IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES AND GENE SETS USING A MODIFIED Q-VALUE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Ekua Fesuwa Bentil

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

June 2014

Fargo, North Dakota

North Dakota State University

Graduate School

Title

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES AND GENE SETS USING A MODIFIED Q-VALUE

By

Ekua Fesuwa Bentil

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr

Chair

Dr. Ronald Degges

Dr. Ann – Marie Fortuna

Approved:

June 17, 2014
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Gene expression technologies allow expression levels to be compared across treatments for thousands of genes simultaneously. Statistical methods exist for identifying differentially expressed (DE) genes and gene sets while controlling multiple testing error. Most methods do not take into account the distribution of effect sizes or the overrepresentation of observed patterns. This paper compares a recently proposed modified q-value method that takes into account such patterns to a traditional q-value method for experiments with three treatments. The results of simulation studies performed suggest that the proposed methods improve upon the traditional method in the identification of DE genes in certain settings, but are outperformed by the traditional method in other settings. Analysis of data sets from real microarray.

ACKNOWLEDGEMENTS

I would like to thank the almighty God for seeing me through my Master's education. I would like to also express my sincere gratitude to my advisor, Dr. Megan Orr, for her continuous support, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Masters.

Besides my advisor, I would like to thank the rest of my committee: Dr. Ronald Degges and. Dr. Ann-Marie Fortuna, for their encouragement, insight and comments.

I wish to thank my mom and aunty, Ms. Philomena Bruce and Mrs. Jemima Otoo respectively, and my fiancé, Dr. Ruben Kotoka for their love, encouragement, time and prayers. I owe them everything and wish I could show them just how much I love and appreciate them.

I also want to thank my family and friends for their unconditional support.

DEDICATION

I dedicated this thesis to my mom and late grandparents.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

DE……..Differentially Expressed

EE……...Equally Expressed

GSEA…….Gene Set Enrichment Analysis

SAFE……..Significance Analysis of Functional Categories

ANCOVA…...Analysis of Covariance

FWER…….Family Wise Error Rate

FDR……False Discovery Ratio

ANOVA…….Analysis of Variance

BH……..Benjamin and Hochberg

pFDR……..Positive False Discovery

GEO……...Gene Expression Omnibus

MSigDB…….Molecular Signatures Database

GO……..Gene Ontology

pFDR……..Positive False Discovery Rate

TMT……...Trimethyltin

GEO……...Gene Expression Omnibus

DDE……...Declared Differentially Expressed

SAM……...Significance Analysis of Microarray

CHAPTER 1: INTRODUCTION

1.1. Background

Microarray, and other gene expression, technology is one of the fastest growing technologies used in the field of genetic, biological and medical research (Macgregor *et al.*, 2002; Petricoin *et al.*, 2002). These technologies facilitate the simultaneous measure of thousands of genes to provide gene expression information at the genome level. In medical research, microarray experiments provide a better insight in identifying clinical markers used in the diagnosis and treatment of disease (Cojocaru *et al.*, 2001). These techniques also aid in the identification of new genes, their functions and expression levels under different conditions. In studying the correlations between therapeutic responses to drugs and genetic profiles of subjects, analysis of genes from a diseased and a normal cell help in the identification of biomedical constitution of proteins synthesized by the diseased genes. These results can be used to synthesize drugs which fight these proteins and reduce their effect in a diseased cell (Petricoin *et al.*, 2002). Several types of microarray technology have been proposed, including Spotted Microarrays (DeRisi, 1996) and Oligonucleotide Microarrays (Lockhart *et al*, 1996).

In many cases, researchers want to compare gene expressions of two or more treatments (or groups) to determine which genes are differentially expressed (DE), i.e., have different mean expression levels across treatments. However, the emphasis in many applications of gene expression data analysis has moved from single gene analysis to gene set testing. This change is due to the reason that many diseases are associated with a modest regulation in a set of related genes rather than a single gene (Subramanian *et al.*, 2005). Gene set testing is expected to overcome some limitations of single gene testing in the areas of interpretation of multiple hypothesis testing and inconsistencies in lists of genes identified as DE across independent

studies (Liat Ein-Dor *et al*., 2006). Also, individual gene analysis may not be effective in measuring significant effects on pathways controlled by gene sets and not a single gene. For example, "an increase of 20% in all genes encoding members of a metabolic pathway may alter the flux through the pathway, which may be more important than a 20-fold increase in a single gene" (Subramanian *et al.*, 2005).

1.2. Research Objective

This research is specific to gene expression data sets with more than two treatments where the treatments can be ranked. Examples include experiments where experimental units receive different doses of a drug or experiments where gene expression levels are measured at different points in time.

The goals of this study are:

(1) Determine if taking into account "overrepresented" gene expression patterns across treatments improves identification of differentially expressed genes. This will be accomplished by conducting a simulation study to determine under which experimental settings taking into account overrepresentation of these patterns improves identification of DE genes compared to a traditional method and by reanalyzing data generated by real gene expression experiments.

(2) Determine a method for identifying differentially expressed gene sets that also takes into account this overrepresentation of gene expression patterns across treatments and evaluate this method. Similar to goal (1), data from real gene expression experiments will be analyzed to determine if taking into account overrepresented patterns in the data improves upon the traditional method.

The rest of this thesis is organized as follows. Statistical methods used in single gene and gene sets analysis, multiple hypothesis testing with emphasis on false discovery rate are reviewed in Chapter 2. Methods and Materials used in the analysis are described in Chapter 3. Results of the simulation study and real data analysis are discussed in Chapter 4, while Chapter 5 provides the overall conclusions and recommendations for future work.

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

This chapter gives a brief overview of common methods used in gene testing and gene set testing. Review of common multiple testing procedures, including methods that are used in the research paper, is also presented. The objectives of this research paper are also described. Finally, the summary of the thesis structure is stated at the end of the chapter.

2.2. Gene Testing

Gene testing (or gene selection) refers to the procedures used to identify or compare gene expression levels across different conditions and can aid in identifying diagnostic or prognostic biomarkers, classifying diseases, monitoring the response to treatments, and understanding the mechanisms involved in the genesis of disease developments (Adi L. Tarca *et al*, 2006). These methods can be grouped into two categories: parametric and nonparametric methods.

2.2.1. Parametric Methods

A commonly used parametric method for detecting DE genes is the two sample t-test and its variations. Thomas *et al.* (2001) suggested estimating the Z-score of each gene, that is, the mean difference between two conditions divided by the pooled standard error, after correcting the sample heterogeneity using a regression approach. The corresponding *p*-values are computed under asymptotic normality and all genes corresponding to *p*-values less than a chosen cut off point are declared to be DE.

Newton *et al.* (2001) proposed a hierarchical model for the gene expression levels based on the assumption that the distribution of the mRNA intensity levels is Gamma-distributed. The posterior odds of change are calculated, and a gene is considered as DE if the odds are too large or too small. Kerr *et al*. (2000) recommended the use of ANOVA (analysis of variance) by fitting

a single model to all of the data in the microarray experiment which includes gene effect, array effect, and their interaction effect, and also assumes equal variance among genes. As an alternative of fitting a single model for the entire experiment, Smyth (2004) proposed fitting a linear model to the expression levels for every gene. For a microarray experiment, the total number of genes analyzed is large enough that the information contained in other genes can be helpful in better estimating of the variance of individual genes. Hence, Smyth (2004) assumed a prior inverse gamma distribution for the variances of the genes in the data set. Because the parameters of this distribution are unknown, they are estimated from the expression values in the data set.

2.2.2. Nonparametric Methods

The fundamental idea of the nonparametric methods is based on the assumption that the data do not follow a normal distribution, a key assumption in most parametric procedures, which may result in invalid results if parametric methods are used.

A common and classical non-parametric procedure used to analyze each gene is the Wilcoxon sign-rank test or Wilcoxon rank sum test (also known as the Mann-Whitney test). This procedure first sorts and ranks the data. The ranks of different treatment groups are then compared by computing the Wilcoxon statistic and its associated *p*-value are obtained from the Wilcoxon rank sum distribution (Zhang, 2006).

Apart from the Wilcoxon test, other non-parametric procedures have also been recommended. Ben-Dor *et al* (2000) proposed the use of a threshold number of misclassification (TNoM) score to select DE genes. They assumed that DE genes will exhibit significantly different values in different classes, and these differences can therefore be distinguished by a threshold number.

Tusher *et al* (2001) proposed a permutation procedure, called Significance Analysis of Microarrays (SAM), by assigning a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. Then, permutated scores are generated by calculating the score for every possible permutation of the observed data in order to create a null distribution of scores. Expected scores for each permutated data set are determined, and genes are declared to be DE if the absolute difference between the score of the original data and the expected null score exceed a specified threshold.

2.3. Gene Set Testing

Different methods for gene set testing have been developed. These procedures use biological knowledge about sets of related genes – gene sets – and can be classified into two groups: competitive analysis and self-contained analysis (Nam *et al.*, 2008). The competitive approach compares a gene set with its complement (i.e., all genes not in the gene set) in terms of association with the phenotype. Two of the most common approaches include Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) and Significance Analysis of Functional Categories in Gene Expression (SAFE) (Barry *et al.*, 2005). A simpler competitive approach includes Fisher's exact test, which determines which gene sets are overexpressed, i.e., have a higher proportion of DE genes in the gene set of interest compared to the set of genes not in the gene set. The self-contained analysis takes into consideration the association between the phenotype and expression levels in the gene set of interest while ignoring genes that are not in the set. Examples of the self-contained gene set testing include Analysis of Covariance (ANCOVA) (Mansmann *et al.*, 2005) and Global test (Goeman *et al.*, 2005; Goeman *et al.*, 2004). The global and self-contained methods suggest different measures of association

(statistics), but all use the basis of multiple hypothesis testing to identify genes that are DE and the significance of the association of the gene set while controlling multiple testing error.

2.4. Multiple Testing

Among the many challenges raised by the analysis of large data sets is the problem of multiple testing. In microarray and other gene expression analysis, it is not unusual to test thousands of hypotheses simultaneously. Hypothesis tests are not free of error, however, and for every hypothesis test there is a risk of falsely rejecting a null hypothesis that is true, i.e. a Type I error, and of failing to reject a null hypothesis that is false, i.e. a Type II error. Traditionally, Type I errors are considered more problematic than Type II errors. The key goal of multiple testing methods is to control the rate at which Type I errors occur when many hypothesis tests are performed simultaneously.

The Family-Wise Error Rate (FWER) is often the preferred error rate to be controlled. Common procedures for identifying DE genes while controlling the FWER are the Bonferroni (SIMES, 1986) and Holm (Holm, 1979) methods. However, for high-dimensional data in which thousands of hypotheses are being tested simultaneously, the FWER generally results in extremely low statistical power for identifying DE genes. In efforts to improve the power of detecting DE genes while still controlling multiple testing error, the False Discovery Rate (FDR) was developed (Hochberg *et al*., 1995).

2.4.1. False Discovery Rate

Many methods have been developed to overcome the problems that arise from multiple testing, and they all attempt to assign an adjusted *p*-value to each hypothesis test, or reduce the *p*-value threshold. Several traditional methods such as the Bonferroni correction are too

conservative, as it tries to reduce the number of false positives but also considerably reduces the number of true discoveries in many cases.

FDR methods also determine adjusted *p*-values for each hypothesis test. More specifically, the FDR controls the proportion of false discoveries among all tests that are significant and has a greater power to determine truly significant results. This approach was proposed by Benjamini and Hochberg (1995) as a multiple-hypothesis testing error measure to control the proportion of Type I errors among all rejected null hypotheses (Hochberg, 1995). Benjamini and Hochberg (BH) considered the case of testing m null hypothesis, of which m_0 are true. Table 2.1 provides notation for random variables associated with different scenarios in a multiple testing experiment.

Table 2.1

Random Variables Corresponding to the Number of Errors Committed when Testing m Hypothesis

	Declared non – significant	Declared Significant	Total
True null hypothesis	U	V	m_0
Non – true null hypothesis	T	S	$m - m_0$
Total	$m-R$	R	m

BH defined the FDR as the

$$FDR = E\left(\frac{V}{\max(R,1)}\right) \tag{2.1}$$

Sequential *p*-value methods were provided to control the FDR. Let $p_1 \le p_2 \le ... \le p_m$ be the ordered *p*-values and let H_i be the null hypothesis of the i^{th} gene with corresponding *p*-value p_i. Now, let k be the largest i for which

$$p_i \le \frac{i}{m} q^* \tag{2.2}$$

If all H_i, for $i = 1, 2, ..., k$ are rejected, then the above formula controls the FDR at q^* for any independent test statistics and any configuration of false null hypotheses. Also if the test statistics corresponding to true null hypotheses are statistically independent, equation (2.2) controls $FDR \le \left(\frac{m_0}{m}\right) q^* \le q^*$. Figure 2.1 below shows the comparison between the controlling procedures used in FDR and FWER.

