Fall 12-2022

Statistical Methods for Meta-Analysis in Large-Scale Genomic Experiments

Wimarsha Thathsarani Jayanetti
Old Dominion University, wjaya001@odu.edu

STATISTICAL METHODS FOR META-ANALYSIS IN LARGE-SCALE GENOMIC EXPERIMENTS

by

Wimarsha Thathsarani Jayanetti
B.S. April 2014, University of Colombo, Sri Lanka
M.S. August 2018, University of Texas at El Paso

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTATIONAL AND APPLIED MATHEMATICS

OLD DOMINION UNIVERSITY
December 2022

Approved by:

Sinjini Sikdar (Director)

N. Rao Chaganty (Member)

Yet Nguyen (Member)

Hadiza Galadima (Member)

ABSTRACT

STATISTICAL METHODS FOR META-ANALYSIS IN LARGE-SCALE GENOMIC EXPERIMENTS

Wimarsha Thathsarani Jayanetti
Old Dominion University, 2022
Director: Dr. Sinjini Sikdar

Recent developments in high throughput genomic assays have opened up the possibility of testing hundreds and thousands of genes simultaneously. With the availability of vast amounts of public databases, researchers tend to combine genomic analysis results from multiple studies in the form of a meta-analysis. Meta-analysis methods can be broadly classified into two main categories. The first approach is to combine the statistical significance (p-values) of the genes from each individual study, and the second approach is to combine the statistical estimates (effect sizes) from the individual studies. In this dissertation, we will discuss how adherence to the standard null distributional assumptions in both categories of meta-analysis methods can lead to incorrect significance testing results in detecting the true set of significant genes. To overcome this, we will also propose two robust meta-analysis methods that perform empirical modifications of the summary results. In the first part, we will propose a new meta-analysis method combining p-values for a gene from multiple studies with an aim to detect significance in a consistent pattern in a majority of studies. Our proposed method performs an empirical modification of the individual p-values using an empirical Bayes approach before meta-analyzing them. In the second part, we will propose a meta-analysis method combining effect size estimates for a gene from multiple studies with an aim to detect significance in at least one study. Here we perform empirical modification of the z-scores, obtained from effect size estimates of the genes and their standard errors, using the empirical Bayes approach. Through various simulation studies and real genomic data applications, we will show that our proposed meta-analysis methods outperform the existing meta-analysis methods in terms of accurately identifying the truly significant set of genes by reducing false discoveries, especially in the presence of unobserved confounding variables.

Dedicated to my family!

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Sinjini Sikdar, for her guidance, close supervision, and unwavering support throughout the process of my doctoral dissertation. Her insightful observations, prompt feedback, and energizing suggestions really aided me in finishing the projects on time.

I also want to express my gratitude to the other committee members, Dr. N. Rao Chaganty, Dr. Yet Nguyen and Dr. Hadiza Galadima, for their recommendations, feedback, and extra direction which were crucial to the success of these projects.

I extend my gratitude to all faculty, and staff members of the Department of Mathematics and Statistics, who helped in numerous ways to carry out these projects. I am also thankful to the NAVSEA project team and the Old Dominion University Research Foundation for the remarkable experience and financial support. Last but not least, I am grateful to my family for their support and encouragement not only during this study but throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

In this chapter we start with section 1.1 where we describe the biological background of genomic data. In section 1.2, we illustrate several existing meta-analysis methods for genomic data. In section 1.3, we briefly describe about the concept of multiple hypothesis testing and existing correction methods. Finally, an overview of the dissertation is outlined in section 1.4.

1.1 BACKGROUND

1.1.1 DNA

The hereditary substance in humans and almost all other animals is DNA or deoxyribonucleic acid. The DNA of nearly every cell in a person's body is identical. The instructions sent to the cells to grow, develop and function can be encoded in a string - a molecule of DNA, a polymer made up of recurring units called nucleotides. DNAs are nucleic acids. Adenine (A), Guanine (G), Cytosine (C), and Thymine (T) are the four nucleotides found in DNA molecules. The nucleotide A always binds to T in DNA, while C always binds to G. These nucleotides, in a specific sequence, store the information for life. DNA is arranged in a double-helix structure, in which two complimentary polymers interlace and twist to form the classic helical shape (see Figure 1). The information for life is stored in the nucleotides in a precise sequence.

