The Pennsylvania State University
The Graduate School

METHODS IN MULTIPLE TESTING AND META-ANALYSIS WITH APPLICATIONS TO THE ANALYSIS OF GENOMIC DATA

A Dissertation in
The Department of Statistics
by
Yihan Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2014

The dissertation of Yihan Li was reviewed and approved* by the following:

Debashis Ghosh
Professor of Statistics
Dissertation Advisor, Chair of Committee

Le Bao
Assistant Professor of Statistics

Qunhua Li
Assistant Professor of Statistics

Robert F. Paulson
Professor of Veterinary and Biomedical Sciences

Aleksandra Slavkovic
Associate Professor of Statistics
Chair of Graduate Program

*Signatures are on file in the Graduate School.

Abstract

With the rapid development of high-throughput technologies, such as microarrays, new statistical challenges arise in the analysis of genomic data from complex experiments. In this thesis, we mainly focused on problems arising from two challenges in the analysis of genomic data: multiple testing and meta-analysis. Multiple testing is a frequently encountered problem in statistics. The false discovery rate (Benjamini and Hochberg, 1995) is now a widely accepted concept in the context of large scale multiple testing, and the corresponding Benjamini-Hochberg procedure is widely adopted. We considered problems related to "multi-dimensional" multiple testing, when the multiple hypotheses being tested cannot be arranged into a single vector, but rather multi-dimensional structures, such as a matrix in the two-dimensional case. Concrete examples we considered include time-course gene expression studies, multiple comparisons of treatments for gene expression studies, hierarchical tumor classification, etc.. Meta-analysis is also an important topic in statistics. We focused on meta-analysis methods for genomic data with heterogeneity, where the goal is to detect consistent signals across the studies. In this thesis, we explored a weighted multiple testing procedure for meta-analysis, where the significance of the genes are weighted by its degree of heterogeneity; we discussed a two-step hierarchical hypothesis set testing framework for multi-dimensional multiple testing that applies to many of the examples aforementioned; and we developed a meta-analysis method based on weighted ordered p-values that aims at detecting signals in the majority of studies. Both simulation studies and data analysis examples will be used to assess the performances of the methodologies discussed.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I cannot begin to say how fortunate I feel for having Dr.Ghosh as my advisor – I cannot have asked for a better one! He has been extremely helpful in providing guidance to my research for the past four years. In addition, he has been a great morale support, providing encouragements whenever needed. I believe having an awesome advisor plays a great part in me having such an enjoyable study and research experience during my Ph.D. studies. I would also like to thank each and one of my committee members, Dr.Bao, Dr.Li and Dr.Paulson, for their help and support in completing this thesis. Thanks also go to the Statistics department and program, for all the friendly and helpful faculty and staff who make the department feel like home, as well as my wonderful classmates and friends who have made this small town a fun place to live for the past five years. Last but not least, I would like to thank my family, whose love and support I cannot do without, and I would like to thank my boyfriend Michael, whose existence has made the last year of my Ph.D. the best year of all.

Chapter 1

Introduction

1.1 Multiple Testing and False Discovery Rate Control

Multiple testing is a problem that frequently arises in statistics. If multiple hypotheses are tested simultaneously at a certain level, the probability of incorrectly rejecting at least one of the true null hypotheses becomes greater than the nominal level. In other words, the type I error rate of the joint hypotheses testing is inflated compared to single hypothesis testing.

The problem of multiple testing was first brought to broad attention in the 1950's in the context of multiple comparisons of means for ANOVA. The concept of "family-wise error rate" (FWER) was brought up during this time. It is defined as the probability of rejecting at least one of the true null hypothesis out of all the hypotheses being tested. Methods such as Tukey's range test and Scheffé's method were proposed, but these methods are very specific to this particular setting of multiple comparisons problem. Dunn (1961) came up with a solution using the Bonferroni inequalities. Though his method was originally intended for the problem of multiple comparisons of means as well, the method can be readily generalized to any kind of multiple testing problem, and is later well known by the name "Bonferroni correction". Bonferroni correction is a very general, easy to use and widely applicable method. It controls the family-wise error rate at level α when simultaneously testing m hypotheses, by adjusting the level of each of the individual tests to be α/m.

The advantage of the Bonferroni correction is that it does not impose any assumptions at all on the hypotheses being tested, therefore making it a very popular method for doing multiple testing on a small number of hypotheses. The limitations of the Bonferroni correction is that it can quickly become very conservative as the number of hypotheses gets large. Several more powerful methods that still maintain control of the family-wise error rate were proposed around the 1980's. A few well known ones include Holm's (1979) step-down procedure, Simes' (1986) procedure, and Hochberg's (1988) step-up procedure. Holm's procedure is uniformly more powerful than the Bonferroni correction method, and also does not impose any restrictions on the hypotheses. Because of this, some recommend always using the Holm's procedure instead of the Bonferroni correction. Simes' procedure only controls the family-wise error rate when all the null hypotheses being tested are true - this is sometimes called the control of family-wise error rate in the weak sense. In addition, Simes' procedure requires independence or positive dependence structures among the hypotheses being tested. Hochberg's procedure is based on the Simes' procedure, but does maintain control of the family-wise error rate in the strong sense - that is, without requiring all the null hypotheses to be true - but it also relies on independence or positive dependence structures among the hypotheses being tested. More details on these methods will be introduced as they come up in later sections.

The concept of the family-wise error rate as well as methods developed to control the FWER were fairly adequate for multiple testing problems arising in most areas of statistical applications up until the 1990's - such as multiple comparisons of means for ANOVA, or problems where multiple response measurements are being tested simultaneously. The number of hypotheses being tested simultaneously is usually at most on the order of tens. However, with the recent rapid development of "high-throughput" technologies in various areas of science, the problem of multiple testing faces new challenges. For example, technologies such as microarrays measure expression across the whole genome, generating data for tens of thousands of genes at a time, which leads to statistical testing for tens of thousands of hypotheses simultaneously. Other examples include the functional magnetic resonance imaging technology (fMRI) used in neurosciences that produces brain activity data for a large number of locations simultaneously.

As the number of hypotheses being tested simultaneously becomes increasingly large, it is crucial that we consider adjustments for multiple testing in order not to immensely increase the chances of making decision errors. However, it also becomes increasingly unrealistic to continue adopting the criteria of family-wise error rate. When thousands or more of hypotheses are being tested simultaneously, not allowing even one mistake to be made seems to become a somewhat too stringent requirement.

Benjamini and Hochberg (1995) proposed a new concept, namely the false discovery rate (FDR). The idea is that instead of controlling the probability of making any incorrect rejections, we could control the expected proportion of incorrect rejections out of all the rejections we make. This concept makes sense for many statistical applications arising in scientific areas involving multiple testing for extremely large numbers of tests, such as for microarray studies or other genomic data analysis. In these cases, statistical testing often serves as a screening process for finding potentially important features (e.g. genes) out of a very large pool of candidates. Under these situations, it is not very necessary that we try not making any mistakes at all, especially since we would likely reject many hypotheses at the same time. Instead, it makes sense to make sure that out of the features that we do "discover", only a small proportion of them are mistaken.

To state the definition of false discovery rate more formally, we summarize the possible outcomes of testing m hypotheses in Table 1.1 below. This table appears in Benjamini and Hochberg (1995) and is widely used when discussing multiple testing problems.

Table 1.1. Summary of the possible outcomes for simultaneously testing m hypotheses.

	Non-Rejections	Rejections	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
Total	$m - R$	R	m

As summarized in Table 1.1, m_0 is the number of true null hypotheses. The total number of hypotheses m is a known constant, while m_0 is assumed to be an unknown constant. R is the number of rejections we make when testing the m hypotheses, and R is an observed random variable. Since we do not know which null hypotheses are really true, U, V, T and S are all unobserved random variables. The false discovery rate is defined to be the expectation of V/R when $R > 0$, and 0 when $R = 0$. That is,

$$FDR = E\Big(\frac{V}{R \vee 1}\Big),$$

where $R \vee 1 = \max(R, 1)$. On the other hand, as a comparison, remember that the family-wise error rate is defined as the probability of making at least one incorrect rejection. Under the notations in Table 1.1,

$$FWER = P(V \geq 1).$$

As mentioned in Benjamini and Hochberg (1995), two facts about the FDR and the FWER can be shown easily: (1) when all the null hypotheses are true, i.e. when $m_0 = m$, the FDR is the same as FWER; (2) in all other cases, the FDR is smaller than the FWER. In other words, FDR is a less stringent criteria than the FWER by definition.

Benjamini and Hochberg (1995) proposed a procedure for multiple testing that aims at controlling the false discovery rate. Suppose $P_1, ..., P_m$ are the p-values corresponding to the m hypotheses respectively, and let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered p-values. Let $k = \max\{1 \leq i \leq m : P_{(i)} \leq i\alpha/m\}$. Then the procedure rejects the hypotheses corresponding to all $P_{(i)}$ for $i = 1, ..., k$. Benjamini and Hochberg (1995) proved that this procedure controls the false discovery rate at level α under independent test statistics corresponding to the true null hypotheses. We hereby refer to this procedure as the “BH procedure” or “BH method”. It can be easily shown that the BH procedure rejects at least as many hypotheses as the Bonferroni method. Benjamini and Hochberg (1995) showed through simulation studies that the BH procedure has higher power than the Bonferroni method, as well as some other FWER controlling procedures such as Hochberg’s procedure.

Based on Benjamini and Hochberg (1995), there is a lot of literature further investigating the theory of false discovery rate and its different variations, as well as extended versions of the BH procedure. Benjamini and Yekutieli (2001) relaxed the conditions for which the BH procedure controls the FDR. They proved that the BH procedure maintains control of the FDR under positive regression dependency among the test statistics corresponding to the true null hypotheses. They also proposed a modified version of the BH procedure that is more conservative but controls the FDR for all other forms of dependency. Genovese and Wasserman (2002) investigated the operating characteristics of the BH procedure, as well as some of its asymptotic properties. Later, Genovese and Wasserman (2004) developed a more general framework that involves modeling the false discovery proportion V/R as a stochastic process.

Storey (2003) introduced a modified version of the false discovery rate, namely the positive false discovery rate (pFDR), which is defined as $E(V/R \mid R > 0)$, and discussed its advantages and disadvantages compared to the original FDR. The same paper also introduced the concept of "q-value", which represents an "FDR-version" of the p-value. The traditional p-value indicates the potential type I error rate associated with rejecting a hypothesis. Adjusted p-values corresponding to various FWER controlling procedures can also be defined. For example, the Bonferroni adjusted p-value is the original p-value times the number of hypotheses m. Analogous to the p-values, the adjusted p-values indicate the potential family-wise error rate associated with rejecting a hypothesis. Using this same idea, the q-value indicates the potential positive false discovery rate associated with rejecting the hypothesis. A Bayesian approach is adopted for much of the discussions and interpretations of the pFDR and the q-value. Storey (2002) also investigated fixing the rejection region and estimating the false discovery rate and also the positive false discovery rate.

Benjamini and Hochberg (2000) proposed an adaptive version of the BH procedure. The original BH procedure was actually proved to control the FDR at level $m_0\alpha/m$ under independent test statistics. Thus when m_0 is small, the BH procedure is in fact conservative. Benjamini and Hochberg (2000) proposed incorporating the estimate m_0 into the procedure to address this issue. Storey and Tibshirani (2003) proposed a smoother estimate for m_0 for the adaptive BH method, and further promoted the use of q-values, which is now widely adopted in analysis of genomic data.

1.2 Multi-Dimensional Multiple Testing Problems

The concept of false discovery rate and the BH procedure (as well as some of their extended and adaptive versions) are now widely accepted and adopted in large scale multiple testing problems. In areas such as analysis of genomic data, it has become standard practice to apply some form of the BH procedure after obtaining the p-values and use the FDR as a criteria for selecting significant genes. Under the framework of false discovery rates, the problem of large scale multiple testing for a single dimension of hypotheses has been pretty well solved.

However, in some more complicated problem settings, we may want to test a "multi-dimensional set of hypotheses" that would not make sense if arranged into the form of a single vector of interchangeable hypotheses. The idea of a multi-dimensional set of hypotheses might sound perplexing at first, but it is actually quite common, such as in many popular research areas in bioinformatics. Consider the following examples: (1) in the case of meta-analysis of microarray studies (or any other type of large scale study), we would have a hypotheses for each gene and each study; (2) in a time-course gene expression study we would have a hypotheses for each gene and each time point; (3) in an fMRI study of the brain where subjects participate in a series of cognitive tasks, we may have a hypotheses associated with each brain location and each task; (4) in a microarray experiment (or other studies with a large number of features tests) where many treatments are compared against a single control, we would have a hypotheses for each gene and each treatment. In all these cases, the set of hypotheses considered are naturally arranged in two dimensions - one dimension would be the gene/brain location or other type of feature being tested, and the other dimension would be the studies/time points/tasks/treatments. It is not sensible to treat all the hypotheses in these two-dimension matrices as simply a long vector of hypotheses - it is technically doable, but we would be losing important information on how the hypotheses relate to each other. We could make independent inferences on each of these hypotheses individually, but a lot of times we are ultimately interested in making inferences about the features (genes, brain locations, etc.), and thus it is important to preserve the structure of the original multi-dimensional set up of the hypotheses.

The set up of multi-dimensional hypotheses brings on new challenges to the area of multiple testing. In the two-dimensional case, one approach would be to consolidate the results from one of the dimensions, so that we end up with a one-dimensional problem. This approach seems more applicable to the case of meta-analysis, where it makes sense to form a single hypotheses for each gene by consolidating results from multiple studies. In other cases, another approach would be to explore new forms of the FDR and corresponding procedures that are adapted to the multi-dimensional structure. Whatever the approach, previous results on the FDR and FDR controlling procedures cannot be directly applied to the multi-dimensional case.

Many literature in recent years address this problem through more concrete cases - although many do not explicitly mention the multi-dimensional nature of the set up. Pyne *et al.* (2006) considered meta-analysis based on the control of the false discovery rate. Benjamini and Heller (2008) considered the problem of testing for partial conjunction of hypotheses with control of an overall false discovery rate, which can be applied to many of the setups of two-dimensional hypotheses mentioned earlier. Natarajan (2012) considered the problem of intersecting independent lists of genes, such as top gene lists from multiple studies, and investigated the control of FDR for the intersections lists. Guo *et al.* (2010) as well as Sun and Wei (2011) both considered the problem of multiple testing for time-course microarray experiments. Benjamini *et al.* (2009) and Heller and Yekutieli (2012) considered the false discovery rate approach to replicability analysis for multiple genome-wide association studies.

We considered a few problems related to multi-dimensional multiple testing. Most of the problem settings and applications are on analysis of genomic data, in particular microarray data, although much of the methodology can be applied to any similar large scale data analysis. In Chapter 2, we consider meta-analysis of genomic data while incorporating information on the heterogeneity among studies by adopting a weighted p-value approach, while maintaining control of the false discovery rate for the genes. In Chapter 3, we talk about the situation where the hypotheses can be arranged into a hierarchical structure. We adopt the concept of "overall false discovery rate" from Heller *et al.* (2009), as well as the structure of their procedure that controls the overall false discovery rate, and develop a general framework for hierarchical hypotheses set testing. We apply this framework to several practical problems, including time-course microarray experiments, multiple comparisons of treatment means for gene expression studies, and selecting important genes in tumor classification.

1.3 Meta-Analysis of Genomic Data

As mentioned earlier, meta-analysis of genomic data can be considered a special case of multi-dimensional multiple testing problems. However, meta-analysis in general is an interesting area of statistical methods in itself. Classical methods such as Fisher's combined probability test (Fisher, 1925) and Stouffer's Z-test (Stouffer *et al.*, 1949) are widely used. Many weighted and generalized versions of these classical p-value combining methods have also been developed throughout the years, such as Mosteller and Bush (1954), Lipták (1958), Lancaster (1961), among others. Other than p-value combining methods, another class of methods integrate studies by combining effect sizes, such as Rhodes *et al.* (2002), Choi *et al.* (2003) and others.

As meta-analysis becomes an increasingly popular tool, new challenges arise, especially when the goal of meta-analysis starts diversifying. Many of the traditional meta-analysis methods aim at gaining power in detecting any signal among the studies by pooling them. Nowadays, it is very common to have the goal of detecting consistent signals across studies instead. This is especially true with meta-analysis of genomic studies. Due to technical variations in high-throughput technologies, it is common to observe heterogeneity across studies. In this case, it is of much more importance to find genes that show consistent signals across studies, as opposed to genes that are found significant only due to a single study. In recent years many new methods have been developed for meta-analysis of genomic data. Some directly addressed the issue of heterogeneity across studies, such as Choi *et al.* (2003), Shabalin *et al.* (2008) and Scharpf *et al.* who modeled the between-study variation, or Lai *et al.* (2007) and Lu *et al.* (2010) who proposed measures of concordance across studies. Others developed methods for the goal of detecting consistent signals across studies, such as a non-parametric approach based on rankings by Hong *et al.* (2006), a framework for testing partial conjunction hypothesis by Benjamini and Heller (2008), and most recently a method based on ordered p-values by Song and Tseng (2013).

In this thesis, both Chapter 2 and Chapter 4 consider the problem of meta-analysis. The method in Chapter 2 focuses on incorporating the information of heterogeneity in the form of weights to influence the significance of genes when conducting meta-analysis. The method takes into account both the issue of heterogeneity in meta-analysis and the aspect of multiple testing at the same time. On the other hand, Chapter 4 focuses on a more basic meta-analysis problem of detecting consistent signals across studies, which does not have to be related to a multiple testing setting. The meta-analysis method proposed in Chapter 4 is a weighted p-value combination method in which the weights are based on the order of the p-values across the studies. Even though we illustrated the method in the analysis of genomic data, the method based on weighted ordered p-values can be applied to any meta-analysis setting where the goal is to look for signals in the majority of studies.

Chapter 2

A Weighted Multiple Testing Procedure for Meta-Analysis

2.1 Introduction

In recent years, with the extensive usage of microarray and other high-throughput technologies in biomedical research, there has been a rapid growth in the amount of publicly available datasets. A few widely used public internet repositories are the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), the EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and the Stanford Microarray Database (SMD, http://smd.stanford.edu/). These resources allow researchers to further exploit the information in these data, especially in the form of meta-analysis. How to effectively integrate information of microarray datasets from multiple studies is becoming an increasingly important problem. The other area of genomics that has increasingly relied on the use of meta-analysis has been genomewide association studies (GWAS); some seminal studies in this field are Scott et al. (2007) and Willer et al. (2008).

Returning to the microarray example, the most common type of analysis is the detection of differentially expressed genes, especially for the case of two groups of samples, namely treatment and control. Combining information from multiple datasets is expected to increase the power for differential expression analysis. Many methods for meta-analysis addressing this type of problem have been proposed and reviewed in recent years. An incomplete list includes the p-value combining approach of Rhodes et al. (2002), the GeneMeta method by Choi *et al.* (2003), RankProd method by Hong *et al.* (2006), the metaArray approach of Choi *et al.* (2007), a hierarchical model put forth by Scharpf et al. (2009), and the mDEDS algorithm of Campaign and Yang (2010). In this latter paper, a comparison between several meta-analysis approaches was considered as well.

One important issue that has received limited discussion in most of this literature is the assessment of the concordance of data among different studies and the integration of this information into further analysis. This problem is very well-understood in the classical meta-analysis problem and tests for between-study heterogeneity have been accordingly developed; see Normand (1999) for a discussion. However, it becomes conceptually problematic to extend this approach directly to the genomic meta-analysis problem, as one has to perform G tests of heterogeneity and then determine based on the result of the tests whether or not meta-analysis is feasible for every single gene. It will be the case that there will be some genes that will show significant evidence for between-study heterogeneity. Thus, the meta-analysis will be done on a subset of genes, contrary to the approaches described in the previous paragraph. In addition, there are issues of "pre-testing" and model selection that arise that complicate the analysis of interpretation of such a meta-analysis.

The evidence for between-study heterogeneity in these high-throughput genomic data settings is growing. One potential cause is errors in mapping the proper gene to the microarray annotation (Dai et al., 2005). This can be viewed as a technical mistake that leads to between study variation. A more biologically oriented cause is that samples being profiled actually represent subtypes of different disease groups. This has been mostly famously explored in breast cancer (Sorlie et al., 2001), prostate cancer (Tomlins et al., 2005) and leukemia (Vardiman et al., 2002). Thus, the presence of subtypes of samples leads to the between-study heterogeneity. A final reason is the fact that due to the diversity of labs generating these data, there will inherently exist between-lab variation and more generally, batch effects. A recent review by Leek et al. (2010) demonstrates the existence of batch effects in a variety of high-throughput genomic datasets.

Two questions then naturally arise from these findings. First, how does one assess between-study variation using high-throughput genomic data? Second, how can one incorporate the between-study variation into the analysis. Measures of reproducibility have been proposed by Parmigiani et al. (2004), Lee et al. (2004) and Li et al. (2011). Implicitly, once these measures are calculated, one would then calculate a summary measure on genes that were "sufficiently" reproducible. Another approach would be to model the between-study variation; this has been done by Shabalin et al. (2008) and Scharpf et al. (2009).

Lai *et al.* (2007) developed a framework for integrating two studies for differential expression analysis that entails assessment of global concordance. By contrast, the approach developed in this article will utilize measures of gene-wise or local concordance. Other methods have been proposed to assess the global concordance of studies, such as the concordance correlation coefficient (CCC) by Miron *et al.* (2006). Lu *et al.* (2010) proposed a multi-class correlation measure to seek for genes of concordant inter-class patterns across studies. The between-study variation can also be modelled as variance components or other parameters in the joint modelling frameworks of GeneMeta (Choi et al., 2003) and Scharpf et al. (2009). These algorithms will tend to be more computationally intensive than what is proposed in this article.

In this paper, we focus on meta-analysis of microarray datasets for differential expression analysis. We take the viewpoint that the major goal in these studies is selection and prioritization of genes for further validation studies. We exploit the weighted hypothesis testing framework proposed by Genovese et al. (2006) and propose an "assumption weighting" methodology. To be specific, we will use weights that assess between-study heterogeneity and incorporate them into a multiple testing procedure. We also note in passing that the use of weights for characterizing the assumptions needed for meta-analysis is a novel application of this weighted hypothesis testing framework. We proposed four different weighting schemes, including two based on an assessment of the concordance of the test statistics of different studies – the I^2 index (Higgins and Thompson, 2002), and two based on a newly proposed measure of expression correlation between studies. The performance of these methods are assessed through simulation studies as well as applications to a set of stem cell studies.

2.2 The Weighted Testing Procedure for Incorporating Heterogeneity into Meta-Analysis

2.2.1 Weighted hypothesis testing framework

Before describing the proposed methodology, we briefly review the weighted hypothesis testing framework of Genovese *et al.* (2006) for multiple hypothesis testing. It is summarized in Box 1. The method incorporates prior information about the hypothesis in the form of p-value weights for each hypothesis, while maintaining control of the false discovery rate for multiple hypothesis testing. The procedure is as follows: first assign weights $W_i > 0$ to the ith null hypothesis ($1 \leq i \leq n$) such that $\bar{W} = 1$; then compute $q_i = p_i/W_i$ for each i, where p_i is the un-weighted p-value for testing the ith hypothesis; and last, if it is desired to control the false discovery rate at a specific α, apply the Benjamini and Hochberg (1995) procedure at level α to the q_i's. In practice, if we simply wish to pick out the top most significant genes, we can directly use the ordering of the q_i's and omit the B-H procedure in the last step. Genovese et al. showed that if the assignment of weights is positively associated with the null hypothesis being false, then the procedure improves power, and that even if the assignment of weights is poor, power is only reduced slightly.

Box 1. Weighted B-H procedure of Genovese *et al.* (2006)

(1) Let $p_1, \ldots, p_n$ denote the n p-values with associated weights $W_1, \ldots, W_n$, such that $\bar{W} = 1$.

(2) Define $q_i = p_i/W_i$, and let $q_{(1)} \leq q_{(2)} \leq \cdots \leq q_{(n)}$ denote the ordered values.

(4) Find $\hat{k} = \max\{1 \leq i \leq n : q_{(i)} \leq i\alpha/n\}$.

(5) If $\hat{k}$ exists, then reject null hypotheses corresponding to $q_{(1)} \leq \cdots \leq q_{(\hat{k})}$. If the set in (4) is empty, reject nothing.

2.2.2 Proposed methodology

We consider S studies to be combined. Within each study, there are two groups of samples (control and treatment), and the goal is to find genes that are differentially expressed between the two groups. Let p_{gs} be the p-value for the two sample t-test between treatment and control for gene g $(1 \leq g \leq G)$, study s $(1 \leq s \leq S)$. Here, as well as in following steps, we used the basic two sample t-test to obtain the t-statistic and p-values. However, this is can be straight-forwardly generalized to other variations of the t-test. Some tests developed in recent years might be more preferable in the analysis of gene expression data, such as the moderated t-statistic by Smyth *et al.* (2003).

By Fisher's combined probability test, we can combine the p-values from the S studies to obtain a single test statistic for each gene: $X_g^2 = -2\sum_s \log{(p_{gs})}$. Under the null hypothesis that the gene is not differentially expressed in any of the studies, and that the studies are independent, X_g^2 has a chi-squared distribution with $2S$ degrees of freedom. Thus for each gene we can obtain a Fisher's p-value, $p_g = P(\chi^2_{2S} > X_g^2)$, for testing differential expression between the two groups combining all the studies. However, Fisher's combined test does not specifically account for heterogeneity among the studies.Thus a significant Fisher's p-value does not necessarily reflect a consensus result. A small Fisher's p-value could be driven by a single study with an extremely small p-value; or even if many of the studies had small p-values, it could be the case that differential expression are of different directions among the studies, thus making it hard to interpret the combined test.

To account for this problem, we will adjust the Fisher's p-value by using the methodology summarized in Box 1. In our case, the prior information is the degree of heterogeneity among the studies. Let U be a measure of heterogeneity among the studies, such that large values of U represent less heterogeneity. Denote by U_g the heterogeneity measure for gene g. Define weights $W_g = U_g/\bar{U_g}$, where $\bar{U_g} = G^{-1}\sum_g U_g$, so that the weights have mean 1. Then the p-value weighting method gives us the adjusted p-value for each gene $q_g = p_g/W_g$. By this construction, genes with homogeneous expression patterns among the studies are assigned larger weights, resulting in smaller adjusted p-values, which are more likely to be found significant. By contrast, genes displaying heterogeneity are down-weighted, resulting in larger adjusted p-values and are less likely to be found significant. We shall explore and compare several different measures of heterogeneity to incorporate the weighted multiple testing procedures.

2.2.3 Using I^2 as weight

2.2.3.1 I^2: definition and weighting scheme

One approach is to assess the concordance of the test statistics of the different studies. A common measure of this kind used in meta-analysis is the Q-statistic (Cochran, 1954). To adjust for the dependency of the Q-statistic on the number of studies, an I^2 index (Higgins and Thompson, 2002) was proposed based on the Q-statistic. We shall adopt the I^2 index as one method of weighting.

Let T_{gs} be the t-statistic for the two sample t-test between control and treatment, for gene g $(1 \leq g \leq G)$, study s $(1 \leq s \leq S)$. The Q-statistic for gene g is defined as $Q_g = \sum_s (T_{gs} - \bar{T}_g)^2$, where $\bar{T}_g = S^{-1} \sum_s T_{gs}$. The I^2 index compares the Q-statistic with its expected value assuming homogeneity (which is $S - 1$). The I^2 index for gene g is

$$I_g^2 = \frac{Q_g - (S-1)}{Q_g}$$

if $Q_g > S - 1$ and defined to be 0 if $Q_g \leq S - 1$. Therefore a large I_g^2 corresponds to higher heterogeneity for that gene, in which case we wish to down-weight.

We define $U_{1g} = (I_g^2 + 10^{-5})^{-1}$ and the corresponding weights $W_{1g} \equiv U_{1g}/\bar{U}_{1g}$ so that higher heterogeneity corresponds to lower weights. Adding 10^{-5} in the denominator is just for technical purposes so that the weights are well defined when $I^2 = 0$.

2.2.3.2 Accounting for direction of change

The I^2 index measures the heterogeneity of the test statistics for different studies, but it does not explicitly take into account the direction of the change, i.e. in our case the sign of the t-statistic. Define $C_g = min\{\sum_s I_{\{T_{gs}>0\}}, \sum_s I_{\{T_{gs}<0\}}\}/S$ for gene g. Thus $0 \leq C_g \leq 0.5$. Small values of C_g indicate that most studies display a change of the same direction between treatment and control, while large values of C_g indicate a relatively stronger disagreement on the direction of change among the studies for that particular gene. Define $U_{2g} = W_{1g}/(C_g + 0.01)$ and the corresponding weights $W_{2g} \equiv U_{2g}/\bar{U}_{2g}$. Genes that show discordant directions of change would be further down-weighted in adjusting the p-values. Adding 0.01 in the formula of U_{2g} is to avoid the denominator being 0.

2.2.4 Using correlation as weight

2.2.4.1 Correlation measurement between a pair of studies

Another intuitive way of quantifying heterogeneity is to assess the expression correlation of a gene between studies. We start with assessing the correlation of expression values between two studies for a given gene. For simplicity of notation, we drop the subscripts g and s, and let x_{ki} $(1 \leq i \leq n_k)$ and y_{ki} $(1 \leq i \leq m_k)$ be the expression value of gene g for class k $(k = 0, 1)$, sample i of the first study and the second study respectively. For the case $n_k = m_k$ $(k = 0, 1)$, a naive way of assessing the expression correlation between the two studies is to directly take the sample correlation of $(x_{01}, \cdots, x_{0n_0}, x_{11}, \cdots, x_{1n_1})$ and $(y_{01}, \cdots, y_{0m_0}, y_{11}, \cdots, y_{1m_1})$. However, this approach is not adaptable to cases where $n_k \neq m_k$, and it also ignores the exchangeability of samples within a group.

We shall introduce a new way of computing the correlation addressing these two issues. In order for the new method to still fit into the general framework of sample correlation, we need to construct paired samples from the two studies. The construction of paired samples are based on the following two considerations: (1) Since we are interested in the comparison between treatment and control, treatment samples from the two studies should be paired together, and control samples from the two studies should be paired together, while it doesn't make sense to pair a treatment sample from one study with a control sample from another study. That is to say sample pairs should be created between $(x_{k1}, \cdots, x_{kn_k})$ and $(y_{k1}, \cdots, y_{km_k})$ for the same k. (2) The ordering of samples within a group should not matter. That is to say, if we have a permutation of the samples $(x_{k1}, \cdots, x_{kn_k})$, the results should remain the same. Based on these two considerations, we create paired samples by taking all possible pairs between $(x_{01}, \cdots, x_{0n_0})$ and $(y_{01}, \cdots, y_{0m_0})$, and all possible pairs between $(x_{11}, \cdots, x_{1n_1})$ and $(y_{11}, \cdots, y_{1m_1})$. Thus the resulting sample vectors of the two studies are in the form of:

$$\begin{array}{cccccccc} (\mathbf{x_{01}}, & \mathbf{x_{02}}, & \cdots & \mathbf{x_{0n_0}}, & \mathbf{x_{11}}, & \mathbf{x_{12}}, & \cdots & \mathbf{x_{1n_1}}) \\ (\mathbf{y_0}, & \mathbf{y_0}, & \cdots & \mathbf{y_0}, & \mathbf{y_1}, & \mathbf{y_1}, & \cdots & \mathbf{y_1}) \end{array}$$

where $\mathbf{x_{ki}} = (x_{ki}, \cdots, x_{ki})$ is a vector of length m_k, and $\mathbf{y_k} = (y_{k1}, \cdots, y_{km_k})$. Both resulting sample vectors are of length $n_0m_0 + n_1m_1$. We shall use the sample correlation of these two sample vectors, denoted by ρ, as our measurement of the expression correlation between the two studies. Intuitively, differential expression of the same direction in both studies will lead to a positive ρ. Similarly, differential expression in opposite directions in the two studies will lead to a significantly negative ρ; no differential expression in one or both studies will lead to an insignificant ρ.

Thus we may calculate a correlation measurement $\rho_{gss'}$ for each gene g and each pair of studies s and s' (there are $S(S-1)/2$ pairs of studies in total). We use these ρ's to construct another measurement of heterogeneity U_{3g} for each gene. Let U_{3g} be the absolute value of the mean $\rho_{gss'}$ for all pairs of studies raised to the power of 10, that is

$$U_{3g} = \left| \frac{2}{S(S-1)} \sum_{1 \leq s < s' \leq S} \rho_{gss'} \right|^{10} .$$

Defining U_{3g} in this way, we will expect large values of U_{3g} if most of the studies display differential expression in the same direction; we will expect small values of U_{3g} if differential expression only occur in a small fraction of the studies, in which case most of the $\rho_{gss'}$ are small, or if the studies are not consistent in the direction of differential expression, in which case the positive ρ's and negative ρ's will be washed out when we take the mean. Thus large values of U_{3g} correspond to homogeneous patterns of differential expression among the studies. Raising the mean correlation to the power of 10 is used to amplify the effects of weighting, as we found that using versions of U_{3g} without raising the power did not lead to change in the weights relative to an unweighted scheme (data not shown). As before, let $W_{3g} = U_{3g}/\bar{U}_{3g}$ be the weights, and obtain the adjusted p-values by dividing the original Fisher's p-values by the weights.

2.2.4.2 Correlation measurement among S studies

Previously, we assessed the correlation of expression levels among S studies by breaking down the studies into pairs. However, we may adopt the idea we used for deriving correlation for a pair of studies to directly construct a measurement of correlation among all the studies.

Similar to before, we shall construct paired samples. Again, it only makes sense to pair up samples coming from the same class (i.e., both samples belong to control group or both samples belong to treatment group), but now the two samples may come from any two studies among all the studies. Thus we pool all the control samples from the S studies and take all possible pairs of samples from this pool, and do the same thing for the treatment samples.

Suppose for study s there are n_{s0} replicates for the control group and n_{s1} replicates for the treatment group. Then there are a total of $N_0 = \sum_s n_{s0}$ control samples and $N_1 = \sum_s n_{s1}$ treatment samples from all the studies. Denote by x_{0i} $(1 \leq i \leq N_0)$ the expression values for the control samples from all the studies, and x_{1i} $(1 \leq i \leq N_1)$ the expression values for the treatment samples from all the studies. Create two sample vectors as follows:

$$\begin{array}{llllllll}(\mathbf{x_{01}}, & \mathbf{x_{02}}, & \cdots & \mathbf{x_{0N_0}}, & \mathbf{x_{11}}, & \mathbf{x_{12}}, & \cdots & \mathbf{x_{1N_1}}) \\ (\mathbf{x^*_{01}}, & \mathbf{x^*_{02}}, & \cdots & \mathbf{x^*_{0N_0}}, & \mathbf{x^*_{11}}, & \mathbf{x^*_{12}}, & \cdots & \mathbf{x^*_{1N_1}})\end{array}$$

where $\mathbf{x_{ki}} = (x_{ki}, \cdots, x_{ki})$ is a vector of length $N_k - 1$, and $\mathbf{x^*_{ki}} = (x_{k1}, \cdots, x_{k(i-1)}, x_{k(i+1)}, \cdots, x_{kN_k})$ is a vector of all the samples from group k excluding x_{ki}. Both resulting sample vectors are of length $N_0(N_0-1)+N_1(N_1-1)$. Let λ be the sample correlation between these two sample vectors. We shall use λ as a measurement of expression correlation among all studies. It can be seen from the construction of λ that consistent differential expression will lead to large values of λ, while inconsistent differential expression or non-differential expression will lead to small values of λ.

Denote by λ_g the λ value computed for each gene. Let $U_{4g} = |\lambda_g|^{10}$ be a fourth measure of heterogeneity among studies, with large values of U_{4g} corresponding to less heterogeneity, and construct weights $W_{4g} = U_{4g}/\bar{U}_{4g}$. The correlation is raised to the power of 10 to be consistent with the mean correlation case.

2.3 Some Theoretical Justification

Genovese *et al.* (2006) proved that their weighted BH procedure as described in Box 1 controls the FDR at the desired level under certain general assumptions of the distributions of the p-values and the weights. They investigated both the finite sample case and asymptotic results. One of the main assumptions for the weights is that the weights are independent of the p-values given the hypotheses. It is questionable whether this assumption holds true in our case, since the four weighting schemes we proposed are all based on empirical weights that are computed from the same data that the p-values are computed from. In this section, we provide some theoretical justification for the weighted BH procedure for weights that may not be independent of the p-values. We focus on the asymptotic behavior of the procedure's control of the FDR. The assumptions and the steps of proof follow closely that of Genovese *et al.* (2006), with necessary adjustments to some of the assumptions.