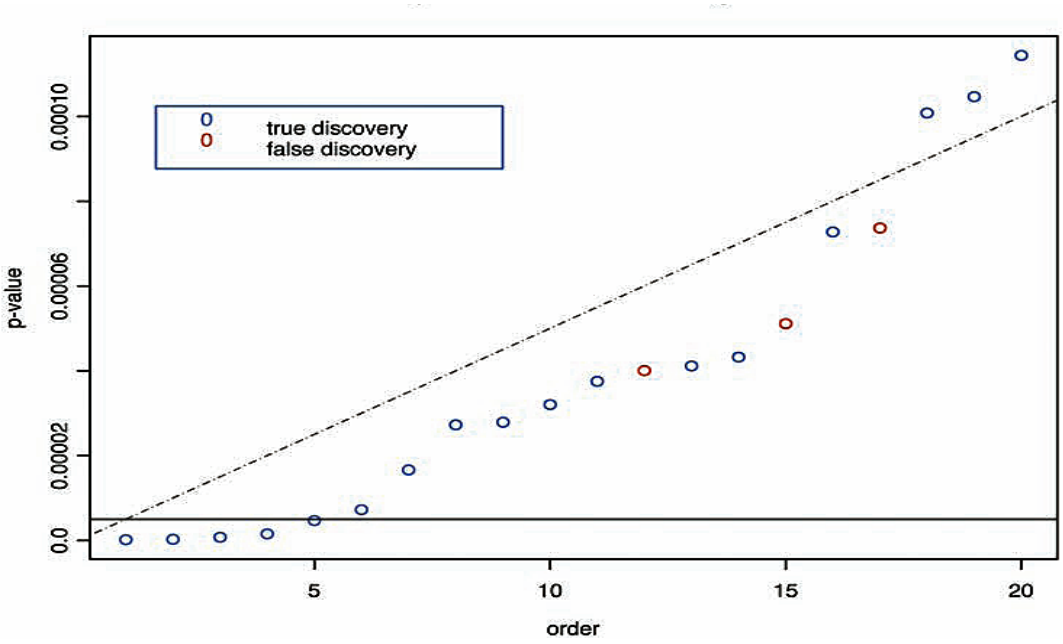

Figure 2.1. Comparison of the controlling procedures of FDR and FWER (Lazar, 2012).

Figure 2.1 above is a plot of the first 20 ordered *p*-values for a gene expression experiment, with the order indicator on the x-axis and *p*-values on the y-axis (Lazar, 2012). The horizontal solid line represents the Bonferroni correction method (controlling procedure for FWER) and the dashed line represents the FDR-controlling procedure. Points that fall below the line for a given method are considered to be significant by that method. From this plot, it is clear that using the FDR controlling procedures allows for more tests to be identified as significant compared to the Bonferroni correction method. Thus, although FDR-controlling methods allow for more type I errors or false discoveries than the FWER, it generally results in higher power in detecting genes with false null hypotheses.

Storey (2002) pointed out the weaknesses in controlling the FDR which was proposed by BH and suggested that the FDR should be calculated as

$$pFDR = E\left(\frac{V}{R}\middle| R > 0\right) \tag{2.3}$$

where $pFDR$ is the positive false discovery rate (Storey, 2002). He later developed the *q*-value, a natural pFDR analogue of the *p*-value, as a hypothesis testing error measure for each of the observed statistics with respect to pFDR (Storey, 2002). The *q*-value for an observed statistic $T = t$ with it rejection region Γ was defined as

$$q_{(j)} = \min\left\{\frac{p_{(r)}\hat{m}_0}{r} : r = k, ..., m\right\} \tag{2.4}$$

where $p_{(r)}\hat{m}_0$ is an estimate of the number of false discoveries and r is the total number of genes declared to be DE if all genes with *p*-values less than or equal to p_r are declared DE. $\hat{m}_0$ is the estimate of the number of EE genes in a data set, and calculated using a method proposed by

Storey *et al* (2003). This procedure involves first ordering all the *p*-values and estimating $\hat{m}_0(\lambda)$ for a range of λ between 0 and 1, where

$$\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} \{p_j > \lambda\}}{(1-\lambda)} \tag{2.5}$$

Then, a natural cubic spline is fit to the points $(\lambda, \hat{m}_0(\lambda))$. Finally, this function is evaluated at $\lambda = 1$ to obtain the final estimate of m_0 (John D. Storey, 2003).

Orr *et al* (2014) suggested that when asymmetry in the distribution of test statistics is observed in two-sample gene expression experiments, the estimation of FDR using the *q*-value method might be improved if this asymmetry is taken into consideration. Consider performing m hypothesis tests in the two treatment case $(t = 1, 2)$, with the null hypothesis for the j^{th} gene being $H_j : \mu_{1j} = \mu_{2j}$ against a two-sided alternative, where μ_{tj} is the population treatment mean expression for gene $j = 1, ..., m$. For each gene, an appropriate *t*-test statistic t_j should be computed with its corresponding two-sided *p*-value obtained. The *p*-values should then be partitioned into two subsets based on the signs of the corresponding test statistics, $\{p_k^{(1)} : k = 1, ..., m_1\}$ and $\{p_k^{(2)} : k = 1, ..., m_2\}$, that represent the subsets of *p*-values corresponding to genes with negative and positive test statistics, respectively (Orr *et al*., 2014). Then the *q*-values for each subset are estimated separately as

$$q_{(k)}^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)} \hat{m}_0 / 2}{r} : r = k, ..., m_1 \right\} \tag{2.6},$$

and

$$q_{(k)}^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)}\, \hat{m}_0/2}{r} : r = k, \ldots, m_2 \right\} \qquad (2.7).$$

The two-sample case was extended to experiments with three treatment groups in cases where the treatments can be ranked. Examples include experiments in which the treatments correspond to different points in time or different doses of a drug. This procedure is described in detail in Section 3.2.2.

CHAPTER 3: METHODS AND MATERIALS

3.1. Introduction

This chapter is devoted to the methodology of the study and the description of real data sets that will be analyzed.

3.2. Methods of Gene Testing

This research focuses on gene expression experiments with more than two treatment groups. Thus, for each gene, we are interested in testing the null hypothesis

$$H_j : \mu_{1j} = \mu_{2j} = ... = \mu_{ij} \tag{3.1}$$

against the alternative that not all population treatment means are equal. In the null hypothesis above, μ_{ij} represents the population mean expression value for the j^{th} gene in the i^{th} treatment.

If the null hypothesis H_j is true, then gene *j* is EE and if false, then gene *j* is said to be DE. Moreover, if H_j is rejected, then gene *j* is DDE.

The moderated one-way analysis of variance *F*-test (Smyth, 2004) will be performed to obtain p_j, the *p*-value corresponding to testing H_j, for each gene. This test assumes the following model for each gene:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \text{ for } i = 1,...,t \text{ ; } \quad j = 1,...,m \text{ ; and } k = 1,...,n_i \text{ ,} \tag{3.2}$$

where y_{ijk} and ε_{ijk} are the expression value and random error, respectively, for the j^{th} gene from the k^{th} experimental unit in the i^{th} treatment, and μ_{ij} is defined above. Additionally, ε_{ijk} is assumed to be normally distributed with mean 0 and variance σ_j^2. The moderated *F*-test assumes and inverse gamma distribution for the population variances of the genes expressions. More specifically,

$$\frac{1}{\sigma_j^2} \sim Gamma\left(\frac{d+d_0}{2}, \frac{ds_j^2 + d_0 s_0^2}{2}\right), \tag{3.3}$$

where d is the degrees of freedom associated with estimating s_j, the sample pooled variance for the j^{th} gene. The constants d_0 and s_0 are unknown, so they are estimated using the data from the microarray experiment. Once this is done, the final estimate of variance for the j^{th} gene is estimated as

$$\tilde{s}_j^2 = \frac{ds_j^2 + d_0 s_0^2}{d + d_0} \tag{3.4}$$

This value replaces the sample pooled variance, s_j^2 (or mean squared error), in the traditional F-test statistic, and the associated p-values for each gene are calculated.

3.2.1. Traditional and Improved Methods for Computing *q*-values

Once the p-values are obtained after performing the moderated F-test on each gene, q-values will be calculated using two methods. The first method will be referred to as the "traditional method", and is the method proposed by Storey (2002). This method is described in section 2.4.1 and the improved method for estimating q-values which was proposed by Orr *et al* (2014, under review) will be used to identify DE genes.

3.2.2. Improved Method for Computing *q*-values

The second method that will be used to calculate q-values will be referred to as the "proposed methods" and is the method proposed by Orr *et al*. (2014). Recall from section 2.4.1 that this method proposed calculating q-values separately for the subset of p-values corresponding to positive test statistics and the subset of p-values corresponding to negative test statistics. Extending this method to gene expression experiments with three treatments, in cases where the treatments can be ranked, is described here.

The proposed methods begins by estimating m_0 using the methods described in Storey (2003) and section 2.41. Thus, $\hat{m}_0$ is the same for both the proposed and traditional methods.

There are $3! = 6$ possible observed orderings of the sample treatment means in gene expression experiments with three samples:

$$\overline{y}_{j1} < \overline{y}_{j2} < \overline{y}_{j3} \text{ (Monotone increasing in ranked treatments),} \tag{3.5}$$

$$\overline{y}_{j1} > \overline{y}_{j2} > \overline{y}_{j3} \text{ (Monotone decreasing in ranked treatments),} \tag{3.6}$$

and

$$\left.\begin{array}{l} \overline{y}_{j1} < \overline{y}_{j3} < \overline{y}_{j2} \\ \overline{y}_{j2} < \overline{y}_{j3} < \overline{y}_{j1} \\ \overline{y}_{j2} < \overline{y}_{j1} < \overline{y}_{j3} \\ \overline{y}_{j3} < \overline{y}_{j1} < \overline{y}_{j2} \end{array}\right\} \text{ (Non-monotone in ranked treatments)} \tag{3.7}$$

The first two orderings are monotone in the ranked treatments while the last four are non-monotone.

For an EE gene, any of the six observed orderings in the sample means are equally likely, therefore the probabilities of obtaining a non-monotone and monotone (either increasing or decreasing) ordering in the ranked treatments are $4/6$ and $2/6$, respectively. Thus, the first partitioning rule includes partitioning the *p*-values into two subsets, $\left\{p_k^{(1)} : k = 1,...,m_1\right\}$ and $\left\{p_k^{(2)} : k = 1,...,m_2\right\}$, corresponding to genes which have ordered sample treatment means that are non-monotone and monotone (increasing or decreasing) in the ranked treatments, respectively. It follows that the *q*-values for each subsets can be estimated separately as

$$q_k^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)} \hat{m}_0 \left(4/6\right)}{r} : r = k, ..., m_1 \right\} \tag{3.8}$$

and

$$q_k^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)} \hat{m}_0 \left(2/6\right)}{r} : r = k, ..., m_2 \right\}. \tag{3.9}$$

The second partitioning rule includes partitioning the set of *p*-values into three subsets, $\left\{ p_k^{(1)} : k = 1, ..., m_1 \right\}$, $\left\{ p_k^{(2)} : k = 1, ..., m_2 \right\}$ and $\left\{ p_k^{(3)} : k = 1, ..., m_3 \right\}$, corresponding to genes which have ordered sample treatment means that are non-monotone, monotone increasing and monotone decreasing in the ranked treatment, respectively. For an EE gene, the probabilities of obtaining a non-monotone, monotone increasing and monotone decreasing ordering in the ranked treatments are $4/6$, $1/6$ and $1/6$, respectively. Therefore, the *q*-values for each subset can be estimated separately as

$$q_k^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)} \hat{m}_0 \left(4/6\right)}{r} : r = k, ..., m_1 \right\} \tag{3.10}$$

$$q_k^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)} \hat{m}_0 \left(1/6\right)}{r} : r = k, ..., m_2 \right\}, \tag{3.11}$$

and

$$q_k^{(3)} = \min\left\{ \frac{p_{(r)}^{(3)} \hat{m}_0 \left(1/6 \right)}{r} : r = k, ..., m_3 \right\}. \tag{3.12}$$

Both of these partitioning rules will be used to estimate *q*-values in the data sets analyzed in Chapter 4. Additionally, the proportion of EE genes for the set of all genes will be estimated as

$$\hat{\pi}_0 = \frac{\hat{m}_0}{m}, \tag{3.13}$$

and for each subset of *p*-values that will be created using the partitioning rules, the proportion of EE genes is estimated as

$$\hat{\pi}_0^{(i)} = \frac{g_i \hat{m}_0}{m_i}, \tag{3.14}$$

where g_i is the probability of observing the i^{th} specified gene expression pattern in the sample means.

3.3. Method of Gene Set Testing

The database of gene sets used in this research was taken from the Molecular Signatures Database (MSigDB). This collection contained 1454 Gene Ontology (GO) sets consisting of genes annotated by the same GO terms (Subramanian *et al.*, 2005; Vamsi K Mootha, 2003). Gene set testing will be performed on the real data sets, described in sections 3.5 and 3.6, using both the traditional and proposed *q*-value methods. For each method, *q*-values will be calculated, and the subset of genes with *q*-values less than a predetermined cutoff will be DDE. Then, for each gene set, a Fisher's exact test will be performed to test for the "over-enrichment" of the gene set, i.e., if the proportion of genes in the gene set that are DDE is greater than the proportion of genes not in the gene set that are DDE. Storey's *q*-value method will then be applied to the resulting *p*-values to determine a final set of over-enriched gene sets.