1.1.2 GENOME

A genome is an organism's complete DNA sequence that contains all of its genetic information. The genome contains all the instructions for creating and maintaining an organism. Genomes come in a variety of sizes and configurations, and they vary between species.

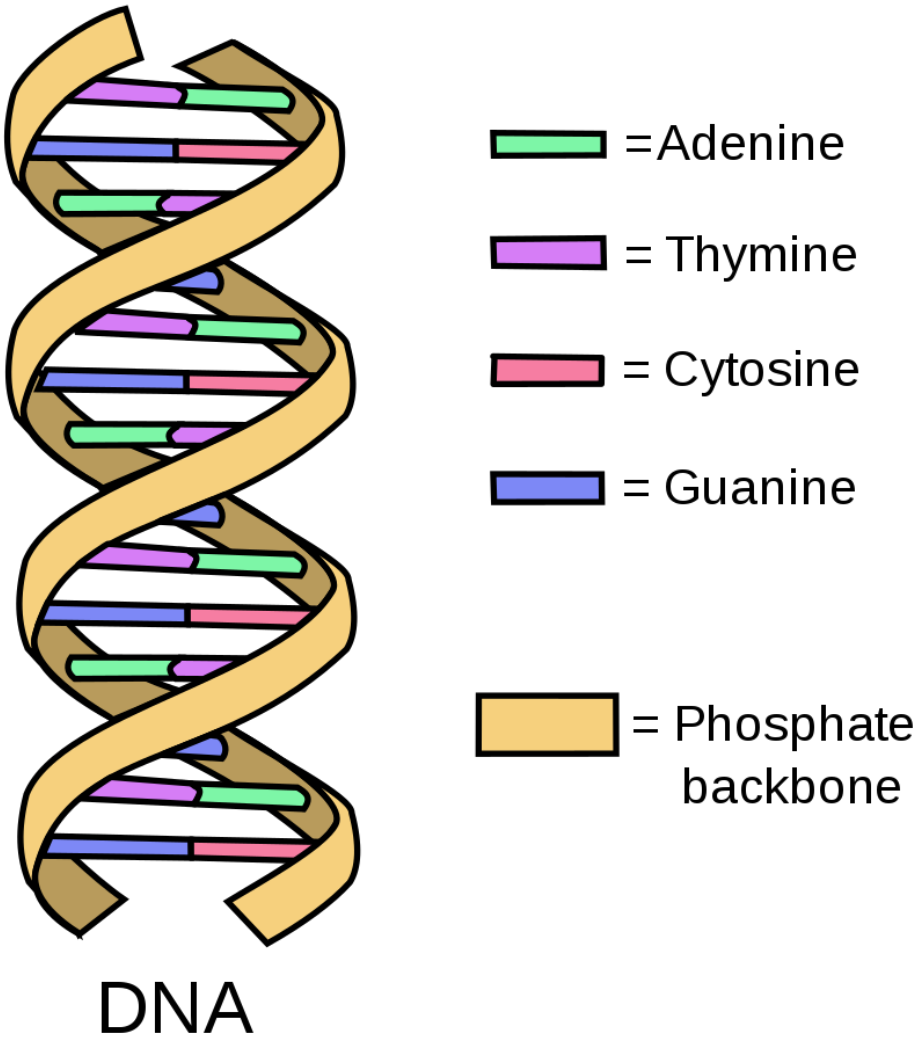

Fig. 1. Double-helix structure of DNA. Adapted from "Building Blocks of the Genetic Code" by the American Society of Human Genetics, 2019 (https://www.ashg.org/discover-genetics/building-blocks/).

1.1.3 GENE

The precise information that encodes for physical products of genetic information is contained in certain areas of the genome. A "gene" is a section of the genome that contains this information. Genes are lengthy strands of DNA that are the fundamental units of inheritance in all living organisms. Genes also provide information about building proteins.

1.1.4 PROTEIN

A protein is made up of a chain of amino acids linked by peptide bonds. Proteins are necessary for life to exist. All living cells rely on them for growth, repair, function, and structure. Because a gene can encode information for proteins and other functional molecules, it is a crucial notion in genomic biology.

1.1.5 CENTRAL DOGMA OF BIOLOGY

The process of converting DNA instructions into a functional product is known as the "Central Dogma". Francis Crick, the discoverer of the structure of DNA, initially proposed it in 1958. The central dogma explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein (see Figure 2). In this process, the DNA is replicated in order to transfer information to new cells. If activated, the genes are then transcribed into messenger ribonucleic acids (mRNAs) in the nucleus and then translated into proteins in the cytoplasm. Information is transferred between information-carrying polymers such as DNA, RNA, and proteins in this process.