Define hypotheses indicator variables $(H_1, ..., H_m)$, where $H_i = 0$ (or $= 1$) if the ith null hypothesis is true (or false), and assume $H_i \sim \text{Ber}(a)$. Let $(P_1, ..., P_m)$ and $(W_1, ..., W_m)$ denote the corresponding p-values and weights. An implicit assumption here is that the weights are positive. Since the p-values and the weights are not necessarily independent, we model the joint distribution of (P, W). Assume that the (P_i, W_i)'s are identically distributed, and their distribution given H_i is

$$(P_i, W_i) \mid H_i = 0 \sim M^0(p, w), \quad (P_i, W_i) \mid H_i = 1 \sim M^1(p, w).$$

With slight abuse of notation, let $M^0(w)$ and $M^1(w)$ denote the marginal distribution of W_i given $H_i = 0$ and $H_i = 1$ respectively. That is, assume

$$W_i \mid H_i = 0 \sim M^0(w), \quad W_i \mid H_i = 1 \sim M^1(w).$$

Let $M^0_w(p)$ and $M^1_w(p)$ denote the distributions of P_i given $W_i = w$ and $H_i = 0$ or $H_i = 1$ respectively. That is, assume

$$P_i \mid W_i = w, H_i = 0 \sim M^0_w(p), \quad P_i \mid W_i = w, H_i = 1 \sim M^1_w(p).$$

Following Genovese *et al.* (2006), assume that the marginal distribution of P_i given H_i is

$$P_i \mid H_i = 0 \sim U \text{ and } P_i \mid H_i = 1 \sim F,$$

where U denotes the Uniform(0,1) distribution and F is a probability distribution on $(0, 1)$ that is stochastically smaller than the Uniform. Thus the marginal distribution of P_i is $G = (1 - a)U + aF$. Genovese *et al.* (2006) assumed that P_i is independent of W_i given H_i. By modeling the joint distribution of (P_i, W_i), we do not have this assumption in general. However, we do assume that the conditional independence holds under the null. In other words, we assume that P_i and W_i are independent when $H_i = 0$, but we do not assume independence of P_i and W_i when $H_i = 1$. Under these assumption, we have

$$M_w^0(p) = U \text{ for } \forall\, w \ \text{ and } \int M_w^1(p) dM^1(w) = F.$$

As discussed in Genovese *et al.* (2006), the original BH procedure is essentially applying a threshold T_{BH} to the p-values, such that

$$T_{BH} = \sup\{t : t/\hat{G}_m(t) \leq \alpha\},$$

where $\hat{G}_m(t)$ is the empirical distribution function of the p-values. In other words, the BH procedure is based upon the distribution of the p-values $G(t) = \text{pr}(P_i \leq t)$. Similarly, the weighted BH procedure is essentially based upon the distribution of the weighted p-values, which we denote as $D(t) = \text{pr}(P_i/W_i \leq t)$. The weighted BH procedure is essentially applying the threshold T_{wBH} to the weighted p-values, where

$$T_{wBH} = \sup\{t : t/\hat{D}_m(t) \leq \alpha\}.$$

Now we shall discuss the asymptotic behavior of the procedure. Following Genovese *et al.* (2006), let $C(t) = D(t)/t$ and $\hat{C}_m(t) = \hat{D}_m(t)/t$. Also redefine the asymptotic version and empirical version of the threshold T_{wBH} as

$$t_* = \sup\{t : C(t) \geq \frac{1}{\alpha}\} \ \text{ and } \ T_m = \sup\{t : \hat{C}_m(t) \geq \frac{1}{\alpha}\}.$$

Similar to Lemma 1 in Genovese *et al.* (2006), we have the following result.

Lemma 2.1. *If $M_w^0(p) = U$ and $M_w^1(p)$ is strictly concave on $[0,1]$ for $\forall$ w, then (i) G is strictly concave on $[0,1]$; (ii) D is strictly concave on $[0,1]$; and (iii) C is monotone decreasing on $(0,1)$.*

Proof. Since $G = (1-a)U + aF$, and $F = \int M_w^1(p)dM^1(w)$, (i) follows immediately.

For $D(t)$ we have

$$\begin{aligned} D(t) &= (1-a)\cdot \mathrm{pr}(P/W \leq t \mid H=0) + a \cdot \mathrm{pr}(P/W \leq t \mid H=1) \\ &= (1-a)\int \mathrm{pr}(P \leq wt \mid W=w, H=0)dM^0(w) \\ &\qquad + a\int \mathrm{pr}(P \leq wt \mid W=w, H=1)dM^1(w) \\ &= (1-a)\int M_w^0(wt)dM^0(w) + a\int M_w^1(wt)dM^1(w) \\ &= (1-a)\int wt\; dM^0(w) + a\int M_w^1(wt)dM^1(w) \\ &= (1-a)\mu_0 t + a\int M_w^1(wt)dM^1(w), \end{aligned}$$

where $\mu_0 = E(W \mid H=0)$. Hence, for $0 \leq \lambda \leq 1$,

$$\begin{aligned} D\{(1-\lambda)t_0 + \lambda t_1\} &= (1-a)\mu_0\{(1-\lambda)t_0 + \lambda t_1\} + a\int M_w^1\big(w\{(1-\lambda)t_0 + \lambda t_1\}\big)dM^1(w) \\ &> (1-\lambda)D(t_0) + \lambda D(t_1). \end{aligned}$$

This proves (ii).

Finally, notice that $D(0) = 0$, and by the concavity of $D(t)$, for $1 > t_1 > t_0 > 0$,

$$C(t_1) = \frac{D(t_1)}{t_1} = \frac{(1 - t_0/t_1)D(0) + (t_0/t_1)D(t_1)}{t_0} \leq \frac{D(t_0)}{t_0} = C(t_0),$$

which proves (iii). □

Remember that in defining the weights, the W_i's are constrained to have mean 1. So although the W_i's are assumed to be identically distributed, they are not independent. However, we can always write $W_i = U_i/\bar{U}_m$ for independent and identically distributed positive random variables $U_1, ..., U_m$. In fact, this is how the weights are usually defined in practice (such as in our four weighting schemes).

Define $\tilde{W}_i = U_i/EU_1$. Then the $\tilde{W}_i$'s defined this way are independent and identically distributed. Let $\tilde{D}(t) = \mathrm{pr}(P_i/\tilde{W}_i \leq t)$ be the distribution function of $P_i/\tilde{W}_i$ and $\tilde{D}_m(t)$ be its empirical distribution function. Also remember that $\hat{D}_m(t)$ is the empirical distribution function corresponding to $D(t) = \mathrm{pr}(P_i/W_i \leq t)$. We have the following results.

Lemma 2.2. $\sup_t |\hat{D}_m(t) - \tilde{D}_m(t)| \to 0$ *and* $\sup_t |\hat{D}_m(t) - \tilde{D}(t)| \to 0$ *almost surely.*

Proof. By the Glivenko-Cantelli theorem, since $P_i/\tilde{W}_i$ are independent and identically distributed, we have $\sup_t |\tilde{D}_m(t) - \tilde{D}(t)| \to 0$. By the strong law of large numbers, since the U_i's are independent and identically distributed, we have $|\bar{U}_m - EU_1| \to 0$ almost surely. It follows that, for any fixed $\varepsilon > 0$,

$$\begin{aligned}\hat{D}_m(t) = \frac{1}{m}\sum_i I\Big\{\frac{P_i}{W_i} \leq t\Big\} &= \frac{1}{m}\sum_i I\Big\{\frac{P_i}{\tilde{W}_i} \leq t\frac{EU_1}{EU_1 + (\bar{U}_m - EU_1)}\Big\}\\ &\leq \frac{1}{m}\sum_i I\Big\{\frac{P_i}{\tilde{W}_i} \leq t\frac{EU_1}{EU_1 - \varepsilon}\Big\} \leq \mathrm{pr}\Big(\frac{P_i}{\tilde{W}_i} \leq t\frac{EU_1}{EU_1 - \varepsilon}\Big) + \varepsilon,\end{aligned}$$

for large enough m, uniformly in t. Similarly, for large enough m and uniformly in t we have

$$\hat{D}_m(t) \geq \frac{1}{m}\sum_i I\Big\{\frac{P_i}{\tilde{W}_i} \leq t\frac{EU_1}{EU_1 + \varepsilon}\Big\} \geq \mathrm{pr}\Big(\frac{P_i}{\tilde{W}_i} \leq t\frac{EU_1}{EU_1 + \varepsilon}\Big) - \varepsilon.$$

Combining the two inequalities, we have

$$\tilde{D}\Big(t\frac{EU_1}{EU_1 + \varepsilon}\Big) - \varepsilon \leq \tilde{D}_m\Big(t\frac{EU_1}{EU_1 + \varepsilon}\Big) \leq \hat{D}_m(t) \leq \tilde{D}_m\Big(t\frac{EU_1}{EU_1 - \varepsilon}\Big) \leq \tilde{D}\Big(t\frac{EU_1}{EU_1 - \varepsilon}\Big) + \varepsilon,$$

for large enough m, uniformly in t. The claim then follows from the arbitrariness of ε and the continuity of $\tilde{D}(t)$. □

Lemma 2.3. $|D(t) - \tilde{D}(t)| \to 0$ *for any* t.

Proof. Notice that $D(t)$ depends on m implicitly (because $W_i = U_i/\bar{U}_m$ depends on m). For given t, and any $\varepsilon > 0$, we can find δ such that

$$\begin{aligned} D(t) = \mathrm{pr}\Big(\frac{P_i}{W_i} \le t\Big) &= \mathrm{pr}\Big(\frac{P_i}{\tilde{W}_i} \le t\frac{EU_1}{EU_1 + (\bar{U}_m - EU_1)}\Big) \\ &= \tilde{D}\Big(t\frac{EU_1}{EU_1 + (\bar{U}_m - EU_1)}\Big) \le \tilde{D}\Big(t\frac{EU_1}{EU_1 - \delta}\Big) \le \tilde{D}(t) + \varepsilon, \end{aligned}$$

for large enough m. Similarly, we can show that $D(t) > \tilde{D}(t) - \varepsilon$. By arbitrariness of ε, we have $|D(t) - \tilde{D}(t)| \to 0$ for any t. □

Let FDP(T) denote the false discovery proportion corresponding to threshold T. The false discovery rate for threshold T is then $E(\mathrm{FDP}(T))$. Using results from Lemma 2.1, 2.2 and 2.3, we have the following Theorem.

Theorem 2.1. *Under the assumptions of Lemma 2.1, $|T_m - t_*| \to 0$ almost surely and $E(FDP(T_m)) \le \alpha + o(1)$.*

Proof. Fix $b > 0$. By the definition of t_* and the fact that $C(t)$ is decreasing (from Lemma 2.1), we have $C(t_* + b) < 1/\alpha$. Fix $\varepsilon > 0$ such that $C(t_* + b) + \varepsilon/(t_* + b) < 1/\alpha$. Then for any given $t > t_* + b$, we have

$$\begin{aligned} \hat{C}_m(t) = \frac{\hat{D}_m(t)}{t} &\le \frac{D(t) + |\hat{D}_m(t) - \tilde{D}(t)| + |\tilde{D}(t) - D(t)|}{t} \le \frac{D(t) + \varepsilon}{t} \\ &= C(t) + \frac{\varepsilon}{t} \le C(t_* + b) + \frac{\varepsilon}{t_* + b} < \frac{1}{\alpha}, \end{aligned}$$

almost surely for large enough m. By the definition of T_m and the arbitrariness of t, we have $T_m < t_* + b$. Similarly, we can show that $T_m > t_* - b$. Hence we have $|T_m - t_*| \le b$ almost surely for all large m. The first claim then follows from the arbitrariness of b.

Now,

$$\mathrm{FDP}(T_m) = \frac{m^{-1}\sum_i (1 - H_i) I\{P_i/W_i \le T_m\}}{m^{-1}\sum_i I\{P_i/W_i \le T_m\}} = \frac{\hat{V}_m(T_m)}{\hat{D}_m(T_m)}.$$

We have

$$|\hat{D}_m(T_m) - D(t_*)| \le |\hat{D}_m(T_m) - \tilde{D}(T_m)| + |\tilde{D}(T_m) - \tilde{D}(t_*)| + |\tilde{D}(t_*) - D(t_*)| \to 0$$

almost surely, by Lemma 2.2, Lemma 2.3 and the continuity of $\tilde{D}(t)$. Similarly, we have $|\hat{V}_m(T_m) - V(t_*)| \to 0$ almost surely, where $V(t) = E((1-H_i)I\{P_i/W_i \leq t\})$. This leads to $|\mathrm{FDP}(T_m) - V(t_*)/D(t_*)| \to 0$ almost surely, and dominated convergence yields $|E(\mathrm{FDP}(T_m)) - V(t_*)/D(t_*)| \to 0$. Now,

$$\begin{aligned} V(t_*) &= E((1-H_i)I\{P_i/W_i \leq t_*\}) = E(I\{P_i/W_i \leq t_*\} \mid H_i = 0)\mathrm{pr}(H_i = 0) \\ &= \mathrm{pr}(I\{P_i/W_i \leq t_*\} \mid H_i = 0)\mathrm{pr}(H_i = 0) = (1-a)\int M_w^0(t_* w)dM^0(w) \\ &= (1-a)t_* E(W_i \mid H_i = 0). \end{aligned}$$

Notice that $1 = E(W_i) = (1-a)E(W_i \mid H_i = 0) + aE(W_i \mid H_i = 1)$. Since W_i is positive, we have $(1-a)E(W_i \mid H_i = 0) \leq 1$, which leads to $V(t_*) \leq t_*$. So we have $V(t_*)/D(t_*) \leq t_*/D(t_*) \leq \alpha$, where the second inequality results from the definition of t_*.

Combining the facts that $|E(\mathrm{FDP}(T_m)) - V(t_*)/D(t_*)| \to 0$ and $V(t_*)/D(t_*) \leq \alpha$ we have $E(\mathrm{FDP}(T_m)) \leq \alpha + o(1)$. □

Theorem 2.1 shows that the expected false discovery proportion (i.e. the false discovery rate) corresponding to the cutoff defined by the weighted BH procedure is controlled at level α asymptotically, without assuming independence between the weights and the p-values under the alternative hypothesis. This provides some justification for our procedures.

2.4 Simulation Studies

We conducted simulation studies to assess the performance of our weighted p-value methods in detecting differential expression across multiple studies. We compared results from the four different weighting schemes, as well as from existing meta-analysis methods RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003).

The simulation study focuses on assessing the methods' abilities of efficiently selecting and prioritizing genes that demonstrate consistent differential expression patterns across the studies. The key issue here is "consistency" across studies, and it consists of two aspects: 1) the proportion of studies that display significant differential expression out of all the studies; 2) whether gene expression change occurs in the same direction for the studies that do show differential expression. We conducted two sets of simulation studies, simulation-I and II, focusing on these two aspects respectively.

The detailed simulation setup is described as follows. For both sets of simulations we simulated a scenario of ten studies to be combined and a total of 10000 genes. Each study consists of a treatment group and a control group. The sample sizes were randomly generated for each study and each group, ranging from 4 to 15. For simulation-I, we simulated 5 categories (scenarios) of differential expression: differential expression in all 10 studies, in 8 studies, 6 studies, 4 studies and 2 studies. 500 genes were assigned to each of these categories, and the rest of the genes were assigned to be not differentially expressed in any of the studies. For simulation-II, we simulated two groups of categories (scenarios) of differential expression. For the first group, differential expression occurs in all 10 studies, but we split this group further into 3 categories: all 10 studies show differential expression of the same direction; 7 out of 10 studies show differential expression in one direction, while the rest 3 show differential expression in the other direction; 5 out of 10 studies show differential expression in one direction, while the rest 5 in the other direction. For the second group, differential expression occurs in 6 out of 10 studies, and we also split this group into 3 categories: all 6 studies show differential expression of the same direction; 4 out of 6 studies show differential expression in one direction, while the rest 2 show differential expression in the other direction; 3 out of 6 studies show differential expression in one direction, while the rest 3 in the other direction. 500 genes were assigned to each of the 6 categories described above, and the rest of the genes were assigned to be not differentially expressed in any of the studies. The simulation setups described above are summarized in Table 2.1.

We use a random effects linear model to model the gene expression measurements. The expression intensity of the ith sample from group k $(k = 0, 1)$ of study s for gene g was simulated from $x_{gsik} \sim N(\mu_{gsk}, \sigma^2_{err})$. The mean expression level of gene g for the control group of study s is modeled as $\mu_{gs0} = \mu + \alpha_g + \beta_s + (\alpha\beta)_{gs}$, where μ represents the overall mean expression level, $\alpha_g \sim N(0, \sigma^2_{gene})$ represents the gene effect, $\beta_s \sim N(0, \sigma^2_{study})$ represents the study effect, and $(\alpha\beta)_{gs} \sim N(0, \sigma^2_{int})$ represents the gene-study interaction. For non-differentially expressed genes, the mean expression level for the treatment group is the same as the control group, i.e., $\mu_{gs1} = \mu_{gs0}$. For differentially expressed genes, we model the difference of mean expression between treatment and control group as $\mu_{gs1} - \mu_{gs0} = \delta + \nu_g + \epsilon_{gs}$, where δ represents the overall mean difference, $\nu_g \sim N(0, \sigma^2_{diff})$ represents the gene effect of the difference, and $\epsilon_{gs} \sim N(0, \sigma^2_{derr})$ represents the gene-study interaction of the difference.

Table 2.1. Summary of the category setups for simulation-I and II.

Simulation I

Categories	1	2	3	4	5	6
# of DE studies	10	8	6	4	2	0
Direction of DE	same	same	same	same	same	same

Simulation II

Categories	1	2	3	4	5	6
# of DE studies	10	10	10	6	6	6
Direction of DE	same	7 ↑ 3 ↓	5 ↑ 5 ↓	same	4 ↑ 2 ↓	3 ↑ 3 ↓

We conducted the simulation using two different sets of parameter choices. In order to get an idea of the general magnitude of parameters in real data, we fitted the above model on a set of microarray gene expression data from a series of stem cell studies (stemcell1, stemcell2, stemcell3 and stemcell4). The first set of parameter choices is based on estimates from the stem cell study data, with $\mu = 5$, $\sigma^2_{gene} = 2.5$, $\sigma^2_{study} = 0.7$, $\sigma^2_{int} = 0.5$, $\sigma^2_{err} = 0.3$, $\delta = 0.8$, $\sigma^2_{diff} = 0.15$, and $\sigma^2_{derr} = 0.3$. We also used a second set of parameter choices with a larger gene effect but smaller effects for the other terms, with $\mu = 5$, $\sigma^2_{gene} = 6.25$, $\sigma^2_{study} = 0.49$, $\sigma^2_{int} = 0.25$, $\sigma^2_{err} = 0.16$, $\delta = 0.8$, $\sigma^2_{diff} = 0.0016$, and $\sigma^2_{derr} = 0.0256$.

Seven methods are tested out for simulation-I and II respectively: Fisher's method, our weighted methods with four different weighting schemes, RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003). For Fisher's method and our weighted methods, we applied the Benjamini-Hochberg (1995) procedure to the p-values/weighted p-values to control for the false discovery rate. For RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003), we used the inherent options provided in their R functions to control for the false discovery rate. For all the methods, the resulting lists of significant genes are obtained controlling for false discovery rate to be less than 0.05.

To assess the performance of the methods, we calculated the proportion of rejections (i.e. proportion of genes found to be significant under FDR<0.05) for each category in simulation-I and II. That is, for each category, we count the number of rejections (significant genes) in that category and divide it by the total number of genes in that category. Both simulations were repeated 50 times, and the final results were averaged over the replicates. Since our goal is to select genes that display consistent differential expression behavior across studies, we would expect to see a higher proportion of rejections for those categories whose genes were simulated to be differentially expressed in a larger number of studies; vice versa, we would expect a lower proportion of rejections for those categories whose genes are only differentially expressed in very few studies; also, we would expect a lower proportion of rejections for those categories whose genes were differentially expressed in different directions across studies.

To better visualize the results, we plotted the proportion of rejections against categories, for both simulation-I and II, and both parameter settings (see Figure 2.1). For simulation-I, the number of studies in which the genes differentially express decreases from 10 to 0 throughout categories 1 to 6 (see Table 2.1). Thus we would expect the proportion of rejections to decrease throughout categories 1 to 6. The faster the proportion of rejections drop, the more sensitive the method is to inconsistent expression patterns across studies. As we can see from the top two plots in Figure 2.1, for both parameter settings, our weighted hypothesis testing methods, especially the two weighting schemes based on correlation, show a relatively higher sensitivity to inconsistent expression patterns across studies. Out of the seven methods, Fisher's method appears to be the least sensitive to inconsistent expression patterns, rejecting a large proportion of genes even when they are only differentially expressed in 2 out of 10 studies (category 5). RankProd (Hong *et al.*,2006) comes next in comparison. The proportion of rejections for RankProd is significantly less than Fisher's method for genes that only differentially express in a small number of studies. Next in comparison is our weighted methods with weights based on the I^2 statistic. The methods that appear to be most sensitive to inconsistent expression patterns across studies are GeneMeta (Choi *et al.*,

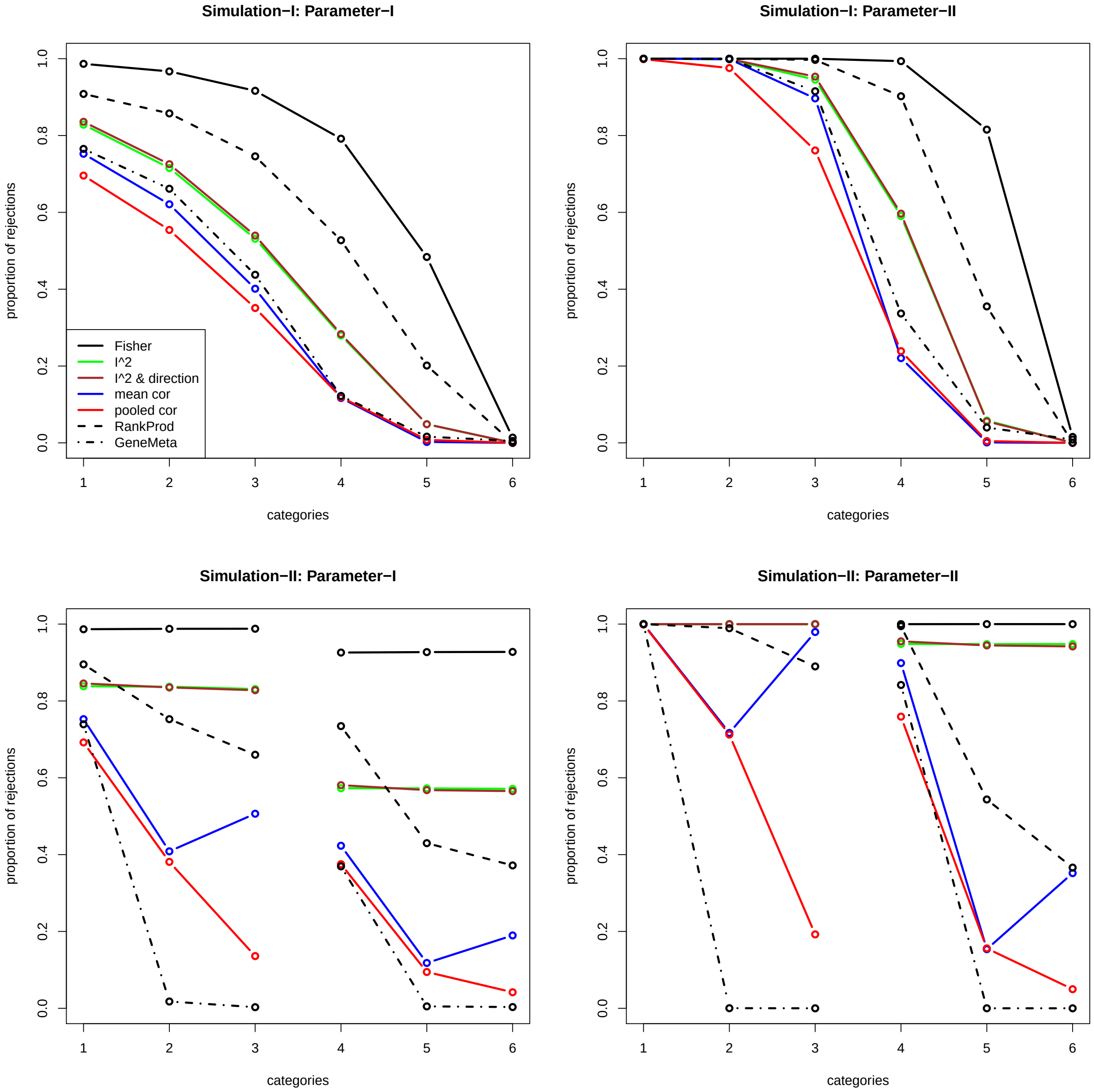

Figure 2.1. Plots showing the proportion of rejections for each category for simulation-I and II, for two parameter settings respectively.

2003) and our weighted methods with weights based on correlation. The weighted method based on pooled correlation seems to be the most sensitive, with proportion of rejections starting to drop for genes that differentially express in 6 out of 10 studies (category 3), and significantly dropping for genes that differentially express in 4 out of 10 studies (category 4). The above observations are more obvious in the case of the second parameter setting, where all seven methods start off rejecting almost all the genes that differentially express in all 10 studies, and end up rejecting almost none of the genes that differentially express in none of the studies, but the proportion of rejections drop at different rates for the seven methods throughout the categories, with the pooled correlation weighted method most sensitive to inconsistent expression patterns. For the first parameter setting, it seems that methods that are more sensitive to inconsistent expression patterns tend to have relatively less power in detecting genes that do differentially express in all the studies. In this case, the scientist would need to consider the trade-off.

Note that category 6 in simulation-I can been seen as an examination of the “specificity” of the methods. Ideally, under any criterion, genes that do not differentially express in any of the studies should not be rejected. We can see from the plots that the proportion of rejections for category 6 is very close to 0 for all the methods under both parameter settings. As the plot only shows the average result across the 50 simulation runs, we also checked the actual rejection proportions for each run to make sure that they were consistently low across simulations. In fact, our two weighted methods based on correlation performed the best in this aspect, not rejecting a single gene in category 6 in all 50 simulation runs for both parameter settings.

For simulation-II, we would expect a lower proportion of rejections for categories 2 and 3 compared to 1, and categories 5 and 6 compared to 4, since genes in categories 2, 3, 5 and 6 differentially express in different directions across the studies, especially for categories 3 and 6, which are the extreme cases where differential expression occurs in one direction for half the studies and the other direction for the other half of the studies (see Table 2.1). Results for simulation-II are shown in the bottom two plots in Figure 2.1. Fisher’s method and the two weighted methods based on I^2 do not appear to be sensitive to different directions of DE across studies. For these three methods, the proportion of rejections are almost the same for genes that differentially express in the same number of studies, regardless of whether the direction of change is the same across the studies. For RankProd (Hong *et al.*,2006), GeneMeta (Choi *et al.*, 2003) and the two weighted methods based on correlation, the proportion of rejections drop for genes that differentially express in different directions across studies. GeneMeta seems to be the most sensitive, with proportion of rejections close to 0 for all the categories with inconsistent DE direction. Our pooled correlation weighted method comes up next in comparison, with proportion of rejections dropping significantly for categories with inconsistent DE direction, especially for categories 3 and 6, where the inconsistency of DE direction is the most extreme.

Summarizing the results from simulation-I and II, our weighted method with pooled correlation weighting scheme and GeneMeta (Choi *et al.*, 2003) seem to be the two more competitive methods out of the seven methods tested, with regard to our goal of selecting genes that display consistent differential expression patterns, where "consistency" is evaluated from two aspects mentioned previously addressed by the two simulations respectively. Notice that under our simulations, it is expected that GeneMeta performs well, since the model we used to simulate the data is essentially the same model that GeneMeta uses to analyze the data. As discussed above, our weighted method with pooled correlation weighting scheme shows competitive results compared with GeneMeta, but our method does not rely on assumptions such as normality of the data, or more importantly, the structure of the model.

To get a little sense of how the methods perform with non-normal data, we experimented simulating the expression data with ϵ_{gs} generated from the log of a gamma distribution. The resulting plots look very similar to those in Figure 2.1. Power is generally lower for all the methods and categories, but the trends of the plots and the relative comparisons between the methods remain almost the same.

During implementation of our weighted methods, we also find that they are significantly more efficient in terms of computing time, compared to RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003), both of which involve permutations in the analysis process. In an experiment of the time costs of the methods, we implemented the three methods on a set of real data (described in the next section of the paper), and the results are summarized in Table 2.2, where the time for our weighted method indicates the time used to achieve results for all four weighting schemes.

Table 2.2. Summary of the time costs of the three methods (in seconds).

	User time	System time	Elapsed time
Weighted methods	175.701	0.430	176.449
RankProd	1174.512	26.272	1203.214
GeneMeta	274.749	13.801	289.317

2.5 Data Analysis: Meta-Analysis of a Set of Stem Cell Studies

We applied the weighted p-value methods to a set of microarray data studies from four stem cell papers (Chin *et al.*, 2009, Guenther *et al.*, 2010, Newman and Cooper, 2010 and Chin *et al.*, 2010). We focus on differential expression analysis between human induced pluripotent stem (hiPS) cells and human embryonic stem (hES) cells. Some of the studies contain other samples such as human fibroblasts, but we only used samples from hiPS cells and hES cells. Studies that did not have at least two samples for each group (hiPS and hES) were excluded due to the inability to perform the t-test, leaving us 9 studies in total. All the studies used the Affymetrix Human Genome U133 Plus 2.0 Array platform, which contains 54675 features (probesets). For all the studies we directly used the data preprocessed by the original contributors and did not perform any additional normalization, except for taking the log for data that were not already on the log2 scale.

We performed a two sample t-test between the hiPS cell samples and the hES cell samples for the 9 studies and applied the Fisher's combined probability test as well as the weighted p-value methods using the four different weighting schemes. To adjust for multiple hypothesis testing, we applied the Benjamini-Hochberg (1995) procedure to the unweighted Fisher's p-values as well as the four kinds of weighted p-values, controlling the false discovery rate at the 0.05 level. 16508 out of 54675 features showed up to be significant by the unweighted Fisher's method. This seems to be an abnormally high proportion of significant features, and may be due to the fact that Fisher's test is prone to be driven by significant results of single studies. On the other hand, for the weighted methods, the weighting scheme based on I^2 index yields 777 significant features; the weighting scheme based on I^2 index and accounting for direction yields 670 significant features; the one based on mean correlation has 778 features showing up significant and the one based on pooled correlation has 196 features showing up significant.

We also performed the RankProd method by Hong *et al.* (2006) and the GeneMeta method by Choi *et al.* (2003) on this set of data. Based on controlling the false discovery rate at the 0.05 level, the RankProd method found 2893 significant up-regulated (iPS compared to hES) features and 1996 significant down-regulated features, while the GeneMeta method found 2021 significant features for a two-sided test. Notice that the different methods resulted in significant gene lists of different sizes. Fisher's method resulted in an abnormally large number of significant features. Our weighted methods result in relatively smaller lists of significant features compared to RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003). The weighted method based on pooled correlation seems to be the most conservative. This observation agrees with the simulation results.

In order for the selected features lists from different methods be more comparable in size, we created lists of the top 2000 features for all the methods, disregarding the false discovery rate, by selecting the features with the smallest p-values/weighted p-values/q-values for each of the methods. Figure 2.2 is a venn diagram displaying the overlap of the top 2000 features found respectively by using our pooled correlation weighted method, the RankProd method and the GeneMeta method. The fact that a significant number of features were only selected by one of the methods shows that these methods rank features from different perspectives, and that it maybe useful to select candidate genes using a combination of these methods, so as to capture genes that are interesting from different aspects.

To further assess the top features lists produced by different methods, pathway analysis was done based on functional annotation clustering analysis using DAVID, which is available at http://david.abcc.ncifcrf.gov/home.jsp. Table 2.3 shows the results for the top features lists by our pooled correlation weighted method, RankProd (Hong *et al.*,2006) and GeneMeta (Choi *et al.*, 2003). The first table shows the top pathways/functions for the 253 features that were selected by these three methods simultaneously. The next three tables show the top pathways/functions for the features that were selected by only one of the methods. We see that each method has its unique contributions of top functions, although the top functions for the pooled correlation weighted method and RankProd seem

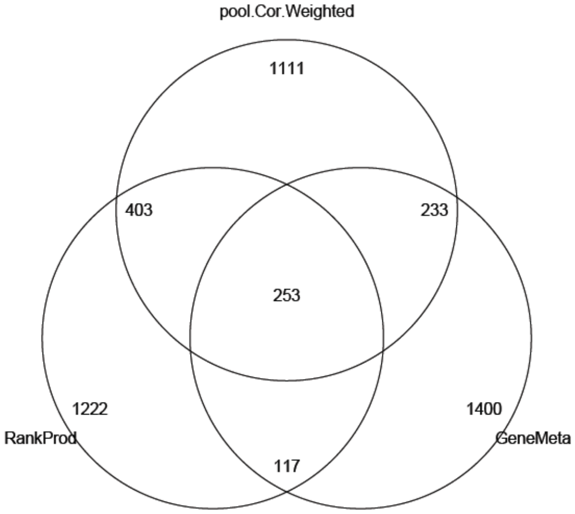

Figure 2.2. Venn diagram showing the overlapping of significant features found by performing different methods on the stem cell data.

to be relatively more related, and they also seem to be more related to the top functions for the intersection of the three methods. The enrichment scores for the top functions by the pooled correlation weighted method and by RankProd also appear to be higher compared to GeneMeta. A higher enrichment score indicates that the top features lists contains features that aggregate into certain functions, as opposed to individual features that are unrelated. Thus results with higher enrichment score may make more sense and be more easily interpretable.

Since different methods have different assumptions and focus on different aspects in selecting significant features, it may be useful to use a combination of methods when doing meta-analysis for real data. Compared to GeneMeta, which relies heavily on assumptions of the distribution and structure of the data, and RankProd, which is less sensitive to inconsistent differential expression patterns across studies, our weighted method is a fast and convenient alternative that aims at selecting consistently differentially expressed genes without imposing too many assumptions on the data.

2.6 Discussion

While there have been many papers and methodologies written on the meta-analysis of high-dimensional genomic data, more recent efforts have focused on assessing reproducibility and concordance of these datasets. We have attempted to address these issues using the weighted hypothesis testing framework of Genovese et al. (2006). This seems very natural for our problem in that the weights represent relative measures of "strength of evidence" for whether or not the data are combinable, and genes for which this assumption does not hold are down-weighted. The assumption weighting methodology itself is quite straightforward to implement; R scripts implementing the proposed methodology are available. While we have used Fisher's method for combining p-values, other methods for combining p-values could be used as well (Hedges and Olkin, 1985).

The application of our methodology to the stem cell dataset found that there was substantial evidence of gene-specific heterogeneity across the datasets. One of the advantages of our methodology, relative to those of Miron et al. (2006) and Lai et al. (2007), is the assessment of local (i.e. gene-specific) heterogeneity, as opposed to the global heterogeneity assessment of those procedures. It appears that there is a bias-variance tradeoff between the two approaches and is something that we plan to investigate in future work.

Table 2.3. Functional annotation clustering results (DAVID) for the top features lists by our pooled correlation weighted method, RankProd and GeneMeta.

Top features by the intersection of the three methods	Enrichment Score
Extracellular region, signal peptide	4.65
Signal peptide, glycoprotein	3.04
Cell migration	2.63
Skeletal, face, head development	2.54
Cell adhesion	2.54
Extracellular matrix	1.91
Endochondral bone morphogenesis	1.64
Negative regulation of DNA binding	1.64
Blood vessel development	1.57
Top features by our weighted method only	Enrichment Score
Blood vessel development	8.24
Embryonic development	6.15
Embryonic morphogenesis	5.44
Cell migration	5.01
Cell adhesion	4.71
Extracellular matrix	4.36
Regulation of cell migration	4.34
Mesenchymal cell development	4.21
Regulation of biosynthetic process and transcription	4.09
Tube and respiratory system development	3.42
Embryonic skeletal system development	3.4
Top features by RankProd only	Enrichment Score
Transcription regulation, zinc finger	9.96
zinc finger	7.1
Tube and respiratory tube development	4.67
Positive regulation of transcription	4.4
Blood vessel development	4.2
Pattern specification process	4
Neuron development, cell morphogenesis	2.82
Negative regulation of transcription	2.77
Pathways in cancer	2.65
Sensory organ development	2.61
Top features by GeneMeta only	Enrichment Score
Nucleotide binding	5.18
Helicase, ATP-binding	2.96
Protein localization	2.73
GTPase regulator activity	2.21
Hydrolase, protease	2.09
Metal-binding, zinc finger	2.01
Chromatin regulator	1.86

Chapter 3

A Two-Step Hierarchical Hypothesis Set Testing Framework

3.1 Introduction

Suppose we want to test for m sets of hypotheses $H(1), ..., H(m)$ simultaneously. In each set of hypotheses $H(i)$, there's a screening hypothesis, denoted by $H_0(i)$, that will be tested first; and the rest of the hypotheses in the set $H_1(i), ..., H_{n(i)}(i)$, which we will refer to as the individual hypotheses, will be tested simultaneously if and only if $H_0(i)$ is rejected. Our interest is in controlling the proportion of false rejections on the level of the hypothesis sets. We define a hypothesis set to be rejected if the screening hypothesis in the set is rejected (which is equivalent to at least one hypothesis in the set is rejected); and we define a hypothesis set to be falsely rejected if at least one true null hypothesis in the set (including the screening hypothesis) were incorrectly rejected. In this case, the overall false discovery rate (OFDR), as introduced in Benjamini and Heller (2008) and further discussed in Heller *et al.* (2009), is the expected proportion of falsely rejected hypothesis sets out of all the rejected hypothesis sets. We give the formal definition below. Throughout this chapter, our goal would be to develop procedures for the hierarchical hypothesis set testing situation described above, while controlling for the overall false discovery rate (OFDR).