3.4. Description of Data Set Simulation

In order to compare the performance of the proposed methods (Orr *et al*, 2014) to the traditional method (Storey, 2002) for estimating *q*-values, data sets with independent normally distributed data will be randomly generated. For each data set, gene expression values will be randomly drawn from a total of m = 10,000 genes, and expression values from a given gene is independent of expression values from all other genes. More specifically,

$$y_{ijk} \sim N\left(\mu_{ij}, \sigma_j^2\right), \tag{3.15}$$

and

$$\sigma_j^2 \sim Inv\Gamma\left(a, b\right) \tag{3.16}$$

where y_{ijk} is defined as in Section 2.2. The population variance for each gene will be drawn from an inverse gamma distribution because the empirical distribution of the sample variances has been shown to closely resemble such distribution in many microarray data sets (Smyth, 2004). The parameters a and b of the inverse gamma distribution used to generate the variances will be estimated using the procedures proposed by Smyth (2004) from data described in Lattanzi *et al* (2007). Multiple simulation settings will be employed in order to evaluate the performances of the methods under different conditions. Two sample sizes, $n = \{4,10\}$ for the number of units in each treatment, and three different values for the number of EE genes, $m_0 = \{9000, 7000, 5000\}$ out of the total genes to be analyzed $m = 10000$, will be used. Three different vectors representing the proportion of DE genes (that is, genes whose ranked treatment means are non-monotone, monotone increasing and monotone decreasing, respectively) will be used and are as follows: $\pi = \{(4/6, 1/6, 1/6), (1/3, 1/3, 1/3), (1/3, 1/2, 1/6)\}$. This will result in eighteen different simulation settings.

For each simulated data set, the moderated *F*-test will be performed to calculate a *p*-value for each gene. *Q*-values will then be calculated using the proposed and traditional methods. Finally, the number of DE genes DDE (S) and the "observed FDR" (V/R), or proportion of EE genes among all DDE genes, will be calculated for each data set. If no genes are DDE for a given data set, V/R will be recorded as 0.

3.5. Description of Real Data Set: TMT in Rats

The first data set that will be analyzed was generated from a gene expression experiment described by Lattanzi *et al* (2007). This experiment was performed to identify genes associated with the presence of trimethyltin (TMT) in rats. Nine samples made up of three treatment groups were obtained using Affymetrix Genechip microarrays. Three rats were assigned to each treatment group: control (0 $\mu mol/L$), $1\mu mol/L$ and $5\mu mol/L$ concentration of TMT, and m = 12,159 gene expression values were measured in each rat. The data set is available at the Gene Expression Omnibus (GEO) with accession number GSE5073.

3.6. Description of Real Data Set: Deferasirox in Leukemia Patients

The second data set that will be reanalyzed was generated from a microarray experiment conducted by Junko *et al*. (2009) to evaluate the effect of deferasirox (ICL670) in human myeloid leukemia cells, and identify molecular pathways responsible for anti-proliferative effects on leukemia cells using gene expression profiling. A total of six samples consisting of three treatments, the control ($0\mu M$), $110\mu M$ and $50\mu M$ of ICL670, were obtained using Affymerix GeneChips (U133 Plus 2.0). Gene expressions values on m = 42,440 genes were measured for each experimental unit. The data set is also available at the Gene Expression Omnibus (GEO) with accession number GSE11670.

3.7. Data Preparation

Before data is used for further analysis, probe sets without a gene symbol will be removed. Also gene symbols that had "_Predicted", "_Mapped" or "///" attached to then will be erased, leaving the actual gene symbol for the analysis.

3.8. Summary

Data from both the simulation studies and real gene expression experiments will be analyzed using the statistical software R. Moderated *F*-tests will be performed for each gene in the data set to obtain an associated *p*-value. Then, *q*-values for each gene will be calculated using the traditional and proposed methods. For each method, genes with *q*-values less than or equal to a desired significance level α will be identified as DE. For the real microarray data sets, gene sets that are over expressed will be determined using Fisher's exact tests.

CHAPTER 4: RESULTS OF SIMULATION STUDIES AND REAL DATA ANALYSIS

4.1. Introduction

In this chapter, simulated gene expression data sets with independent normally distributed data will be analyzed to compare the performances of both the traditional and proposed methods for calculating *q*-values. Additionally, real gene expression data sets will be analyzed using both the traditional and proposed methods for identifying DE genes and gene sets

4.2. Results - Simulation Studies

For each of the 18 simulation settings, 100 gene expression data sets were randomly generated. Table 4.1 below presents the mean S and mean V/R for each simulation setting. The corresponding standard errors for the mean S and the mean V/R are reported in the parenthesis.

Table 4.1

The Mean S and Mean V/R for the Traditional and Proposed Q-value Methods with it Associated Standard Errors in Parenthesis for Each Simulation Setting using Independent Normally Distributed Data

			Mean S			Mean V/R		
			Traditional	Proposed		Traditional	Proposed	
n	m_0	π_i		I	II		I	II
4	9000	π_1	2.69 (0.33)	5.46 (0.49)	7.43 (0.51)	0.031 (0.007)	0.055 (0.009)	0.071 (0.010)
		π_2	6.51 (0.68)	31.48 (1.64)	34.64 (1.54)	0.078 (0.019)	0.054 (0.005)	0.060 (0.005)
		π_3	5.38 (0.54)	29.62 (1.54)	50.07 (1.93)	0.050 (0.012)	0.049 (0.004)	0.052 (0.003)
	7000	π_1	163.93 (3.90)	219.93 (3.78)	222.60 (3.72)	0.050 (0.002)	0.049 (0.001)	0.050 (0.001)

Table 4.1. *The Mean S and Mean V/R for the Traditional and Proposed Q-value Methods with it Associated Standard Errors in Parenthesis for Each Simulation Setting using Independent Normally Distributed Data (continued).*

			Mean S			Mean V/R		
			Traditional	Proposed		Traditional	Proposed	
n	m_0	π_i		I	II		I	II
	7000	π_2	326.85 (5.05)	671.75 (4.19)	672.06 (4.21)	0.048 (0.001)	0.048 (0.001)	0.048 (0.001)
		π_3	317.86 (4.97)	668.11 (4.56)	753.51 (4.29)	0.050 (0.001)	0.049 (0.001)	0.049 (0.001)
	5000	π_1	980.15 (7.51)	1062.55 (7.37)	1063.43 (7.37)	0.044 (0.001)	0.045 (0.001)	0.045 (0.001)
		π_2	1551.85 (8.14)	2027.12 (6.46)	2027.48 (6.59)	0.047 (0.001)	0.047 (0.001)	0.047 (0.001)
		π_3	1561.720 (9.30)	2042.80 (7.58)	2113.74 (6.92)	0.047 (0.001)	0.048 (0.000)	0.048 (0.001)
10	9000	π_1	542.14 (1.67)	546.36 (1.59)	546.46 (1.58)	0.050 (0.001)	0.050 (0.001)	0.050 (0.001)
		π_2	686.56 (1.40)	710.68 (1.24)	710.73 (1.23)	0.049 (0.001)	0.050 (0.001)	0.050 (0.001)
		π_3	689.26 (1.59)	712.19 (1.45)	714.65 (1.53)	0.051 (0.001)	0.050 (0.001)	0.050 (0.001)
	7000	π_1	2128.87 (2.43)	2123.87 (2.29)	2123.96 (2.29)	0.050 (0.000)	0.050 (0.001)	0.050 (0.001)
		π_2	2488.97 (2.48)	2479.63 (2.18)	2479.60 (2.18)	0.050 (0.000)	0.051 (0.000)	0.051 (0.000)
		π_3	2483.96 (2.16)	2471.88 (1.86)	2468.96 (1.77)	0.050 (0.000)	0.050 (0.000)	0.050 (0.000)
	5000	π_1	4006.12 (3.39)	3986.43 (3.20)	3986.67 (3.22)	0.049 (0.000)	0.049 (0.000)	0.049 (0.000)
		π_2	4461.38 (2.08)	4394.81 (2.21)	4394.87 (2.20)	0.050 (0.000)	0.050 (0.000)	0.050 (0.000)
		π_3	4459.18 (2.23)	4391.03 (2.19)	4382.05 (2.28)	0.050 (0.000)	0.050 (0.000)	0.050 (0.000)

In Table 4.1, higher values of mean S correspond to better performance in identification of differentially expressed genes. The proposed methods outperformed the traditional method in settings with $n = 4$. Additionally in the settings with $n = 4$ and $\pi = \pi_3$, the second portioning rule was a better option to be used than the first partitioning rule, since it provided a higher mean S. For setting with $n = 10$, the traditional method either performs similarly or outperforms the proposed methods, except when, $m_0 = 9000$ in which cases the proposed methods outperformed the traditional method.

Also in Table 4.1, mean V/R values close to $\alpha = 0.05$ indicate adequate control of FDR at the 5% nominal level. Both the traditional and proposed methods appear to adequately control FDR.

4.3. Real Data Analysis I – Presence of Trimethyltin in Rat

The data from the gene expression experiment described in Lattanzi (2007) is reanalyzed using the traditional and the proposed methods. The description of the data set is given in section 3.5.

For each gene, the null hypothesis

$$H_j : \mu_{1j} = \mu_{2j} = \mu_{3j}, \tag{4.1}$$

for $j = 1, 2, ..., 12159$ is tested using the moderated *F*-test.

From the *p*-values obtained in this analysis, the estimated number of EE genes is, $\hat{m}_0 = 7867.592$ corresponding to an estimated proportion of EE genes of $\hat{\pi}_0 = 0.647$.

Using the traditional method, the *q*-value for each gene is estimated as

$$q_{(j)} = \min\left\{ \frac{p_{(r)}\left(7867.592\right)}{r} : r = 1,...,12159 \right\} \tag{4.2}$$

For the first partitioning rule (two subsets of *p*-values), the numbers of *p*-values associated with genes that have sample treatment means that are non-monotone and monotone in the ranked treatments are $m_1 = 7323$ and $m_2 = 4836$ with the estimated proportion of EE genes $\hat{\pi}_0^{(1)} = 0.716$ and $\hat{\pi}_0^{(2)} = 0.542$, respectively. Using (3.8) and (3.9), the *q*-values for the subset of *p*-values associated with non-monotone orderings of the sample means and *p*-values associated with monotone orderings of the sample means are calculated separately as:

$$q_{(k)}^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)}\left[7867.592\left(4/6\right)\right]}{r} : r = 1,...,7323 \right\} \tag{4.3}$$

and

$$q_{(k)}^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)}\left[7867.592\left(2/6\right)\right]}{r} : r = 1,...,4836 \right\} \tag{4.4}$$

For the second partitioning rule, *q*-values are calculated separately for three subsets of *p*-values. The numbers of *p*-values associated with genes that have sample treatment means with non-monotone orderings, monotone increasing orderings, and monotone decreasing orderings are $m_1 = 7323$, $m_2 = 1878$ and $m_3 = 2958$ with estimated proportions of EE genes $\hat{\pi}_0^{(1)} = 0.716$, $\hat{\pi}_0^{(2)} = 0.698$ and $\hat{\pi}_0^{(3)} = 0.443$, respectively. Using (3.10), (3.11) and (3.12), the q-values are calculated for each subset separately as:

$$q_{(k)}^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)}\left[7867.592\left(4/6\right)\right]}{r} : r = 1,...,7323 \right\}, \tag{4.5}$$

$$q_{(k)}^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)}\left[7867.592\left(1/6\right)\right]}{r} : r = 1,...,1878 \right\}, \tag{4.6}$$

and

$$q_{(k)}^{(3)} = \min\left\{ \frac{p_{(r)}^{(3)}\left[7867.592\left(1/6\right)\right]}{r} : r = 1,...,2958 \right\} \tag{4.7}$$

The table 4.2 below shows the number of genes DDE controlling FDR at $\alpha = 0.05$, $\alpha = 0.10$ and $\alpha = 0.20$ by both methods.

Table 4.2

Number of Genes DDE by the Traditional and Proposed Methods for Estimating Q-values at Three Significance Level (α = 0.05, α = 0.10, α = 0.20)

	Number of genes DDE		
Significance Level	Traditional Method	Proposed Methods	
		I	II
0.05	363	433	410
0.10	742	853	851
0.20	1647	1808	1824

From the table above it can be seen that the proposed methods identify more genes as DE than the traditional method, regardless of α. For the proposed methods, the different partitioning rules result in similar numbers.

Figure 4.1 presents the histogram of observed *p*-values corresponding to analysis by the traditional *q*-value method in which no partitioning is used. Figures 4.2 and 4.3 present histograms corresponding to analysis by the proposed methods using the first and second partitioning rules, respectively. In each histogram, the estimated proportion of EE genes for each subset is plotted as a dashed horizontal line.

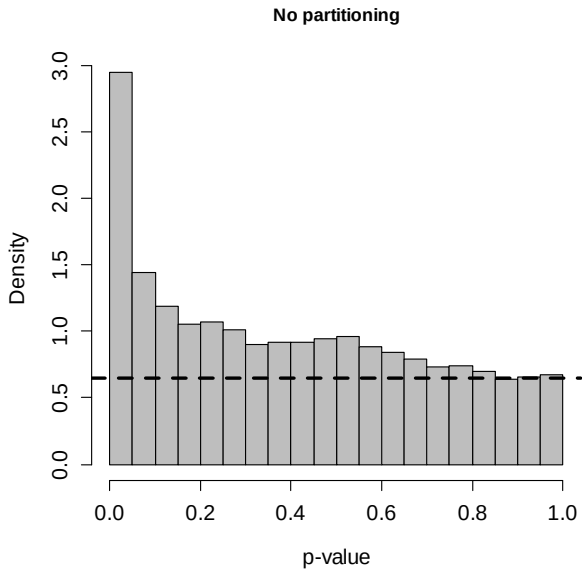

Figure 4.1. Distribution of *p*-values when no partitioning is used.