1.1.6 GENE EXPRESSION

Gene expression is the process by which information encoded in DNA is transformed into instructions for creating proteins or other molecules which is summarized in central dogma of biology. This process enables a cell to respond to changes in its environment. It serves as an on/off switch for controlling when proteins are produced as well as a volume control for increasing or decreasing the number of proteins produced. The condition of the system at any particular time is determined by the gene expression measure.

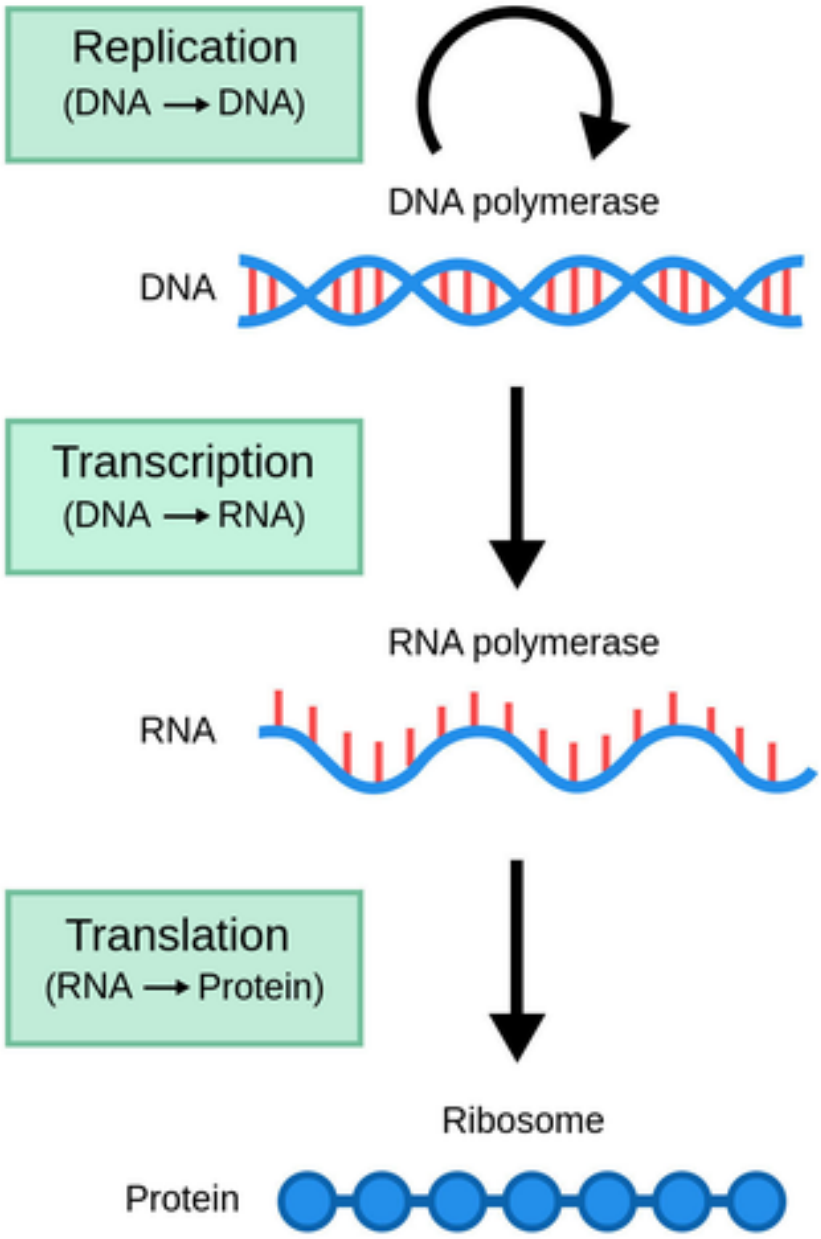

Fig. 2. Central dogma of biology. Adapted from "Central dogma of molecular biology" by the Labster Theory, 2021 (https://theory.labster.com/central_dogma_molecular_biology_pre/).