Definition 3.1. *A hypothesis set i $(i = 1, ..., m)$ is said to be rejected if $H_0(i)$ is* *rejected. A hypothesis set i is said to be falsely rejected if it is rejected and at least one hypothesis in the set $H_0(i), H_1(i), ..., H_{n(i)}(i)$ is falsely rejected. The overall false discovery rate is defined as*

$$OFDR = E\Big(\frac{V}{R \vee 1}\Big),$$

where $R \vee 1 = max(R, 1)$, R is the total number of hypothesis sets rejected out of m, and V is the total number of hypothesis sets that are falsely rejected out of m.

We allow a flexible relationship between the screening hypothesis and the individual hypotheses within a hypothesis set. We list and comment on a few possible scenarios below as examples, some of which will be further explored through simulation studies or data analysis examples in following sections. The applicability of the general procedure we will be introducing in section 3.2 is not limited to these cases, since essentially we do not impose any restrictions on the setup of the hypothesis sets.

(i) The screening hypothesis is a meta-analysis of the individual hypotheses (typically related through p-values directly).

(ii) The screening hypothesis a multivariate form of the individual hypotheses or some other form of overall summary of the individual hypotheses (not related through test statistics or p-values directly).

(iii) The screening hypothesis is not directly related to the individual hypotheses, but the problem or situation calls for testing in two (or more) stages.

For scenario (i), depending on different types of meta-analysis, there are different sub-scenarios: (a) the screening null hypothesis is the conjunction of the individual null hypotheses, i.e. the screening hypothesis tests for whether at least one individual null hypothesis is false; (b) the screening null hypothesis is a partial conjunction of the individual null hypotheses, i.e. the screening hypothesis tests for whether at least $u(i)$ out of $n(i)$ of the individual null hypotheses are false (see Benjamini and Heller (2008) for more details on testing for partial conjunction hypotheses); (c) the screening hypothesis tests for whether all individual null hypotheses are false. Case (a) is probably the more commonly conducted type of meta-analysis. However, even within case (a), depending on the meta-analysis method used, the dynamics between the screening hypothesis and the individual hypotheses can still be different. We will discuss this case more detail in section 3.3.

For scenario (ii), we briefly describe two possible cases: (a) in case of a multivariate response, the screening hypothesis tests for differences between treatment groups (or some other hypothesis) on the multivariate level, while the individual hypotheses test for the same differences (or hypotheses) on the univariate responses; (b) in case of an ANOVA problem, the screening hypothesis tests for whether the treatment factor is significant at all, and the individual hypotheses test for pairwise comparisons between the factor levels. The problem setting in Heller *et al.* (2009) falls under case (a), where the screening hypothesis compares the joint expression of genes within a gene set across treatments, and the individual hypotheses compares expressions of individual genes across treatments. Case (b) has been explored by Jiang and Doerge (2006). However, in Jiang and Doerge (2006) their goal was to control the FDR on the level of the individual hypotheses, in contrast with our goal of controlling the OFDR on the hypothesis set level. We will discuss case (b) and conduct comparisons with Jiang and Doerge (2006) in section 3.4.

In both scenario (i) and (ii), the need for a screening hypothesis comes from a statistical point of view. Scenario (iii) refers to situations where we simply need or want to conduct tests in a hierarchical manner - moving on to the next step only if the hypothesis in the previous step is rejected. In this case, the screening hypothesis merely "screens" whether we should continue to the next step, and might not actually relate to subsequent hypotheses in content directly. One such example would be hierarchical classification, for instance cancer classification, which we will explore through data analysis in section 3.6.

3.2 The Two-Step Hierarchical Testing Framework

Now we shall introduce our general two-step hierarchical hypothesis set testing procedure. Let $p_j(i)$ be the unadjusted p-value for individually testing the jth hypothesis in the ith set, where $i = 1, ..., m$ and $j = 0, 1, ...n(i)$. The general procedure is as follows:

Procedure 3.1.

(1) Apply the BH procedure at level α to the m p-values corresponding to the screening hypotheses $p_0(1), ..., p_0(m)$. Let R be the number of rejected screening hypothesis.

(2) For each rejected hypothesis set $H(i)$, test for additional hypotheses $H_1(i), ..., H_{n(i)}(i)$ simultaneously, applying a p-value adjusting procedure on $p_1(i), ..., p_{n(i)}(i)$ such that the family-wise error rate (FWER) of these $n(i)$ tests are controlled at level $R\alpha/m$.

This procedure is a generalization of the procedure proposed by Heller *et al.* (2009) used for testing differential expression for gene sets. They discussed possibilities of testing for the screening hypothesis in different ways and using alternative family-wise error rate controlling procedures in the second step under their problem setting. Here we formally build a general framework that allows for flexibility and adaptability, not only in the techniques used in the two steps, but also in the situations and problem settings that the procedure can be applied on. We have discussed this in the previous section and will further illustrate this point through a series of different applications in subsequent sections.

Procedure 3.1 controls the OFDR at level α, under the condition that the p-values of the individual hypotheses in each hypothesis set are independent from all other screening hypothesis p-values. The proof follows directly from the proof in Heller *et al.* (2009) of a similar claim on their procedure. We state the theorem formally below and provide the proof.

Theorem 3.1. *Procedure 3.1 controls the overall false discovery rate (OFDR) at level α assuming that for each hypothesis set $H(i)$, the p-values $p_0(i), p_1(i), ..., p_{n(i)}(i)$ are independent of all the other screening p-values, $p_0(1), ..., p_0(m)$ excluding $p_0(i)$.*

Proof. Recall that $OFDR = E(V/(R \vee 1))$, where R is number of hypothesis sets rejected, and V is the number of hypothesis sets falsely rejected. For hypothesis set i $(i = 1, ..., m)$, define $R(i) = 1$ if hypothesis set i is rejected and 0 otherwise, and define $V(i) = 1$ if it is falsely rejected and 0 otherwise. Let I_0 be the index set for hypothesis sets that have at least one true null hypothesis. Then

$$\begin{aligned} OFDR = E\Big(\frac{V}{R \vee 1}\Big) &= \sum_{k=1}^{m} E\Big(\frac{\sum_{i=1}^{m} V(i)}{k} \mid R = k\Big) P(R = k) \\ &= \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} P(V(i) = 1,\ R = k) = \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} P(V(i) = 1,\ C_k^{(i)}). \end{aligned}$$

Here, $C_k^{(i)}$ denotes the event that $p_{0(k-1)}^{(i)} \le k\alpha/m,\ p_{0(k)}^{(i)} > (k+1)\alpha/m, ..., p_{0(m-1)}^{(i)} > \alpha$, where $p_{0(1)}^{(i)} \le \cdots \le p_{0(m-1)}^{(i)}$ are the ordered screening p-values excluding $p_0(i)$. Basically, it denotes that apart from $p_0(i)$, $k-1$ other p-values were rejected. Then

$$\begin{aligned} OFDR &= \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} P(R(i) = 1,\ \text{type-I error for hypothesis set } i,\ C_k^{(i)}) \\ &= \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} P(p_0(i) \le k\alpha/m,\ \text{type-I error for hypothesis set } i,\ C_k^{(i)}) \\ &= \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} P(p_0(i) \le k\alpha/m,\ \text{type-I error for hypothesis set } i) P(C_k^{(i)}), \end{aligned}$$

where the last equality follows from the assumption that $p_0(i), ..., p_{n(i)}(i)$ are independent of all the other screening p-values. Now, for $\forall\ i \in I_0$, either the screening hypothesis $H_0(i)$ is true, or at least one of individual hypotheses $H_1(i), ..., H_{n(i)}(i)$ is true. If $H_0(i)$ is true, then

$$\begin{aligned} &P(p_0(i) \le k\alpha/m,\ \text{type-I error for hypothesis set } i \mid H_0(i) \text{ is true}) \\ &= P(p_0(i) \le k\alpha/m \mid H_0(i) \text{ is true}) \le k\alpha/m, \end{aligned}$$

assuming that $p_0(i)$ is distributed as Uniform$(0, 1)$ under the null. If $H_0(i)$ is not true, then at least one of $H_1(i), ..., H_{n(i)}(i)$ is true, and we have

$$\begin{aligned}
&P(p_0(i) \leq k\alpha/m, \text{ type-I error for hypothesis set } i \mid H_0(i) \text{ not true})\\
&\leq P(\text{type-I error for hypothesis set } i \mid H_0(i) \text{ not true})\\
&= P(\text{type-I error for testing } H_1(i), ..., H_{n(i)}(i) \mid H_0(i) \text{ not true}) \leq k\alpha/m.
\end{aligned}$$

The last inequality follows from the fact that the FWER for testing $H_1(i), ..., H_{n(i)}(i)$ is controlled at $k\alpha/m$ in the second step of the procedure. Thus for $\forall\ i \in I_0$ we have $P(p_0(i) \leq k\alpha/m,$ type-I error for hypothesis set $i) \leq k\alpha/m$. This leads to

$$OFDR \leq \sum_{k=1}^{m} \sum_{i \in I_0} \frac{1}{k} \frac{k\alpha}{m} P(C_k^{(i)}) = \sum_{i \in I_0} \frac{\alpha}{m} \sum_{k=1}^{m} P(C_k^{(i)}) = \sum_{i \in I_0} \frac{\alpha}{m} \leq \alpha.$$

□

3.3 Application to Multidimensional Directional Decisions Problem

3.3.1 Discussions about the problem setting

This section is motivated by Guo *et al.* (2010), who are concerned with identifying genes with significant expression patterns over ordered categories. The problem arises in time-course or dose-response experiments, where gene expression is recorded for a number of ordered categories, and the goal is to identify genes that have some pattern of expression change between successive categories. If for each gene we take the expression change between each pair of adjacent categories, then the problem becomes detecting whether the expression changes are non-zero, and subsequently, what is the direction of the change. The term "multidimensional" refers to the fact that there is more than one parameter (namely expression change) to be tested for each gene when there are more than two ordered categories - thus we are making multiple directional decisions for each gene. For the rest of this section, we shall adopt the notations used in Guo *et al.* (2010).

Following Guo's paper, suppose we have gene expression data over $q+1$ ordered categories. Then for each gene j ($j = 1, ..., m$), the vector representing the change of expression between each pair of adjacent categories can be denoted by $\boldsymbol{\delta}_j = (\delta_{1j}, ..., \delta_{qj})'$. Guo *et al.* proposed a two step testing strategy for this problem. The first step is to test $H_{0j} : \boldsymbol{\delta}_j = \mathbf{0}$ against $\delta_{\mathbf{j}} \neq \mathbf{0}$, i.e. whether there is any expression change at all for each gene j. If we reject the null hypothesis of zero-change for gene j, we proceed to test for each δ_{ij} $(i = 1, ..., q)$, to decide which of them are nonzero and what their signs are. Guo *et al.* (2010) hope to control errors arising in both the first and second step of testing. An error occurs in the first step if we reject the null hypothesis of a gene that really has $\boldsymbol{\delta}_j = \mathbf{0}$. Errors occur in the second step if we make incorrect directional decisions about a gene, that is, we claim δ_{ij} to be positive when in fact $\delta_{ij} \leq 0$, or we claim δ_{ij} to be negative when in fact $\delta_{ij} \geq 0$. They promoted the concept of mixed directional FDR (mdFDR), which is essentially the sum of the FDR's generated by the two steps of testing, namely the traditional FDR and the pure directional FDR (dFDR), and is formally defined as

$$mdFDR = FDR + dFDR = E\Big(\frac{V}{R \vee 1}\Big) + E\Big(\frac{S}{R \vee 1}\Big) = E\Big(\frac{V + S}{R \vee 1}\Big),$$

where R is the number of rejected hypotheses among $H_1, ..., H_m$, $R \vee 1 = \max(R, 1)$, V is the number of falsely rejected true null hypotheses among $H_1, ..., H_m$, and S is number of hypotheses among $H_1, ..., H_m$ that were correctly rejected but for which at least one directional decision on its components was incorrect.

The important thing to note about the definition of the mixed directional FDR (mdFDR) is that the denominator is the same for its two components. The first component, namely the FDR, is very straightforward, since it is defined as the traditional FDR among the m genes. For the second component, namely the pure directional FDR (dFDR), the number of directional decisions made (or the number of hypotheses tested in the second step) is actually $m \times q$. However, the denominator for calculating the dFDR is m, instead of $m \times q$. In other words, for any incorrect directional decisions we make, we hold the corresponding gene accountable. It doesn't matter how many directional mistakes we make for a gene - so long as we make one directional mistake - the gene is counted as falsely rejected. And what we care about is how many genes we made incorrect decisions about, not how many incorrect decisions we made in total. Note also that we only count directional errors if the gene was correctly rejected in the first step. If a false rejection was already made in the first step for a gene, it would be counted in the FDR part, and would not be recounted in the dFDR part. Last but not least, note that we would only have the chance to make a mistake on a gene if it were rejected in the first step. Genes that were not rejected in the first step obviously cannot contribute to the FDR; they also cannot contribute to the dFDR since they weren't even given the chance to go through the second round of testing for directional decisions.

It is apparent that OFDR $\leq$ mdFDR. This is because OFDR only takes into account type-I errors for the individual hypotheses in the second step and not directional errors. Thus while the general Procedure 3.1 guarantees the control of the OFDR under independence conditions (by Theorem 3.1), it does not automatically guarantee the control of the mdFDR. However, we can easily extend the proof of Theorem 3.1 to obtain the following results.

Corollary 3.1. *Under the assumptions of Theorem 3.1, Procedure 3.1 controls the mixed-directional FDR (mdFDR) at level α if the family-wise error rate controlling procedure used in step (2) maintains control of both type-I and directional errors.*

Although Corollary 3.1 is just a direct extension of Theorem 3.1, it provides great practical benefits. This is because many commonly used family-wise error rate controlling procedures have been shown to maintain control of directional errors under certain conditions (Finner, 1999). For instance, Finner (1999) showed that the Bonferroni procedure, Holm's procedure, as well as Hochberg's procedure are all able to control both type-I and directional errors family-wise, for multiple two-sided tests involving independent T-statistics. This covers the most common two-sided tests as well as some of the most common family-wise error rate controlling procedures. Thus Corollary 3.1 allows us to extend Procedure 3.1 to many situations where directional decisions are needed, such as the scenario discussed in this section, and ensures control of the mdFDR in addition to the OFDR.

Through this series of observations, it becomes clear to us that the problem in Guo *et al.* (2010) falls under the framework of the two-step hierarchical testing of multiple hypothesis sets. To be more explicit, each gene corresponds to a hypothesis set with a total of $q + 1$ hypotheses, where the screening hypothesis tests for $\delta_{\mathbf{j}} \neq \mathbf{0}$ and the q individual hypotheses in the set test for $\delta_{ij} \neq 0$ for $i = 1, ..., q$. As in our framework, the individual hypotheses are only tested if the screening hypothesis is rejected. The set up of this problem falls under scenario (i) mentioned in section 3.1 - the screening hypothesis is a meta-analysis of the individual hypotheses. In other words, the screening hypothesis for a gene is the intersection of the individual hypotheses, i.e. it tests for whether any of the components of $\boldsymbol{\delta}_j$ is non-zero, and thereby answers the question of whether that gene has any non-constant expression pattern at all. For this specific problem, it makes sense that if we reject the screening hypothesis then we should reject at least one individual hypothesis - it would be hard to interpret a result stating that a gene has non-constant pattern and yet unable to claim any of the components to be non-zero. This property is called "consonance". As mentioned in Holm (1979), a test procedure is called consonant if it avoids the situation where a hypothesis is rejected but we fail to reject any other hypothesis implied by it. Therefore any test procedure developed for this problem should be consonant. As an example, meta-analysis method based on the Bonferroni procedure is consonant. Note that meta-analysis methods used for testing the conjunction of null hypotheses do not necessarily need to be consonant. In general, methods based on ordered p-values are usually consonant - since rejection of the conjunction null is usually based on rejecting the smallest individual p-value. On the other hand, methods based on aggregating information from all the individual p-values may not be consonant, such as the Fisher's combined test. These tests are usually more powerful, exactly because they utilize the fact that information on any of the individual hypotheses may not be enough to reject the null, but an aggregation of information would help reject the conjunction of nulls. They can be used for problems that do not call for the consonance property, but they are not suitable for the problem setting in this section. Whether the consonance property is necessary or not entirely depends on what makes sense for a particular problem. In some of the following sections we will talk about applications where we do not specifically require this property for our procedures.

3.3.2 Testing procedures

3.3.2.1 Original procedures

Two testing procedures for this problem were proposed in Guo *et al.* (2010). Both are two-step procedures with the first step testing for the conjunction null, and the second step testing for the individual components. For ease of discussion and comparison, we restate the two procedures below. Following the notations in Guo *et al.* (2010), let $\tilde{P}_{ij}$ be the p-values for testing each individual hypothesis $\delta_{ij} \neq 0$, for $i = 1, ..., q$, $j = 1, ..., m$. And let $\tilde{P}_{(1)j} \leq \cdots \leq \tilde{P}_{(q)j}$ be the ordered versions of $\tilde{P}_{ij}$, $i = 1, ..., q$, for a fixed j.

Procedure 3.2. *(Guo et al., 2010)*

(1) Based on the Bonferroni test, let $P_j = q\tilde{P}_{(1)j}$ for $j = 1, ..., m$. Apply the BH method at level α on $P_1, ..., P_m$ to test for $H_{01}, ..., H_{0m}$ simultaneously. Let R be the number of rejected hypothesis out of m.

(2) For every $i = 1, ..., q$ and $j = 1, ..., m$ with $\tilde{P}_{ij} \leq R\alpha/(qm)$, declare $\delta_{ij} >$ or < 0 according to the signs of the corresponding test statistics.

Procedure 3.3. *(Guo et al., 2010)*

(1) Based on the Simes test, let $P_j = \min_{1 \leq i \leq q}\{q\tilde{P}_{(1)j}/i\}$ for $j = 1, ..., m$. Apply the BH method at level α on $P_1, ..., P_m$ to test for $H_{01}, ..., H_{0m}$ simultaneously. Let R be the number of rejected hypothesis out of m.

(2) For every $j = 1, ..., m$, let $R_j = \max\{i : \tilde{P}_{(i)j} \leq iR\alpha/(qm)\}$ if the maximum exists and 0 otherwise. For every i and j with $\tilde{P}_{ij} \leq R_j R\alpha/(qm)$, declare $\delta_{ij} >$ or < 0 according to the signs of the corresponding test statistics.

Two points are worth noting here:

(i) The second step in both procedures call for testing individual p-values for all i and j, which seems to be different from our general Procedure 3.1. However, by further examination, it is not hard to see that by construction of the P_j and the cutoff in the second step, only the genes that were found significant in the first step have a chance to have significant components in the second step. In other words, it would be equivalent to define Procedures 3.2 and 3.3 such that the second step only tests for the $\tilde{P}_{ij}$'s such that P_j is rejected in the first step.

(ii) Following the previous observation, it is easy to see that if P_j is found significant in the first step, then at least one $\tilde{P}_{ij}$ will be rejected in the second step. This is because the second step is basically based on the same multiple testing method as the first step (either Bonferroni or Simes), except adjusting the cutoff to be $\frac{R}{m}\alpha$, which is essentially the cutoff for rejecting genes in the first step. For example, in Procedure 3.2, for each j found significant in the first step, the second step tests for q hypotheses simultaneously, and adjusts the p-values using the Bonferroni method with the level controlled at $\frac{R}{m}\alpha$, leading to a cutoff of $\frac{R}{qm}\alpha$ for each individual p-value. This shows that Procedures 3.2 and 3.3 are consonant.

By the observations in the previous paragraph, we can see that the structure of Procedures 3.2 and 3.3 look the same as the structure of our general Procedure 3.1: An FDR controlling procedure is applied in the first step; only genes found significant in the first step are tested in the second step; for each significant gene, the individual hypotheses in the second step are tested simultaneously with adjusted level $\frac{R}{m}\alpha$. However, one very important detail needs to be checked to see whether these two procedures indeed fall under our general framework. In the second step of the general Procedure 3.1, not only is the level adjusted to be $\frac{R}{m}\alpha$, the crucial point is that the family-wise error rate (FWER) needs to be controlled at that level. For Procedures 3.2, since the second step uses the Bonferroni method to simultaneously test the q individual hypotheses, the family-wise error rate is indeed controlled at $\frac{R}{m}\alpha$. However, this is not the case for Procedures 3.3. This is because Simes test is only meant to test the conjunction null, but not meant to make statements about the individual hypotheses (see Simes, 1986). When used to test the conjunction of nulls, the Simes test controls the family-wise error rate in the weak sense (that is, under the null that all the individual nulls are true); it does not, however, control the family-wise error rate in the general sense. The method for testing the q individual hypotheses at level $\frac{R}{m}\alpha$ in the second step of Procedure 3.3 was actually explicitly discussed in Simes (1986), and as Simes pointed out, there is no formal basis for rejecting individual hypotheses using that method, even under conditions where the Simes test is justified to test the overall conjunction null. In summary, step two in Procedure 3.3 does not control the FWER at the desired level.

Now it is clear to us that this multidimensional directional decisions problem falls under our general hierarchical hypothesis set testing framework and that Procedure 3.2 conforms to our general Procedure 3.1, but that Procedure 3.3 does not. It follows from Theorem 3.1 and Corollary 3.1 that Procedure 3.2 by Guo *et al.* (2010) controls the OFDR and mdFDR at level α under certain independence conditions, while Procedure 3.3 is not justified to control the OFDR or mdFDR. Not surprisingly, these results agree with Guo *et al.* (2010). They proved that Procedure 3.2 controls the mdFDR, and they discussed the fact that Procedure 3.3 does not control the mdFDR, despite having the potential to be more powerful, since the Simes test is more powerful than the Bonferroni test.

3.3.2.2 New procedures

So far it seems that all we did was adapt the work of Guo *et al.* (2010) into our framework. However, the advantage of looking at this problem through the light of our framework is that we can now easily understand why Procedure 3.3 does not work, and more importantly, how we can come up with new procedures for this problem that do control the OFDR and mdFDR, and at the same time, are more powerful than Procedure 3.2.

As discussed in the last section, Procedure 3.3 fails to control the OFDR because the second step does not control the FWER for testing individual hypotheses. A naive modification of Procedure 3.3 would be to simply substitute the second step with a procedure that does control the FWER. Since the Simes test does not control the FWER, a more conservative option such as the Bonferroni test might be an option. In this case, we would still use Simes test for the first step, but then switch to Bonferroni's test for the second step. It follows from our general framework that this modified procedure would control the OFDR. It is also potentially more powerful, since Simes test is more powerful than Bonferroni's test in the first step. However, an important flaw of this potential procedure is that it does not have the consonant property. Since the two steps in the procedure are based on two different tests and are not inherently connected, it is not guaranteed that any individual hypotheses would be rejected following the rejection of the gene in the first step. Since Bonferroni's method is more stringent than Simes, it is very possible that we will claim genes to have significant expression patterns, and yet cannot claim significant directional changes in any of its components.

Thus the key to a more powerful new procedure that yet controls the OFDR, is a multiple testing procedure that controls the FWER when testing for individual hypotheses, and that is more powerful than the Bonferroni method. Both the first and second step of the procedure needs to be based on the same method, so that the consonant property can be conserved. Two common multiple testing methods that satisfy these criteria are the Holm's method (Holm, 1979) and the Hochberg's method (Hochberg, 1988). Both these methods can be directly applied in the second step of the procedures to simultaneously test for the q individual hypotheses. For the first step, we can also easily construct p-values for testing the conjunction of nulls based on these two methods, following the idea that we reject the conjunction of nulls at a certain level if at least one individual null can be rejected at that level.

Before we proceed to construct our new procedures, we'll now review the Holm and Hochberg procedures, and how to test for conjunction of nulls based on these two methods. Both methods are based on the ordered p-values, which in our context are $\tilde{P}_{(1)j} \leq \cdots \leq \tilde{P}_{(q)j}$ (for a fixed j). Holm's method, also referred to as the Holm's step-down procedure, rejects the individual hypotheses sequentially. In our context, it starts from $i = 1$, and rejects $\tilde{P}_{(i)j}$ if $\tilde{P}_{(i)j} \leq \frac{\alpha}{q-i+1}$ and that all the previous $\tilde{P}_{(k)j}$ for $k < i$ have been rejected. In other words, it rejects starting from the smallest p-value until for some i, $\tilde{P}_{(i)j} \leq \frac{\alpha}{q-i+1}$ is no longer satisfied. To use Holm's method for testing the conjunction of nulls, we simply need to see whether the smallest p-value will be rejected by Holm's. That is, let the p-value for testing the conjunction of nulls be $P_j = q\tilde{P}_{(1)j}$. Notice that this is the same as the Bonferroni's method for testing conjunction of nulls. However, Holm's method is more powerful than the Bonferroni's method in further testing of the individual hypotheses, because it adjusts the cutoff by $\frac{1}{q-i+1}$, instead of $\frac{1}{q}$ as in Bonferroni. Similar to Bonferroni's method, Holm's method controls the FWER without imposing any assumptions on the p-values. Hochberg's method, also referred to as the Hochberg's step-up procedure, uses the same set of adjustments for the cutoffs, but does not require rejecting the hypotheses sequentially, thus achieving higher power. To be more specific, Hochberg's method finds the largest i such that $\tilde{P}_{(i)j} \leq \frac{\alpha}{q-i+1}$ holds true, and rejects all $\tilde{P}_{(k)j}$ such that $k \leq i$, regardless of whether $\tilde{P}_{(k)j} \leq \frac{\alpha}{q-k+1}$ holds true for every $k \leq i$. To use Hochberg's method for testing the conjunction of nulls, we need to see whether $\tilde{P}_{(i)j} \leq \frac{\alpha}{q-i+1}$ holds true for at least some i (not necessarily $i = 1$). If we let the p-value for testing the conjunction of nulls be $P_j = \min_{1 \leq i \leq q}\{(q+1-i)\tilde{P}_{(i)j}\}$, then the rejection of P_j is equivalent to the rejection of at least one $\tilde{P}_{(i)j}$. The price to be paid for the gained power by Hochberg's method is that it controls the FWER only under independence of the p-values or certain positive dependence structures - same as the conditions for the Simes method. Sarkar and Chang (1997) showed that the Simes method, hence the Hochberg's method, controls the FWER when the multiple p-values come from multivariate test statistics $X_1, ..., X_n$ that have joint probability density of the form $\int \prod_{i=1}^{n} f(x_i, z)g(z)dz$, for some probability densities $f(x, z)$ and $g(z)$, such that $f(x, z)$ is TP_2 in (x, z) [1]. As Sarkar and Chang (1997) pointed out, statistics with such probability density are positively dependent, and many multivariate distributions in the context of multiple testing problems can be represented in the that form, including common distributions such as the equicorrelated multivariate normal with positive correlation. They also showed that the results hold for larger classes of positive dependence structures such as the positive regression dependence introduced by Lehmann (1966).

Now we shall formally state the two new methods we propose for the multidimensional directional decisions problem. Both follow the structure of the general Procedure 3.1.

Procedure 3.4.

(1) Based on the Holm's method, let $P_j = q\tilde{P}_{(1)j}$ for $j = 1, ..., m$. Apply the BH method at level α on $P_1, ..., P_m$ to test for $H_{01}, ..., H_{0m}$ simultaneously. Let R be the number of rejected hypothesis out of m.

(2) For every $j = 1, ..., m$, let $R_j = \max\{1 \leq i \leq q : \tilde{P}_{ij} \leq R\alpha\{m(q+1-$

[1] A nonnegative bivariate function $f(x, y)$ is said to be TP_2 in (x, y) if $f(x, y)f(x', y') \geq f(x, y')f(x', y)$ for all $x < x'$ and $y < y'$.

$l)\}^{-1}$, for $l = 1, ..., i\}$, if the maximum exists; otherwise $R_j = 0$. For every i and j with $\tilde{P}_{ij} \leq R\alpha\{m(q+1-R_j)\}^{-1}$ (or equivalently $\tilde{P}_{ij} \leq \tilde{P}_{(R_j)j}$), declare $\delta_{ij} > 0$ or < 0 according to the signs of the corresponding test statistics.

Procedure 3.5.

(1) Based on the Hochberg's method, let $P_j = \min_{1 \leq i \leq q}\{(q+1-i)\tilde{P}_{(i)j}\}$ for $j = 1, ..., m$. Apply the BH method at level α on $P_1, ..., P_m$ to test for $H_{01}, ..., H_{0m}$ simultaneously. Let R be the number of rejected hypothesis out of m.

(2) For every $j = 1, ..., m$, let $R_j = \max\{1 \leq i \leq q : \tilde{P}_{(i)j} \leq R\alpha\{m(q+1-i)\}^{-1}\}$, if the maximum exists; otherwise $R_j = 0$. For every i and j with $\tilde{P}_{ij} \leq R\alpha\{m(q+1-R_j)\}^{-1}$ (or equivalently $\tilde{P}_{ij} \leq \tilde{P}_{(R_j)j}$), declare $\delta_{ij} > 0$ or < 0 according to the signs of the corresponding test statistics.

We have stated the second step of Procedures 3.4 and 3.5 so that the individual hypotheses are tested for every gene j. This is in accordance with Procedures 3.2 and 3.3 in Guo *et al.* (2010). As we have discussed before, by construction of the procedures, it is equivalent if we only test for the genes found significant in the first step, in accordance with the general Procedure 3.1. By Theorem 3.1, Corollary 3.1 and theory for Holm's method and Hochberg's method, we have the following results:

Corollary 3.2. *Procedure 3.4 controls the OFDR and mdFDR under independence between the genes, and without any assumptions about the individual hypotheses within a gene.*

Corollary 3.3. *Procedure 3.5 controls the OFDR and mdFDR under independence between the genes and also independence (or certain positive dependence structures as required by Hochberg's method) between the individual hypotheses within a gene.*

Regarding the power of these procedures, since Hochberg's method is uniformly more powerful than Holm's method, which is uniformly more powerful than Bonforroni's method, we have the following result:

Corollary 3.4. *Procedure 3.5 is at least as powerful as Procedure 3.4, which is at least as powerful as Procedure 3.2.*

3.3.3 Simulation studies

We conducted simulation studies to compare the new Procedures 3.4 and 3.5 with Procedure 3.2 from the original paper by Guo *et al.* (2010). We did not include Procedure 3.3 in the simulations since it does not control the mdFDR. We do include a one-step BH method in the study, to illustrate the differences between a simple one-step method and the two-step methods. The one-step BH method simply applies the BH procedure at level α to all the $m \times q$ $\tilde{P}_{ij}$'s at the same time.

Before we proceed to the simulations studies, we point out two interesting properties of the simple BH method in comparison with Procedures 3.2 and 3.4. For the simple BH method, define a gene j to be rejected, if at least one of its components $\tilde{P}_{ij}$ is rejected.

Proposition 3.1. *The simple BH method rejects at least as many genes as Procedure 3.2 or 3.4.*

Note that Procedure 3.2 or 3.4 rejects exactly the same genes since their first step is actually identical. To prove Proposition 3.1, suppose $\tilde{P}_{(1)j}$ is the smallest out of the $\tilde{P}_{ij}$'s for each fixed j. Let $\tilde{P}_{(1)(1)} \leq \tilde{P}_{(1)(2)} \leq \cdots \leq \tilde{P}_{(1)(m)}$ be the ordered version of $\tilde{P}_{(1)1}, \tilde{P}_{(1)2}, ..., \tilde{P}_{(1)m}$. For Procedures 3.2 and 3.4, the p-value in the first step for testing each gene is $P_j = q\tilde{P}_{(1)j}$. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered versions of $P_1, ..., P_m$. Suppose Procedures 3.2 and 3.4 reject R genes out of m in the first step. This leads to the fact that $P_{(R)} = q\tilde{P}_{(1)(R)} \leq \frac{R}{m}\alpha$. Now, in order to implement the simple BH method, all the $\tilde{P}_{ij}$'s are ranked together from 1 to $m \times q$. Specifically, suppose $\tilde{P}_{(1)(R)}$ is ranked number K in all $\tilde{P}_{ij}$'s. Then, the simple BH method will reject $\tilde{P}_{(1)(R)}$ if $\tilde{P}_{(1)(R)} \leq \frac{K}{qm}\alpha$. Now, since $\tilde{P}_{(1)(R)}$ is ranked number R among $\tilde{P}_{(1)1}, \tilde{P}_{(1)2}, ..., \tilde{P}_{(1)m}$, we know that we must have $K \geq R$. We also know that $q\tilde{P}_{(1)(R)} \leq \frac{R}{m}\alpha$, which leads to $\tilde{P}_{(1)(R)} \leq \frac{R}{qm}\alpha \leq \frac{K}{qm}\alpha$. Thus we have proved that if a gene is rejected by Procedure 3.2 or 3.4, then it is guaranteed to be rejected by the simple BH method.

To bring this a step further, we also have the following result:

Proposition 3.2. *The simple BH method rejects as least as many individual hypotheses as Procedure 3.2.*

To prove Proposition 3.2, remember that Procedure 3.2 rejects any $\tilde{P}_{ij}$ such that $\tilde{P}_{ij} \leq \frac{R}{qm}\alpha$, where R is the number of genes rejected in the first step. We have just shown that the number of genes rejected by the simple BH method, say R_g, is at least R. It is also apparent that the number of individual hypotheses rejected by the simple BH method, say R_{BH}, is at least R_g. Any $\tilde{P}_{ij}$ is rejected by the simple BH method if $\tilde{P}_{ij} \leq \frac{R_{BH}}{qm}\alpha$. Since $R \leq R_{BH}$, the cutoff for rejecting individual hypotheses by Procedure 3.2 is no larger than the cutoff by the simple BH method. Thus we have proved that the simple BH method rejects at least as many individual hypotheses as Procedure 3.2.

These interesting properties may lead us to think that the simple BH method is more powerful than Procedure 3.2 or even Procedure 3.4. However, before we start comparing the power, it is crucial to first ensure the control of the mdFDR. The simple BH method obviously controls the FDR with respect to the individual hypotheses. However, it is not clear to us as to how it performs with respect to the mdFDR (or OFDR). We shall investigate this in the following simulation studies.

We shall follow the main simulation study set up in Guo *et al.* (2010). We simulate a setting with $m = 1000$ genes, and 6 time points. The gene expression vectors $\mathbf{Z}_i$ ($i = 1, ..., 6$) for the 6 time points are simulated from independent m-dimensional multivariate normal distributions, where $Z_{ij} \sim N(\mu_{ij}, 1)$ ($j = 1, ..., m$) and have a common correlation ρ. ρ is set to be 0, 0.2, 0.5 or 0.8 for four separate simulation studies respectively. Thus expressions from different time points are independent, but expressions could be positively correlated across genes. Let $\delta_{ij} = (\mu_{i+1,j} - \mu_{ij})/\sqrt{2}$ for $i = 1, ..., 5$. Out of the m $\boldsymbol{\delta}_j$'s, m_0 were set to a zero vector, and the δ_{ij}'s in 50%, 25% and 25% of the remaining $m - m_0$ $\boldsymbol{\delta}_j$'s were randomly generated (uniformly) from the intervals $(-0.75, 0.75)$, $(-4.25, -2.75)$ and $(2.75, 4.25)$ respectively. The test statistic for testing each δ_{ij} is $T_{ij} = (Z_{i+1,j} - Z_{ij})/\sqrt{2}$ and the corresponding p-value is computed by $\tilde{P}_{ij} = 2\{1 - \Phi(|Tij|)\}$, where $\Phi(\cdot)$ is the standard normal CDF. Simulation results are averaged across 1000 replications. The level α is set to be 0.05. Notice that even though theory on all the procedures were developed under independence between genes, we also investigate the cases where genes are positively correlated in our simulation studies.