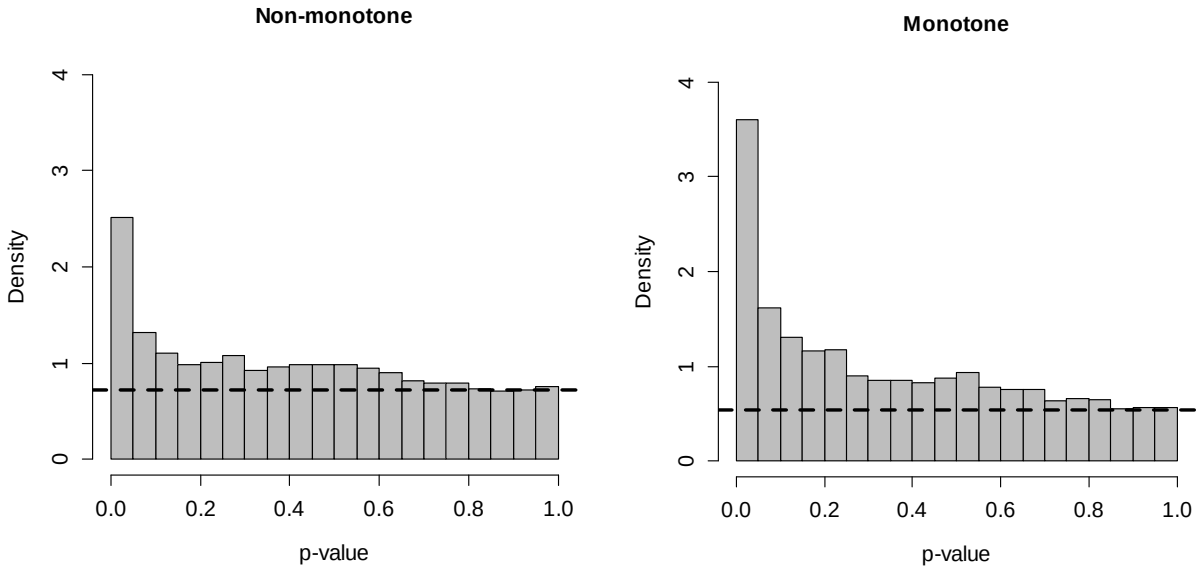

Figure 4.2. Histogram of two subsets of *p*-values using the first partitioning rule.

In figure 4.2 and 4.3, the distributions of *p*-values corresponding to genes with sample treatment means that exhibit monotonicity in the ranked treatments are stochastically smaller than the distribution of *p*-values for genes that do not exhibit monotonicity. This indicates that a higher proportion of genes are DE among the genes that show monotonicity in the ranked treatment means than genes that do not exhibit monotonicity. This also indicates that the proposed methods might be preferred to the traditional method and is a possible reason why the proposed methods identify more genes as DE than the traditional method.

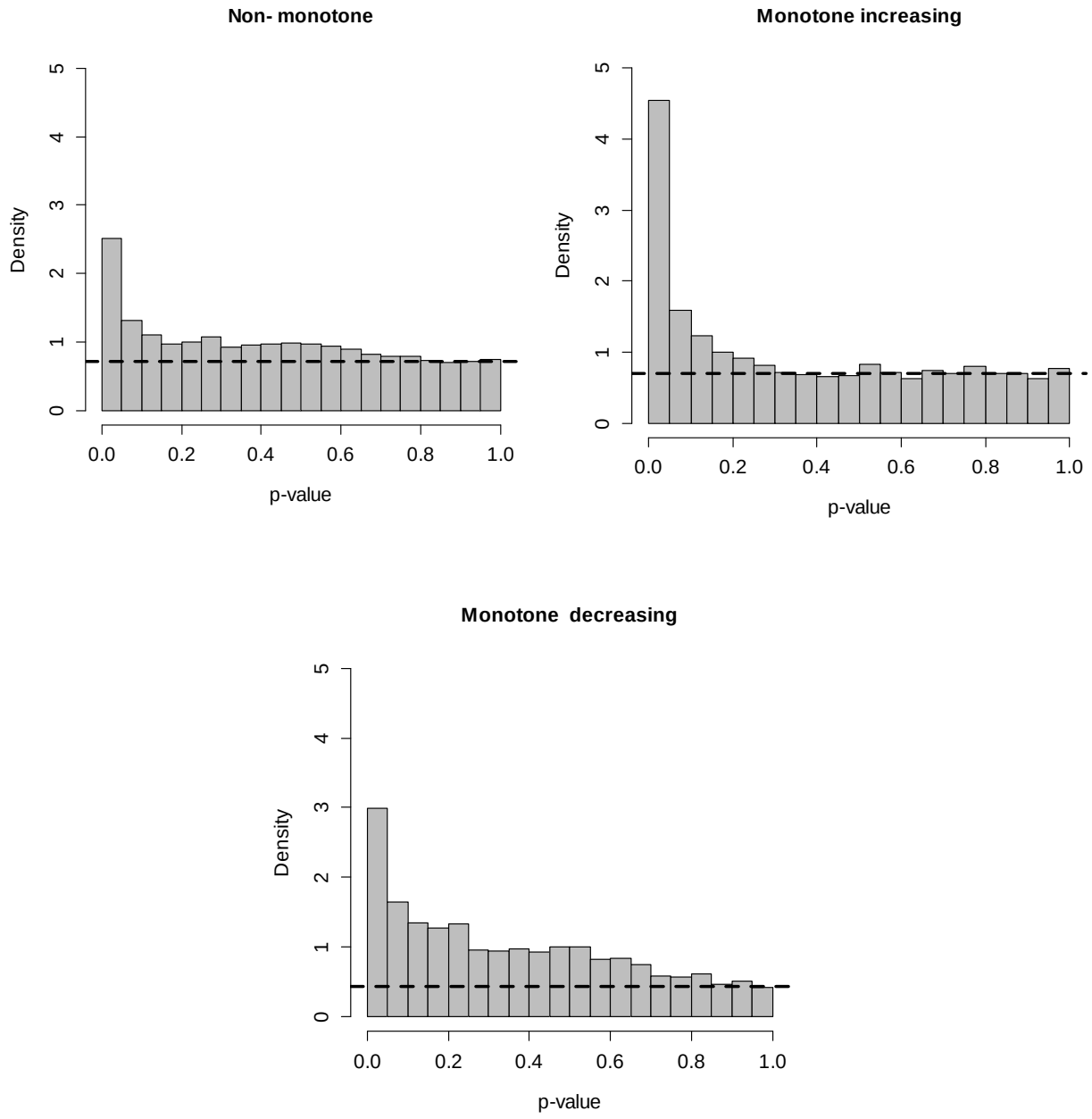

Figure 4.3. Histogram of the three subsets of *p*-values for the second partitioning rule.

Figure 4.4 and 4.5 present the scatter plots of the *q*-values corresponding to analysis by the traditional *q*-value method in which no partitioning is used versus the proposed methods, partitioning rule I and the traditional method versus proposed methods, partitioning rule II.

Figure 4.6 presents a scatter plot corresponding to analysis by the proposed methods, partitioning rule I versus portioning rule II.

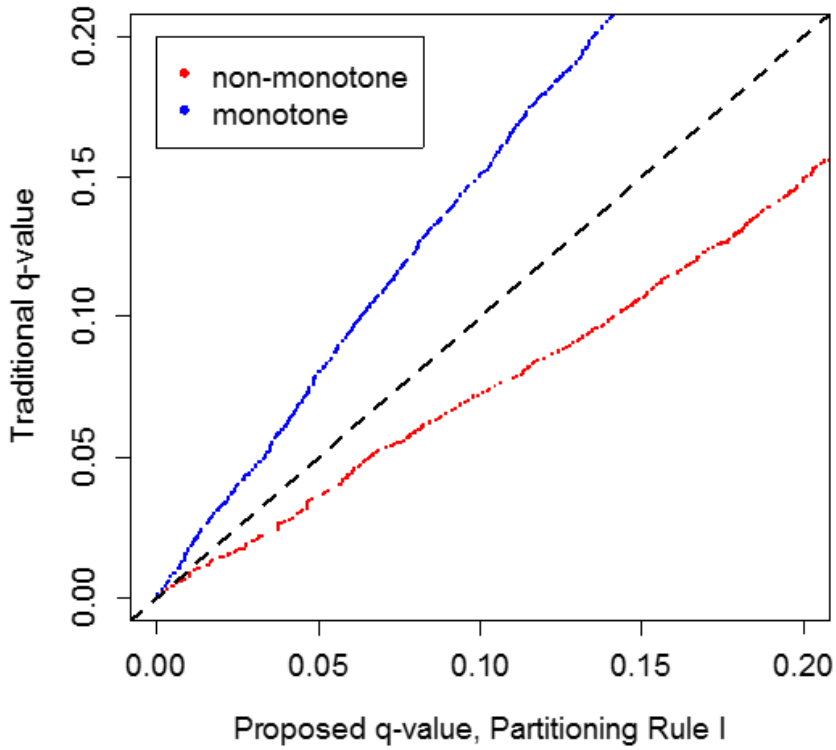

Figure 4.4. Traditional *q*-value versus Partitioning Rule I.

In the figure above, it can clearly be seen that the proposed method using partitioning rule I produces smaller *q*-values than the traditional method for genes that exhibit monotonicity in the rank treatment means and larger *q*-values than the traditional method for genes that exhibit non-monotonicity in their ranked treatment means. The observed proportions of genes in the ordered sample treatment means that are non-monotone and monotone are 0.602 and 0.398, respectively. For EE genes, the expected proportion of genes that are non-monotone is 0.667, but the observed proportion of all genes exhibiting non-monotonicity is lower. Similarly the expected proportion of EE genes that are monotone are 0.333, which are lower than the observed proportion all genes exhibiting monotonicity. This indicates that genes with ordered sample means that are monotone

are overrepresented and result in lower q-values and, potentially, improved identification of DE genes.

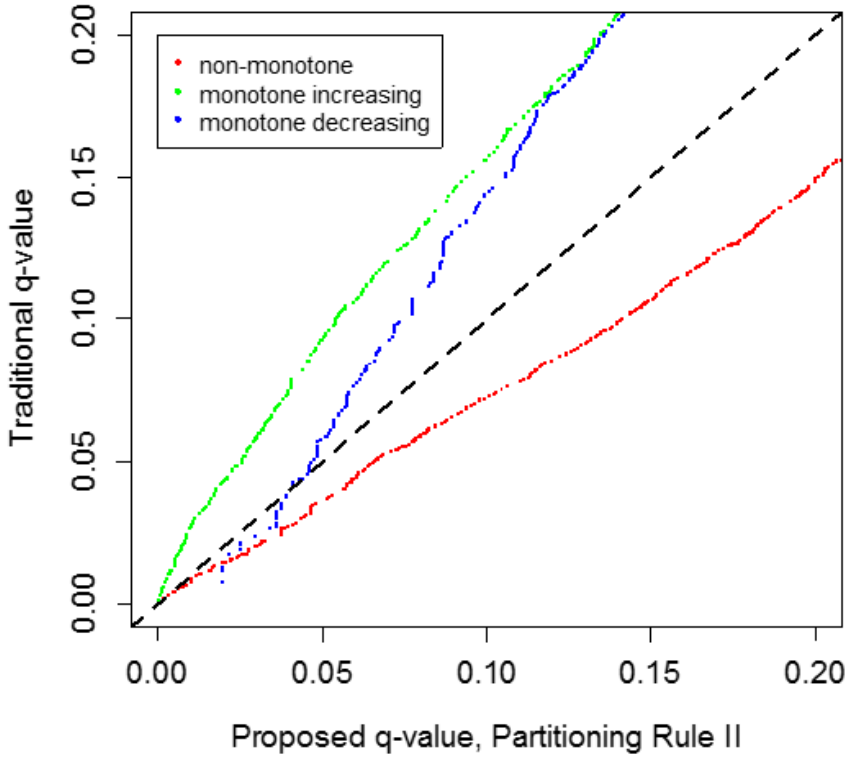

Figure 4.5. Traditional q-value versus Partitioning Rule II.

Figure 4.5 also shows that when partitioning rule II is used, genes which have ordered sample means that are monotone increasing or monotone decreasing produce lower q-values than the traditional method, but produce larger q-values than the traditional method for genes that have ordered sample treatment means that are non-monotone. The observed proportions of genes with ordered sample treatment means that are non-monotone, monotone increasing and monotone decreasing are 0.602, 0.154 and 0.243, respectively. These expected proportions for EE genes are 0.667, 0.167, and 0.167, respectively. This indicates that, genes with ordered sample means that are monotone decreasing are overrepresented, resulting in smaller q-values and improves identification of DE genes. Additionally, genes with ordered sample means that are

monotone increasing appear to be less “underrepresented”, resulting in lower *q*-values when compared to the traditional method.

