: 1584883278.

[22] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.,

and Mesirov, J.P. (2005), "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences, USA*, 102, 15545-15550.

[23] Tian, L., Greenberg, S.A., and Kong, S.W., Altschuler, J., Kohane, I.S., Park, P.J. (2005), "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences, USA*, 102, 13544-13549.

[24] Tsai, C.A., and Chen, J.J. (2009), "Multivariate analysis of variance test for gene set analysis," *Bioinformatics*, 25, 897-903.

[25] Tusher, V.G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences, USA*, 98, 5116-5121.

[26] Westfall, P.H., and Young, S.S. (1993), "Resampling-based multiple testing: examples and methods for P-value adjustment," John Wiley and Sons, ISBN: 0471557617.