Figure 3.1 shows the control of the mdFDR by Procedures 3.2 (Bonferroni), 3.4 (Holm) and 3.5 (Hochberg). The three kinds of FDR shown in the plots are: the original FDR (with respect to the screening hypotheses), the pure directional

FDR, and the mixed directional FDR (overall FDR). This is in accordance with the definition of the mdFDR in Guo *et al.* (2010). As shown in the four plots in Figure 3.1, the FDR and the dFDR add up to the mdFDR. Notice that Figure 3.1 shows that the mdFDR is not only controlled at the desired level ($\alpha = 0.05$) under independence between genes, it is also controlled under positive correlation between genes. In fact, it seems that the mdFDR is controlled at an even lower level when the correlation between genes is high. Results from the three procedures are plotted in different colors. However, we can hardly distinguish the difference in the plots. It seems that all three procedures behave almost identically as far as the control of the FDR, dFDR and mdFDR.

Figure 3.2 shows the behavior of the mdFDR by the simple BH method. In addition to the three kinds of FDR plotted in Figure 3.1, we also plotted the FDR with respect to all the $m \times q$ individual hypotheses. Not surprisingly, the FDR with respect to the individual hypotheses is controlled by the simple BH method at the desired level ($\alpha = 0.05$). However, Figure 3.2 shows that the simple BH method is not able to control the mdFDR at the desired level, as we had suspected earlier. As we can see from the plots, this is mainly because the simple BH method fails to control the FDR with respect the screening hypothesis (or genes), possibly a result from rejecting too many genes (as discussed earlier).

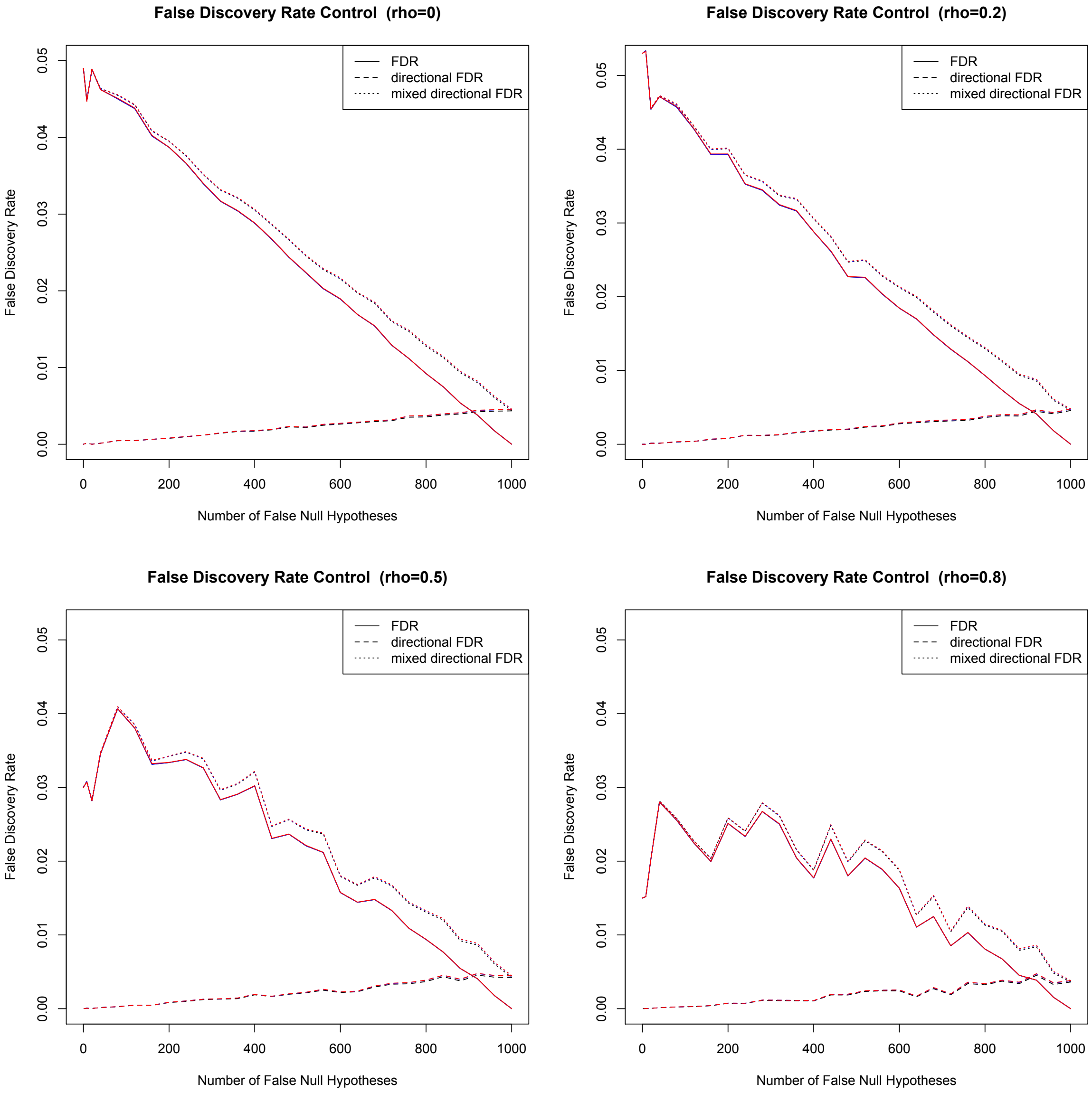

Figure 3.1. Plots showing the control of the mdFDR for Procedures 3.2, 3.4 and 3.5. The four plots correspond to the four different correlation settings. The x-axis is $m - m_0$. Procedures 3.2 (Bonferroni), 3.4 (Holm) and 3.5 (Hochberg) are plotted in black, blue and red respectively (but they are essentially indistinguishable in the plots).

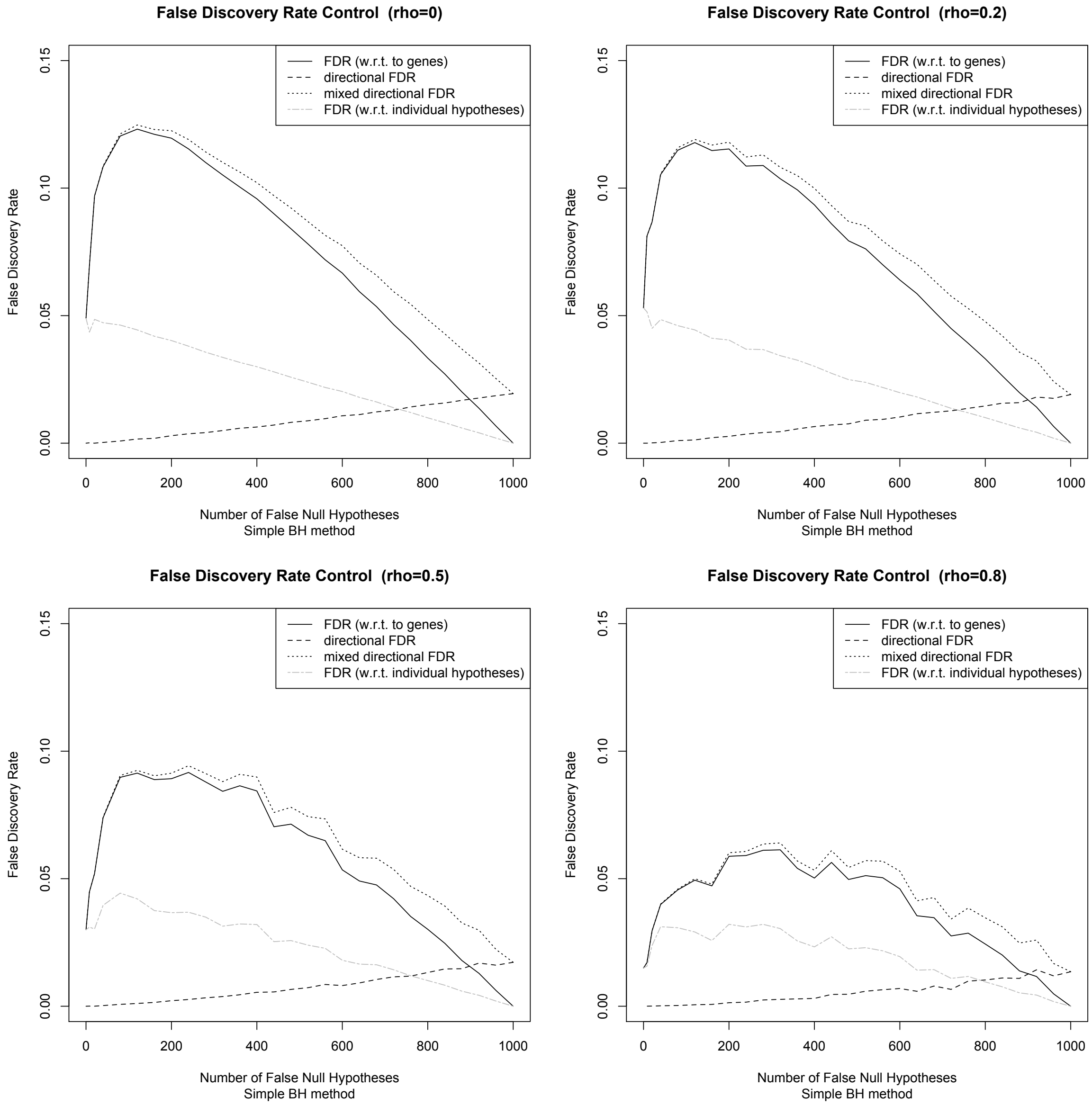

Figure 3.2. Plots showing the control (or the lack thereof) of the mdFDR for the simple BH method. The four plots correspond to the four different correlation settings. The x-axis is $m - m_0$.

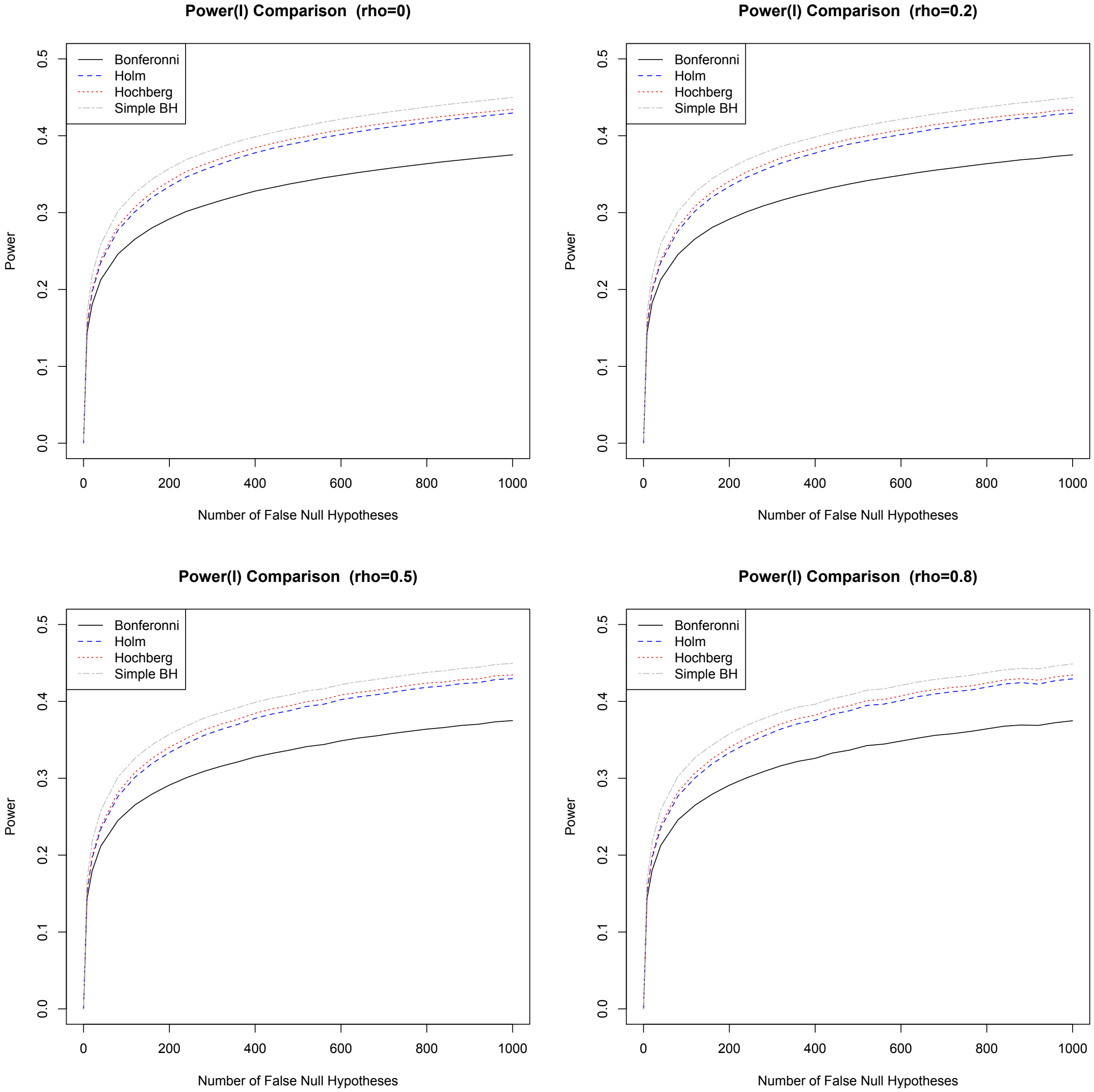

Figure 3.3. Plots showing the power with respect to the individual hypotheses. The four plots correspond to the four different correlation settings. The x-axis is $m - m_0$.

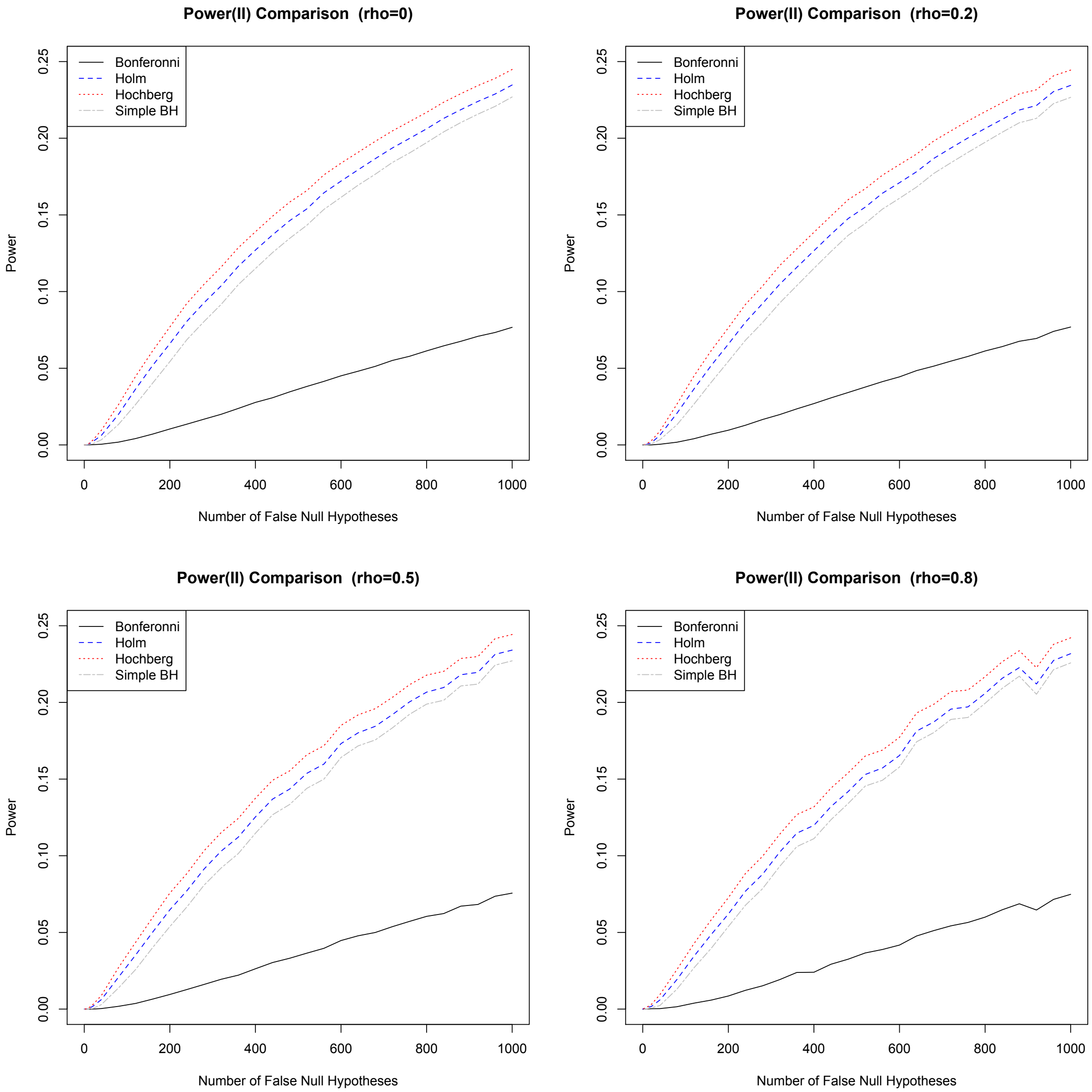

Figure 3.4. Plots showing the power with respect to the genes. The four plots correspond to the four different correlation settings. The x-axis is $m - m_0$.

Simulation results shown in Figure 3.1 and Figure 3.2 are overall not surprising, since most of them are in accordance with theory about the procedures' control of the different kinds of FDR's. What would be more interesting is the comparison of power for the different methods. Because of the multidimensional nature of the problem, there are several ways in which we can define "power". Guo *et al.* (2010) did not explicitly mention their definition of power that they used in there simulations. However, through comparison of simulation results, it seems that they defined power as the number of correct directional decisions out of all the non-zero δ_{ij}'s. Basically, it is the power with respect to the individual hypotheses. We shall name this definition "power (I)". However, since we are interested in controlling the overall FDR with respect to the genes, it makes sense for us to also examine the power with respect to the genes. Due to the fact that we are testing multiple hypotheses for each gene, there are many options as to how we can define the "gene-wise" power. Recall that for the OFDR, we define a gene to be falsely rejected if we make any false rejections out of all the hypotheses tested for that gene. We adopt a similar approach to defining the gene-wise power. We define a gene to be correctly rejected, if and only if the decisions on each hypotheses for that gene is correct. Of course, it only makes sense to look at the genes for which the screening hypothesis is indeed false (i.e. genes that do have expression pattern). This means that in order for a non-null gene to be correctly rejected, we need to: (1) correctly reject the screening hypothesis; (2) correctly reject all the individual hypotheses that are indeed false; (3) not reject any individual hypotheses for which the null is indeed true. We shall name this definition "power (II)".

Figures 3.3 and 3.4 respectively show the comparisons of the two definitions of power for the different methods. Since the denominators in the two definitions of power are different, Figures 3.3 and 3.4 are not directly comparable against each other, but it is still worth noting that power (II) (Figure 3.4) is overall much lower than power (1) (Figure 3.3) for any method. This is not surprising, because power (II) is a somewhat more stringent measure, since it requires making "absolutely" correct decisions on the genes. In both Figures 3.3 and 3.4, we can see that Procedure 3.2 (Bonferroni) has considerably less power than Procedures 3.4 (Holm) and 3.5 (Hochberg) that we proposed. Procedures 3.4 (Holm) and 3.5 (Hochberg) perform relatively similarly, with Procedure 3.5 (Hochberg) having slightly higher power than Procedure 3.4 (Holm). The behavior of the simple BH method is interesting. For power (I) (Figure 3.3), it performs the best out of all four methods. This seems to concord with the discussion we had earlier about the simple BH method being able to reject as least as many genes as Procedures 3.2 and 3.4 and at least as many individual hypotheses as Procedure 3.2. But remember that this comes at the price of not being able to control the mdFDR as shown in Figure 3.2. For power (II) (Figure 3.3), the simple BH method still tops Procedure 3.2 (Bonferroni) but is not as good as Procedures 3.4 (Holm) and 3.5 (Hochberg). This concords with previous discussions and observations that the simple BH method might be rejecting more hypotheses at the expense of making more mistakes - and for power (II), any mistake within a gene will rule the gene out from the correctly rejected genes.

In summary, simulation studies show that our newly proposed Procedures 3.4 and 3.5 do indeed control the mdFDR as claimed under independence between genes, and also appears to control the mdFDR under positive dependence between genes. Our new procedures show considerably higher power than the original procedure in Guo *et al.* (2010). Procedure 3.5 has the highest power out of the three two-step procedures, while Procedure 3.4 is a close second. If the dependencies between the individual hypotheses within a gene is unknown, then it is recommended to use Procedure 3.4, since it does not impose any restrictions on the dependency structures and has very little loss of power compared to Procedure 3.5, which does require certain dependency structures between the individual hypotheses in order to control the mdFDR.

Thus far, we have successfully applied our two-step hierarchical hypothesis set testing framework on the multidimensional directional decisions problem, and have used the theory of our general two-step testing procedure to construct new procedures for the problem that control the mixed directional FDR at the desired level while achieving higher power. Following this general framework and procedure, other potentially more powerful new procedures can be constructed for different dependency structures among the individual hypotheses or other special cases, as long as the procedure is based on multiple testing methods that control the FWER.

3.4 Application to Multiple Comparisons for a Large Number of Tests

3.4.1 Problem setting and testing procedures

Another rather straightforward problem setting where the two-step hierarchical testing idea can be applied is the analysis of variance (ANOVA) followed by multiple comparisons between the treatments. Multiple comparison procedures as followup tests for the ANOVA have been studied extensively, but when the problem is put under the grand scheme of large-scale hypotheses testing (as with genomic data), new challenges arise. Here, our focus is not on the multiple comparisons problem after performing a single ANOVA, rather, we want to investigate procedures that perform a large number of tests of ANOVA followed by multiple comparisons, and where the number of tests performed is far greater than the number of multiple comparisons in each ANOVA.

This problem of multiple comparisons for a large number of tests is encountered very frequently in analysis of genomic data (such as microarray data). Often, scientists will want to investigate differential expression across several treatment groups. If there are only two treatment groups, then a simple two-sample t-test would suffice for each gene, and multiple testing adjustments, such as the BH procedure, can be directly applied to the p-values across all the genes. When more than two treatment groups are present, the problem becomes a bit more complicated. We can easily apply an ANOVA to each gene followed by multiple comparisons for the treatments groups, however, it is not straightforward as to how we can apply adjustments for multiple testing across the genes since we are performing more than one test for each gene.

Jiang and Doerge (2006) considered this problem and proposed a two-step procedure. The idea is to select the genes for which the treatment effect is significant in the first step, and perform pairwise comparisons between the treatments in the second step only for the significant genes. More specifically, suppose there are m genes and C pairwise comparisons for each gene (e.g. in the case of four treatment groups, $C = 6$). Their goal is to control the FDR over the $m \times C$ tests. They called this the “overall FDR”. However, to avoid confusion with the OFDR we have defined earlier, we shall call their "overall FDR" the "final FDR". In their paper, they discussed using fixed rejection regions in the two steps respectively and estimating the final FDR, as well as using FDR controlling procedures for the two steps respectively and estimating the final FDR. They also made recommendations on the levels to use for FDR controlling in the two steps in order to control the final FDR at a certain level. Since we hope to control some type of FDR in the end, we mainly focus on their procedure where they apply FDR controlling procedures in the two steps at certain levels to control the final FDR. Further discussions and comparisons would be based on this procedure in Jiang and Doerge (2006) that we restate as Procedure 3.6 below.

Similar to the notations in the last section, let p_j $(j = 1, ..., m)$ denote the p-value for the global F-test of treatment effect for gene j. Following Jiang and Doerge (2006), pairwise comparisons for a given gene are performed by t-tests between each pair of treatment groups. Let p_{ij} denote the p-value for the ith pairwise comparison $(i = 1, ..., C)$ for gene j.

Procedure 3.6. *(Jiang and Doerge, 2006)*

(1) Apply an FDR controlling procedure at level α_1 on the m p_i's corresponding to the global F-tests for the m genes. Let R be the number of significant genes out of m.

(2) For the R significant genes from the first step, perform C pairwise comparisons for each gene. Apply an FDR controlling procedure at level α_2 on the $C \times R$ p_{ij}'s, and declare the significance of pairwise comparisons accordingly. Pairwise comparisons for the in-significant genes are concluded to be in-significant.

They showed through simulation studies that if the sum of α_1 and α_2 is no greater than 0.05, then the final FDR is approximately controlled at the 0.05 level. They recommended combinations such as $\alpha_1 = 0.04$, $\alpha_2 = 0.01$ or $\alpha_1 = 0.03$, $\alpha_2 = 0.02$ for the two steps.

It is not hard to see that our general two-step hypothesis set testing framework is suited for this problem as well. Using terminology from the general framework, tests performed on each gene are a hypothesis set. For each gene, the global F-test for the ANOVA serves as the screening hypothesis. Multiple comparisons would be performed if and only if the F-test for that gene is significant. The multiple comparison tests between each pair of treatment groups are the individual hypotheses in the set. As discussed in Section 3.1, this setting falls under scenario (ii), where the screening hypothesis serves as some kind of overall summary of the individual hypotheses. This scenario is a little bit different from the multidimensional directional decisions problem discussed in Section 3.3. In that problem, the screening null hypothesis is the conjunction of the individual null hypotheses. Here, the screening hypothesis is related to the individual hypotheses, but not as directly. As a consequence, it is not essential that the consonant property is conserved - that is, it is not unreasonable to have situations where we reject a gene (claiming there is difference among the treatment groups), yet do not find any pairs of treatments significantly different.

Following the general Procedure 3.1, we propose a new two-step procedure for this problem based on the Holm's method, similar to Procedure 3.4.

Procedure 3.7.

(1) Apply an FDR controlling procedure at level α on the m p_i's corresponding to the global F-tests for the m genes. Let R be the number of significant genes out of m.

(2) For the R significant genes from the first step, perform C pairwise comparisons for each gene. For each gene j that is significant in step one, let $R_j = \max\{1 \leq i \leq C : \; p_{(l)j} \leq R\alpha\{m(C+1-l)\}^{-1}, \text{ for } l = 1, ..., i\}$, if the maximum exists; otherwise $R_j = 0$. For $p_{ij} \leq p_{(R_j)j}$, conclude that the ith pairwise comparison for gene j is significant. Pairwise comparisons for the in-significant genes are concluded to be in-significant.

By Theorem 3.1, Procedure 3.7 controls the OFDR at level α under independence between the genes.

Both Procedures 3.6 and 3.7 are two-step procedures where the first step focuses on finding the significant genes, and the second step proceeds with pairwise comparisons for the significant genes. In fact, the first steps of the two procedures are very similar - both use FDR controlling procedures to adjust for multiple testing. However, their are many important differences between the two procedures.

In step one, we can see that the level at which the FDR is controlled is different for the two procedures. Procedure 3.7 directly uses the level α, which is the level at which we wish to control the overall FDR (OFDR) in the end; Procedure 3.6 uses level α_1, which needs to be smaller than α, if we wish to control the final FDR under level α. Step two is where the most crucial differences between the two procedures lie. Procedure 3.6 applies an FDR controlling procedure simultaneously on the pairwise comparisons of all the significant genes; Procedure 3.7 applies a FWER controlling procedure on the pairwise comparisons for each significant gene respectively. In a way, step two of the two procedures are completely different - they differ in the criteria of multiple testing adjustments (FDR controlling or FWER controlling), and they also differ on how the multiple testing adjustments are applied (for each gene respectively or for all genes simultaneously).

There is a good reason why Procedures 3.6 and 3.7 differ so much. Despite addressing the same problem, the two procedures were developed with very different goals in mind. While both procedures aim at controlling some type of FDR in the end, the final FDR in Jiang and Doerge (2006) and the OFDR in our general framework are completely different concepts. Using terminology from our general framework, the final FDR in Jiang and Doerge (2006) is the FDR with respect to all the individual hypotheses for all the genes. That is, it is the expected proportion of falsely rejected pairwise comparisons out of all the rejected pairwise comparisons. This concept of FDR is no different than the one we would consider if we just used a simple one-step method by applying the BH procedure on all the $m \times C$ pairwise comparisons simultaneously. On the other hand, the OFDR that we wish to control is with respect to the genes. That is, it is the expected proportion of falsely rejected genes out of all the rejected genes, where a gene is falsely rejected if at least one hypothesis for that gene is falsely rejected (see Definition 3.1). The concept of OFDR makes sense for this problem setting, because ultimately we want to be confident about the conclusions we make on the genes, despite the multiple pairwise comparisons associated with each gene.

Another potential difference between Procedures 3.6 and 3.7 lies in the FDR controlling methods used the two procedures. In Procedures 3.6 and 3.7, we did not explicitly state what FDR controlling procedures to use. For Procedure 3.7 that we proposed, the default FDR controlling procedure to use in step one is the BH procedure - similar to Procedures 3.4 and 3.5 in Section 3.3. However, for both steps in Procedure 3.6, Jiang and Doerge (2006) used adaptive BH procedures that utilize the estimation of the proportion of true nulls (i.e., m_0/m, in the case of step one). The adaptive BH procedures are more powerful than the original BH procedure, but there is a lack of formal theory on the control of the FDR by the adaptive procedures, since the accuracy of the estimated proportion of true nulls cannot be guaranteed. Nonetheless, when the proportion of true nulls is in fact small, the gain of power by using the adaptive BH procedures can be huge. The main reason behind this is that the original BH procedure actually controls the FDR at the level $m_0\alpha/m \leq \alpha$. When m_0/m is small, the BH procedure is in fact very conservative. The idea behind the adaptive BH procedures is essentially to apply the BH procedure at level $m_0\alpha/m$, so that the FDR is controlled at $\frac{m_0}{m}\frac{m}{m_0}\alpha = \alpha$ - but since the true value of m_0 is unknown, they use an estimation. Further discussions on the effects of using the adaptive BH method versus the original BH method will be presented in the next section.

3.4.2 Simulation studies

We conducted simulation studies to compare Procedures 3.6 and 3.7. The simulation scheme follows that of Jiang and Doerge (2006). We simulate a setting of $m = 1000$ genes, and 3 treatment groups, resulting in $C = 3$ pairwise comparisons among the treatment groups. Let R_1 be the proportion of genes for which there is a treatment effect. Among the genes with treatment effects, let R_2 be the proportion of genes that are not differentially expressed between two treatments, but differentially expressed under the third treatment; the rest of the genes with treatment effects are differentially expressed under all three treatments. More specifically, the mean expression for the three treatments are set to be $(0,0,0)$ for $(1-R_1)\times m$ genes; $(4,2,0)$ for $R_1\times(1-R_2)\times m$ genes; $(4,0,0)$ for $0.5\times R_1\times R_2\times m$ genes; and $(2,0,0)$ for $0.5 \times R_1 \times R_2 \times m$ genes. We consider all combinations of $R_1 = 0.1, 0.2, 0.3, 0.4, 0.5$ and $R_2 = 0, 0.2, 0.4, 0.6, 0.8, 1$. Assume a sample size of 6 for each treatment group. Expression values are generated from normal distributions with standard deviation 1, and means following the scheme described above. Global F-tests from ANOVA are used to test for the overall treatment effects for each gene. T-tests are used to test for the pairwise comparisons for each treatment pair in the genes with significant treatment effects. α is set to be 0.05. α_1 and α_2 in Procedure 3.6 are set to be 0.04 and 0.01 respectively. Final results are averaged over 1000 replications.

Regarding the issue of choosing between the original BH method and the adaptive BH methods, we applied both in our simulation studies. The adaptive BH method used utilizes the estimation of the proportion of true nulls by Storey and Tibshirani's smoother estimate (2003), consistent with that in Jiang and Doerge (2006). Remember that Jiang and Doerge (2006) used the adaptive BH method in their procedures by default, while we used the original BH method in our procedures by default. However, in order for Procedures 3.6 and 3.7 to be more comparable, we either used the original BH method in both procedures or the adaptive BH method in both procedures, and only compare the results from using the same BH method.

Recall that Procedures 3.6 and 3.7 are aimed at controlling two different types of FDR - the FDR with respect to all the $m \times C$ pairwise comparisons, and the OFDR with respect to the m genes. For a comprehensive comparison of the two procedures, we present results for the two types of FDR for both procedures. As to the comparison of power, we again run into the issue of different ways of defining power. We follow the discussions in Section 3.3.3 regarding this issue, and again adopt two definitions of power, namely power (I) - the power with respect to all the pairwise comparisons, and power (II) - the power with respect to genes.

Let's first take a look at the results from using the adaptive BH method for both procedures, which are summarized in Tables 3.1-3.4. Table 3.1 shows the FDR with respect to all the pairwise comparisons for both procedures and all combinations of R_1 and R_2. This FDR is the one that Jiang and Doerge (2006) are trying to control. As seen from the table, Procedure 3.6 by Jiang and Doerge (2006) indeed controls the FDR at the desired level (0.05) for all different settings of R_1 and R_2. On the other hand, our proposed Procedure 3.7 also controls this FDR, but at a much lower level. This indicates that Procedure 3.7 might be more conservative than Procedure 3.6 in rejecting individual pairwise comparisons. Table 3.2 shows the OFDR with respect to the genes for the two procedures. Overall, neither procedure is able to strictly control the OFDR at the 0.05 level. This is not very surprising for Procedure 3.6, since it was not designed to control the OFDR. For Procedure 3.7, failure to control the OFDR is due to the use of the adaptive BH method - which does not guarantee stringent control of the FDR - instead of the original BH method.

Interestingly, If we compare Tables 3.1 and 3.2, we can see that the two procedures produce similar OFDR's, but the FDR's are very different. More specifically, for Procedure 3.6, the OFDR did not increase much compared to the FDR; but for Procedure 3.7, the FDR is drastically lower than the OFDR. Looking back at the two procedure, we see that the first steps of the two procedures are very similar. In fact, the only difference in the first step is the level at which the BH method is applied - Procedure 3.6 uses a slightly more conservative level than Procedure 3.7 in the first step. As a results, the two procedures end up with a very similar set of significant genes from the first step, with Procedure 3.7 picking out a few more. Since the two procedures reject similar genes in the first step, they end up with similar OFDR's. However, there is a drastic difference between the two procedures as to how they select the individual pairwise comparisons. Notice that if the treatment effect is truly significant for a gene, then at least two out of the three pairwise comparisons are truly significant. Also remember that for both procedures, pairwise comparisons are only tested for the significant genes. This means that if most of the genes that were picked out in the first step are correct, then most pairwise comparisons we test for in the second step are truly significant. When the proportion of true nulls is low, the adaptive BH method has a great advantage - it can be a lot more powerful than the original BH method, while still maintaining reasonable control of the FDR. Procedure 3.6 uses the adaptive BH method in its second step for selecting the pairwise comparisons. As a result, Procedure 3.6 rejects a lot more pairwise comparisons than Procedure 3.7 in the second step, despite picking out a few less genes in the first step. This leads to a much higher FDR with respect to pairwise comparisons for Procedure 3.6 compared to Procedure 3.7. However, this will probably also lead to higher power with respect to the pairwise comparisons for Procedure 3.6, as we shall see in Tables 3.3 and 3.4.