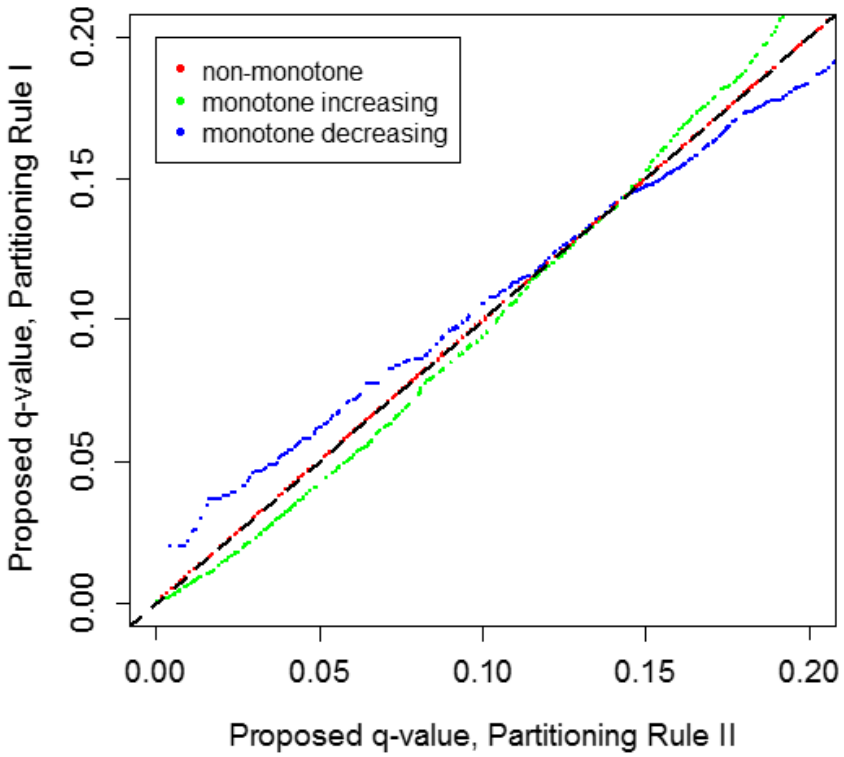

Figure 4.6. Partitioning Rule I versus Partitioning Rule II.

Figure 4.6 shows that, for small *q*-values, partitioning rule II produces smaller *q*-values for genes that have ordered sample means that are monotone decreasing in the ranked treatment means than partitioning rule I and larger *q*-values for genes that are monotone increasing in the ranked treatment means. The *q*-values for partitioning rule I and partitioning rule II are the same for genes that exhibit non-monotonicity. The smaller *q*-values observed in genes that exhibit monotone decreasing patterns can again be explained by the overrepresentation of genes in this group.

Based on the histogram of *p*-values, scatterplots of *q*-values and the observed proportions of genes exhibiting specific patterns, the second partitioning rule is to be used for this data. Since

the distribution of *p*-values was stochastically smaller than when first partitioning rule and no partitioning rule are used. Also, it was observed from the scatterplots of *q*-values for the second partitioning rule versus the traditional method, that where monotonicity is present resulted in smaller *q*-values and the observed proportion of genes that are monotone decreasing was higher than the expected proportion of EE genes that monotone decreasing.

4.3.1. Results – Over Expressed Gene Sets

A total of 1454 gene sets were analyzed to detect which gene set are overexpressed using the traditional and proposed methods, in conjunction with Fisher's exact tests, while controlling FDR at significance levels of 0.05, 0.10 and 0.20. No gene sets were declared to be overexpressed at significance levels of 0.05 or 0.10 using any of the methods. Five gene sets were identified to be overexpressed by the proposed methods and eight gene sets were identified by the traditional method, both at significance level 0.20.

4.4. Real Data Analysis II – Effect of Deferasirox in Leukemia Patients

The data from the gene expression experiment described in Junko (2009) is reanalyzed using the traditional and the proposed methods. The description of the data set is given in section 2.6.

For each gene, the null hypothesis

$$H_j : \mu_{1j} = \mu_{2j} = \mu_{3j} \tag{4.8}$$

for j = 1, 2,…, 42440 is tested using the moderated *F*-test.

From the *p*-values obtained in this analysis, the estimated number of EE genes is $\hat{m}_0 = 22593.14$ corresponding to $\hat{\pi}_0 = 0.532$. Using the traditional method, the *q*-value for each gene is estimated as;

$$q_{(j)} = \min\left\{ \frac{p_{(r)}\left(22593.14\right)}{r} : r = 1,...,42440 \right\} \tag{4.9}$$

Because this experiment, like the experiment described in Section 4.1, had three treatments, the same six observed orderings in equations (3.5), (3.6) and (3.7) of the sample treatment means are possible.

For the first partitioning rule (two subsets of *p*-values), the numbers of *p*-values associated with genes that have sample treatment means that are non-monotone and monotone in the ranked treatments are $m_1 = 24881$ and $m_2 = 17559$ with the estimated proportion of EE genes of $\hat{\pi}_0^{(1)} = 0.605$ and $\hat{\pi}_0^{(2)} = 0.429$ respectively. Using (3.8) and (3.9), the *q*-values for the subset of *p*-values associated with non-monotone orderings of the sample means and *p*-values associated with monotone orderings of the sample means are calculated separately as

$$q_{(k)}^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)}\left[22593.14\left(4/6\right)\right]}{r} : r = 1,...,24881 \right\} \tag{4.10}$$

and

$$q_{(k)}^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)}\left[22593.14\left(2/6\right)\right]}{r} : r = 1,...,17559 \right\} \tag{4.11}$$

For the second partitioning rule, *q*-values are calculated separately for three subsets of *p*-values. The numbers of *p*-values associated with genes that have sample treatment means with non-monotone orderings, monotone increasing orderings, and monotone decreasing orderings are

$m_1 = 24881$, $m_2 = 9259$ and $m_3 = 8300$ with estimated proportion of EE genes $\hat{\pi}_0^{(1)} = 0.605$, $\hat{\pi}_0^{(2)} = 0.407$ and $\hat{\pi}_0^{(3)} = 0.454$, respectively. Using (3.10), (3.11) and (3.12), the *q*-values are calculated for each subset separately as

$$q_{(k)}^{(1)} = \min\left\{ \frac{p_{(r)}^{(1)}\left[22593.14\left(4/6\right)\right]}{r} : r = 1,...,24881 \right\}, \tag{4.12}$$

$$q_{(k)}^{(2)} = \min\left\{ \frac{p_{(r)}^{(2)}\left[22593.14\left(1/6\right)\right]}{r} : r = 1,...,9259 \right\}, \tag{4.13}$$

and

$$q_{(k)}^{(3)} = \min\left\{ \frac{p_{(r)}^{(3)}\left[22593.14\left(1/6\right)\right]}{r} : r = 1,...,8300 \right\} \tag{4.14}$$

Table 4.3 below shows the number of genes DDE controlling FDR at $\alpha = 0.05, \alpha = 0.10$ and $\alpha = 0.20$ by both methods.

Table 4.3

Number of Genes DDE by the Traditional and Proposed Methods for Estimating Q-values at Three Significance Level (α = 0.05, α = 0.10, α = 0.20)

	Number of DDE		
Significance Level	Traditional Method	Proposed Methods	
		I	II
0.05	4442	5013	5025
0.10	8482	8459	8463
0.20	14540	14375	14432

From the table above, it can be seen that the traditional method identified more DE genes for significance levels of 0.10 and 0.20 compared to the proposed methods, but the proposed methods identified more genes as DE compare to the traditional method at significance level 0.05. An increase in the significance level was used to determine if it will results in lower likelihood of a false negative.

Figure 4.7 presents the histogram of observed *p*-values corresponding to analysis by the traditional *q*-value method in which no partitioning is used. Figures 4.8 and 4.9 present histograms corresponding to analysis by the proposed methods using the first and second partitioning rules, respectively. In each histogram, the estimated proportion of EE genes for each subset is plotted as a dashed horizontal line

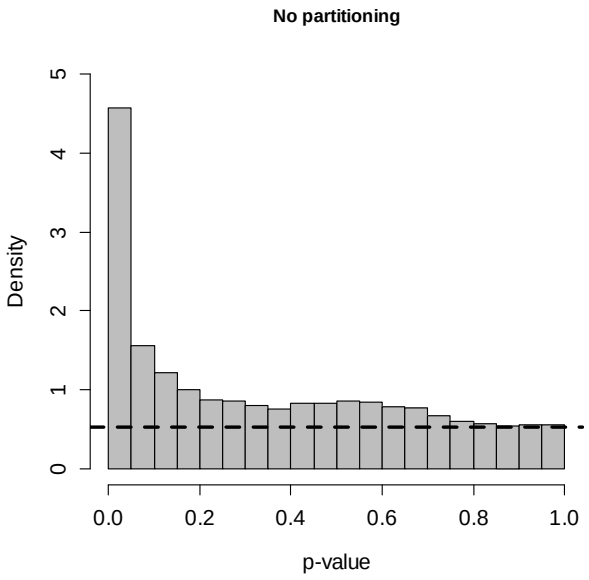

Figure 4.7. Distribution of *p*-values when no partitioning is used.

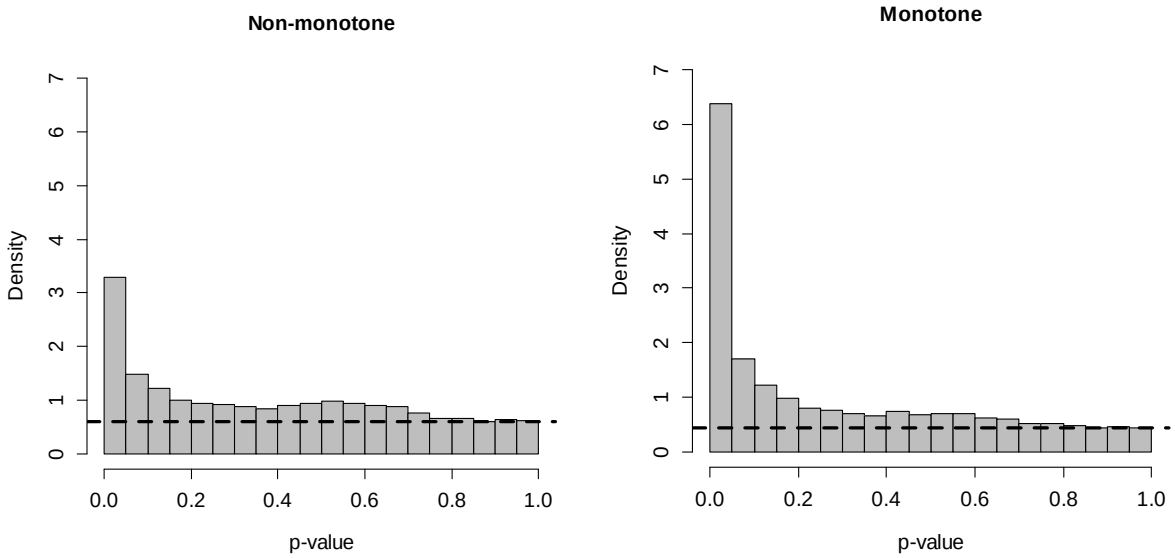

Figure 4.8. Histogram of two subsets of *p*-values using the first partitioning rule.

In figure 4.8 and 4.9, the distributions of *p*-values corresponding to genes with sample treatment means that exhibit monotonicity in the ranked treatments are stochastically smaller than the distribution of *p*-values for genes that do not exhibit monotonicity. Similar to data set analyzed in section 4.3, this indicates that a higher proportion of genes are DE among the genes that shows monotonicity in the ranked treatment means than the genes that do not exhibit monotonicity and that the proposed methods might be preferred to the traditional method and is a possible reason why the proposed methods identify more genes as DE than the traditional method.

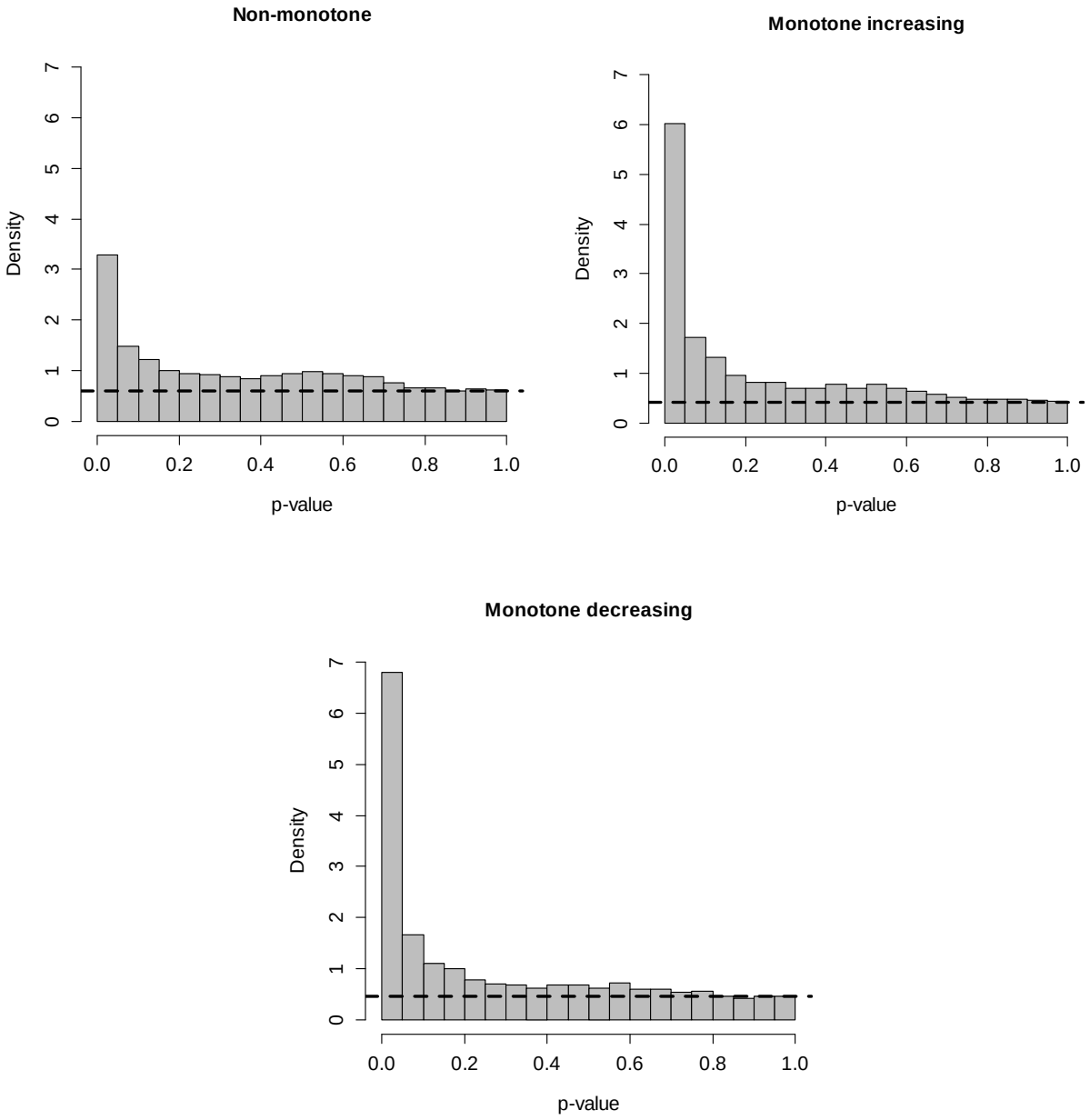

Figure 4.9. Histogram of the three subsets of *p*-values for the second partitioning rule.

Figure 4.10 and 4.11 present the scatter plots of the *q*-values corresponding to analysis by the traditional *q*-value method in which no partitioning is used versus the proposed methods, partitioning rule I and the traditional method versus proposed methods, partitioning rule II.

Figure 4.12 presents a scatter plot corresponding to analysis by the proposed methods, partitioning rule I versus portioning rule II.

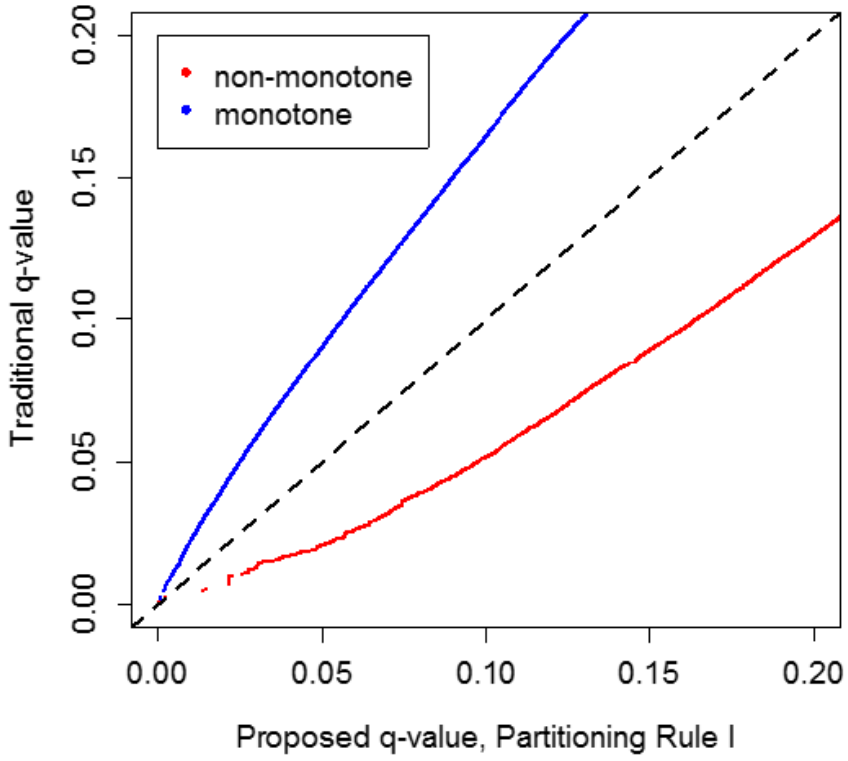

Figure 4.10. Traditional *q*-value versus Partitioning Rule I.

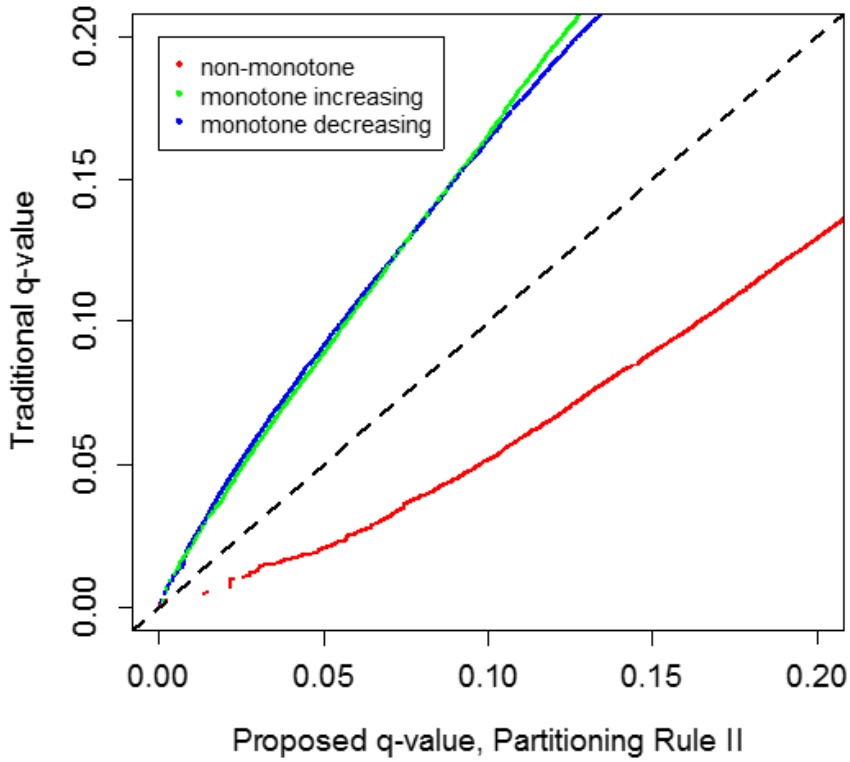

Figure 4.11. Traditional *q*-value versus Partitioning Rule II.

In the figure 4.10 and 4.11 above, it can clearly be seen that partitioning rule I and II produces smaller *q*-values than the traditional method for genes that exhibit monotonicity in the rank treatment means and larger *q*-values than the traditional method for genes that exhibit non-monotonicity in their ranked treatment means. The observed proportion of genes in the ordered sample treatment means that are non-monotone for both partitioning rules is 0.587. For EE genes, the expected proportion of genes that are non-monotone is 0.667, but the observed proportion of all genes exhibiting non-monotonicity is lower. For partitioning rule I the observed proportion of genes in the ordered sample treatment means that are monotone is 0.414. Likewise the expected proportion of EE genes that are monotone is 0.333, which is lower than the observed proportion of all genes exhibiting monotonicity. For partitioning rule II, the observed

proportion of genes in the ordered sample treatment means that are, monotone increasing and monotone decreasing were 0.218, and 0.196 respectively. Similarly, the expected proportion of EE genes that are monotone increasing and monotone decreasing is 0.167, which is lower than the observed proportion of all genes that are monotone increasing and monotone decreasing. This indicates that, for both partitioning rules genes with ordered sample means that exhibits monotonicity are overrepresented and result in lower *q*-values and, potentially, improved identification of DE genes.

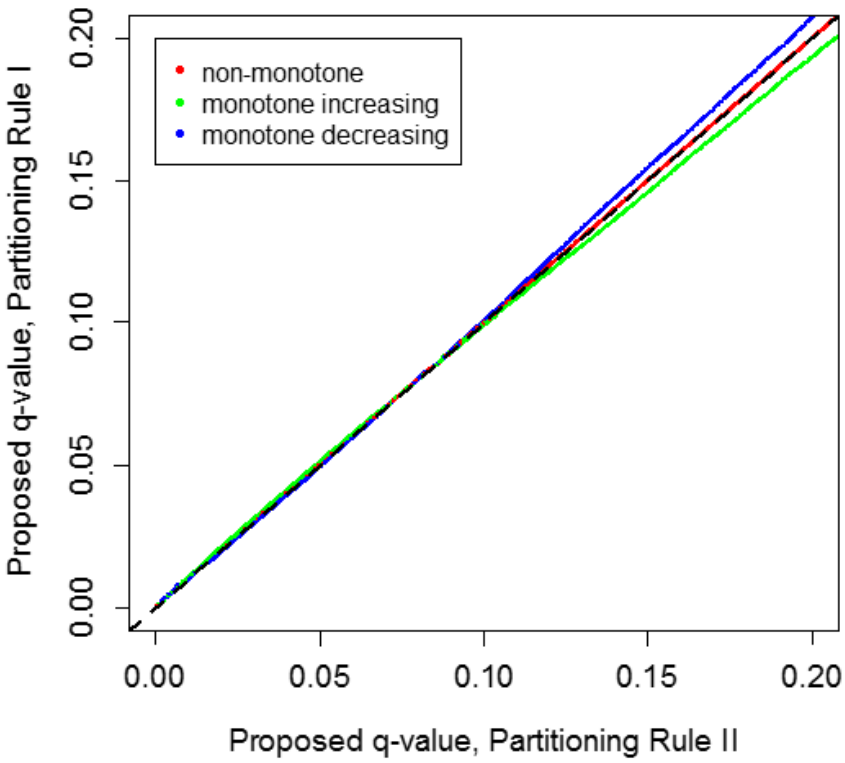

Figure 4.12. Partitioning Rule I versus Partitioning Rule II.

Figure 4.12 shows that, for small *q*-values, partitioning rule II produces smaller *q*-values for genes that have ordered sample means that are monotone decreasing in the ranked treatment means than partitioning rule I and larger *q*-values for genes that are monotone increasing in the ranked treatment means. The *q*-values for partitioning rule I and partitioning rule II are the same

for genes that exhibit non-monotonicity. The smaller *q*-values observed in genes that exhibit monotone increasing patterns can again be explained by the overrepresentation of genes in this group.

Based on the histogram of *p*-values, scatterplots of *q*-values and the observed proportions of genes exhibiting specific patterns, the second partitioning rule is to be used for this data. Since the distribution of *p*-values was stochastically smaller than when first partitioning rule and no partitioning rule are used. Also, it was observed from the scatterplots of *q*-values for the second partitioning rule versus the traditional method, that where monotonicity is present resulted in lower *q*-values and the observed proportion of genes that are monotone increasing was higher than the expected proportion of EE genes that monotone increasing.

4.4.1. Results – Over Expressed Gene Sets

Total of 1454 gene sets were analyzed to detect which gene set are overexpressed using the traditional and proposed methods, in conjunction with Fisher's exact tests, while controlling FDR at significance levels of 0.05, 0.10 and 0.20. Table 4.3 below gives a summary of the gene set analysis.

Table 4.4

Number of Gene Sets Identified to be Overexpressed by both the Traditional and Proposed Methods

	Number of Gene Sets		
Significance Level	Traditional Method	Proposed Methods	
		I	II
0.05	40	37	36
0.10	29	34	34
0.20	5	11	13

From the table above, it can be seen that the proposed methods identified more gene sets that are overexpressed at significance level 0.10 and 0.20 than the traditional method. The traditional method only identified more overexpressed gene sets than the proposed methods at 0.05 significance level.

CHAPTER 5: CONCLUSION RECOMMENDATION AND FUTURE WORK

5.1. Conclusion

In this research, existing and proposed methods were used to detect differentially expressed genes while controlling false discovery rate for microarray experiments in which the treatments can be ranked. The performances of these methods were evaluated using both simulated and real data.

The proposed methods for estimating the FDR by first partitioning the *p*-values into subsets based on observed patterns in the sample data, and then calculating *q*-values separately for each subset was shown to have advantages over the traditional *q*-value method in simulation settings with small sample size ($n = 4$). In the simulation settings, settings with $n = 10$ and $m_0 = 7000$ or $m_0 = 5000$, the proposed methods was outperformed by the traditional *q*-value method. Both the proposed and traditional methods adequately controlled the FDR at 5% significance level.

In the analysis of real gene expression data, the proposed methods generally declared more genes to be DE than the traditional *q*-value method, regardless of α. An exception for this include the data from the second gene expression experiment that was analyzed. For significance levels of 0.10 and 0.20, the traditional method declared more genes to be DE.

Also, the scatter plots showed that if the overrepresentation of gene expression patterns across treatment is taken into account, this can lead to an improvement in the identification of DE genes. Thus, using the proposed methods over the traditional method is recommended in these cases. Additionally, when monotonicity is overrepresented in one direction but not another, it's suggested that partitioning rule II be used over partitioning rule I.

For gene set testing, the traditional method identified more gene sets to be over expressed than the proposed methods. As the significance level was increased from 0.05 to 0.20, the number of gene sets identified to be overexpressed decreased.

5.2. Recommendations

The following recommendations are offered for related research in identification of differentially expressed genes:

(1) The proposed method is generally only recommended for analysis in gene expression experiments with small samples sizes.

(2) If monotonicity in the ordered sample means is overrepresented in both directions (increasing and decreasing), the proposed method should be used using the first partitioning rule.

(3) If monotonicity in the ordered sample means is overrepresented in only one direction (increasing or decreasing), the proposed method should be used using the first partitioning rule.

5.3. Future Work

It is noticeable that in this research, the proposed methods identified fewer overexpressed gene sets than the traditional method. Also, there was a significant decrease in the number of overexpressed gene sets identified as the significance level was increased. Therefore, it is desirable to develop a gene set testing method to identify overexpressed gene sets which does not depend on identifying genes using single gene testing.