Table 3.1. Simulation results showing the FDR with respect to all the pairwise comparisons. (Procedures are implemented using the adaptive BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0309	0.0413	0.0411	0.0391	0.0387	0.0364
0.2	0.0291	0.0405	0.0367	0.0356	0.0341	0.0321
0.3	0.0270	0.0379	0.0342	0.0324	0.0302	0.0268
0.4	0.0261	0.0359	0.0319	0.0282	0.0251	0.0224
0.5	0.0251	0.0344	0.0295	0.0247	0.0211	0.0187

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0066	0.0069	0.0070	0.0072	0.0069	0.0071
0.2	0.0061	0.0066	0.0078	0.0085	0.0092	0.0101
0.3	0.0057	0.0066	0.0080	0.0093	0.0109	0.0123
0.4	0.0053	0.0067	0.0080	0.0098	0.0120	0.0145
0.5	0.0046	0.0064	0.0084	0.0105	0.0134	0.0167

Table 3.2. Simulation results showing the OFDR with respect to the genes. (Procedures are implemented using the adaptive BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0494	0.0642	0.0675	0.0640	0.0611	0.0556
0.2	0.0453	0.0672	0.0631	0.0612	0.0574	0.0553
0.3	0.0431	0.0654	0.0616	0.0593	0.0570	0.0544
0.4	0.0429	0.0638	0.0601	0.0570	0.0557	0.0539
0.5	0.0428	0.0636	0.0597	0.0575	0.0554	0.0537

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0536	0.0510	0.0510	0.0512	0.0531	0.0517
0.2	0.0532	0.0506	0.0529	0.0545	0.0545	0.0551
0.3	0.0518	0.0526	0.0538	0.0560	0.0578	0.0589
0.4	0.0524	0.0532	0.0549	0.0566	0.0601	0.0627
0.5	0.0523	0.0541	0.0568	0.0602	0.0642	0.0672

Table 3.3. Simulation results showing power (I) with respect to all the pairwise comparisons. (Procedures are implemented using the adaptive BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.9779	0.9384	0.8719	0.7941	0.7216	0.6507
0.2	0.9775	0.9532	0.8914	0.8311	0.7807	0.7368
0.3	0.9770	0.9583	0.9028	0.8518	0.8121	0.7811
0.4	0.9761	0.9603	0.9106	0.8669	0.8311	0.8033
0.5	0.9754	0.9636	0.9175	0.8770	0.8422	0.8184

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.5662	0.5620	0.5574	0.5547	0.5540	0.5543
0.2	0.6727	0.6659	0.6615	0.6561	0.6518	0.6456
0.3	0.7367	0.7304	0.7244	0.7163	0.7097	0.7033
0.4	0.7809	0.7744	0.7675	0.7602	0.7522	0.7428
0.5	0.8143	0.8076	0.8015	0.7945	0.7859	0.7768

Table 3.4. Simulation results showing power (II) with respect to the genes. (Procedures are implemented using the adaptive BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.9352	0.8646	0.7658	0.6811	0.6372	0.6266
0.2	0.9328	0.8847	0.7856	0.7167	0.6941	0.7013
0.3	0.9313	0.8905	0.7948	0.7347	0.7204	0.7364
0.4	0.9285	0.8901	0.8004	0.7495	0.7349	0.7498
0.5	0.9263	0.8935	0.8059	0.7563	0.7405	0.7571

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.1770	0.2426	0.3089	0.3772	0.4454	0.5124
0.2	0.3126	0.3643	0.4220	0.4785	0.5390	0.5976
0.3	0.4115	0.4563	0.5024	0.5477	0.5971	0.6486
0.4	0.4905	0.5250	0.5631	0.6028	0.6418	0.6826
0.5	0.5538	0.5818	0.6130	0.6457	0.6774	0.7110

Tables 3.3 and 3.4 show the results for power (I) and (II) for the two procedures using the adaptive BH method. For both definitions of power, Procedure 3.6 is consistently more powerful than Procedure 3.7 for all combinations of R_1 and R_2. As discussed in the last paragraph, this is due to the advantage of using the adaptive BH method for the second step in Procedure 3.6.

Next let's look at the results from using the original BH method for both procedures, which are shown in Tables 3.5-3.8. Before comparing the results between Procedures 3.6 and 3.7, let's first take a look at how switching from the adaptive BH method to original BH method affects the two procedures respectively. Comparing Tables 3.5-3.8 with Tables 3.1-3.4, we see that the results for Procedure 3.7 does not change much. This shows that the adaptive BH method did not assert much influence on the first step of gene selection compared to the original BH method. This is possibly because the significant genes and non-significant genes are relatively well separated in this simulation study. However, the results for Procedure 3.6 are heavily influenced by the BH method we choose to use. The FDR with respect to the individual pairwise comparisons is drastically lower when the original BH method is used, while the power is also much lower compared to using the adaptive BH method. This agrees with the discussion we had earlier about how the adaptive BH method results in rejecting many more pairwise comparisons in the second step of Procedure 3.6.

Now, comparing the two procedures, we see from Tables 3.5 and 3.6 that both procedures control the FDR as well as OFDR at the 0.05 level. For both definitions of power, we see from Tables 3.7 and 3.8 that Procedure 3.6 has slightly higher power when R_1 is small, while Procedure 3.7 has higher power when R_1 gets larger.

In summary, for this particular problem and simulation setting, the comparisons between Procedure 3.6 by Jiang and Doerge (2006) and our Procedure 3.7 largely depends on the type of BH method used in the procedures, as well as what FDR criteria we wish to use. Procedure 3.7 has the advantage of proved control of the OFDR, and also higher power under certain parameter settings, when the original BH method is adopted. Procedure 3.6 seems to work particularly well when the adaptive BH method is adopted, as it was proposed in Jiang and Doerge (2006). However, the control of the FDR is not a strictly proven result for Procedure 3.6, and we have little knowledge on its behavior with respect to the OFDR.

In general, Procedure 3.6 by Jiang and Doerge (2006) seems to be well suited for the problem of pairwise comparisons for a large number of tests. Moreover, the procedure can be used for the general problem of hierarchical multiple hypothesis set testing. It is tempting to think that Procedure 3.6 might be a better method overall compared to procedures developed based on our general Procedure 3.1. However, the success of Procedure 3.6 most probably results from some particular properties of this problem setting. As discussed before, in this problem setting, if a screening hypothesis is indeed false, then at least a certain (usually fairly large) proportion of the corresponding individual hypotheses are also false. This is not the general case for a hierarchical multiple hypothesis set testing problem. Also, Procedure 3.6 relies on the adaptive BH method for better performance, which leaves a question mark on its control of the FDR, especially the OFDR. We shall explore some of these issues in the next section.

Table 3.5. Simulation results showing the FDR with respect to all the pairwise comparisons. (Procedures are implemented using the original BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0138	0.0153	0.0154	0.0164	0.0180	0.0187
0.2	0.0087	0.0090	0.0103	0.0108	0.0118	0.0128
0.3	0.0060	0.0064	0.0071	0.0076	0.0086	0.0093
0.4	0.0043	0.0047	0.0051	0.0057	0.0065	0.0074
0.5	0.0031	0.0036	0.0040	0.0044	0.0051	0.0058

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0064	0.0066	0.0067	0.0068	0.0066	0.0067
0.2	0.0056	0.0062	0.0072	0.0080	0.0085	0.0091
0.3	0.0050	0.0059	0.0071	0.0082	0.0196	0.0109
0.4	0.0045	0.0057	0.0069	0.0084	0.0104	0.0125
0.5	0.0038	0.0053	0.0070	0.0088	0.0112	0.0137

Table 3.6. Simulation results showing the OFDR with respect to the genes. (Procedures are implemented using the original BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0400	0.0371	0.0370	0.0377	0.0401	0.0401
0.2	0.0347	0.0319	0.0341	0.0346	0.0352	0.0362
0.3	0.0294	0.0286	0.0293	0.0301	0.0313	0.0320
0.4	0.0253	0.0249	0.0250	0.0259	0.0271	0.0281
0.5	0.0213	0.0208	0.0214	0.0223	0.0228	0.0238

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.0481	0.0458	0.0457	0.0458	0.0471	0.0461
0.2	0.0427	0.0405	0.0424	0.0442	0.0440	0.0445
0.3	0.0365	0.0366	0.0382	0.0400	0.0420	0.0428
0.4	0.0317	0.0325	0.0340	0.0362	0.0392	0.0413
0.5	0.0269	0.0281	0.0311	0.0340	0.0370	0.0400

Table 3.7. Simulation results showing power (I) with respect to all the pairwise comparisons. (Procedures are implemented using the original BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.6643	0.6586	0.6518	0.6421	0.6287	0.6083
0.2	0.6628	0.6619	0.6641	0.6636	0.6643	0.6619
0.3	0.6628	0.6645	0.6677	0.6711	0.6755	0.6830
0.4	0.6627	0.6653	0.6697	0.6756	0.6820	0.6909
0.5	0.6641	0.6663	0.6714	0.6786	0.6866	0.6979

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.5653	0.5604	0.5547	0.5507	0.5486	0.5464
0.2	0.6710	0.6626	0.6567	0.6489	0.6425	0.6339
0.3	0.7342	0.7252	0.7166	0.7056	0.6957	0.6860
0.4	0.7775	0.7680	0.7578	0.7468	0.7345	0.7207
0.5	0.8105	0.8000	0.7898	0.7787	0.7643	0.7495

Table 3.8. Simulation results showing power (II) with respect to the genes. (Procedures are implemented using the original BH method.)

Procedure 3.6

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.2315	0.3027	0.3731	0.4437	0.5109	0.5752
0.2	0.2293	0.3040	0.3852	0.4608	0.5398	0.6154
0.3	0.2274	0.3059	0.3854	0.4654	0.5460	0.6283
0.4	0.2286	0.3054	0.3872	0.4682	0.5485	0.6305
0.5	0.2294	0.3068	0.3870	0.4692	0.5507	0.6337

Procedure 3.7

R_1	$R_2 = 0.0$	0.2	0.4	0.6	0.8	1.0
0.1	0.1760	0.2410	0.3063	0.3737	0.4408	0.5061
0.2	0.3100	0.3603	0.4171	0.4719	0.5314	0.5891
0.3	0.4073	0.4495	0.4938	0.5377	0.5860	0.6369
0.4	0.4843	0.5164	0.5525	0.5904	0.6280	0.6684
0.5	0.5464	0.5712	0.6002	0.6311	0.6611	0.6947

3.5 An Integrated Simulation Study

As discussed at the end of the last section, since Procedure 3.6 by Jiang and Doerge (2006) can also be generalized to be applied to any problem under the framework of hierarchical hypothesis set testing, it would interesting to compare Procedure 3.6 with our procedures in some settings other than the one discussed in Section 3.4.

As an example, Procedure 3.6 can be modified to fit the problem of multidimensional directional decisions, as described in Section 3.3. We simply need to change the global F-test in step one to a test of the conjunction of nulls for the individual hypotheses. The test of the conjunction of nulls can be based on either the Bonferroni, Holm or Hochberg method, as in Procedures 3.2, 3.4 and 3.5. Since Hochberg's method is the most powerful out of the three, we might as well adopt the Hochberg's method. The following procedure is a modified version of Procedure 3.6 for the multidimensional directional decisions problem, with the first step based on Hochberg's method. Since Jiang and Doerge (2006) originally intended for the adaptive BH method to be used in their procedures, we shall use that in this modified procedure. The notations follow from Section 3.3.

Procedure 3.8.

(1) Based on the Hochberg's method, let $P_j = \min_{1\le i\le q}\{(q+1-i)\tilde{P}_{(i)j}\}$ for $j = 1, ..., m$. Apply the adaptive BH procedure at level α_1 on $P_1, ..., P_m$ to test for $H_{01}, ..., H_{0m}$ simultaneously. Let R be the number of rejected hypothesis out of m.

(2) For the R significant genes from the first step, apply the adaptive BH procedure at level α_2 on the $q \times R$ $\tilde{P}_{ij}$'s, and declare the significance of the δ_{ij}'s accordingly. The signs of the δ_{ij}'s are decided declared according to the signs of the corresponding test statistics. δ_{ij}'s for the in-significant genes are concluded to be not significantly different from 0.

As mentioned in Section 3.4, the recommended values for α_1 and α_2 when $\alpha = 0.05$ are $(0.04, 0.01)$ or $(0.03, 0.02)$. We shall adopt the combination of $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$.

We would like to investigate how the procedures compare under a situation where only a small number of individual hypotheses are truly significant when the corresponding screening hypothesis is truly significant. Guo *et al.* (2010) actually talked about this particular setting under the framework of the multidimensional directional decisions problem. In fact, this is the setting where their proposed procedure based on the Simes test (Procedure 3.3) fails to control the mdFDR (or OFDR). So for this section, we shall return to the problem setting of Section 3.3, and adopt this particular simulation setting in Guo *et al.* (2010) to do an integrated simulation study, comparing Procedure 3.2 by Guo *et al.*, (2010), our Procedures 3.4 and 3.5, and Procedure 3.8 modified from Jiang and Doerge (2006), as well as the simple BH method.

The simulation scheme is similar to that described in Section 3.3.3. We again simulate $m = 1000$ genes and 6 time points. The correlation between genes, ρ, is set to be 0. The main difference is in the δ_{ij}'s. In Section 3.3.3, all δ_{ij}'s are randomly generated from different uniform distributions. Thus the probability of $\delta_{ij} = 0$ is 0. In that case, some δ_{ij}'s might be small, but essentially, if a gene is significant, then the corresponding δ_{ij}'s are almost surely non-zero. Now, we try to simulate another extreme situation, where only one δ_{ij} is truly non-zero if the corresponding gene is significant. In Guo *et al.* (2010), they set $\delta_{1j} = 100$. We are interested in the case where overall the effect size is relatively small, so we generate δ_{1j} randomly from uniform $(0.3, 3)$ (which is of similar magnitude to the uniform distributions used in Section 3.3.3). The μ_{ij}'s and $\mathbf{Z}_i$'s are generated the same way as Section 3.3.3, and the T_{ij}'s and $\tilde{P}_{ij}$'s are computed the same way too. The level α is set to be 0.05. The results are averaged over 2500 replications.

Figure 3.5 shows the results for different kinds of FDR. The top left plot shows the control of the mdFDR for Procedures 3.2, 3.4 and 3.5, in colors black, blue and red respectively. Procedures 3.4 and 3.5 has slightly higher dFDR and thus also slightly higher mdFDR compared to Procedure 3.2. Compared to the top left plot in Figure 3.1 in Section 3.3.3, we see that the directional FDR in this case is drastically higher, and hence the mdFDR is much higher too, especially when the number of truly significant genes is large. However, the mdFDR is still controlled at the desired level (0.05) for these three procedures - as it is proved to be. The top right plot shows the control of the mdFDR for the simple BH method. It seems that the simple BH method performs not much differently than the first three two-step procedures. As shown in the plot, the FDR with respect to the individual hypotheses is also controlled, as expected. The bottom left plot shows the FDR's for Procedure 3.8. The directional FDR increases quickly as the number of truly significant genes increases, causing the mdFDR to exceed the 0.05 level. Interestingly, the FDR with respect to the individual hypotheses is also not controlled by Procedure 3.8, and is in most cases even higher than the mdFDR. The bottom right plot shows the comparison of mdFDR (or OFDR) for all the procedures. It is easy to see the contrast between the procedures in this plot. Procedure 3.8 is the only procedure that fails to control the mdFDR. All other procedures have very similar behavior.

Figure 3.6 shows the comparison of power for all the procedures. In this case, the results for power (I) and (II) are almost identical. For both definitions of power, all procedures except 3.8 behave very similarly. Compared to the other procedures, Procedure 3.8 has slightly lower power most of the times, with the exception when the number of truly significant genes is high.

To summarize the results, we see that Procedure 3.8 is not guaranteed to control the OFDR in general. Moreover, Procedure 3.8 also fails to control the FDR with respect to the individual hypotheses, which is the type of FDR Jiang and Doerge (2006) set out to control. This shows that the empirical results in Jiang and Doerge (2006) cannot be readily generalized. The idea of a two-step adaptive BH procedure is appealing, and it has been shown to work well for the pairwise comparisons problem. But we should be cautious when trying to apply this procedure to other situations.

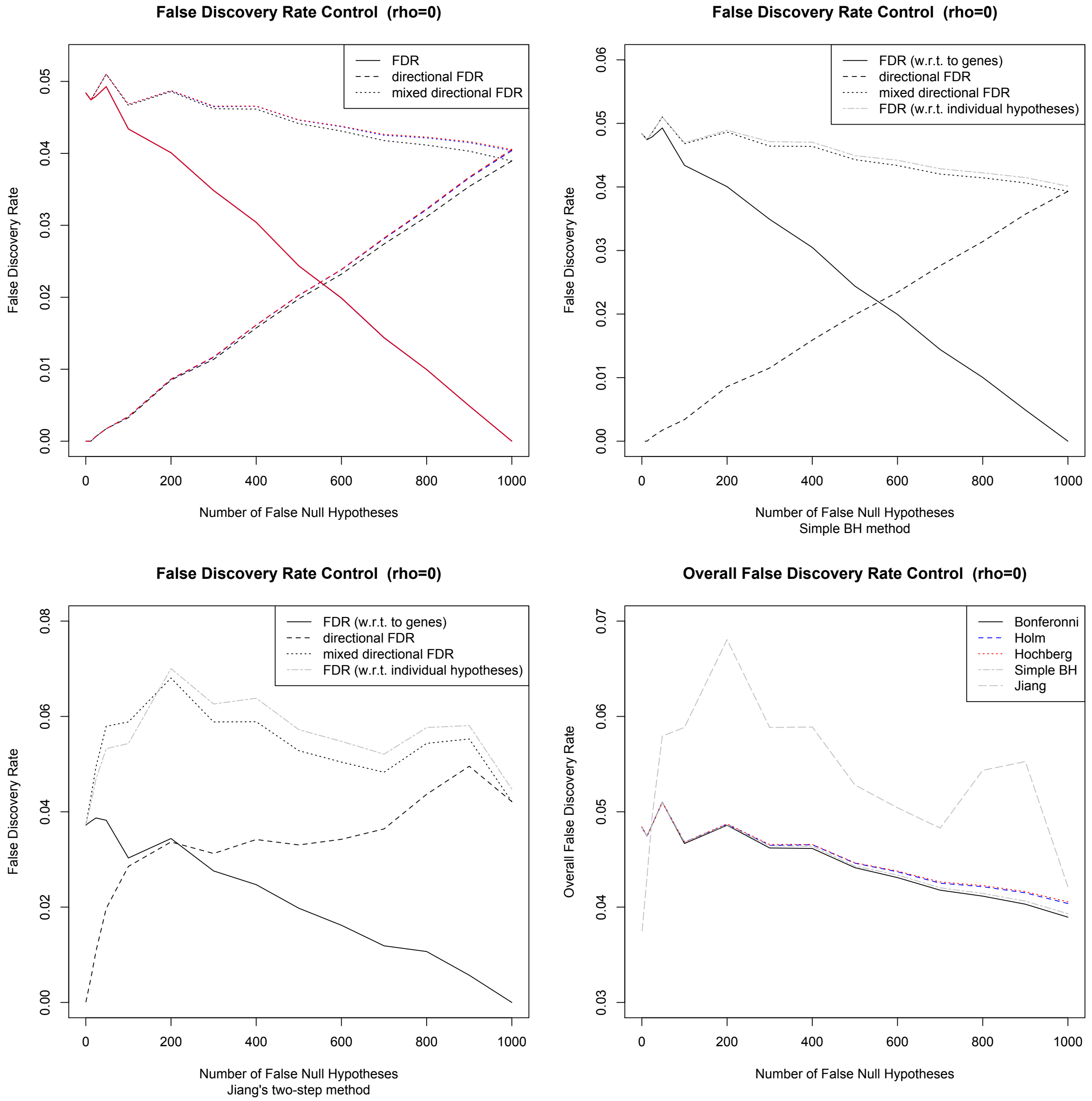

Figure 3.5. Plots showing the control or lack of control of the mdFDR for different procedures. The top left plots shows the control of the mdFDR for Procedures 3.2, 3.4 and 3.5. The top right plot show the control of the mdFDR for the simple BH method. The bottom left plot shows the lack of control of the mdFDR for Procedure 3.8. The bottom right plot compares the mdFDR (or OFDR) for all the procedures.

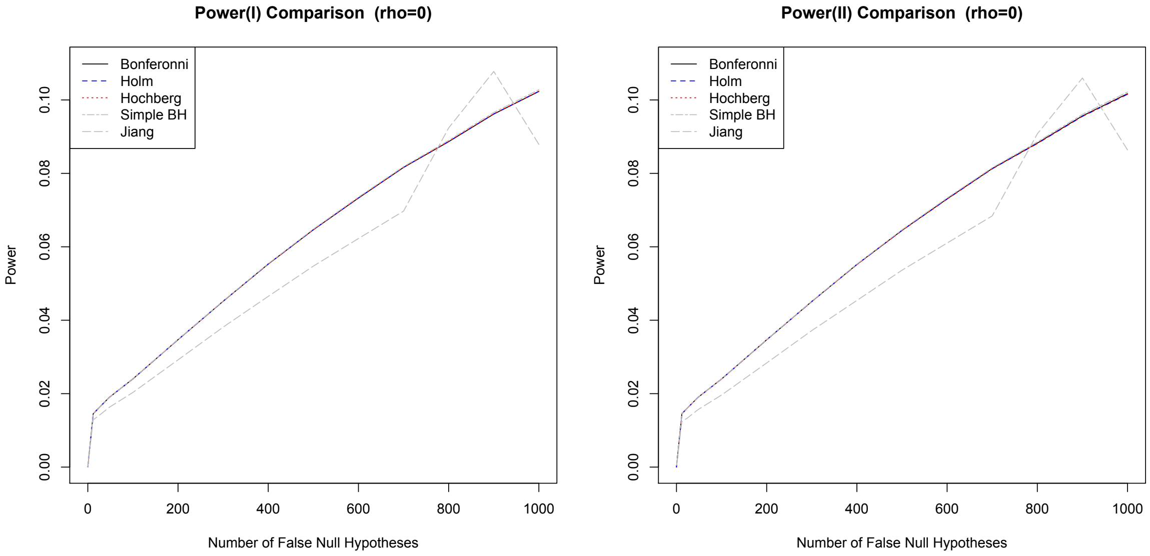

Figure 3.6. Plots comparing power (I) and power (II) for all the procedures. The x-axis is $m - m_0$.

3.6 Data Analysis: Identifying Important Genes in Tumor Classification

Another scenario for which hierarchical hypothesis set testing may be useful is when we are conducting tests between classified groups that are arranged in a hierarchical structure. One concrete example of this is tumor classification. There is a large number of literature concerning tumor classification using gene expression data. The underlying problem is the identification of genes that are differentially expressed between tumor and normal samples, as well as genes that are differentially expressed between different tumor types.

It is natural to adopt the hierarchical hypothesis set testing framework when doing some kind of meta-analysis of gene expression profiles of samples from multiple tumor types. If we are more concerned with differential expression occurring in tumor tissue compared to normal tissue, we can first screen for genes that differentially express between tumor and control samples for all the tumor types combined, and then further screen for genes that show differential expression between tumor and control samples for each specific type of tumor. If we are more concerned with differential expression between different tumor types, we can arrange the numerous tumor types into a tree-based classification structure, and test for differential expression between groups of tumor types for each layer of the classification tree step-by-step. We shall conduct data analysis for both cases in the following two subsections respectively. The tumor classification structure as well as the data will be based on a paper by Shedden *et al.* (2003) on tumor classification.

3.6.1 Identifying genes that are differentially expressed between different tumor classes or types

In Shedden *et al.* (2003) a total of 14 types of tumors are considered - breast, prostate, lung, colon, lymphoma, melanoma, bladder, uterus, leukemia, kidney, pancreas, ovary, mesothelioma and CNS. A pathological tree-based framework for the classification of these 14 types of tumor is provided. Tumors are first classified as “solid” class or “hematolymphoid” class. Two types of tumor belong to the “hematolymphoid” class, namely lymphoma and leukemia. The “solid” class includes three types of tumor - CNS, mesothelioma and melanoma, as well as another subclass of tumor "epithelial". The subclass "epithelial" is further divided into "mullerian", which includes tumors from ovary and uterus, and "non-mullerian", which includes all the other tumor types within the 14 that have not yet been mentioned. This classification tree structure is shown in Figure 3.7 as in Shedden *et al.* (2003).

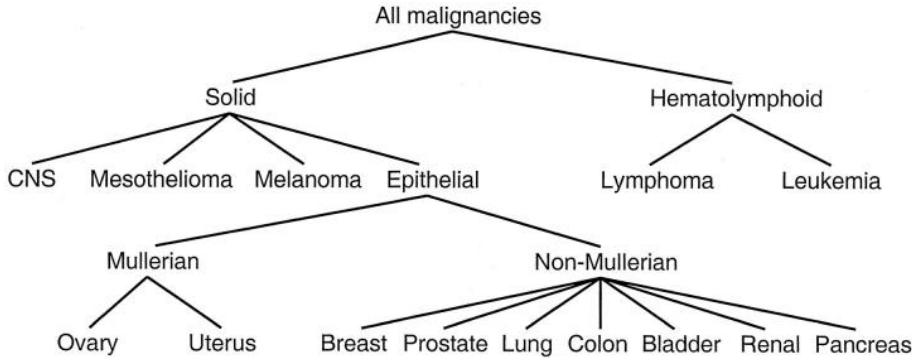

Figure 3.7. Pathological tree-based framework for tumor classification from Shedden *et al.* (2003).

Shedden *et al.* (2003) considered three microarray data sets that have been used previously in studies for cancer diagnosis. We shall adopt the most comprehensive one out of the three for our analysis. The data originated from a paper on multi-class cancer diagnosis by Ramaswamy *et al.* (2001) and can be obtained from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. The main data set (a combination of the training and test data sets in Ramaswamy *et al.*, 2001) contains 190 tumor samples that belong to the 14 types mentioned previously. Gene expression data were obtained using Affymetrix HuGeneFL and Hu35KsubA microarrays. Following Shedden *et al.* (2003), we eliminated the Affymetrix control genes and logarithmically transformed the raw data from Ramaswamy *et al.* (2001) by taking $log[\max(X, 0) + 50]$.

Our goal is to identify genes that differentially express between different classes or types of tumor. To apply our general two-step hierarchical testing framework, we only consider the first two layers of classification in Figure 3.7 for now. As is the case in previous sections, we test multiple hypotheses for each gene. In other words, there is a hypothesis set associated with each gene. In the first step, we select genes that differentially express between the tumor classes "solid" and "hematolymphoid". That is, for each gene, the screening hypothesis tests for whether the gene differentially expresses between the two tumor classes in the first layer. If a gene "passes" the screening test, then in the second step we test for whether it differentially expresses between each pair of subclasses or types within the "solid" class or the "hematolymphoid" class. To be more specific, in the second step, we have 7 individual hypotheses for each gene, which tests for whether the gene differentially expresses in: (1) lymphoma versus leukemia; (2) CNS versus mesothelioma; (3) CNS versus melanoma; (4) CNS versus subclass "epithelial"; (5) mesothelioma versus melanoma; (6) mesothelioma versus "epithelial"; and (7) melanoma versus "epithelial". We do not consider comparing a subclass/type from "solid" with another subclass/type from "hematolymphoid" since the gene is already significantly differentially expressed between the two classes. By construction, the 7 individual hypotheses in the second step are not independent of each other. Hence, we adopt the Holm's method for controlling the family-wise error rate for each hypothesis set in the second step, since Holm's method works without assumptions on the dependencies between the individual hypotheses. We formally state our procedure for this problem below.

Procedure 3.9.

(1) Let P_{0j} be the p-value for testing whether gene j differentially expresses between "solid" class and "hematolymphoid" class for $j = 1, ..., m$. Apply the BH method at level α on $P_{01}, ..., P_{0m}$. Let R be the number of rejected genes out of m.

(2) For every gene j that was rejected in step (1), let P_{ij} $(i = 1, ..., 7)$ be the p-values for testing whether the gene differentially expresses between the 7 pairs of tumor subclasses or types as described previously. Let $R_j = \max\{1 \leq i \leq 7 : \ P_{(l)j} \leq R\alpha\{m(7+1-l)\}^{-1}, \text{ for } l = 1, ..., i\}$, if the maximum exists; otherwise $R_j = 0$. Following the Holm's method, declare the individual hypothesis corresponding to P_{ij} to be significant if $P_{ij} \leq P_{(R_j)j}$.

We applied Procedure 3.9 to the data set from Ramaswamy *et al.* (2001). P-values to test for differential expression between two groups are obtained by two-sample T-tests. The level α, for which the overall FDR is controlled at, is taken to be 0.05. Out of the 16004 transcripts present in the data set, 10440 were found to be significant in step one ("solid" versus "hematolymphoid"). The number of significant transcripts for the 7 individual hypotheses in step two are as follows: 3799 for lymphoma versus leukemia; 4005 for CNS versus mesothelioma; 5096 for CNS versus melanoma; 5158 for CNS versus "epithelial"; 1253 for mesothelioma versus melanoma; 1138 for mesothelioma versus "epithelial"; 2095 for melanoma versus "epithelial".

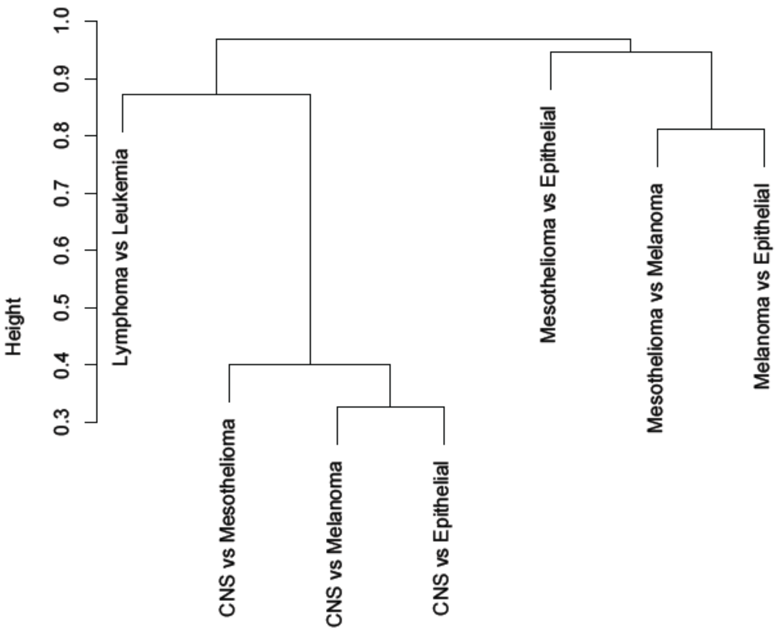

Figure 3.8. Clustering results for significantly differentially expressed genes for the 7 tumor subclass/type pairs.

In order to get an idea of how the 7 pairwise comparisons relate to each other, we did a clustering of the 16004×7 decision matrix. It gives us a summary of how similar the 7 lists of significant genes are from each other. The clustering result is shown in Figure 3.8. From the plot we can see that the three comparisons involving CNS are clustered together and show the smallest differences with each other. This indicates that these three comparisons results in similar gene lists. In other words, there is a lot of overlap in the genes that were found to significantly differentially express in CNS versus the other three types/subclasses of tumor. These three lists also have some overlap with the significant gene list for lymphoma versus leukemia. There are also some similarities in the genes that show significant differential expression in pairwise comparisons between mesothelioma, melanoma and “epithelial”. On the other hand, the significant gene list for lymphoma versus leukemia is quite different from the gene lists involving CNS, and are both very different from the genes lists involving comparisons between mesothelioma, melanoma and “epithelial”.

For the four types/subclasses of tumor in “solid”, we now have significant gene lists for each pairwise comparison. It would be interesting to look at the overlap of genes showing up in pairwise comparisons involving the same tumor type/subclass. For example, the intersection of the significant genes from CNS versus mesothelioma, CNS versus melanoma and CNS versus “epithelial” would be the genes that differentially express between CNS and every other type/subclass within the “solid” class. These genes can be considered unique to CNS tumor in some way. So we created 5 intersection gene lists, that correspond to (1) lymphoma versus leukemia; (2) CNS versus the others belonging to “solid”; (3) mesothelioma versus the others belonging to “solid”; (4) melanoma versus the others belonging to “solid”; and (5) subclass “epithelial” versus the others belonging to “solid”. We did pathway analysis on these 5 lists as well as the list corresponding to “solid” versus “hematolymphoid”. Pathway analysis was done using functional annotation clustering in DAVID, which is available at http://david.abcc.ncifcrf.gov/home.jsp. The processing limit of DAVID is 3000 genes per list. For our gene lists that exceed this limit, we include only the top 3000 genes in the list. The ranking of the significant genes in a list is based on the effect size of the two-sample T-test between the two groups. For the “intersection lists”, the ranking of genes are based on the sum of the effect sizes from each original list. The full set of DAVID results of top pathways along with the enrichment scores for each of the 6 lists mentioned above can be found in Appendix A.

3.6.2 Identifying genes that are differentially expressed between tumor and normal tissues for different tumor types

Identifying genes that differentially express between tumor and normal tissues is a very common problem. However, when there are multiple classes or types of tumor involved, the problem becomes more interesting. In this case, we can do a meta-analysis of the comparisons between tumor and normal tissues for the different tumor types. In the data from Ramaswamy *et al.* (2001), in addition to the 190 tumor samples, there are also 90 control samples of normal tissue coming from 12 locations corresponding to 12 out of the 14 tumor types mentioned previously. The two tumor types that do not have corresponding control samples are melanoma and mesothelioma - we exclude these two types of tumor from the following analysis.

We again apply our two-step hierarchical hypothesis set testing idea to this problem. In this case, the problem is somewhat similar to the multidimensional directional decisions problem in Section 3.3, because the first step in the procedure would serve as a meta-analysis of the following individual hypotheses in the second step. For each of the 12 tumor types that have corresponding controls, we perform a two-sample T-test for the specific type of tumor versus its corresponding control samples. In the first step of our procedure, we would perform a meta-analysis on these 12 tests for each gene. In other words, the screening null hypothesis would be that there is no difference in the expression between tumor and normal samples for that gene for any of the 12 tumor types. The BH method would be applied in the first step to the screening p-values for each gene. For the genes that we reject in the first step, we proceed to the second step to look at the 12 individual tests for each tumor type. We control the family-wise error rate for these 12 individual tests.

There are many choices of meta-analysis methods for the first step. Procedures 3.2, 3.4 and 3.5 can all be used in this case (with slight modification of notations). Although, Procedure 3.5, which is based on the Hochberg's method, might not be very suitable, since the 12 individual hypotheses corresponding to the 12 tumor types are not independent. On the other hand, we are able to choose from a larger selection of meta-analysis methods compared to Section 3.3, because we do not necessarily need to preserve the "consonance" property. In other words, it is not unreasonable to claim a gene to be significantly differentially expressed between tumor and normal samples overall, even if we cannot make the claim for any particular type of tumor. In that case, the result indicates that we do not have enough information to claim significant differential expression for any type of tumor, but when we pool the information, we are able to reject the null hypothesis that the gene has the same expression in tumor and normal samples for all tumor types. This enables us to use meta-analysis methods in the first step that are not directly related to the FWER controlling methods for the individual hypotheses in the second step. For example, we can use Fisher's combined probability test in the first step to test for the conjunction of nulls, and switch to Holm's method in the second step to control the family-wise error rate. We state two procedures below. Procedure 3.10 is based on Holm's method for both the first and second step, similar to Procedure 3.4. Note that the first step would be the same as using Bonferroni's method, as discussed in Section 3.3.2. Procedure 3.11 uses the Fisher's combined test in the first step, while still using Holm's method for the second step. The second step of the two procedures are the same, but Procedure 3.10 is potentially more powerful than Procedure 3.11 because Fisher's method is usually a more powerful meta-analysis method and would likely result in more significant genes in the first step.

Procedure 3.10.

(1) Let P_{ij} be the p-value for testing whether gene j ($j = 1, ..., m$) differentially expresses between tumor and normal samples for the i's type of tumor for $i = 1, ..., 12$. Based on the Holm's method, let $P_j = 12P_{(1)j}$ for $j = 1, ..., m$. Apply the BH method at level α on $P_1, ..., P_m$. Let R be the number of rejected hypotheses out of m.

(2) For every gene j that was rejected in step (1), let $R_j = \max\{1 \leq i \leq 12 : P_{(l)j} \leq R\alpha\{m(12+1-l)\}^{-1}$, for $l = 1, ..., i\}$, if the maximum exists; otherwise $R_j = 0$. Following the Holm's method, declare the individual hypothesis corresponding to P_{ij} to be significant if $P_{ij} \leq P_{(R_j)j}$.

Procedure 3.11.

(1) Let P_{ij} be the p-value for testing whether gene j ($j = 1, ..., m$) differentially expresses between tumor and normal samples for the i's type of tumor for $i = 1, ..., 12$. For each j, let P_j be the p-value for the Fisher's combined probability test for the P_{ij}'s ($i = 1, ..., 12$). Apply the BH method at level α on $P_1, ..., P_m$. Let R be the number of rejected hypothesis out of m.

(2) For every gene j that was rejected in step (1), let $R_j = \max\{1 \leq i \leq 12 : P_{(l)j} \leq R\alpha\{m(12+1-l)\}^{-1}$, for $l = 1, ..., i\}$, if the maximum exists; otherwise $R_j = 0$. Following the Holm's method, declare the individual hypothesis corresponding to P_{ij} to be significant if $P_{ij} \leq P_{(R_j)j}$.