REFERENCES

Adi L. Tarca P., Roberto Romero, MDa, and Sorin Draghici, PhD. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics & Gynecology, 195*(2), 373 - 388.

Barry W. T., Nobel A. B., & Wright F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics, 21*(9), 1943-1949. doi: 10.1093/bioinformatics/bti260

Ben-Dor A B. L., Friedman N, Nachman I, Schummer M, Yakhini Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology, 7*(3-4), 559 583. doi: doi:10.1089/106652700750050943

Cojocaru G. S., Rechavi G., & Kaminski N. (2001). The use of microarrays in medicine. *Isr Med Assoc J, 3*(4), 292-296.

DeRisi J., Penland,L., Brown,P.O., et al. (1996). Use of a cDNAmicroarray to analyse gene expression patterns in human cancer. *Nature Genetics, 14*(4), 457 - 460.

Goeman J. J., Oosting J., Cleton-Jansen A.-M.*, et al.* (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics, 21*(9), 1950-1957. doi: 10.1093/bioinformatics/bti267

Goeman J. J., van de Geer S. A., de Kort F.*, et al.* (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20*(1), 93-99. doi: 10.1093/bioinformatics/btg382

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

Holm S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65 - 70.

John D. Storey a. R. T. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440-9445.

Junko H. Ohyashiki C. K., Ryoko Hamamura, Seiichi Okabe, Tetsuzo Tauchi and Kazuma Ohyashiki. (2009). Blackwell Publishing Asia The oral iron chelator deferasirox represses signaling through the mTOR in myeloid leukemia cells by enhancing expression of REDD1. *Cancer Science, 100*(5), 970 - 977. doi: 10.1111/j.1349-7006.2009.01131.x

Lazar N. (2012). The Big Picture: Multiplicity Control in Large Data Sets Presents New Challenges and Opportunites. *CHANCE, 25*(2), 37-40.

Liat Ein-Dor O. Z., Eytan Domany. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America, 103*(15), 5923 - 5928. doi: 10.1073/pnas.0601231103

Lockhart D. J., Dong,H.L., Byrne,M.C., Follettie,M.T., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology, 14*(13), 1675 - 1680.

M. Kathleen Kerr M. M. a. G. A. C. (2000). Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology, 7*, 819 - 837.

Macgregor P. F., & Squire J. A. (2002). Application of Microarrays to the Analysis of Gene Expression in Cancer. *Clinical Chemistry, 48*(8), 1170-1177.

Mansmann U., & Meister R. (2005). Testing Differential Gene Expression in Functional Groups Goeman's Global Test versus an ANCOVA Approach. *Methods of Information in Medicine, 44*(3), 449-453.

Nam D., & Kim S.-Y. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics, 9*(3), 189-197. doi: 10.1093/bib/bbn001

Newton MA K. C., Richmond CS, Blattner FR, Tsui KW. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology, 8*(1), 37 - 52.

Orr M., Liu, P, & Nettleton, D. (2014). An Improved Method for Computing Q-values when the Distribution of Effect Sizes is Asymmetric, submitted to *Bioinformatics*.

Petricoin E. F., 3rd, Hackett J. L., Lesko L. J.*, et al.* (2002). Medical applications of microarray technologies: a regulatory science perspective. *Nat Genet, 32 Suppl*, 474-479. doi: 10.1038/ng1029

SIMES R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*(3), 751-754. doi: 10.1093/biomet/73.3.751

Smyth G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*(1).

Storey J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistics Society, 64*(3), 479 - 498.

Subramanian A., Tamayo P., Mootha V. K.*, et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. doi: 10.1073/pnas.0506580102

Thomas Jeffrey G. J. M. O., Stephen J. Tapscott, and Lue Ping Zhao. (2001). An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research, 11*, 1227 - 1236.

Tusher V. G., Tibshirani R., & Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, 98*(9), 5116-5121. doi: 10.1073/pnas.091062498

Vamsi K Mootha C. M. L., Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics, 34*, 267-273. doi: 10.1038/ng1180

Wanda Lattanzi C. B., Carlo Gangitano and Fabrizio Michetti. (2007). Hypoxia-like transcriptional activation in TMT-induced degeneration: microarray expression analysis on PC12 cells. *Journal of Neurochemistry, 100*, 1688–1702.

Zhang A. (2006). *Advanced Analysis of gene expression microarray data*: World Scientific Publishing Co, Pte. Ltd.

APPENDIX

A1. Simulation Code

```
library(limma)
library(qvalue)
library(pscl)

source("incr_decr_fn.R")

ni <- 4                 ### number of samples per treatment
m0 <- 7000              ## number of EE genes
m <- 10000              ## total number of genes
m1 <- m - m0            ## number of DE genes
pi <- c(4/6,1/6,1/6)    ### proportion of monotone and non-monotone
Nm1 <- m1 * pi   ### proportion of DE genes for non-monotone and monotone (increase
        & decrease)


### Defining the variance

d0 <- 2.15
s20 <- 0.037
vars <- rigamma(n = m, alpha = d0/2, beta = d0*s20/2)



### Simulating the data sets

sim <- function(x){

        dat <- matrix(NA, nrow = m0, ncol = 3*ni)    matrix of EE genes

        for(i in 1 : m0) {
                sdi <- sqrt(vars[i])
                dati <- rnorm(n = 3*ni, mean = 0, sd = sdi)
                dat[i,] <- dati
        }

        for(i in 1 : m1) {        ### treatment means relationship matrix (user definition)

                mat1 <- c(rep(c(0,0,1,0,0,1,0,1,0,0,1,0,1,0,0,1,0,0,0,2,1,1,0,2,2,0,1,1,2,0))
                mat2 <- c(rep(c(0,1,2),500))
                mat3 <- c(rep(c(0,-1,-2),500))
                means <- matrix(c(mat1, mat2, mat3), nrow = m1, ncol = 3, byrow =
                TRUE)
```

```
        }

        dat1 <- matrix(NA, nrow = m1, ncol = 3*ni)          ### matrix of DE genes

        for(i in 1 : m1) {

                DEms <- means[i,]
                sdi <- sqrt(vars[i])
                dat1i <- rnorm(n = ni, mean = DEms[1]*sdi, sd = sdi)
                dat2i <- rnorm(n = ni, mean = DEms[2]*sdi, sd = sdi)
                dat3i <- rnorm(n = ni, mean = DEms(Petricoin et al.)*sdi, sd = sdi)
                datai <- c(dat1i, dat2i, dat3i)
                dat1[i,] <- datai

        }

        data <- rbind(dat,dat1)

        }


dataset <- lapply(1:50, sim)            ### Number of simulated data sets


### Analzes using moderated F test, traditional q-value and the proposed methods

        RVSVR <- data.frame()
        RVSVR1 <- data.frame()
        RVSVR2 <- data.frame()

        for(i in 1 : length(dataset)) {

                aa <- as.data.frame(dataset[i])
                grps <- as.factor(c(rep(1,4),rep(2,4),rep(3,4)))        ### User define (depending
                on the  number of treatments per group)

        ##microarray analysis -- moderated F-test

        design <- model.matrix(~grps+0)  ##design matrix
        colnames(design)=c("t1","t2","t3")
        contr.mat <- makeContrasts(t1-t2, t1-t3, t2-t3, levels=design) ##contrasts of interest
```

```
##Perform moderated F-test

fit1 <- lmFit(aa,design)
fit2 <- contrasts.fit(fit1,contr.mat)
fit3 <- eBayes(fit2)
pvs <- fit3$F.p.value  ##ANOVA p-values



qv <- qvalue(pvs)
pi0hat <- qv$pi0 ##estimate of proportion of EE genes
m0hat <- m*pi0hat ##estimate of the number of EE genes
qvals <- qv$qvalues ##q-values for traditional method



R<-sum(qvals <= 0.05)   ### Number of genes DDE using traditional method
V <- sum(qvals[0:m0] <= 0.05) ### Number of EE genes DDE
S <- sum(qvals[(m0 + 1) : m] <= 0.05)  ### Number of DE genes DDE
VR <- V / max(R,1)



##analysis using partitions of two subsets of p-values

yb1 <- apply(aa[,grps==1], 1, mean)
yb2 <- apply(aa[,grps==2], 1, mean)
yb3 <- apply(aa[,grps==3], 1, mean)
ybars <- cbind(yb1, yb2, yb3)



ybind <- apply(ybars, 1, incr_decr_fn)
sum(ybind == 0) ##not monotone
sum(ybind == 1) ##increasing
sum(ybind == 2) ##decreasing



##Analysis using partitions of two subsets of p-values

pv1 <- pvs[ybind == 0]  ##p-values with non-monotone means
pv2 <- pvs[ybind != 0]  ##p-values with monotome means (increasing or
                          decreasing)
```

```
m1 <- length(pv1)  ##number of genes with non-monotone means
m2 <- length(pv2)  ##number of genes with monotone means



m0hat1 <- m0hat*4/6  ##estimate of number of EE genes with non-monotone
                       means
m0hat2 <- m0hat*2/6  ##estimate of number of EE genes with monotone means



pi0hat1 <- m0hat1/m1  ##estimate of proportion of EE genes among genes with
                        non-monotone means
pi0hat2 <- m0hat2/m2  ##estimate of proportion of EE genes among genes with
                        monotone means

##calculate FDR for genes with non-monotone means

prank1 <- rank(pv1)
fdrh1 <- pv1*m0hat1/prank1
qval1 <- sapply(prank1, function(x) min(fdrh1[prank1 >= x]))



##calculate FDR for genes with monotone means

prank2 <- rank(pv2)
fdrh2 <- pv2*m0hat2/prank2
qval2 <- sapply(prank2, function(x) min(fdrh2[prank2 >= x]))



qvals2 <- rep(NA, m)
qvals2[ybind == 0] <- qval1
qvals2[ybind != 0] <- qval2



R1 <- sum(qvals2 <= 0.05)   ### Number of genes DDE using traditional method
V1 <- sum(qvals2[0:m0] <= 0.05) ### Number of EE genes DDE
S1 <- sum(qvals2[(m0 + 1) : m] <= 0.05)  ### Number of DE genes DDE
VR1 <- V1 / max(R1,1)
```

```
##Analysis using partitions of three subsets of p-values

        pv1 <- pvs[ybind == 0]  ##p-values with non-monotone means
        pv2 <- pvs[ybind == 1]  ##p-values with monotone increasing means
        pv3 <- pvs[ybind == 2]  ##p-values with monotone decreasing means



        m1 <- length(pv1)  ##number of genes with non-monotone means
        m2 <- length(pv2)  ##number of genes with monotone increasing means
        m3 <- length(pv3)  ##number of genes with monotone decreasing means



        m0hat1 <- m0hat*4/6  ##estimate of number of EE genes among genes with non-
                               monotone means
        m0hat2 <- m0hat*1/6  ##estimate of number of EE genes among genes with
                               monotone increasing means
        m0hat3 <- m0hat*1/6  ##estimate of number of EE genes among genes with
                               monotone decreasing means


        pi0hat1 <- m0hat1/m1  ##estimate of proportion of EE genes among genes with
                              non-monotone means
        pi0hat2 <- m0hat2/m2  ##estimate of proportion of EE genes among genes with
                              monotone increasing means
        pi0hat3 <- m0hat3/m3  ##estimate of proportions of EE genes among genes with
                              monotone decreasing  means


##calculate FDR

##calculate FDR for genes with non-monotone means

        prank1 <- rank(pv1)
        fdrh1 <- pv1*m0hat1/prank1
        qval1 <- sapply(prank1, function(x) min(fdrh1[prank1 >= x]))



##calculate FDR for genes with monotone increasing means

        prank2 <- rank(pv2)
        fdrh2 <- pv2*m0hat2/prank2
        qval2 <- sapply(prank2, function(x) min(fdrh2[prank2 >= x]))
```

```
##calculate FDR for genes with monotone decreasing means

        prank3 <- rank(pv3)
        fdrh3 <- pv3*m0hat3/prank3
        qval3 <- sapply(prank3, function(x) min(fdrh3[prank3 >= x]))



        qvals3 <- rep(NA, m)
        qvals3[ybind == 0] <- qval1
        qvals3[ybind == 1] <- qval2
        qvals3[ybind == 2] <- qval3



        R2 <-sum(qvals3 <= 0.05)   ###  Number of genes DDE using traditional method
        V2 <- sum(qvals3[0:m0] <= 0.05) ### Number of EE genes DDE
        S2 <- sum(qvals3[(m0 + 1) : m] <= 0.05)  ### Number of DE genes DDE
        VR2 <- V2 / max(R2,1)



        RVSVR <- rbind(RVSVR,c(S, VR))          ### S and VR using the traditional
                                                    method
        colnames(RVSVR) <- c("S", "VR")



        RVSVR1 <- rbind(RVSVR1,c(S1, VR1))      ### S and VR using the proposed
                                                    method (2 subsets of p-values)
        colnames(RVSVR1) <- c("S", "VR")



        RVSVR2 <- rbind(RVSVR2,c(S2, VR2))      ### S and VR using the proposed
                                                    method (3 subsets of p-values)
        colnames(RVSVR2) <- c("S", "VR")

    }


## The mean and standard errors of S and V/R
```

```
MStd <- round(apply(RVSVR, 2, function(x) c(mean(x), sqrt(var(x) / length(x)))),digits = 3)

MStd1 <- round(apply(RVSVR1, 2, function(x) c(mean(x), sqrt(var(x) / length(x)))),digits = 3)

MStd2 <- round(apply(RVSVR2, 2, function(x) c(mean(x), sqrt(var(x) / length(x)))),digits = 3)
```