We applied Procedures 3.10 and 3.11 to the data set from Ramaswamy *et al.* (2001). As a comparison, we also applied the simple BH method that we have mentioned before, which is simply applying the BH procedure to all the 16004×12 individual hypotheses. Table 3.9 summarizes the number of significant genes found by each of the three methods, for each tumor type respectively, as well as the total number of significant individual hypotheses and total number of significant genes found. Notice that for Procedure 3.11 the genes that are found to be significant in the second step is only a subset of the genes found significant in the first step. Because Procedure 3.11 does not preserve the "consonant" property, genes that were found significant in the first step are not guaranteed to be significant for any of the 12 individual hypotheses, in which case we say that gene was not found significant in the second step. Procedure 3.10 does not have this problem - all the genes found significant in the first step would have at least one individual hypothesis rejected in the second step.

From Table 3.9 we can see that Procedure 3.10 is the most conservative, while Procedure 3.11 always claims the most number of significance. The simple BH method falls in between. Although, remember that both Procedures 3.10 and 3.11 guarantees the control of the OFDR, while the simple BH method does not. In this case, the OFDR is the expected proportion of genes found significant in the first step for which we make at least one incorrect rejection out of the total of 13 tests we perform on it (12 for each tumor type and one for meta-analysis). From the "overall" table in Table 3.9 we see that a huge number of genes were claimed to be significant in the first step by Procedure 3.11. However, about half of them did not have any significant individual hypotheses, and thus were not counted as significant in the second step. The fact that about 89% of the genes were rejected by the Fisher's combined test even after apply the BH method is somewhat hard to interpret. However, it does result in a larger pool of genes to be entered into the second step for further screening. The second step of Procedure 3.11 is as strict as the second step of Procedure 3.10 - causing a huge reduction of the number of genes found significant in the second step - but still results in about 45% more significant genes than Procedure 3.10 in the second step. For some of the tumor types, Procedure 3.11 more than doubles the number of significant genes found compared to Procedure 3.10.

Table 3.9. Summary of the number of significant genes found by the three methods.

By tumor type:

	Procedure 3.10	Procedure 3.11	simple BH
Breast	352	755	476
Prostate	14	30	20
Lung	70	188	101
Colon	535	1012	674
Lymphoma	1174	1724	1391
Bladder	1679	2622	2028
Uterus	85	182	112
Leukemia	962	1310	1095
Kidney	226	393	276
Pancreas	418	1212	667
Ovary	252	446	317
CNS	866	1443	1056

Overall:

	Procedure 3.10	Procedure 3.11	simple BH
sig. individual hypotheses	6633	11317	8213
sig. genes in step two	5098	7414	NA
sig. genes in step one	5098	14232	6130

We performed pathway analysis on the resulting gene lists using functional annotation clustering in DAVID (http://david.abcc.ncifcrf.gov/home.jsp). Analysis was done respectively for the 13 lists of genes by Procedure 3.10 and another 13 lists by Procedure 3.11 - 12 for each type of tumor and one for meta-analysis. For lists that exceed the 3000 number limit, we used the top 3000 genes ranked by the Fisher's combined test p-value. The full set of DAVID results of top pathways along with the enrichment scores for the 26 lists can be found in Appendix A.

We also did cluster analysis for the 12 lists for each of the tumor types, similar to what we did in the last subsection. We did separate cluster analyses for the lists produced by Procedures 3.10 and 3.11. The results are shown in Figures 3.9 and 3.10. From the two figures we see that "lung" and "prostate" seem to be left out. This could be because the number of significant genes are very small for these two types of tumors compared to the others. Although the number of significant genes differ quite a bit for the different types of tumor, this does not seem to directly influence the clustering results for the other types of tumor other than "lung" and "prostate". Despite the differences between Figures 3.9 and 3.10, "lymphoma", "CNS" and "leukemia" are grouped together as a result of both procedures. By comparing the y-axis of the two figures, we can see that the distances between clusters are generally slightly smaller for the results from Procedure 3.11 compared to those from Procedure 3.10. This is probably because more genes are included in the lists produced by Procedure 3.11.

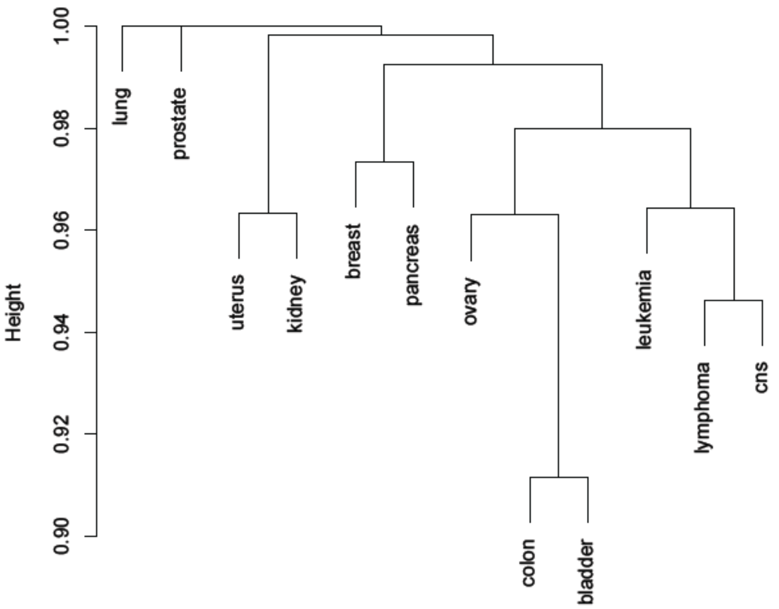

Figure 3.9. Clustering results for significantly differentially expressed genes between tumor and normal samples for the 12 tumor types by Procedure 3.10.

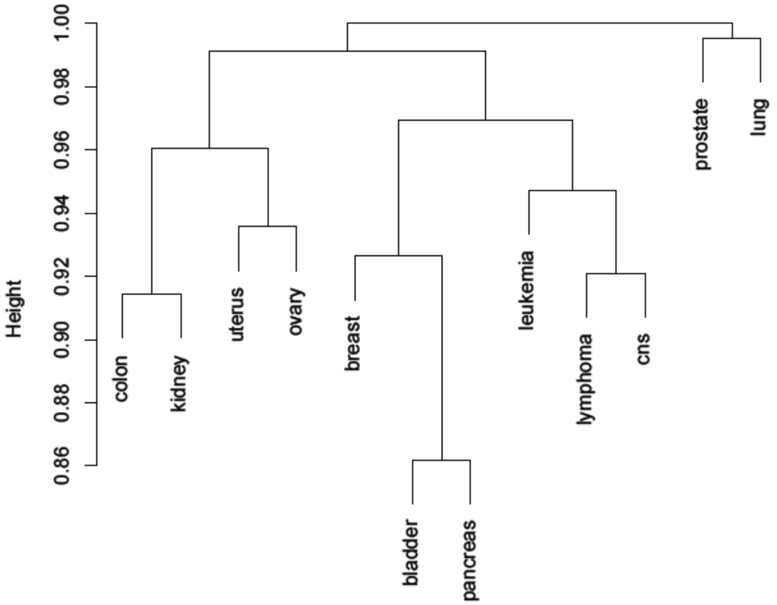

Figure 3.10. Clustering results for significantly differentially expressed genes between tumor and normal samples for the 12 tumor types by Procedure 3.11.

3.7 Discussion

In large-scale experiments, such as microarray gene expression studies, as the problem and the designs become more complicated, new issues in multiple testing arise. For instance, in microarray time-course or dose-response experiments, in addition to considering tens of thousands of genes simultaneously, multiple hypotheses are often being tested for each gene. As a result, the problem of multiplicity becomes multi-dimensional. Traditional concepts of type I error control and methods for large-scale multiple testing (e.g. the FDR and the Benjamini-Hochberg procedure) can still be used, but may not be optimal for these more complex designs. Hence, it is important to consider new measures of type I error and develop statistical methods for these multi-dimensional multiple testing problems.

The methodology in this chapter provides one way of approaching these problems. We have formulated certain types of multi-dimensional multiple testing problems as multiple hypothesis set testing. In the case of microarray time-course/dose-response experiments, we consider each gene to be associated with a hypothesis set, where the multiple individual hypotheses in the set test for differential expression among a number of different time points or dose levels. We have adopted the concept of the Overall FDR, which is a measure of the FDR on the hypothesis set level. By doing so, we aim at controlling the false discoveries on the gene level, which increases the interpretability of the results, compared to focusing on the FDR of all the individual hypotheses. We discussed a general two-step hierarchical testing procedure for multiple hypothesis set testing, which is proved to control the OFDR under independence across the hypothesis sets. We also extended the general procedure to enable directional decisions for two-sided tests and discussed the control of the mdFDR under certain conditions. We then suggested three specific procedures for microarray time-course/dose-response experiments. These procedures not only allow us to test for differential expression across multiple time points or dose levels, but are also capable of identifying the direction of expression change, while still maintaining control of the OFDR and mdFDR. We evaluated the performance of the proposed procedures under both independence and dependence between genes and compared the power with previous methods. Finally, the methodology is applied to analyze data from a microarray dose-response study to identify genes that are differentially expressed at low concentrations of estrogen in breast cancer cells.

The key point in the hypothesis set testing framework is that the two-dimensional multiplicity is transformed into a hierarchical structure. Hypotheses are tested in the unit of sets in the first step. This is realized by the formulation of a screening hypothesis for each set. The first step of our procedures deals with the hypothesis sets much like dealing with a traditional multiple testing problem. By applying the Benjamini-Hochberg procedure to the screening hypotheses, we are able to adjust for part of the multiplicity on the hypothesis set level. Additional type I errors (and sometimes directional errors) that can potentially occur while making inference for the individual hypotheses in each set are controlled in the second step by applying family-wise error rate controlling procedures. Together, the OFDR (or mdFDR) is controlled at the hypothesis set level.

Although we only discussed a few applications, the proposed methodology is widely applicable. The framework of multiple hypothesis set testing is very flexible and can be easily adapted to many large-scale multiple testing problems with complex designs. On the other hand, it would be interesting to develop more powerful procedures for each specific type of problem. For example, if a large proportion of individual hypotheses are expected to be significant given the significance of the hypothesis set, then we can potentially improve power by incorporating adaptive multiple testing methods into the procedure. Much future work can be done on adapting the hierarchical hypothesis set testing framework and procedures to different multi-dimensional multiple testing problems.

Chapter 4

Meta-Analysis Based on Weighted Ordered P-values

4.1 Introduction

Meta-analysis has long been used to integrate data and/or results from multiple studies targeting the same (or similar) question(s). It is commonly used in many areas of statistical applications such as clinical studies and psychology experiments. In recent years, meta-analysis has been frequently adopted in genomic data analysis, due to the fast development of high-throughput technology and the vast amounts of data available in public databases. Our discussions will mainly be in the context of genomic data, even though most meta-analysis techniques are not restricted to any particular area of study.

Many meta-analysis methods have been developed throughout the years. Roughly speaking, there are two main approaches to meta-analysis methods (Song and Tseng, 2013): the first directly combines the p-values from the studies; while the second attempts to model the data or the effect sizes from the combined studies. The former includes methods such as the Fisher's combined probability test (Fisher, 1925) and the Stouffer's Z-test (Stouffer et al., 1949), as well as weighted variations of these classical tests. The latter includes a variety of fixed effects and random effects models, such as in Choi et al. (2003). Each approach has its advantages and disadvantages. The p-value combining methods are relatively flexible in that they require minimal information and assumptions from the studies. We will mainly focus on methods that directly combine the p-values.

Most of the traditional meta-analysis methods (e.g. Fisher's and Stouffer's methods) aim at testing the alternative hypothesis that at least one of the studies is non-null. While this aligns with earlier goals of meta-analyses to gain power in detecting signals by combining multiple studies, it is frequently not the case in meta-analyses nowadays. In meta-analysis of genomic studies for example, the goal is often to identify genes that are differentially expressed in a consistent pattern across multiple studies. The extreme of this would be to test for the alternative that the null can be rejected for all the studies. A solution to this extreme alternative dates back to the maxP method by Wilkinson (1951). However, the maxP method is often considered too conservative. A recent approach by Phillips and Ghosh (2013) improves the power of testing for this disjunction of nulls when the rejection of all p-values associated with a gene is required. In practical meta-analysis settings of combining studies, the goal of rejecting all studies may still be considered extreme. Ideally, we would want to target at detecting consistent signals across studies while avoiding being overly exclusive. This issue has gained attention in recent years, and a number of authors have tried to address this problem. Benjamini and Heller (2008) discussed a framework for testing partial conjunction hypotheses, where they test for the alternative that at least u out of n null hypotheses are false against the partial conjunction null that no more than $n-u+1$ of the null hypotheses are true. Song and Tseng (2013) proposed the rth ordered p-value (rOP) method that aims at testing the alternative hypothesis that there is signal in at least a given percentage of studies. Other methods exist that address this problem from different approaches, such as RankProd by Hong et al. (2006) that looks for consistently highly ranked genes. The weighted testing procedure discussed in Chapter 2 which weights genes by its expression consistency across studies is another approach for this problem.

We consider the problem of detecting signals in the majority of studies. Our approach adopts one aspect of the rOP method of Song and Tseng (2013) in that we also consider ordered p-values. But instead of using a single rth ordered p-value as the statistic (as rOP does), we combine all or a subset of the ordered p-values while weighting them based on their order (weighted ordered p-values, WOP). P-values closer to the median are highly weighted and the smallest/largest p-values are down-weighted. The idea is that among the collection of p-values, the median p-values are likely to be a better reflection of the behavior of the majority of studies than the smallest or largest p-values. Olkin and Saner (2001) discussed a trimmed Fisher's procedure that leaves out a number of the smallest and/or largest p-values from the calculation of Fisher's statistic to remove the effect of possible aberrant extremes. In our consideration, we still keep the smallest/largest p-values because they do carry certain information, but we down-weight them because they may be relatively less relevant when considering the "majority" of studies. To reflect the up-weighting of the medians and down-weighting of the extremes, we calculated our weights based on the binomial distribution. To summarize the weighted ordered p-values, we considered two classical p-value combining methods: Fisher's statistic and Stouffer's Z-test. We also consider incorporating the order-based weights into Lancaster's statistic (Lancaster, 1961), which is a generalized Fisher's method based on sums of chi-square statistics with varying degrees of freedom (more details on Lancaster's statistic will be discussed in Section 4.3). In general, other summary statistics can be used under this framework as well. By incorporating more than one order statistic into our summary statistic, our method can be considered an expansion of the rth ordered p-value (rOP) method proposed by Song and Tseng (2013). In general, both the rOP method and the original summary statistic we use (e.g. Fisher's or Stouffer's method) are special cases under the WOP framework - one having all the weight on a single ordered p-value and the other having evenly distributed weights.

While many weighted variations of Fisher's statistic and Stouffer's statistic have been developed throughout the years, most of them distribute weights according to the sample sizes and/or effect sizes of the studies, or other similar considerations (e.g. Mosteller and Bush (1954), Lipták (1958), Won et al. (2009), Makambi (2003), among others). Li and Tseng (2011) proposed an interesting adaptively weighted statistic, where the weights are used to maximize the significance of the summary statistic. However, to the best of our knowledge, none of the weighting schemes are based on the ordered p-values, which is what makes our method unique. Xie et al. (2011) discussed a meta-analysis approach using confidence distributions, where they incorporated the use of medians and kernel functions, but their approach is under a completely different framework from ours.

We conducted simulation studies to evaluate the properties of our WOP methods. We demonstrate the strengths of various versions of the WOP methods compared with the unweighted summary statistics, as well as with the rOP method (Song and Tseng, 2013). The proposed methodology is then applied to the meta-analysis of a set of gene expression data from multiple stem cell studies.

4.2 Hypothesis Settings for Meta-Analysis

Before performing any meta-analysis, it is always important to figure out the goal of combining multiple studies. When a single hypothesis test is conducted, it is clear what the null and alternative hypotheses are. In meta-analysis we usually combine studies designed to test the same set of null and alternative hypotheses. However, the null and alternative hypotheses of the meta-analysis test are not always obvious and largely depends on the researcher's goals. Here we consider the example of meta-analysis of differential expression studies of genomic data. Sometimes a gene is of interest as long as it is differentially expressed in at least one study, while other times we hope to target genes that are differentially expressed in all studies. Li and Tseng (2011) and Song and Tseng (2013) both provided extensive discussions on the different scenarios that lead to different hypothesis settings. Following their notation, let θ_{gk} be the true effect size for gene g $(1 \leq g \leq G)$ in study k $(1 \leq k \leq K)$. As in Song and Tseng (2013), for a given $1 \leq r \leq K$, a general hypothesis setting for the meta-analysis of the K studies for a given gene g can be formulated as:

$$HS_r : \left\{ H_0 : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) = 0 \ \text{ versus } \ H_a^{(r)} : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) \geq r \right\}.$$

When $r = 1$, HS_1 is the classical setting of testing for non-zero effect size in at least one study against the conjunction of nulls. HS_1 is the hypothesis setting that Fisher's method, Stouffer's Z-test and many other traditional methods test for. When $r = K$, HS_K tests for the alternative that all the studies have non-zero effect size. For instance, the maxP method (Wilkinson, 1951) tests for HS_K. When $1 < r < K$, HS_r provides a compromise between the two aforementioned hypothesis settings, and tests for at least a pre-specified number of non-zero effects. For a given r, the rOP (rth ordered p-value) method by Song and Tseng (2013) is used to test for HS_r.

Benjamini and Heller (2008) considered the problem of testing the same alternative as HS_r but under the framework of partial conjunction nulls. The difference is that, instead of using a unified null ($H_0 : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) = 0$) for testing different alternatives as in HS_r, they considered partial conjunction nulls in the form of $H_0^{(r)} : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) < r$ for each r. Thus the general hypothesis setting for a given gene has the form:

$$HS_r^{(partial)} : \left\{ H_0^{(r)} : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) < r \text{ versus } H_a^{(r)} : \sum_{k=1}^{K} I(\theta_{gk} \neq 0) \geq r \right\}.$$

The $HS_r^{(partial)}$ hypothesis setting has the advantage of a clear partition of the whole parameter space into the null and the alternative spaces, without leaving any "grey areas" that do not belong to either the null or alternative hypotheses as the HS_r setting does. However, it requires the framework developed by Benjamini and Heller (2008) to deal with the partial conjunction of nulls. The HS_r setting is easier to work with because of the unified simple conjunction of nulls. It can be interpreted as testing against the same null space that all the effects are zero while focusing on different subsets of the non-null space.

In this paper, we aim at testing for non-zero effect sizes in the majority of studies against the null that the effects sizes are zero in all studies. Thus we are testing against the conjunction of nulls while trying to focus on a certain subset of the non-null space. In general, the hypothesis for our meta-analysis approach can be considered to fall under the HS_r setting. Song and Tseng (2013) suggested a few data-driven methods for selecting r. To prevent any potential issues of post hoc choices of r, we choose to fix r before any analysis is conducted. While it is hard to specify what the term "majority" exactly means, as a general rule, we choose to test the hypothesis setting HS_m, where $m = \lceil K/2 \rceil$, $\lceil x \rceil$ being the smallest integer no less than x. Essentially we are aiming at the alternative that at least half of the studies have non-zero effect sizes. We shall compare our methods with the rOP method for the same hypothesis setting, which is the mth ordered p-value.

Under our weighted ordered p-values framework, we also explored methods for other hypothesis settings with r ranging from $m+1$ to K, while comparing with the corresponding rOP method. But in general we hope to provide a simple to use method without having to put too much effort in selecting a particular r, and therefore we will focus mostly on testing HS_m.

4.3 The Weighted Ordered P-values (WOP) Method

4.3.1 The WOP framework

We first describe the general framework for the weighted ordered p-values (WOP) method. Suppose we have p-values $p_1, p_2, \cdots, p_K$ for testing the hypothesis of interest for each of the K studies. Let $p_{(1)}, p_{(2)}, \cdots, p_{(K)}$ be the ordered p-values. Now consider a set of weights $w_1, w_2, \cdots, w_K$ associated with the corresponding ordered p-values. Summary statistics of weighted ordered p-values can be expressed in the following general form:

$$T = \sum_{i=1}^{K} w_i H(p_{(i)}).$$

As mentioned by previous authors, many traditional p-value combination methods can be expressed in the general form of $T' = \sum_{i=1}^{K} w_i' H(p_i)$ (for example, see Zaykin, 2011). For instance, Stouffer's Z-test takes $H(\cdot)$ to be the inverse normal function, while Fisher's method has $H(p_i) = -2\log(p_i)$. The difference between T' and T, albeit subtle in notation, is the essence of the WOP framework. In the WOP framework, the weight w_i is associated with the ith ordered p-value $p_{(i)}$. In other words, the ranking of a p-value in relation to the p-values from the other studies determines its weighting. In traditional weighted p-value combining methods, the weight w_i' is associated with the ith study; i.e. the weight assigned to a p-value is associated with the characteristics of that particular study, be it the sample size, the effect size, or other features.

The weighted ordered p-values (WOP) idea can also be incorporated into other weighted p-value combining methods that are not in the previously described general form. For instance, Lancaster's statistic (Lancaster, 1961) is a generalization of Fisher's method in the form of $T'_L = \sum_{i=1}^{K} [\chi^2_{n'_i}]^{-1}(1-p_i)$ with the distribution $\chi^2_{\sum n'_i}$ under the null, where $[\chi^2_{n'_i}]^{-1}$ is the inverse chi-square transformation with n'_i degrees of freedom. The varying degrees of freedom n'_i in Lancaster's statistic act like weights for each study. Using the weighted ordered p-values idea, we can let n_i be associated with the order-based weights w_i. Basically, let n_i $(i = 1, \cdots, K)$ be integers that are proportional to w_i $(i = 1, \cdots, K)$, where w_i's are the weights corresponding to the ordered p-value $p_{(i)}$'s. Thus the WOP version of the Lancaster's statistic is $T_L = \sum_{i=1}^{K} [\chi^2_{n_i}]^{-1}(1 - p_{(i)})$.

Allowing the weights to depend on the ordering of the p-values opens up a whole new arena of considerations when combining multiple p-values. We can consider giving more weights to the p-values that are closer to the center of the distribution of the p-values since they might hold more credibility, and at the same time down-weight the outlying p-values. We can consider giving more weight to a particular ordered p-value, depending on the hypothesis setting. For instance, if the entirety of weights is placed on the rth ordered p-value, the statistic degenerates to the rOP method. However, using the WOP framework, if we highly weight the rth ordered p-value but still distribute some weight to the other p-values, the method can be viewed as a more robust version of the rOP method for testing HS_r.

We shall develop a few specific methods under the WOP framework. We consider weights based on the binomial distribution, which will be described in more detail in section 4.3.2. As for the p-value combining methods, we shall focus on Fisher's method, with $H_F(p_{(i)}) = -2\log(p_{(i)})$, and Stouffer's method, with $H_Z(p_{(i)}) = \phi^{-1}(1 - p_{(i)})$, $\phi(\cdot)$ being the standard normal distribution function. We also consider Lancaster's method with $T_L = \sum_{i=1}^{K} [\chi^2_{n_i}]^{-1}(1 - p_{(i)})$. Implementations of the proposed methods will be discussed in section 4.3.3.

4.3.2 Binomial weights and half-binomial weights

In this section, we discuss two possible weighting schemes for the WOP framework. We mainly consider testing the alternative hypothesis that the effect sizes are non-zero in the majority of studies, which is the hypothesis setting HS_m. We will also briefly discuss extending the weighting schemes to testing other hypothesis settings HS_r, for $m < r \leq K$.

Inspired by the rOP method, which uses the rth ordered p-value for testing the hypothesis setting HS_r, we consider placing the highest weight on the median p-values for testing HS_m. This makes intuitive sense, since if a consensus does exist among the studies, we have reason to believe that the behavior of the majority of studies should be best captured by the p-values that are closer to the center of the distribution. Since we do not insist on non-zero effect sizes for every single study, we consider down-weighting the largest p-values among the studies. On the other hand, p-value combining methods such as Fisher's method are known to be very sensitive to single extremely small p-values, thus the smallest p-values should also be down-weighted, to avoid a small number of extremely small p-values biasing the results of the majority of studies. In summary, we would like our weighting scheme w_i, as a function of i, to reflect a unimodal shape, with the highest weights being w_m (when K is odd) or w_{m-1} and w_m (when K is even), and such that w_i decreases as i goes to 1 or K.

To reflect the above properties of the weights, we constructed the weights based on the binomial distribution. Let $f(x;n,p)$ be the probability mass function of the binomial distribution $B(n,p)$, for $x = 0, 1, \cdots, n$. We define the binomial weighting scheme such that

$$w_i^b = f(i-1; K-1, 0.5), \quad i = 1, 2, \cdots, K.$$

In the binomial weighting scheme, all the weights are non-zero, thus every p-value contributes to the combined statistic. To further reduce the influence of the smallest p-values on the summary statistic, we may argue that only $p_{(m)}, p_{(m+1)}, \cdots, p_{(K)}$ matters in testing the alternative that at least m studies have non-zero effect size. With these considerations, we define what we call the half-binomial weighting scheme such that

$$w_i^{hb} = \begin{cases} 0, & i = 1, 2, \cdots, m-1, \\ w_i^b, & i = m, m+1, \cdots, K. \end{cases}$$

We will discuss more on the effects of these two different weighting schemes through simulation studies in section 4.4.

So far we have constructed the binomial weighting scheme and half-binomial weighting scheme for testing the hypothesis setting HS_m. We can extend the ideas

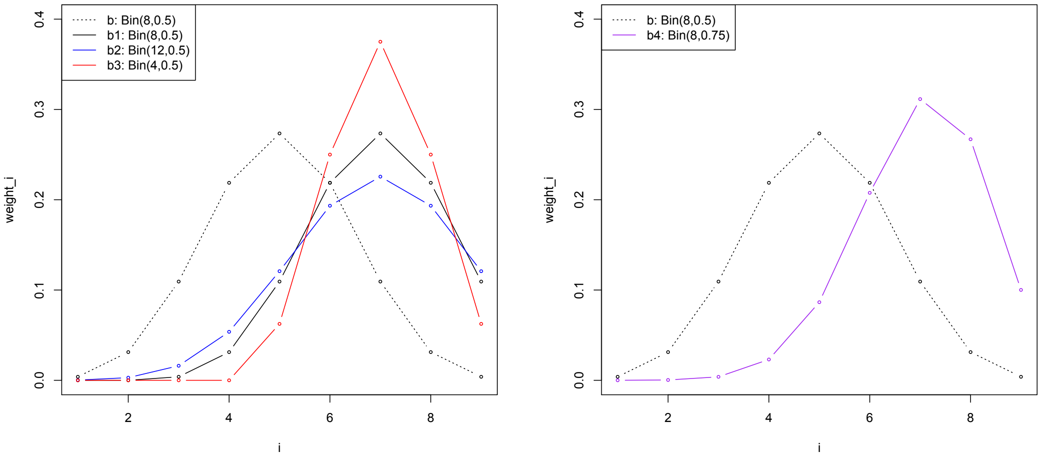

Figure 4.1. Illustration of the four weighting schemes w^{b1}, w^{b2}, w^{b3} and w^{b4} for $m < r \leq K$. The plots reflects an example of $K = 9$. The weighting scheme w^b for testing HS_5 is plotted as a reference. The other weighting schemes are for testing HS_7. The plot on the left shows the three weighting schemes based on binomial distributions with $p = 0.5$. The plot on the right shows w^{b4} which uses an adjusted value of p in the binomial distribution to allow w_7 to have the highest value.

of these two weighting schemes to testing HS_r, for $m < r \leq K$. Instead of placing the highest weights on the medians, the highest weight should now be assigned to $p_{(r)}$. When $r \neq m$, we lose the natural symmetry of the weights. However, we can still base the weights on the binomial distribution. A few possible weighting schemes are defined below:

$$w_i^{b1} = \begin{cases} 0, & i = 1, \cdots, r - m \\ w^b_{i-(r-m)}, & i = r - m + 1, \cdots, K. \end{cases}$$

$$w_i^{b2} = f(i - 1; K + 2(r - m) - 1, 0.5), \qquad i = 1, \cdots, K.$$

$$w_i^{b3} = \begin{cases} 0, & i = 1, \cdots, 2(r - m) \\ f(i - (r - m + 1); K - 2(r - m) - 1, 0.5), & i = r - m + 1, \cdots, K. \end{cases}$$

$$w_i^{b4} = f(i - 1; K - 1, \frac{r - 1}{K - 1}), \qquad i = 1, \cdots, K.$$

The weighting scheme w^{b1} is based on the same binomial distribution as w^b, except that the values are shifted so that the center of the distribution falls on r instead of m. Because of the shift, the first few weights are set to be 0, while the last few values in the probability mass function of the binomial distribution are truncated. The weighting scheme w^{b2} increases the parameter n to $K + 2(r - m) - 1$ in the binomial distribution to ensure that the first few weights are non-zero. The weighting scheme w^{b3} decreases the parameter n to $K - 2(r - m) - 1$ in the binomial distribution so that the distribution is not truncated on the right. The schemes w^{b1}, w^{b2} and w^{b3} are all based on binomial distributions with the parameter $p = 0.5$. They allow w_r to have the highest value by shifting the distribution and/or changing the parameter n in the binomial distribution. The weighting scheme w^{b4} takes a different approach by directly adjusting the parameter p, so that w_r has the highest value without having to shift the distribution. For w^{b4}, the weights are no longer symmetric with respect to w_r. See Figure 4.1 for an illustration of these four weighting schemes. Corresponding half-binomial weights for these four binomial weighting schemes can be easily constructed by setting the weights to be 0 for $i = 1, \cdots, r - 1$.

For WOP based Lancaster's statistic $T_L = \sum_{i=1}^{K} [\chi^2_{n_i}]^{-1}(1 - p_{(i)})$, the different weighting schemes can be readily applied by proportionally transforming the w_i's into integer weight n_i's. To be specific, suppose a weighting scheme is based on the binomial distribution $B(n, 0.5)$, then simply let $n_i = 2^{n+1} w_i$.

4.3.3 Implementation of the WOP method

Under the null hypothesis, the p_k's are assumed to follow uniform (0,1) distributions. For Stouffer's Z-test, $H_Z(p_i) = \phi^{-1}(1 - p_i)$ follows a standard normal distribution. Therefore the traditional weighted Z-test in the form of $\sum_{i=1}^{K} w'_i H_Z(p_i)$ still follows a normal distribution. For Fisher's method, $H_F(p_i) = -2 \log(p_i)$ has a chi-square distribution with two degrees of freedom. Therefore the distribution of the traditional weighted Fisher's test $\sum_{i=1}^{K} w'_i H_F(p_i)$ is essentially weighted sums of exponential distributions. The distribution of weighted sums of exponential variables is not as straightforward, though many authors have researched on both the exact and approximations of this distribution, a summary of which can be found in Olkin and Saner (2001). For original Lancaster's method, the statistic $\sum_{i=1}^{K} [\chi^2_{n_i}]^{-1}(1 - p_i)$ is the sum of independent chi-square variables, therefore following a chi-square distribution with $\sum_{i=1}^{K} n_i$ degrees of freedom.

When we consider weighted ordered p-values in the form of $\sum_{i=1}^{K} w_i H(p_{(i)})$, however, the problem becomes much more complicated. Even for Stouffer's method, the distribution of the sum of weighted ordered normal variables is not readily available. As for Fisher's method, Olkin and Saner (2001) studied the distribution of the trimmed Fisher's statistic, which is $\sum_{i=1}^{K} -2w_i log(p_{(i)})$ for the special case of $w_i = 0$ for $i = 1, \cdots, s_1$ and $i = s_2, \cdots, K$. They transformed the distribution of the sum of ordered chi-squared variables back to a weighted sum of exponential variables without order. But as discussed in their paper, the exact distribution of weighted sums of exponential variables are generally impractical for practitioners, and we would therefore have to use approximation methods. As for WOP based Lancaster's method, the independence between the chi-square variables no longer hold due to the dependence of the degrees of freedom on the order of the p-values, and therefore complicating the distribution of the WOP based Lancaster's statistic.

Considering the complexity of the exact distributions of weighted sums of ordered variables, as well as the fact that the uniformness of the original p-values is not always guaranteed in practice, we recommend obtaining the p-values for the WOP statistics through permutation analysis when the original data from all the studies are available. Let $T_g = \sum_{i=1}^{K} w_i H(p_{g(i)})$ denote the WOP statistic for gene g, where $p_{g(i)}$ is the ith ordered p-value of $p_{g1}, \cdots, p_{gK}$. Permute group labels in each study B times, and recalculate the p-values for the permuted data $p_{gk}^{(b)}$, for $1 \leq k \leq K$, $1 \leq g \leq G$, $1 \leq b \leq B$. Then calculate the WOP statistics for the permuted p-values $T_g^{(b)}$ for $1 \leq g \leq G$, $1 \leq b \leq B$. The p-value for the WOP statistic T_g is then computed as

$$p_g^T = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\{T_g \geq T_{g'}^{(b)}\}}{B \cdot G}$$

for $1 \leq g \leq G$. Once the p-values for the WOP statistics for each gene are obtained, we may apply the Benjamini-Hochberg (BH) method (Benjamini and Hochberg, 1995) on the p_g^T's $(1 \leq g \leq G)$ to account for multiple testing across the genes and control the false discovery rate (FDR).

If the original data are not available and only the p-values for each gene and each study are known, the permutation analysis cannot be applied. In this case, we can simulate the distribution of the WOP statistics numerically, by simulating $U(0,1)$ random variables that represent the p-values under the null distribution. The WOP statistics calculated from the data can then be compared to the numerical distribution to obtain the WOP p-values. We simulated numerical distributions of the WOP statistics for testing HS_m based on the Fisher's, Stouffer's and Lancaster's methods with binomial and half-binomial weighting schemes respectively, for study numbers ranging from 4 to 23. We conducted simulation studies to compare the WOP p-values obtained either though comparing with the numerical distribution or by performing permutation analysis. Results show that the two methods provide perfectly correlated p-values and that the number of rejections obtained from the two methods after applying the Benjamini-Hochberg adjustment are very similar (data not shown). Therefore both methods are reliable choices for obtaining the WOP p-values in practice. The numerical distribution provides an option for when the original data are not available and is also more time-efficient. The permutation analysis can be used if the uniformness of the original p-values are questionable but that the original data is available.

4.3.4 Considerations of One- or Two-Sided Tests

In Section 4.2 we used two-sided alternatives as an example when setting up the hypothesis. The hypothesis setup HS_r can be similarly developed for one-sided alternatives. In fact, the interpretation of the meta-analysis results is easier for one-sided tests, since we do not need to worry about the concordance of the directions of the effect sizes as we do for two-sided tests. Since the WOP methods directly combine the p-values, the direction of the effect sizes are not taken into account for two-sided tests. Thus a significant result from the WOP meta-analysis of two-sided tests indicate that there are at least r studies with non-zero effect size, but without any implications about the concordance or discordance of the directions of the effects. This may not be an issue in the case that the direction of effect is not of great importance. However, in genomic data analysis, it is often desirable to distinguish between up- and down-regulated genes, and a result stating that the gene is differentially expressed across many studies but with possible opposite directions of expression change may be confusing. In these cases, it might be problematic to directly apply the WOP methods to the two-sided p-values. On the other hand, since both up- and down-regulated genes may be of interest at the same time, we cannot pre-specify one particular one-sided test for all genes. For such scenarios we recommend using the test of Pearson (1934) in combination with the WOP methods. To do so, for each gene we need to conduct two WOP meta-analyses on one-sided p-values: one on the left-tailed p-values for all studies, and the other on the right-tailed p-values for all studies. Let p^L_{WOP} and p^R_{WOP} be the WOP meta-analysis p-values for the left-tailed and right-tailed tests respectively. We shall then adopt the idea of Pearson's test and define $p^C_{WOP} = \min\{1, 2\min(p^L_{WOP}, p^R_{WOP})\}$, where the superscript "C" stands for "concordant". As discussed in Owen (2009), the equation for obtaining p^C_{WOP} provides a conservative p-value for Pearson's test. By adopting the Pearson's test, the results are more interpretable. A significant result now indicates that the gene is consistently up- or down-regulated in at least r studies.

4.4 A Simulation Study

We conducted a simulation study to compare the performances of the WOP methods with the original Fisher's and Stouffer's method, as well as with the rOP method by Song and Tseng (2013). We shall also demonstrate the differences between the binomial and half-binomial weighting schemes through the simulation.

We simulate the setting of a meta-analysis of differential expression studies, with 2000 genes and 7 studies. Out of the 2000 genes, 1650 genes are assumed to be not differentially expressed in any study, while 50 genes are assumed to be differentially expressed in $1, 2, \cdots, 7$ studies respectively. The sample sizes for the treatment and control groups are randomly generated for each study, varying from 5 to 20. Gene expression values are randomly generated from normal distributions. Control samples are generated from $N(0, 1)$, as well as treatment samples that are not differentially expressed. Treatment samples that are differentially expressed are generated from $N(1, 1)$. Two-sample T-tests are used to obtain the p-values p_{gk} for each gene and each study. Our WOP methods aim at testing the hypothesis that the gene is differentially expressed in the majority of studies. In this case, $m = 4$, corresponding to the hypothesis setting HS_4. The rOP statistic (Song and

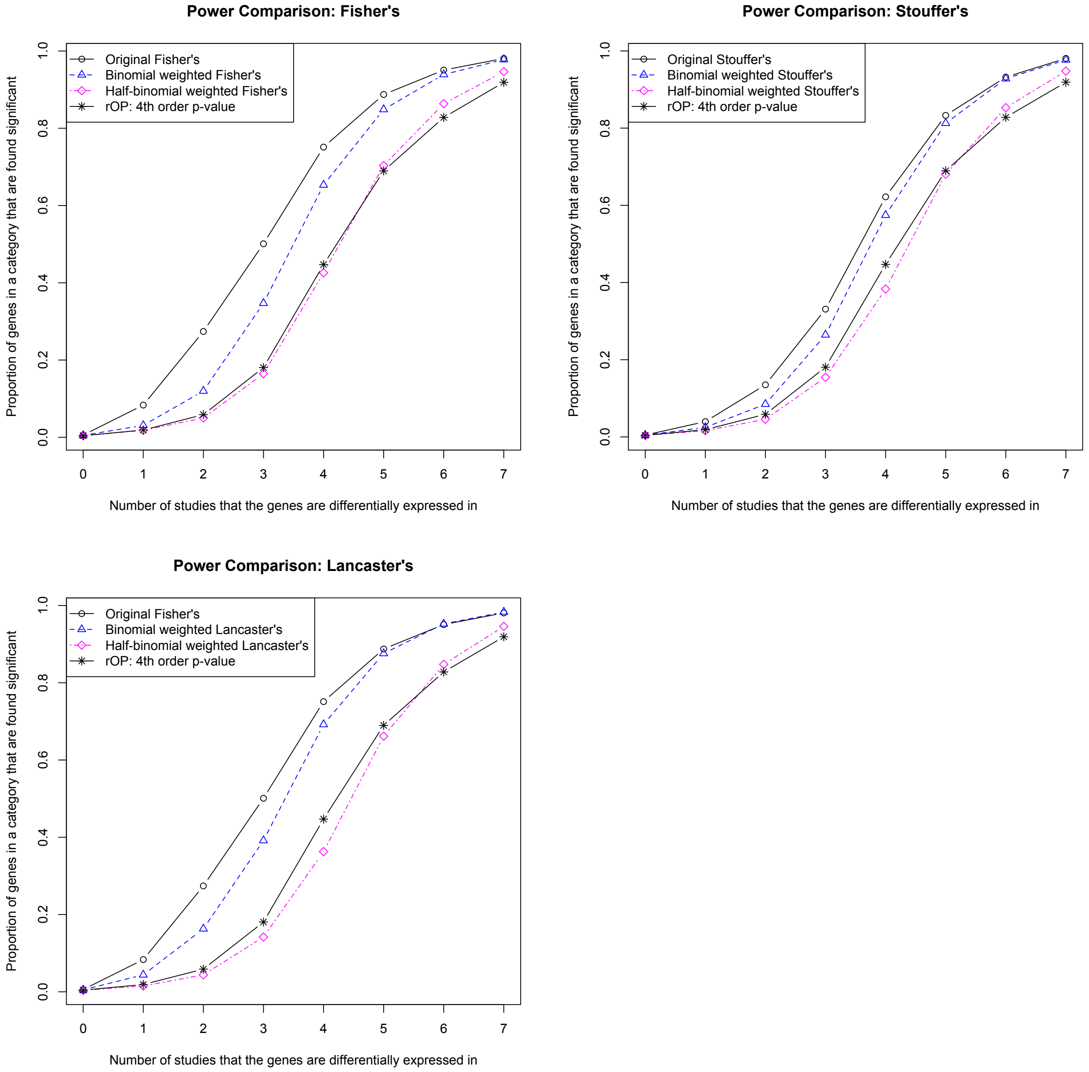

Figure 4.2. Power comparisons for the different methods. The x-axis denotes the 8 categories of genes, categorized by the number of studies that the genes are differentially expressed in. There are 1650 genes in the category 0 (no differential expression in any studies). The rest of the categories contain 50 genes each. For each category, the proportion of genes found significant within that category are plotted for each method.

Tseng, 2013) for testing HS_4 is the 4th ordered p-value. Note that the original Fisher's and Stouffer's method are supposed to test for HS_1. We used permutation analysis to obtain the WOP p-values for binomial and half-binomial weighted Fisher's and Stouffer's statistic. P-values for the rOP method were also computed by permutation analysis as recommended by Song and Tseng (2013). P-values for the original Fisher's and Stouffer's method are computed directly via their respective distributions. To obtain a list of significant genes, the Benjamini-Hochberg procedure is applied to the p-values with the FDR controlled at the 0.05 level. Results are averaged over 100 replications.

Figure 4.2 compares the power of the methods for different categories of genes. Genes are categorized based on the number of studies that they are differentially expressed in (0 to 7). The proportion of genes rejected within each category are plotted for the different methods. As expected, regardless of the method, the proportion of genes rejected within a category increases from 0 to close to 1 as the number of studies that the genes are differentially expressed in increases from 0 to 7 (out of 7). Fisher's method, being a very powerful method for testing HS_1, has the highest proportion of rejections for every category from 1 to 7. Stouffer's method also has the highest proportion of rejections when compared to the corresponding weighted versions and the rOP method. However, keep in mind that our goal in this study is to test for HS_4 instead of HS_1. Rejections in categories 0 to 3 are considered undesirable under this goal. Both the WOP methods and the rOP method reject much smaller proportions of genes in categories 1 to 3 compared to the original Fisher's or Stouffer's method. Less rejections in categories 1 to 3 come at the expense of less power for categories 4 to 7, which is true for both WOP and rOP methods, but particularly so for the rOP method. For the binomial weighted WOP methods, the rejections for categories 1 to 3 are much lower compared to Fisher's or Stouffer's method, but the power for categories 4 to 7 gradually increases to catch up with Fisher's or Stouffer's. When it comes to categories 6 and 7, especially category 7, the binomial weighted WOP methods have virtually the same power as Fisher's and Stouffer's. On the other hand, the rOP method has the lowest power for categories 6 and 7, which are the genes that are differentially expressed in all or almost all the studies. The half-binomial weighted WOP methods have the lowest rejection rates for categories 1 to 3, even lower than the rOP method, but their power surpasses the power of the rOP method when it comes to categories 6 and 7.

To better look at the trade off between rejection rates for categories 0 to 3 versus categories 4 to 7, we plotted the ROC curves for the methods, as seen in Figure 4.3. Since our goal is to test for HS_4, we define rejections in categories 0 to 3 to be

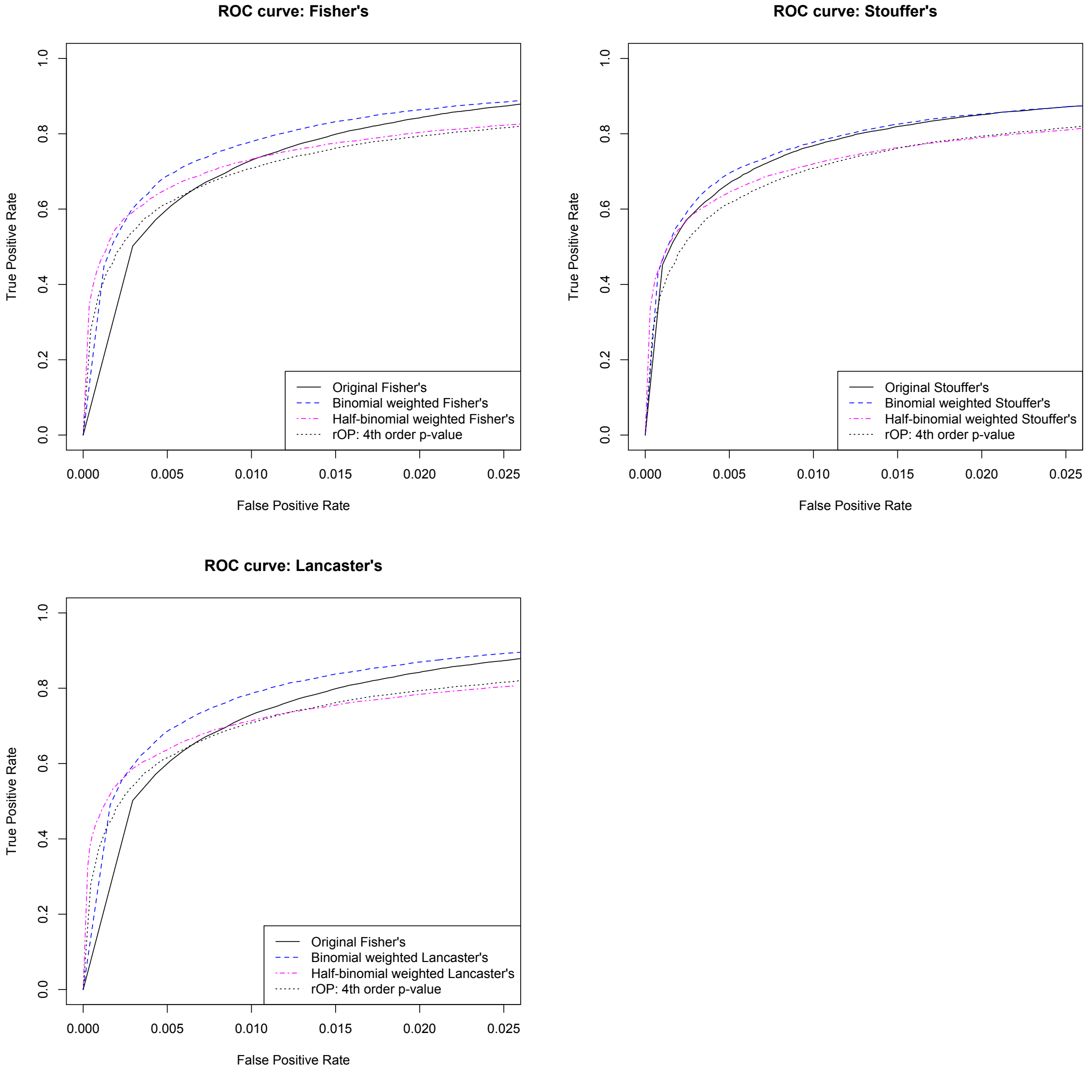

Figure 4.3. ROC curves for the different methods. Rejections of genes differentially expressed in less than 4 studies are considered false positives. Rejections of genes differentially expressed in 4 or more studies are consider true positives.

false positives, while rejections in categories 4 to 7 are considered true positives. From the ROC curves we can clearly see that in terms of the trade off between true and false positives the binomial weighted WOP methods beat the original Fisher's or Stouffer's method, while the half-binomial weighted WOP methods beat the rOP method. We also gain a better insight into the comparison between the binomial and half-binomial weighting schemes. The half-binomial weighting scheme is relatively more conservative, with relatively higher true positive rates at

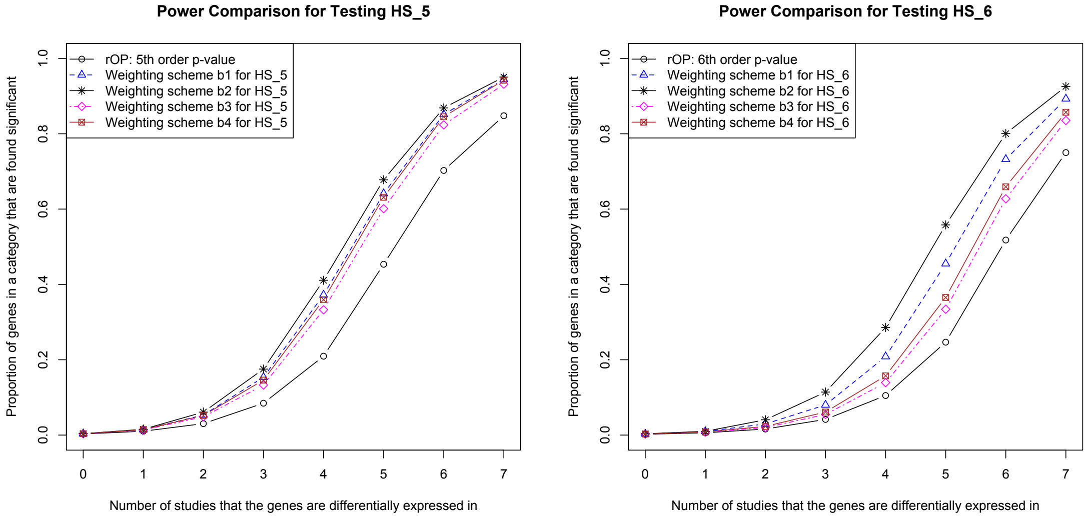

Figure 4.4. Power comparisons for the four weighting schemes w^{b1}, w^{b2}, w^{b3}, w^{b4} and the rOP method for testing HS_5 and HS_6 respectively. The weighting schemes are used in combination with the Fisher's method. The axes are configured the same way as Figure 4.2.

very low false positive rates. On the other hand, the binomial weighting scheme achieves higher power when slightly higher false positive rates are allowed.

Using the same simulation setting, we also explored the performances of the weighting schemes w^{b1}, w^{b2}, w^{b3} and w^{b4}, as well as their corresponding half-binomial versions, for testing the hypothesis settings HS_5 and HS_6. We used Fisher's method in combination with the different weighting schemes and compared the results with using the rOP method for the same hypotheses. Figures 4.4 and 4.5 show comparisons of power for the different methods. Overall, the results reflect the same trends as those discussed previously for testing HS_4. The WOP methods with weighting schemes w^{b1}, w^{b2}, w^{b3} and w^{b4} show higher rejection rates in all categories compared to the corresponding rOP method, but the power increase is more evident for the categories where most studies are differentially expressed. On the other hand, the WOP methods with the half-binomial versions of w^{b1}, w^{b2}, w^{b3} and w^{b4} are again relatively conservative, performing very similarly to the rOP methods. From the definition of the weighting schemes we can see that the half-binomial weighting schemes would be more and more close to the rOP methods as r increases, degenerating to the rOP method when $r = K$. But

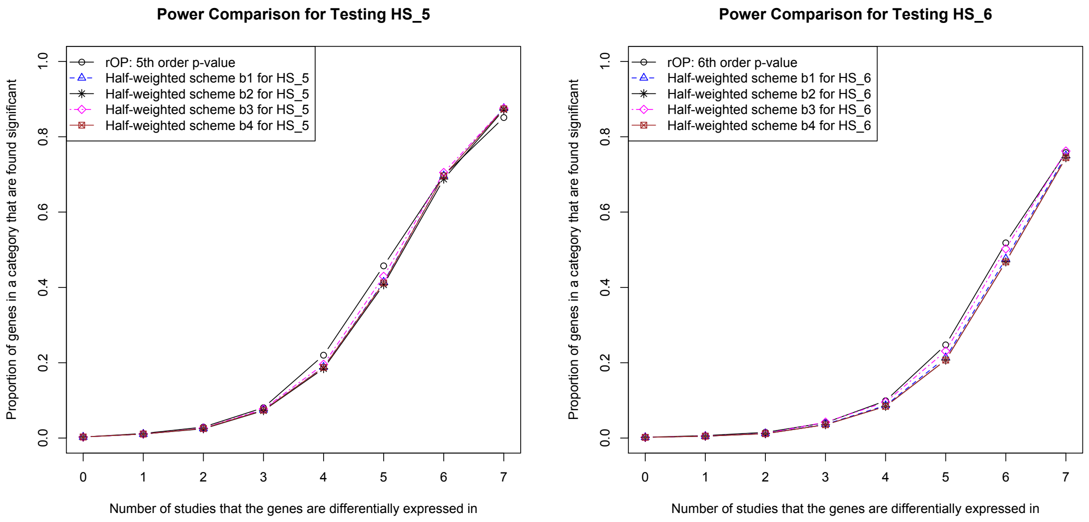

Figure 4.5. Power comparisons for the half-binomial versions of the four weighting schemes w^{b1}, w^{b2}, w^{b3}, w^{b4} and the rOP method for testing HS_5 and HS_6 respectively. The weighting schemes are used in combination with the Fisher's method. The axes are configured the same way as Figure 4.2.

for the case of HS_5 we can still see that the half-binomial weighting schemes have less rejections than the corresponding rOP methods in the first few categories but gradually catching up and surpassing for categories 6 and 7. Among the three weighting schemes, w^{b3} performs closest to the rOP method, while w^{b2} shows the most difference. This is because w^{b3} has the most concentrated weights, which degenerates to the rOP method when testing HS_7, whereas w^{b2} has the most spread out weights. The weighting scheme w^{b4} generally falls between w^{b1} and w^{b3}.

In summary, the binomial weighted WOP methods, can be considered an improvement over the original Fisher's or Stouffer's method for testing differential expression in the majority of studies, with lower rejection rates for genes that are differentially expressed in a small number of studies, and just as high power for genes that are differentially expressed in almost all studies. On the other hand, the half-binomial weighted WOP methods are more robust versions of the rOP method, having similar properties to the rOP method, but even lower rejections rates for categories 1 to 3 and higher power for categories 6 and 7.

4.5 An Application to Meta-Analysis of a Set of Stem Cell Studies

As an application of the proposed methodology, we conduct meta-analysis on a set of microarray data studies from four stem cell papers: Chin et al. (2009), Guenther et al. (2010), Newman and Cooper (2010) and Chin et al. (2010). We aim at looking for genes that are differentially expressed between human induced pluripotent stem (hiPS) cells and human embryonic stem (hES) cells in the majority of studies. Some of the studies contain other samples such as human fibroblasts, but we only used samples from hiPS cells and hES cells. We included studies that had at least two samples for each group (hiPS and hES), giving us a total of 9 studies. All the studies used the Affymetrix Human Genome U133 Plus 2.0 Array platform, which contains 54675 probesets. We directly used the data preprocessed by the original contributors and did not perform any additional normalization, except for taking the log for data that were not already on the log2 scale. We performed a two sample T-test between the hiPS cell samples and the hES cell samples to obtain the original p-values for differential expression for each gene and each study. The hypothesis setting for the meta-analysis is HS_5, i.e. we aim at testing the alternative that the gene is differentially expressed in at least 5 out of the 9 studies. We applied the proposed WOP methods, in particular the binomial and half-binomial weighted Fisher's and Stouffer's statistic to the p-values. The p-values for the WOP statistics are obtained by comparing the statistics to the corresponding numerical distributions. We also applied the original Fisher's and Stouffer's method, and the rOP method (in this case the 5th ordered p-value). The p-values for the rOP method are obtained via its theoretical distribution. To adjust for multiple testing, we applied the Benjamini-Hochberg (1995) procedure afterwards, controlling the false discovery rate at the 0.05 level.

As discussed in Section 4.3.4, there are many options regarding one- and two-sided tests. We can: 1) directly apply the meta-analysis methods to the two-sided p-values to find consistently differentially expressed genes; 2) apply the meta-analysis methods to the right-tailed p-values to find consistently up-regulated genes; 3) apply the meta-analysis methods to the left-tailed p-values to find consistently down-regulated genes; 4) apply the meta-analysis methods to both left-

Table 4.1. Number of significantly differentially expressed probesets for the meta-analysis of stem cell studies by different methods.

Method	Fisher's	Stouffer's	rOP
Original (unweighted)	16508	11969	6330
WOP: Binomial weighted	10309	8927	N/A
WOP: Half-binomial weighted	6170	5805	N/A

Table 4.2. Number of significantly up-regulated probesets for the meta-analysis of stem cell studies by different methods.

Method	Fisher's	Stouffer's	rOP
Original (unweighted)	3868	1302	1074
WOP: Binomial weighted	1744	1310	N/A
WOP: Half-binomial weighted	1046	909	N/A

Table 4.3. Number of significantly down-regulated probesets for the meta-analysis of stem cell studies by different methods.

Method	Fisher's	Stouffer's	rOP
Original (unweighted)	11384	5862	3306
WOP: Binomial weighted	5816	4674	N/A
WOP: Half-binomial weighted	3294	3019	N/A

and right-tailed p-values respectively and adopt the Pearson's test to find consistently up- or down-regulated genes. We applied all four options to the stem cell data to see how these different options affect the results. The results of the number of probesets found significant by the different meta-analysis methods are summarized in Tables 4.1, 4.2, 4.3 and 4.4, corresponding to the four options respectively. Comparing across the four tables we see that Table 4.1 has the largest number of significant probesets. Since the direction of effect sizes are not distinguished by directly using two-sided p-values, the significant probesets in Table 4.1 may potentially have opposite directions of effects across the studies. Tables 4.2

Table 4.4. Number of significantly up- or down-regulated probesets using Pearson's test for the meta-analysis of stem cell studies by different methods.

Method	Fisher's	Stouffer's	rOP
Original (unweighted)	14078	6369	4012
WOP: Binomial weighted	6594	5327	N/A
WOP: Half-binomial weighted	3918	3465	N/A

and 4.3 have relatively smaller numbers of significant probesets since they are focused specifically on either up- or down-regulation. For this particular application, there seems to be much more down-regulation of hiPS versus hES compared to up-regulation. Table 4.4 includes both significantly up- and down-regulated probesets, but the numbers are smaller than the sums of Tables 4.2 and 4.3 because of the adjustment in Pearson's test. Comparing Tables 4.1 and 4.4, the biggest drop in number is for original Stouffer's method. A possible explanation is that, in the case where the probeset is significantly up-regulated in one study while significantly down-regulated in another study, applying the original Stouffer's test on one-sided p-values will cancel out the effect of these two studies for the probeset, since the two corresponding Z-scores will have opposite signs. This cancellation of effects would not happen, for instance, with the original Fisher's method. This makes it such that the original Stouffer's method is relatively more critical when it comes to ensuring the concordance of the direction of effect when applied to one-sided p-values. Figure 4.6 shows a Venn diagram of the probesets found by using the four different options combined with the half-binomial weighted Fisher's method. We see that no probeset is found to be both significantly up- and down-regulated. All the probesets found to be significantly up-regulated are also identified by the Pearson's test, giving us confidence in this relatively small pool of probesets. The set of significantly down-regulated probesets is much larger, most of which are also found by Pearson's test, another subset of which are also found by the two-sided tests, but still some that were not identified by other methods. Both Pearson's test and the two-sided test found probesets that were not identified by either up- or down-regulation, with the two-sided test yielding a much larger number, possibly some of which having discordant direction of effect.

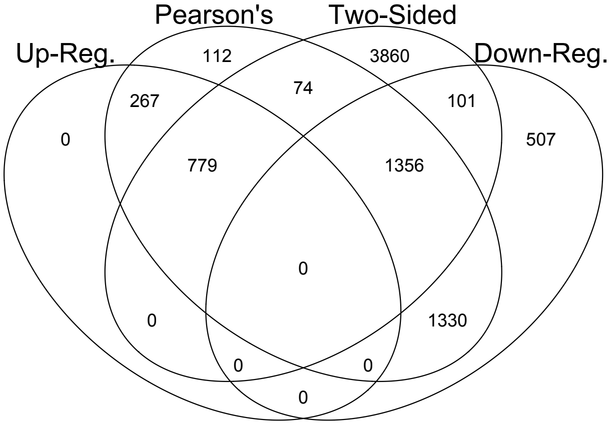

Figure 4.6. Venn diagram for the probesets found significantly differentially expressed (two-sided), significantly up-regulated, significantly down-regulated, and significantly up- or down-regulated (Pearson't test) by the half-binomial weighted Fisher's method.

Next we compare the results across different meta-analysis methods. For each table, the original Fisher's and Stouffer's methods finds the most number of significant probesets. The binomial weighted WOP methods finds much less significant probesets compared to the corresponding unweighted methods, but more than the rOP method. The half-binomial weighted WOP methods finds the least number of significant probesets. This agrees with the results from the simulation studies. The original Fisher's and Stouffer's methods have likely picked out many genes that are differentially expressed in a few, but less than half the studies. The half-binomial weighted WOP methods and the rOP method are relatively more conservative and focused. Figure 4.7 shows a Venn diagram of the probesets found significant by the binomial and half-binomial weighted Fisher's method and the rOP method from Table 4.1. We can see that the probesets detected by both the half-binomial weighted Fisher's method and the rOP method are mostly detected by the binomial weighted Fisher's method as well. However, there are still a number of unique genes that are only detected by either the half-binomial weighted Fisher's

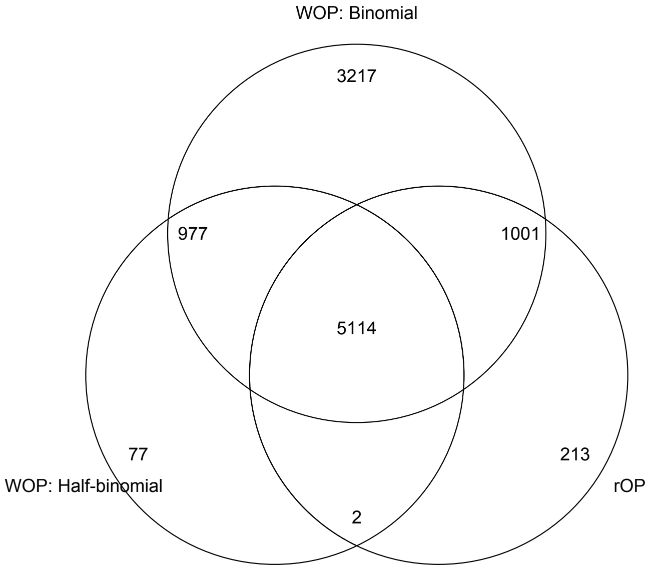

Figure 4.7. Venn diagram for the probesets found significantly differentially expressed (two-sided) by the binomial weighted Fisher's method, the half-binomial weighted Fisher's method and the rOP method.

method or the rOP method. To get an idea of the types of probesets detected by only one of the three aforementioned methods, we randomly selected some of these probesets and plotted the ordered original p-values from the 9 studies for these probesets. As seen in Figure 4.8, probesets exclusively detected by the binomial weighted Fisher's method tend to have very small values for the two or three smallest p-values, but relatively larger values starting the 5th ordered p-value. This shows that the binomial weighted Fisher's method is more prone to influences by the smallest p-values, since it takes into account all the p-values in the statistics. The rOP method, which uses the 5th ordered p-value as the statistic, exclusively identifies probesets that are guaranteed to have relatively small p-values up to the 5th ordered p-value, but tend to have very large values starting the 6th ordered p-value. This shows the sensitivity of the rOP method to the particular value of r chosen. On the other hand, the half-binomial weighted Fisher's method weights in the 5th through the 9th ordered p-values, and thus is able to identify probesets that have relatively small values through larger ordered p-values. In other words, probesets that have relatively small p-values for most of the studies can be exclusively identified by the half-binomial weighted method, even if the smallest p-values are not very small.

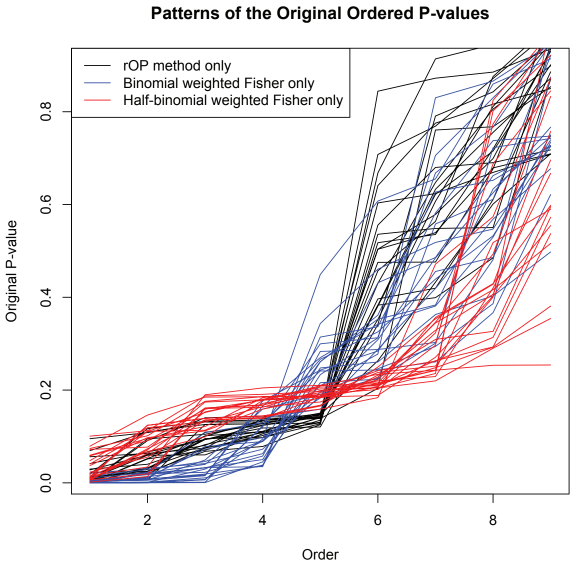

Figure 4.8. Pattern of the original ordered p-values from the 9 studies for probesets detected by one of the three methods only. The x-axis is the order of the p-values from the 9 studies. The y-axis is the p-values. The plot includes a random subset of 20 probesets that are detected exclusively by each of the three methods.

To look at the pathways associated with the significant lists of probesets, we performed functional annotation clustering analysis using DAVID (Huang et al, 2009), which is available at http://david.abcc.ncifcrf.gov/home.jsp. Some of the top functions that show up include metal-binding, nucleoplasm, ubl conjugation, vasculature development and head/face development. We also looked at the pathways for the probesets that were exclusively detected by one of the methods. Functions such as neuron projection, neuron differentiation and development, which are meaningful in the stem cell study setting, were found to be associated with the probesets exclusively identified by the half-binomial weighted Fisher's method. These functions did not show up in pathway analyses of the other lists.

4.6 Further Applications and Comparisons with the rOP Method

To compare the performances of the WOP methods and the rOP method (Song and Tseng, 2013) in real data application, we applied our WOP methods to the three microarray meta-analysis applications in Song and Tseng (2013). The first application consists of comparisons of two subtypes of brain tumors - anaplastic astrocytoma (AA) and glioblastoma multiforme (GBM), from 7 studies. The second application combines 9 studies comparing post-mortem brain tissues between MDD patients and control samples. In the third application, 16 diabetes microarray studies consisting of different organisms and tissues were combined. See Song and Tseng (2013) for more details on the contexts of these three meta-analysis applications.

To ensure that the results are directly comparable, for each meta-analysis we directly used the two-sided p-values for each gene and each individual study calculated by Song and Tseng (2013). In Song and Tseng (2013), permutation analysis is used for the brain cancer studies and the MDD studies, while theoretical distributions are used to obtain results for the diabetes studies. We follow Song and Tseng (2013) and also directly use the two-sided p-values for the permuted datasets provided by Song and Tseng (2013) for the brain cancer studies and the MDD studies. See Song and Tseng (2013) for more details on the preprocessing of the data and the calculation of the original p-values.

We applied our WOP methods, namely the binomial and half-binomial weighted Fisher's and Stouffer's statistic, to the three sets of studies. For the brain cancer studies, we test for HS_4 out of 7 studies. For the MDD studies we test for HS_5 out of 9 studies. For the diabetes studies, we test for HS_9 out of 16 studies. We also applied the corresponding rOP methods for testing the same hypotheses, using $r = m = 4,\ 5$ and 9 respectively for the three meta-analyses. In addition, we applied the rOP methods using the selected r values in Song and Tseng (2013) for comparison. To be specific, Song and Tseng (2013) used $r = 5$ for the brain cancer studies, $r = 7$ for the MDD studies, and $r = 12$ for the diabetes studies. Permutation analysis is used for the brain cancer studies and the MDD studies for all methods. For the diabetes studies, we used our numerically simulated distribution to obtain the p-values for the WOP statistics.

Table 4.5. Number of significant genes for the three meta-analyses by different methods. For the rOP method, “selected r” refers to the choice of r in Song and Tseng (2013) for a particular meta-analysis.

	WOP	WOP	WOP	WOP	rOP	rOP
	Fisher	Fisher	Stouffer	Stouffer	$r = m$	selected r
	binomial	half-binomial	binomial	half-binomial		
Brain Cancer	2477	1887	2261	1805	1921	1469
MDD	1070	930	1152	969	565	617
Diabetes	1333	1016	1277	1004	912	636

Table 4.5 shows the numbers of significant genes using the different methods for the three meta-analyses with the FDR controlled at the 0.05 level. We can observe that the WOP methods using the binomial weighting scheme generally detects more significant genes than the half-binomial weighting scheme. In most cases, the rOP methods (using either $r = m$ or selected r) detect less genes than the WOP methods, although in general closer in number to the WOP methods using the half-binomial weighting scheme. One interesting observation is that for the MDD studies, the rOP method based on $r = m = 5$ detects less genes than based on the selected $r = 7$. This result is counterintuitive, since one would expect that genes that are significant in at least 7 studies would be a subset of the genes that are significant in at least 5 studies. The result could be due to the fact that r is selected in Song and Tseng (2013) to optimize the results and therefore outperforms a general choice of $r = m$. However, this still reflects the fact that the rOP method is sensitive to the choice of r.

To further investigate this problem, we looked at the overlap of the detected genes by the rOP methods using either $r = m$ or selected r, as well as with the detected genes by the half-binomial weighted Fisher's method. See Figure 4.9 for a Venn diagram of the genes detected by the aforementioned three methods for the MDD studies. As shown in Figure 4.9, only 269 genes overlap between the two rOP methods, which is less than half of the genes detected by either method. However, all 269 genes are detected by the half-binomial weighted Fisher's method. In addition, the WOP method also picked up 251 of the genes only detected by rOP

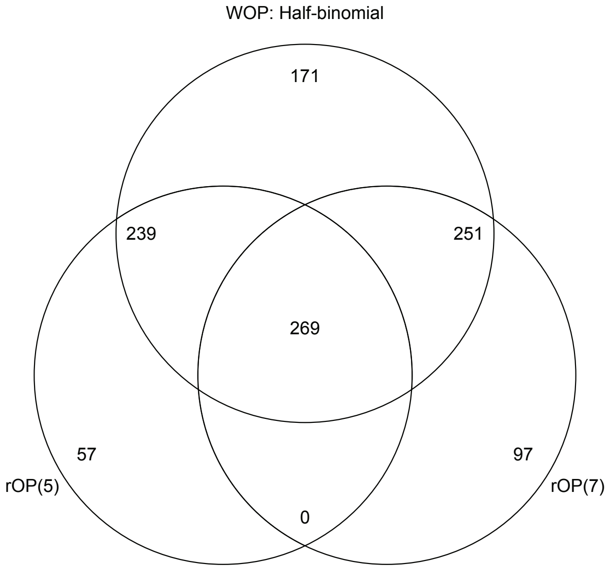

Figure 4.9. Venn diagram for the genes found significant in the meta-analysis of the MDD studies by the rOP method based on $r = m$, the rOP method based on selected r, and the half binomial weighted Fisher's method. In this case, $m = 5$ and the selected $r = 7$.

based on selected r and 239 of the genes only detected by rOP based on $r = m$, which accounts for most of the genes detected by either method. Further, we noticed that even for the brain cancer studies and the diabetes studies, where rOP based on $r = m$ did detect more genes than rOP based on selected r, there are still a large number of genes detected by rOP based on selected r that are not detected by rOP based on $r = m$. On the other hand, most of these genes that are detected by only one of the rOP methods are detected by the half-binomial weighted Fisher's method. From these observations we see that the WOP method is more robust compared to the rOP method for detecting significance in the majority of studies.

4.7 Discussion

Meta-analysis is a useful tool in integrating data from different sources to test a particular hypothesis. While this paper mainly discussed the application of meta-analysis on microarray differential expression studies, other areas of genomic studies have increasingly relied on the use of meta-analysis, such as genome-wide association studies (GWAS). Some seminal studies include Scott et al. (2007) and Willer et al. (2008). Meta-analysis is also frequently used in clinical studies, psychological studies and statistical applications in other social sciences. More and more meta-analyses nowadays aim at detecting consistent findings across a number of studies. While most of the traditional meta-analysis methods test for significance in at least one of the studies, it is important to develop new meta-analysis methods that focus on testing for significance in the majority of studies.

The weighted ordered p-value (WOP) method provides such a framework. It is unique in its use of weights that are based on the order of the p-values. The rOP method by Song and Tseng (2013), which is also based on ordered p-values, can be considered a very special case under the WOP framework, where all the weight is placed on one single ordered p-value.The WOP methods does not require pre-specification of r and is less sensitive to the choice of its value. The half-binomial weighted WOP methods have been shown to be more robust and have better receiver operating characteristics compared to the corresponding rOP method.

The WOP framework has the advantage of being flexible. The framework allows for different weighting schemes and summary statistics to be used. Even though this paper mainly focused on two particular weighting schemes based on the binomial distribution and two summary statistics (Fisher's and Stouffer's statistics), in general, other summary statistics and weighting schemes can be used. Future research can be done to try to optimize the weighting scheme to suit specific meta-analysis purposes.

Chapter 5

Conclusions and Future Work

This thesis touched on a few topics in the wide range of statistical methodology used for analyzing large scale genomic data. In the past decade, high-throughput technology has really revolutionized our capabilities in exploring the expression of genes and its associations with mutations, diseases and other important biological implications. With the vast amounts of genomic data generated and the possible problems we might tackle with these data comes a whole new world of statistical challenges.

The topics in this thesis are mainly grounded in one of the basic problems in genomic data analysis – differential expression analysis. While many methods have already been established for normalization of array data and basic differential expression analysis, and the use of the false discovery rate (FDR) has set widely accepted standards for type-I error control in the case of large scale testing, there are still many improvements to be made when it comes to more complicated designs. We mainly considered problems where there are multiple p-values associated with each gene. To be specific, our motivating examples include problems such as meta-analysis of differential expression analyses, or time-course experiments where we are interested in differential expression between multiple time points.