A2. Gene Expression Analysis Code (Traditional and Proposed Methods)

```
library(qvalue)
source("incr_decr_fn.R")



aa<-read.csv("new_gse5073.csv",header=T)          ###read in data

data<-aa[,3:11]          ####data containing expression values only

m<- dim(aa)[1]

x<-as.factor(c(1,2,3,2,2,3,3,1,1))                ###treatment levels



###Perform ANOVA

dd<-data.frame()

for(i in 1 : nrow(aa)) {

        abc<-as.numeric(data[i,])
        model3<-summary(aov(lm(abc~x)))
        fvalue<-model3[[1]]$'F value'
        pvalue<-model3[[1]]$'Pr(>F)'

dd<-rbind(dd,c(fvalue,pvalue))

colnames(dd) <- c("Fvalue", "NAA", "Pvalue", "NAAA")

}



data1<-dd[,c(1,3)]              ###data containing fvalues and pvalues
data2<-cbind(aa,data1)          ####data containing the expression values, fvals & pvals
```

```
pvalue<-data1[,2]              ###data containig only pvalues
qvalues<-qvalue(pvalue)        ### qvalues of the pvalues


pi0hat <- qvalues$pi0          ##estimate of proportion of EE genes
m0hat <- m*pi0hat              ##estimate of number of EE genes
qvals <- qvalues$qvalues


data3<-cbind(data2,qvals)


sig.genesT1 <- data3[which(qvals<=0.05),]  ###genes DDE using traditional method
nrow(sig.genesT1)

sig.genesT2 <- data3[which(qvals<=0.10),]  ###genes DDE using traditional method
nrow(sig.genesT2)

sig.genesT3 <- data3[which(qvals<=0.20),]  ###genes DDE using traditional method
nrow(sig.genesT3)


##Histogram of p-values with pi0hat line

hist(pvalue, probability=TRUE, col="gray", xlab="p-value", main="No partitioning",

      cex.lab=1.2, cex.axis=1.2, cex.main=1)

abline(h=pi0hat, lty=2, lwd=3)


##analysis using partitions of two subsets of p-values

yb1 <- apply(data[,x==1], 1, mean)
yb2 <- apply(data[,x==2], 1, mean)
yb3 <- apply(data[,x==3], 1, mean)

ybars <- cbind(yb1, yb2, yb3)

ybind <- apply(ybars, 1, incr_decr_fn)
sum(ybind == 0) ##not monotone
```

```
sum(ybind == 1) ##increasing
sum(ybind == 2) ##decreasing
```

##Analysis using partitions of two subsets of p-values

```
pv1 <- pvalue[ybind == 0]        ##p-values with non-monotone means
pv2 <- pvalue[ybind != 0]        ##p-values with monotome means (increasing or
                                  decreasing)

m1 <- length(pv1)                ##number of genes with non-monotone means
m2 <- length(pv2)                ##number of genes with monotone means

m0hat1 <- m0hat*4/6              ##estimate of number of EE genes with non-
                                  monotone means
m0hat2 <- m0hat*2/6              ##estimate of number of EE genes with monotone
                                  means

pi0hat1 <- m0hat1/m1             ##estimate of proportion of EE genes among genes
                                  with non-monotone means
pi0hat2 <- m0hat2/m2             ##estimate of proportion of EE genes among genes
                                  with monotone means

par(mfrow=c(1,2))
```

##Histogram of p-values corresponding to genes with non-monotone means (with pi0hat line)

```
hist(pv1, probability=TRUE, ylim=c(0,4), breaks=20, col="gray", main="Non-
        monotone", xlab="p-value", cex.lab=1.2, cex.axis=1.2, cex.main=1.2)
abline(h=pi0hat1, lty=2, lwd=3)
```

##Histogram of p-values corresponding to genes with monotone means (with pi0hat line)

```
hist(pv2, probability=TRUE, ylim=c(0,4), breaks=20, col="gray", main="Monotone",
        xlab="p-value", cex.lab=1.2, cex.axis=1.2, cex.main=1.2)
abline(h=pi0hat2, lty=2, lwd=3)


##calculate FDR for genes with non-monotone means

prank1 <- rank(pv1)
fdrh1 <- pv1*m0hat1/prank1
qval1 <- sapply(prank1, function(x) min(fdrh1[prank1 >= x]))


##calculate FDR for genes with monotone means

prank2 <- rank(pv2)
fdrh2 <- pv2*m0hat2/prank2
qval2 <- sapply(prank2, function(x) min(fdrh2[prank2 >= x]))



qvals2 <- rep(NA, m)
qvals2[ybind == 0] <- qval1
qvals2[ybind != 0] <- qval2


###Scatter plots
###Traditional vs Partitioning Rule I

colors <- rep("blue", m)
colors[ybind==0] <- "red"

plot(qvals2, qvals, pch=20, xlim=c(0,0.20), ylim=c(0,0.20), cex=0.5, col=colors,
xlab="Proposed q-value, Partitioning Rule I", ylab="Traditional q-value", cex.lab=1.2,
cex.axis=1.2)
abline(a=0, b=1, lty=2, lwd=2)

legend(x=0, y=0.65, xjust=0, yjust=1, legend=c("non-monotone", "monotone"), pch=20,
        col=c("red", "blue"), cex=1.2)


data4<-cbind(data2,qvals2)
```

```
sig.genesI1 <- data4[which(qvals2<=0.05),]      ###genes DDE when p-values
                                                partitioned into two subsets
nrow(sig.genesI1)

sig.genesI2 <- data4[which(qvals2<=0.10),]      ###genes DDE when p-values
                                                partitioned into two subsets
nrow(sig.genesI2)


sig.genesI3 <- data4[which(qvals2<=0.20),]      ###genes DDE when p-values
                                                partitioned into two subsets
nrow(sig.genesI3)
```

##Analysis using partitions of three subsets of p-values

```
pv1 <- pvalue[ybind == 0]       ###p-values with non-monotone means
pv2 <- pvalue[ybind == 1]       ###p-values with monotone increasing means
pv3 <- pvalue[ybind == 2]       ###p-values with monotone decreasing means


m1 <- length(pv1)           ###number of genes with non-monotone means
m2 <- length(pv2)           ###number of genes with monotone increasing means
m3 <- length(pv3)           ###number of genes with monotone decreasing means


m0hat1 <- m0hat*4/6         ###estimate of number of EE genes among genes with non-
                            monotone means
m0hat2 <- m0hat*1/6          ###estimate of number of EE genes among genes with
                            monotone increasing means
m0hat3 <- m0hat*1/6          ###estimate of number of EE genes among genes with
                            monotone decreasing means


pi0hat1 <- m0hat1/m1        ###estimate of proportion of EE genes among genes with
                            non-monotone means
pi0hat2 <- m0hat2/m2        ###estimate of proportion of EE genes among genes with
                            monotone increasing means
pi0hat3 <- m0hat3/m3        ###estimate of proportions of EE genes among genes with
                            monotone decreasing means
```

```
par(mfrow=c(1,3))

##Histogram of p-values corresponding to genes with non-monotone means (with pi0hat line)
        hist(pv1, probability=TRUE, ylim=c(0,4), breaks=20, col="gray", main="Non-
                monotone", xlab="p-value", cex.lab=1.2, cex.axis=1.2, cex.main=1.2)
        abline(h=pi0hat1, lty=2, lwd=3)

##Histogram of p-values corresponding to genes with monotone increasing means (with pi0hat
line)
        hist(pv2, probability=TRUE, ylim=c(0,4), breaks=20, col="gray", main="Monotone
                increasing", xlab="p-value", cex.lab=1.2, cex.axis=1.2, cex.main=1.2)
        abline(h=pi0hat2, lty=2, lwd=3)

##Histogram of p-values corresponding to genes with monotone decreasing means (with pi0hat
line)
        hist(pv3, probability=TRUE, ylim=c(0,4), breaks=20, col="gray", main="Monotone
                decreasing", xlab="p-value", cex.lab=1.2, cex.axis=1.2, cex.main=1.2)
        abline(h=pi0hat3, lty=2, lwd=3)

##calculate FDR
##calculate FDR for genes with non-monotone means
        prank1 <- rank(pv1)
        fdrh1 <- pv1*m0hat1/prank1
        qval1 <- sapply(prank1, function(x) min(fdrh1[prank1 >= x]))
```

```
##calculate FDR for genes with monotone increasing means

        prank2 <- rank(pv2)
        fdrh2 <- pv2*m0hat2/prank2
        qval2 <- sapply(prank2, function(x) min(fdrh2[prank2 >= x]))



##calculate FDR for genes with monotone decreasing means

        prank3 <- rank(pv3)
        fdrh3 <- pv3*m0hat3/prank3
        qval3 <- sapply(prank3, function(x) min(fdrh3[prank3 >= x]))



        qvals3 <- rep(NA, m)
        qvals3[ybind == 0] <- qval1
        qvals3[ybind == 1] <- qval2
        qvals3[ybind == 2] <- qval3




###Scatter plots
###Traditional vs Partitioning Rule II

        colors <- rep("blue", m)
        colors[ybind==0] <- "red"
        colors[ybind==1] <- "green"



        plot(qvals3, qvals, pch=20, xlim=c(0,0.20), ylim=c(0,0.20), cex=0.5, col=colors,
        xlab="Proposed q-value, Partitioning Rule II",ylab="Traditional q-value", cex.lab=1.2,
        cex.axis=1.2)
        abline(a=0, b=1, lty=2, lwd=2)

        legend(x=0, y=0.65, xjust=0, yjust=1, legend=c("non-monotone", "monotone increasing",
                "monotone decreasing"), pch=20, col=c("red", "green","blue"), cex=0.9)


###Partitioning Rule I vs Partitioning Rule II
```

```
plot(qvals2, qvals3, pch=20, xlim=c(0,0.20), ylim=c(0,0.20), cex=0.5, col=colors,
xlab="Proposed q-value, Partitioning Rule II", ylab="Proposed q-value, Partitioning Rule
I", cex.lab=1.2, cex.axis=1.2)
abline(a=0, b=1, lty=2, lwd=2)

legend(x=0, y=0.72, xjust=0, yjust=1, legend=c("non-monotone", "monotone increasing",
        "monotone decreasing"), pch=20, col=c("red", "green","blue"), cex=1)


data5<-cbind(data2,qvals3)

sig.genesII1 <- data5[which(qvals3<=0.05),]        ###genes DDE when p-values
                                                   partitioned into three subsets
nrow(sig.genesII1)

sig.genesII2 <- data5[which(qvals3<=0.10),]        ###genes DDE when p-values
                                                   partitioned into three subsets
nrow(sig.genesII2)

sig.genesII3 <- data5[which(qvals3<=0.20),]        ###genes DDE when p-values
                                                   partitioned into three subsets
nrow(sig.genesII3)


### Data containing probe sets, gene symbol and qvals(traditional and proposed)

new_data<-cbind(data3[,c(1:2,14)],data4[,14],data5[,14])
colnames(new_data)<-c("Name","Symbol","Traditional","Proposed I","Proposed II")
```

A3. Gene Set Code

```
library(qvalue)

setgene<-read.csv("setgene.csv",header=F)          ###gene sets data

data <-read.csv("data.csv")    ### data containing probe set, gene symbols, trad.qvals and
                               improved qvals


m<-length(unique(Symbol))          ###Number of gene symbols

n_set<-as.data.frame(setgene[,1])  ###gene sets names
```

```
trad.DDE1<-unique(Symbol[Traditional <= 0.05])   ###Number of unique symbol using
                                                    a specific method and significance level


###Counts data

ds<-data.frame()

    for(i in 1 : nrow(setgene)){
        s<-setgene[i,-(1)]
        ss<-s[!s == ""]
        n1<-length(ss)
        set<-ss[which(ss %in% trad.DDE1)]
        x1<-length(set)
        n<-length(trad.DDE1)
        x2<-n-x1
        n2<-m-n1
        ds<-rbind(ds,c(x1,x2,n1,n2))
        colnames(ds) <- c("X1", "X2", "N1", "N2")}


###Fisher Exact Test

df<-data.frame()

    for(i in 1 : nrow(ds)){
        mat <- as.numeric(ds[i,])
        mata <- fisher.test(matrix(mat,nrow=2),alternative="two.sided")
        pv <- mata$p.value
        odds <- mata$estimate
        df <- rbind(df,c(pv,odds))
        colnames(df) <- c("pvalue", "oddsR")}


data1<-cbind(ds,df)          ###Data containing the counts, p-value and odds ratio

dat<-data1[,5]               ###p-values

qval<-qvalue(dat)                ###q-values of the p-values
qvals <- qval$qvalues
```

```
trad1<-cbind(n_set,data1,qvals)        ###data containing gene sets names, counts, p-values and q-values

sig.genesetT1<-trad1[trad1$qvals<=0.05 & trad1$oddsR>1,]        ###over expressed gene sets
nrow(sig.genesetT1)
```

A4. Increasing and Decreasing Function

```
incr_decr_fn <- function(ybs){
        ##0: non monotonic
        ##1: increasing
        ##2: decreasing
        lybs <- length(ybs)
        ret <- 0
        if(sum(order(ybs) == (1:lybs)) == lybs) { ret <- 1 }
        if(sum(order(ybs, decreasing=TRUE) == (1:lybs)) == lybs) { ret <- 2 }
        if(length(unique(ybs))==1) { ret <- 0 }

return(ret)}
```