Two Chapters in this thesis (Chapters 2 and 4) considered the problem of meta-analysis. Meta-analysis of genomic data has become increasingly popular with the growing amounts of public data available that are targeting the same questions. By pooling results from several studies, researchers can look for genes that are consistently identified across multiple studies while ruling out genes that only showed up significant in single studies. This brings us to one of the big issue in meta-analysis of genomic data, which is heterogeneity among studies. How to deal with heterogeneity while identifying consistent signal is the main theme of Chapters 2 and 4.

The methods in the two chapters have different focuses though. The weighted multiple testing procedure for meta-analysis in Chapter 2 is a more integrated approach that treats the problem of heterogeneity in meta-analysis and the problem of large scale multiple testing as a whole. The method starts with conducting basic meta-analysis for each gene, and then integrates the information of heterogeneity into the multiple testing step. By up-weighting the genes that display more consistent behavior and down-weighting the genes that show more heterogeneity, the method is able to prioritize genes that are more consistently differentially expressed across studies while controlling the false discovery rate. On the other hand, the weighted ordered p-values (WOP) method in Chapter 4 is more focused on meta-analysis itself, aiming at testing whether a gene is differentially expressed in the majority of studies. Because of this, the method in Chapter 4 is in no way limited to genomic data analysis, or even any large scale data. It can be used in any meta-analysis problem where the goal is to detect signals in the majority of studies. The idea of the method is to summarize the p-values from the multiple studies while weighting them according to their ordering. Since we care about the majority of the studies, the median p-values are given more weight while the outlying p-values are given less weight. The idea of assigning weights according to the order of the p-values is the key to this method and different from previous weighted p-value combination methods.

Both methods in Chapters 2 and 4 rely on the use of weights, albeit in very different ways. For both methods, more work can be done in exploring different weighting schemes. For the weighted multiple testing procedure in Chapter 2, the key is in finding more robust measures of heterogeneity among the studies. As we can see from the simulation studies, different measure of heterogeneity lead to different weights and ultimately different prioritization of genes. Therefore it is important to find a robust measure that best reflects the research interest. For the WOP method in Chapter 4, we explored a few variations of weighting schemes, but all based on the binomial distribution. We chose the binomial distribution because it conforms to our desired shape, its good properties such as symmetry, as well as its simplicity in implementation. However, it would be interesting to explore other forms of weights, and compare their respective properties. Another future working direction would be to research more on the case of testing HS_r where $m < r \leq K$. We explored this case a little bit in Chapter 4, but not extensively. It would be especially interesting to come up with unique weighting schemes to deal with this unsymmetrical situation.

Chapter 3 of this thesis considered the problem of what we call "multi-dimensional" multiplet testing. It specifically deals with the case where not only thousands of genes are being tested at the same time, but multiple tests are being conducted on each gene. It addresses the problem of how to control the false discovery rate in this situation. We adopted the concept of the Overall FDR (OFDR), which is a measure of the FDR on the hypothesis set level. By doing so, we aim at controlling the false discoveries on the gene level, which increases the interpretability of the results, compared to focusing on the FDR of all the individual hypotheses. The key point in the hypothesis set testing framework is that the two-dimensional multiplicity is transformed into a hierarchical structure. Hypotheses are tested in the unit of sets in the first step. This is realized by the formulation of a screening hypothesis for each set. The first step of our procedures deals with the hypothesis sets much like dealing with a traditional multiple testing problem. By applying the Benjamini-Hochberg procedure to the screening hypotheses, we are able to adjust for part of the multiplicity on the hypothesis set level. Additional type I errors (and sometimes directional errors) that can potentially occur while making inference for the individual hypotheses in each set are controlled in the second step by applying family-wise error rate controlling procedures. Together, the OFDR (or mdFDR) is controlled at the hypothesis set level.

The hierarchical hypothesis set testing framework is very flexible. The formulation of the screening hypothesis for each set are determined by the particular problem setting. Even the false discovery rate controlling procedure in the first step can be adjusted according the specific situation. For example, if a large proportion of individual hypotheses are expected to be significant given the significance of the hypothesis set, then we can potentially improve power by incorporating adaptive multiple testing methods into the procedure. Future work can be done in adapting the framework to different multi-dimensional multiple testing problems.

For all three methods discussed in this thesis, most of the theory is based on the independence of p-values across the genes. Previous research have shown that the conditions for the original Benjamini-Hochberg procedure to control the FDR can be loosened to certain positive dependence structures across genes (see e.g. Benjamini and Yekutieli, 2000), and thus providing some intuition that our methods can still be useful under these situations, since most of our methods are directly or indirectly based on the original B-H procedure. However, an area of future work can definitely be to research how our methods would be affected without the independence assumption – how far the conditions can be loosened and whether amendments to the methods need to be made.

Appendix A

Pathway Analysis Results for Section 3.6.1

The following tables display the functional annotation clustering results by DAVID, performed on the 6 significant gene lists in Section 3.6.1. One list corresponds to the screening test - “solid” versus “hematolymphoid”. The other five lists correspond to: (1) lymphoma versus leukemia; (2) CNS versus the others belonging to “solid”; (3) mesothelioma versus the others belonging to “solid”; (4) melanoma versus the others belonging to “solid”; and (5) subclass “epithelial” versus the others belonging to “solid”. Lists (2)-(5) are lists created by taking the intersection of the lists from pairwise comparisons. In the tables, the first column contains the enrichment scores for the groups of features, ranked from high to low, and the second column displays the features.

Table A.1. Functional annotation clustering results (DAVID) for Section 3.6.1.

	Solid versus Hematolymphoid
40.98	Rna binding
40.38	Ribonucleoprotein, ribosome, cytosolic part
30.43	Membrane-enclosed lumen, organelle lumen
28.18	mRNA processing/splicing
23.04	Non-membrane-bounded organelle, cytoskeleton
16.17	Ubl conjugation, isopeptide bond
14.7	Nucleotide binding, RNA recognition motif, RRM
9.15	Transcription factor/activator activity
7.54	Macromolecule complex subunit organization, protein complex assembly
7.32	Leukocyte/lymphocyte activation/differentiation, Immune system development, T cell/B cell activation/differentiation

Table A.2. Functional annotation clustering results (DAVID) for Section 3.6.1.

	CNS versus the others in Solid
18.11	Nuclear lumen, intracellular organelle lumen, membrane-enclosed lumen
11.94	Protein catabolic process, ubl conjugation pathway
11.35	mRNA metabolic process, spliceosome
9.28	Non-membrane-bounded organelle, cytoskeleton
8.71	RNA binding, nucleotide-binding, RNA recognition motif, RRM
8.7	Chromatin modification, chromosome organization, histone modification
8.33	Nucleotide binding, ribonucleotide binding, ATP binding, kinase
7.65	Negative regulation of biosynthetic process/transcription
7.22	Intracellular transport, protein transport, protein localization
6.82	Liqase, small protein conjugation, ubiquitin-protein liqase activitiy
6.44	Transcription factor/activator binding
6.26	Golgi vesicle transport
5.21	Golgi apparatus
4.92	Histone modification, histone acetyltransferase complex

Table A.3. Functional annotation clustering results (DAVID) for Section 3.6.1.

	Melanoma versus the others in Solid
4.51	Isopeptide bond, ubl conjugation
4.48	Negative regulation of biosynthetic process/transcription
3.92	Nuclear lumen, nucleoplasm, organelle lumen
3.27	Nuclear envelope, nuclear pore, organelle envelope
3.12	Nucleotide biding, ATP binding, protein kinase
2.71	Cell death, apoptosis
2.58	Transcription, DNA-binding
2.5	Nuclear chromosome
2.43	DNA damage response, signal transduction, cell cycle checkpoint
2.21	Chromatin organization
1.95	DNA damage response, induction of apoptosis by intracellular signals

Table A.4. Functional annotation clustering results (DAVID) for Section 3.6.1.

	Mesothelioma versus the others in Solid
1.98	Ubl cojugation, isopeptide bond
1.53	EGF-like calcium binding
1.4	Nuclear envelope, nuclear pore, organelle envelope
1.35	bromodomain
1.27	Positive regulation of transcription/RNA metabolic process
1.21	Endomembrane system, intracellular transport, protein localization
1.17	Isopeptide bond, transciption, RNA biosynthetic process
1.08	Activation of immune response, serine proteinase, complement pathway

Table A.5. Functional annotation clustering results (DAVID) for Section 3.6.1.

	Epithelial versus the others in Solid
2.52	Protein amino acid glycosylation
1.33	Transcription
1.07	Endomembrane system, intracellular transport, protein localization
0.99	Intermediate filament protein
0.98	Contractile fiber, muscle protein, muscle system process

Table A.6. Functional annotation clustering results (DAVID) for Section 3.6.1.

	Leukemia versus Lymphoma
31.17	Organelle lumen, nucleoplasm, membrane-enclosed lumen
20.73	Ubl conjugation, isopeptide bond
19.4	mRNA splicing/processing
18.41	Ribonucleoprotein, cytosolic ribosome
16.94	RNA binding, RNA recognition motif, RRM
16.75	Non-membrane-bounded organelle, cytoskeleton
11.7	p-loop, nucleotide binding, gtp binding
10.73	Transcription, dna binding
10.69	Transcription factor/activator activity
10.62	Regulation/induction of apoptosis, regulation of cell death
9.08	Positive regulation of macromolecule metabolic/ biosynthetic process
8.86	Nuclear chromosome
8.67	Negative regulation of macromolecule metabolic/ biosynthetic process
8.48	Atp-binding, phosphotransferase, protein kinase
7.83	DNA repair, response to DNA damage stimulus, cellular response to stress
7.29	Cell fraction, insoluble fraction, membrane fraction
7.06	Regulation of protein metabolic process, liqase activity, proteasome
6.96	Immune system development, T cell/B cell activation/differentiation, leukocyte/lymphocyte activation/differentiation

Appendix **B**

Pathway Analysis Results for Section 3.6.2

The following tables display the functional annotation clustering results by DAVID, performed on a total of 26 significant gene lists from Section 3.6.2. 13 of the lists are produced by Procedure 3.10 and the other 13 by Procedure 3.11. The first list out of the thirteen corresponds to the overall screening test - the meta-analysis of tumor versus control for the 12 types of tumor . The other twelve lists correspond to the 12 tumor types respectively. In the tables, the first column contains the enrichment scores for the groups of features, ranked from high to low, and the second column displays the features.

Table B.1. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Overall (by Procedure 3.10)
7.09	Cytoskeletal protein binding, actin binding
6.23	Negative regulation of biosynthetic/macromolecule metabolic process
5	Intracellular transport, protein localization
4.98	Actin filament-based process, cytoskeleton organization
4.9	Nuclear lumen, organelle lumen
4.81	Cell death, apoptosis
4.55	Actin filament bundle, stress fiber, actomyosin
4.21	Vasculature development, blood vessel development
3.85	Cytoskeleton, non-membrane-bounded organelle
3.77	Rna-binding, RNA recognition motif, RRM

Table B.2. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Overall (by Procedure 3.11)
9.33	Cytoskeletal protein binding, actin binding
6.63	Vasculature development, blood vessel morphogenesis
5.88	Cytoskeleton, non-membrane-bounded organelle
5.65	Actin filament-based process, cytoskeleton organization
5.02	Protein dimerization activity, identical protein binding
4.52	Tube development, lung development
4.47	Nuclear lumen, membrane-enclosed lumen, organelle lumen
3.86	Actin filament bundle, stress fiber, actomysosin
3.51	Intracellular transport, protein localization
3.45	Regulation of cell motion/migration

Table B.3. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Breast (by Procedure 3.10)
36.11	Translational elongation, ribosome, ribonucleoprotein, protein biosynthesis
5.73	Membrane-enclosed lumen, intracellular organelle lumen
4.99	Rna-binding, nucleotide-binding, RNA recognition motif, RRM
4.83	RNA processing, RNA splicing
3.78	Mitochondrion, transit peptide, organelle envelope
3.78	Diamond-blackfan anemia, ribosome biogenesis
3.53	Initiation factor, translation factor
3.13	Isopeptide bond, ubl conjugation
3.1	Mitochondrial inner membrane, oxidoreductase activity, acting on heme group of donors, oxidative phosphorylation, Parkinsons disease, cardiac muscle contraction, Alzheimers disease
2.9	Cellular macromolecular complex, protein complex biogenesis

Table B.4. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Breast (by Procedure 3.11)
47.43	Translational elongation, ribosome, ribonucleoprotein, protein biosynthesis
17.31	Structural molecule activity, intracellular non-membrane-bounded organelle
9.69	Membrane-enclosed lumen, intracellular organelle lumen
7.89	RNA processing, RNA splicing
7.24	Translational initiation, initiation factor
6.58	Ribonucleoprotein complex biogenesis, Diamond-blackfan anemia
5.9	Respiratory chain, mitochondrion inner membrane, Huntingtons disease, Parkinsons disease, Alzheimers disease
5.79	Isopeptide bond, ubl conjugation
5.73	RNA binding, RRM, RNA recognition motif
5.12	Respiratory chain, Mitochondrial inner membrane, oxidoreductase activity, oxidative phosphorylation, Parkisons disease, cardiac muscle contraction, Alzheimers disease
4.22	Posttranscriptional regulation of gene expression

Table B.5. Functional annotation clustering results (DAVID) for Section 3.6.2.

	CNS (by Procedure 3.10)
7.42	Nuclear lumen, organelle lumen, nucleoplasm
5.15	Nucleolus, non-membrane-bounded organelle, cytoskeleton
3.44	Nucleus, transcription, dna-binding
3.06	(positive) regulation of macromolecule biosynthetic process/transcription
2.71	Repeat: WD 1/2/3/4/5/6/7, WD40 repeat
2.5	Chromosome organization, chromatin regulator/modification
2.49	Domain: ARID, AT-rich interaction region
2.43	Zinc finger, PHD-finger
2.42	Regulation of mRNA stability
2.38	Mitochondrial inner membrane, Alzheimers disease, electron transfer, oxidative phosphorylation, inorganic cation/hydrogen ion transmembrane transporter activity, oxidoreductase activity, cardiac muscle contraction, Huntingtons disease, respiratory chain, organelle envelope, Parkinsons disease

Table B.6. Functional annotation clustering results (DAVID) for Section 3.6.2.

	CNS (by Procedure 3.11)
6.58	Nuclear lumen, organelle lumen, nucleoplasm
5.12	non-membrane-bounded organelle, cytoskeleton
4.5	Domain: ARID, AT-rich interaction region
3.75	Inorganic cation/hydrogen ion transmembrane transporter activity, oxidative phosphorylation, Alzheimers disease, electron transfer, Hungtingtons disease, respiratory chain, cardiac muscle contraction, generation of precursor metabolites and energy, oxidoreductase activity, Parkinsons disease, organelle envelope,
3.67	Chromatin regulator/organization
3.57	Chromosome
3.52	mRNA metabolic process/splicing
3.28	(positive) regulation of transcription
3.19	(negative) regulation of macromolecule biosynthetic process/transcription
3.02	Macromolecular complex assembly, protein complex biogenesis

Table B.7. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Colon (by Procedure 3.10)
3.12	Actin cytoskeleton organization, actin filament-based process
2.47	Cell morphogenesis, neuron projection development, axongenesis
2	Peroxisome, microbody
1.87	Cytoskeletal protein binding, actin binding
1.82	Cell death, apoptosis
1.71	Actin filament bundle, stress fiber, actomyosin
1.69	Cell-matrix adhesion, focal adhesion formation
1.67	Neurogenesis, differentiation, developmental protein
1.62	Homeostatic process, metal/chemical homeostasis
1.59	Regulation of phosphorus metabolic process, regulation of kinase activity
1.56	Heart trabecula formation, heart morphogenesis
1.54	Regulation of blood vessel size/tube size, vascular process in circulatory system, smooth muscle contraction, blood circulation, vasoconstriction, regulation of blood pressure
1.51	Smooth muscle contraction, contractile fiber

Table B.8. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Colon (by Procedure 3.11)
3.41	Cell death, apoptosis
2.94	Regulation of tube/blood vessel size, vascular process in circulatory system, circulatory system process, regulation of blood pressure, smooth muscle contraction, vasodilation
2.86	Peroxisome, microbody
2.77	Cytoskeletal protein binding, actin bin ding
2.62	Protein dimerization activity, identical protein binding
2.47	Actin filament-based process, cytoskeleton organization
2.14	Extracellular matrix
2.04	Homeostatic process, chemical/metal homeostasis
2.03	Negative regulation of biosynthetic process, transcription repressor activity
2	Cell fraction, insoluble fraction
2	Cell projection morphogenesis
1.95	Intracellular transport, protein localization
1.91	Response to hormone/steroid/endogenous stimulation

Table B.9. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Lymphoma (by Procedure 3.10)
6.41	Extracellular region, signal peptide, glycoprotein, disulfide bond
5.44	Proteinaceous extracellular matrix
4.76	Response to wounding, inflammatory response, defense response
3.57	Actin-binding, cytoskeletal protein binding
3.55	Extracellular matrix
3.53	Protein dimerization activity, identical protein binding
3.14	Tube/epithelium development, tissue morphogenesis
3.09	Regulation of cell shape/cell morphogenesis
3.08	Cell adhesion
3	Skeletal system/bone development, ossification
2.85	Ubl conjugation, isopeptide bond
2.82	SMAD binding, transmembrane receptor protein serine/threonine kinase signaling pathway, transforming growth factor beta receptor signaling pathway
2.77	Plasma membrane part

Table B.10. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Lymphoma (by Procedure 3.11)
5.14	Inflammatory response, response to wounding, defense response
4.66	Actin-binding, cytoskeletal protein binding
4.57	Extracellular region, signal peptide, glycoprotein, disulfide bond
4.3	Extracellular matrix
3.28	Regulation of cell shape/cell morphogenesis
3.07	Chemokine receptor binding, cytokine, inflammatory response
3	Plasma membrane part
2.93	Skeletal system/bone development, ossification
2.92	SMAD binding, transmembrane receptor protein serine/threonine kinase signaling pathway, transforming growth factor beta receptor signaling pathway
2.88	Extracellular matrix
2.63	Cell adhesion
2.54	Prenylation, methylation
2.54	Protein dimerization activity, identical protein binding

Table B.11. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Kidney (by Procedure 3.10)
3.01	Amine/nitrogen compound/cellular amino acid/carboxylic acid/organic acid biosynthetic process
2.78	(biogenic) Amine biosynthetic process, cellular amino acid derivative metabolic process
2.26	Cell fraction, insoluble fraction, membrane fraction
1.97	Gpi-anchor, phosphatidylinositol linkage, anchored to plasma membrane
1.93	Peroxisome, microbody
1.85	Cofactor binding, NAD or NADH binding, coenzyme binding
1.79	Regulation of cell migration/locomotion
1.63	Metal binding, copper ion binding
1.53	(negative) Regulation of T cell/antigen receptor signaling pathway, regulation of T cell activation/cell communication/immune system process/lymphocyte activation/leukocyte activation
1.5	Methionine metabolic process, sulfur amino acid memtabolic process

Table B.12. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Kidney (by Procedure 3.11)
3.31	NAD or NADH binding, cofactor binding, coenzyme binding
2.35	Amine/nitrogen compound/cellular amino acid/carboxylic acid/ organic acid/glutamine family amino acid biosynthetic process
1.95	(biogenic) Cellular amino acid derivative metabolic/biosynthetic process
1.86	Cell/insoluble/membrane/vesicular fraction, microsome
1.74	Mitochondrion, transit peptide
1.74	Copper ion binding
1.68	NAD(P)-binding, glucose/ribitol dehydrogenase, short-chain dehydrogenase
1.68	Gpi-anchor, membrane protein, phosphatidylinositol linkage
1.68	(negative) regulation of cell migration/locomotion
1.6	Peroxisome, microbody

Table B.13. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Bladder (by Procedure 3.10)
8.68	Membrane-enclosed lumen, organelle/nuclear lumen
6.72	Ribonucleoprotein complex, ribosome
6.47	Non-membrane-bounded organelle, cytoskeleton
5.3	Cytoskeletal protein binding, actin binding
4.99	mRNA metabolic process, RNA splicing
4.69	Intracellular transport, protein localization
4	Rna-binding, rna recognition motif, RRM
3.95	Ubl conjugation pathway, protein catabolic process, liqase
3.91	Translational initiation
3.7	Mitochondrion, transit peptide
3.59	Mitochondrion, organelle envelope, respiratory chain, Alzheimers disease, Parkisons disease, oxidative phosphorylation, Huntingtons disease, NADH dehydrogenase activity
3.34	Transcription factor/repressor/activator activity
3.14	Macromolecular complex assembly, protein complex biogenesis

Table B.14. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Bladder (by Procedure 3.11)
17.95	Ribonucleoprotein complex, ribosome
17.6	Rna-binding
12.1	Membrane-enclosed lumen, organelle/nuclear lumen
9.32	Non-membrane-bounded organelle, cytoskeleton
9.31	Intracellular transport, protein localization
9.22	Ubl conjugation pathway, protein catabolic process
7.69	mRNA metabolic process, RNA splicing
7.07	Mitochondrion, transit peptide
6.66	Rna-binding, rna recognition motif, RRM
6.53	Mitochondrion, organelle envelope, respiratory chain, Alzheimers disease, Parkisons disease, oxidative phosphorylation, Huntingtons disease, NADH dehydrogenase activity
5.33	Cytoskeletal protein binding, actin binding
5.08	Ubl conjugation pathway, liqase
4.87	Translational initiation

Table B.15. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Pancreas (by Procedure 3.10)
3.11	Extracellular matrix
2.72	Isopeptide bond, ubl conjugation
2.6	Vacuole, lysosome
2.39	Antigen processing and presentation of peptide/exogenous antigen, heterodimer, immunoglobulin c1-set, major histocompatibility complex, viral myocarditis, MHC class protein complex, graft-versus-host disease, type I diabetes mellitus, autoimmune thyroid disease, immune response
2.09	Domain:SEA
2.06	Response to nutrient, response to extracellular stimulus
2.02	Proteinase inhibito, Kazal, follistatin-like, N-terminal
1.9	Response to organic substance, response to steroid hormone/hormone/estrogen/endogenous/abiotic stimulus
1.84	mRNA processing/splicing
1.83	RRM, rna binding
1.81	Melanosome, pigment granule, cytoplasmic vesicle
1.79	Cytoskeleton organization, actin filament-based process
1.78	Extracellular matrix, collagen fibril organization, triple helix, hydroxylysine
1.73	Leukocyte transendothelial migration, pathogenic Escherichia coli infection, tight junction

Table B.16. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Pancreas (by Procedure 3.11)
8.59	Response to organic substance, response to hormone/endogenous stimulus
6.95	Extracellular matrix
6.42	Cell adhesion
5.13	Vacuole, lysosome
4.97	Melanosome, pigment granule, vesicle, cytoplasmic vesicle
4.97	Actin cytoskeleton, actin binding
4.63	Extracellular matrix, ECM-receptor interaction
4.36	Ubl conjugation, isopeptide bond
4.2	Response to extracellular stimulus, response to nutrient
3.84	mRNA splicing/processing
3.7	Actin cytoskeleton organization, actin filament-based process
3.62	Cell motion/migration
3.58	Transcription from RNA polymerase II promotor
3.57	Rna-biding, RRM, RNA recognition motif

Table B.17. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Leukemia (by Procedure 3.10)
6.08	Cell cycle, cell division, mitosis, M phase
4.96	Phosphate metabolic process, kinase, nucleotide-binding
4.48	Regulation/induction of apoptosis, regulation/induction of cell death
4.29	Nuclear lumen, organelle lumen, nucleoplasm
4.2	Mitotic cell cycle, interphase
3.68	Cytoskeleton, non-membrane-bounded organelle
3.35	Vacuole, lysosome
3.21	Regulation of cell cycle
3.06	Hemopoietic or lymphoid organ development, immune system development, leukocyte differentiation, T cell differentiation/activation, lymphocyte activation
3.03	Regulation of phosphate metabolic process/kinase activity

Table B.18. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Leukemia (by Procedure 3.11)
5.57	Cell cycle, cell division, mitosis, M phase
4.47	Cytoskeleton, non-membrane-bounded organelle
4.53	Kinase, Phosphate metabolic process, nucleotide-binding
4.32	Cytoskeleton organization, actin filament-based process
4.3	Nuclear lumen, organelle lumen, nucleoplasm
4.21	Cytoskeleton protein binding, actin binding
3.62	Immune system development, Hemopoietic or lymphoid organ development, leukocyte/lymphocyte differentiation/activation, T cell differentiation/activation, B cell activation
3.5	Vacuole, lysosome
3.23	Regulation/induction of apoptosis, regulation of programmed cell death
3.13	Insoluble fraction, membrane fraction

Table B.19. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Ovary (by Procedure 3.10)
2.75	Response to extracellular stimulus/nutrient/vitamin A/retinoic acid
2.43	Regulation of transcription
2.28	Actin binding, cytoskeletal protein binding
2.26	Vasculature/blood vessel development, angiogenesis
2.21	Regulation of transcription
2.1	(negative) regulation of binding
2.02	Tube/tissue/embryonic/epithelial tube morphogenesis, epithelium/tube/mesoderm development
1.98	Tube morphogenesis/development, branching morphogenesis of a tube
1.93	Response to extracellular/steroid hormone/progesterone/endogenous/ mechanical/temperature/abiotic/corticosteroid/glucocorticoid stimulus, response to organic substance/lipopolysaccharide/durg/ toxin/camp/bacterium/radiation, aging

Table B.20. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Ovary (by Procedure 3.11)
3.57	Tube morphogenesis/development, branching morphogenesis of a tube
3	Actin binding, cytoskeletal protein binding
2.94	Protein dimerization. Identical protein binding
2.79	Vasculature/blood vessel development, angiogenesis
2.68	Embryonic development ending in birth or egg hatching, in utero/chordate embryonic development
2.63	Epithelium development, tissue/embryonic morphogenesis
2.47	Regulation of locomotion/cell migration
2.46	Protein phosphatase 1, regulatory subunit 12A/B/ C, eukaryote
2.44	Response to mechanical/abiotic stimulus, respose to organic cyclic substance

Table B.21. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Uterus (by Procedure 3.10)
0.99	Immunoglobulin domain
0.95	Protease, proteolysis, peptidase activity, zymogen, hydrolase
0.92	Transferase, serine protein kinase, threonine protein kinase, atp binding

Table B.22. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Uterus (by Procedure 3.11)
1.57	Serine/threonine protein kinase, Transferase, atp binding
1.33	Cell fraction, membrane fraction, insoluble fraction
1.2	Apical part of cell, actin cytoskeleton, apical plasma membrane, cell cortex, cognition, sensory perception
1.19	Actin cytoskeleton, actin binding
1.13	Disulfide bond, signal peptide, glycoprotein
1.11	Cell fate commitment, endocrine system development
1.11	Immunoglobulin v-set

Table B.23. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Prostate (by Procedure 3.10)
1.72	Regulation of cell communication/signal transduction, disease mutation
1.69	Regulation of cell migration, disulfide bond, glycoprotein, signal peptide
1.44	Response to hormone/endogenous stimulus, regulation of protein kinase activity, regulation of phosphorus metabolic process

Table B.24. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Prostate (by Procedure 3.11)
2.43	Disulfide bond, extracellular region, signal peptide, glycoprotein
2.16	Response to oxygen levels, tissue remodeling, response to hypoxia
1.78	Regulation of cell proliferation/migration/biosynthetic process, regulation of DNA replication, growth factor activity, focal adhesion, response to hormone/endogenous stimulus, wound healing, mammary gland development, pathways in cancer, lung development, prostate cancer, embryonic development ending in birth or egg hatching, enzyme linked receptor protein signaling pathway, regulation of cell death

Table B.25. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Lung (by Procedure 3.10)
1.05	Metal-binding, zinc, ion binding, cation binding
0.87	Cytoplasmic membrane-bounded vesicle
0.84	Endoplasmic reticulum, endomembrane system, organelle membrane

Table B.26. Functional annotation clustering results (DAVID) for Section 3.6.2.

	Lung (by Procedure 3.11)
1.53	DNA binding, transcription activator
1.39	Nucleolus, nuclear lumen, intracellular organelle lumen,
1.39	Mitosis, nuclear division, M phase of mitotic cell cycle, organelle fission

Bibliography

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215-1222.

Benjamini, Y. , Heller, R., Yekutieli, D. (2009). Selective inference in complex research. *The Royal Society of Philosophical Transactions: Series A*, **367**, 1-17.

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 Series B*, **57**, 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

Benjamini, Y. and Yekutieli, D. (2000). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.

Campain, A. and Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, **11**, 408. PubMed PMID: 20678237; PubMed Central PMCID: PMC2922198.

Chin, M. H. et al. (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures, *Cell Stem Cell*, **5**, 111-123.

Chin, M. H. et al. (2010). Molecular analyses of human induced pluripotent stem cells and embryonic stem cells, *Cell Stem Cell*, **7**, 263-269.

Choi, J. K. et al. (2003). Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, **19**, 84-90.

Choi, H. et al. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**, 364.

Cochran, W. G. (1954). The combination of estimates from different experiments, *Biometrics*, **10**, 101-129.

Dai, M. et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, **33**, e175.

Dunn, O. J. (1961). Multiple comparisons among means, *Journal of American Statistical Association*, **56**, 52-64.

Finner, H. (1999). Stepwise multiple testing procedures and control of directional errors, *The Annals of Statistics*, **27**, 274-289.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Genovese, C. R., Roeder, K. and Wasserman, L. (2006). False discovery control with *p*-value weighting, *Biometrika*, **93**, 509-524.

Genovese, C. R. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure, *Journal of Royal Statistical Society: Series B*, **64**, 499-517.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control, *The Annals of Statistics*, **32**, 1035-1061.

Guenther, M. G. et al. (2010). Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell*, **7**, 249-257.

Guo, W., Sarkar, S. K. and Peddada, S. D. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, **66**, 485-492.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic Press: Boston.

Heller, R., Manduchi, E., Grant, G. R. and Ewens, W. J. (2009). A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics*, **25**, 1019-1025.

Heller, R., Yekutieli, D. (2012). Replicability analysis for genome-wide association studies. **arXiv:1209.2829**.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**, 1539–1558.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

Hong, F., et al. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825-2827.

Huang, D. W., Sherman B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols*, **4(1)**, 44–57.

Huedo-Medina, T. B. et al. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, **11**, 193-206.

Jiang, H. and Doerge, R. W. (2006). A two-step multiple comparison procedure for a large number of tests and multiple treatments. *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 28.

Lai, Y. et al. (2007). A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups, *Bioinformatics*, **23**, 1243-1250.

Lancaster, H. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, **3**, 20–33.

Lee, J. K. et al. (2003). Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biology*, **4**, R82.

Leek, J. T. et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, **11**, 733-9.

Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, **37**, 1137-1153.

Li, Q., Brown, J. B., Huang, H. and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, to appear.

Li, Y. and Ghosh, D. (2012). Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *Bioinformatics*, **28**, 807–814.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, **5**, 994–1019.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Közl*, **3**, 171–196.

Lu, S. et al. (2010). Biomarker detection in the integration of multiple multi-class genomic studies, *Bioinformatics*, **26**, 333-340.

Makambi, K. H. (2003). Weighted inverse chi-square method for correlated significance tests. *Journal of Applied Statistics*, **30**, 225–234.

Miron, M. et al. (2006). A methodology for global validation of microarray experiments, *BMC Bioinformatics*, **7**, 333.

Mosteller, F. and Bush, R. R. (1954). Selected quantitative techniques. *Handbook of Social Psychology*. Cambridge, MA: Addison-Wesley.

Natarajan, L., Pu, M., Messer, K. (2012). Statistical tests for the intersection of independent lists of genes: sensitivity, FDR, and type I error control, *The Annals of Applied Statistics*, **6**, 521-541.

Newman, A. M. and Cooper, J. B. (2010). Lab-specific gene expression signatures in pluripotent stem cells, *Cell Stem Cell*, **7**, 258-262.

Normand, S.-L. T. (1999). Tutorial in Biostatistics. Meta-Analysis: Formulating, Evaluating, Combining, and Reporting. *Statistics in Medicine*, **18**, 321 – 359.

Olkin, I. and Saner, H. (2001). Approximations for trimmed Fisher procedures in research synthesis. *Statistical Methods in Medical Research*, **10**, 267–276.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics*, **37**, 3867–3892.

Parmigiani, G. et al. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer, *Clinical Cancer Research*, **10**, 2922-2927.

Pearson, K. (1934). On a new method of determining "goodness of fit". *Biometrika*, **26**, 425–442.

Phillips, D. and Ghosh, D. (2013). Testing the disjunction hypothesis using Voronoi diagrams, with applications to genetics. *Annals of Applied Statistics*, to appear.

Pyne, S., Futcher, B., Skiena, S. (2006). Meta-analysis based on control of false discovery rate: combining yeast ChiP-chip datasets, *Bioinformatics*, **22**, 2516-2522.

Ramaswamy, S. et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of National Academy of Sciences USA*, **98**, 15149-15154.

Rhodes, D. et al. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, **62**, 4427 – 4433.

Sarkar, S. K. and Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, **92**, 1601-1608.

Scharpf, R. B. et al. (2009). A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, **104**, 1295–1310.

Scott, L. J. et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 341-5.

Shabalin, A. A. et al. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, **24**, 1154 – 1160.

Shedden, K. A. et al. (2003). Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *American Journal of Pathology*, **163**, 1985-1995.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751-754.

Song, C. and Tseng, G. C. (2013). Hypothesis setting and order statistic for robust genomic meta-analysis. *Annals of Applied Statistics*, to appear.

Sørlie, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, USA*, **98**, 10869-74.

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

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**, 2013-2035.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, USA*, **100**, 9440-9445.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). *The American soldier: adjustment during army life*. Princeton University Press.

Sun, W. and Wei, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association*, **106**, 73-88.

Tomlins, S.A., et al. (2005). Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer, *Science*, **310**, 644–648.

Vardiman, J. W, Harris, N. L and Brunning, R. D. (2002). The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*, **100**, 2292–2302.

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

Willer, C. J. et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, **40**, 161 – 169.

Won, S., Morris, N., Lu, Q. and Elston, R. (2009). Choosing an optimal method to combine P-values. *Statistics in Medicine*, **28**, 1537–1553.

Xie, M., Singh, K. and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, **106**, 320–333.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, **8**, 1836–1841.

YIHAN LI

412 Kemmerer Road, State College, PA 16801 | (814) 206-6158 | yihanli@psu.edu

EDUCATION

Pennsylvania State University, University Park, PA
Ph.D. in Statistics **2014**
Dissertation: “Methods in Multiple Testing and Meta-Analysis with Applications to the Analysis of Genomic Data”

Peking University, Beijing, China
B.S. in Statistics **2009**

AWARDS

Eberly College of Science Scholarship, Pennsylvania State University **2010 – 2011**
GlaxoSmithKline Scholar Award **2009**

PUBLICATIONS

Li, Y. and Ghosh, D. (2012) Assumption Weighting for Incorporating Heterogeneity into Meta-Analysis of Genomic Data. **Bioinformatics***, 28 (6): 807-814*
Paper presented at Joint Statistical Meeting, San Diego, CA **2012**

Li, Y. and Ghosh, D. (2014) A Two-Step Hierarchical Hypothesis Set Testing Framework, with Applications to Gene Expression Data on Ordered Categories. Submitted.
Paper presented at Joint Statistical Meeting, Montréal, Canada **2013**

Li, Y. and Ghosh, D. (2014) Meta-Analysis Based on Weighted Ordered P-values for Genomic Data with Heterogeneity. Submitted.

TEACHING EXPERIENCE

Pennsylvania State University, University Park, PA
Teaching Assistant & Consultant – STAT 580/581: Statistical Consulting Practicum I/II **2012**

Lecturer – “Bioinformatics and Genomics Retreat Workshop” **2012**

Teaching Assistant – STAT 505: Applied Multivariate Statistical Analysis **2011**

Teaching/Lab Assistant – STAT 200: Elementary Statistics **2009, 2010, 2012**

PROFESSIONAL EXPERIENCE

AbbVie, North Chicago, IL
Intern, Clinical Statistics **2013**
Researched on multiple testing strategies in clinical trials; conducted simulations studies for power and probability of success calculation.

Janssen Research & Development LLC., Tittusville, NJ
Biostatistics Intern **2012**
Developed methodology for enrichment trial design; conducted simulations studies to evaluate new methodology; SAS programming for analyzing and summarizing trial data.

MEMBERSHIPS

American Statistical Association (ASA)
Eastern North American Region/International Biometric Society (ENAR)