1.1.7 HIGH-THROUGHPUT EXPERIMENTAL METHODS IN GENOMICS

High-throughput experimental methods have become critical in genomics research. They can sequence multiple DNA molecules in parallel, allowing them to sequence hundreds of millions of DNA molecules at once. This advantage enables researchers to generate large data sets, allowing them to gain a more comprehensive understanding of various diseases and developmental stages.

Several methods for measuring expression levels have long been available, but only for a few genes at a time. This approach, which was the only one available for a long time, is slow, expensive, and inefficient for large-scale gene screening. Prior to the widespread adoption of microarray technologies, large-scale screenings of gene expression signatures were impossible. Microarrays allow researchers to examine thousands of genes at once. Until the development of sequencing technology (e.g. Next-generation sequencing), microarrays were the standard instrument for the quantification stage. Most biological phenomena involving transcription, gene regulation, or DNA mutation can be detected throughout the entire genome using high-throughput experimental techniques, such as microarrays, which have become the gold standard in genome biology research. The steps in high throughput techniques are as follows and also summarized in Figure 3:

- Extraction: This is the process of extracting the genetic material of interest, such as RNA or DNA.
- Enrichment: In this step, enrichment for the event of interest is done. Enrichment processes are not required in some circumstances, such as whole-genome DNA sequencing where genomic DNA fragments are collected and sequenced.
- Quantification: The enriched material is quantified in this step.

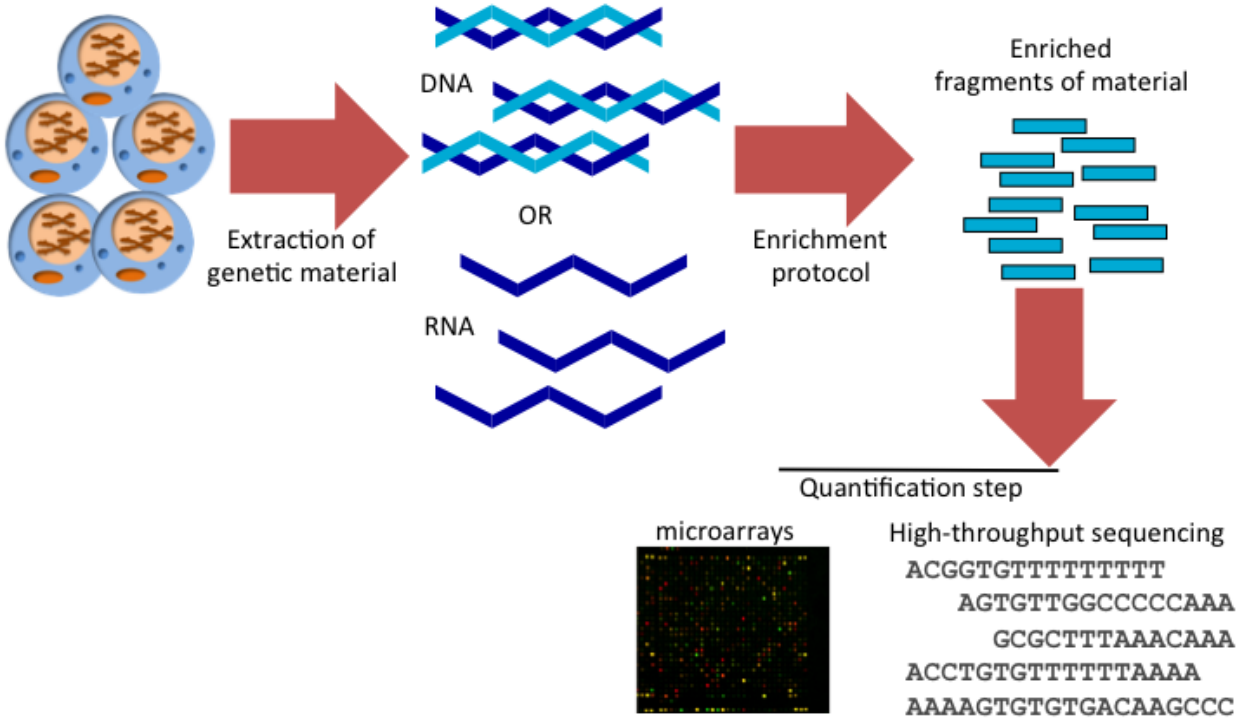

Fig. 3. Three common steps in high throughput techniques. Adapted from Podgórski, K. (2021). Computational Genomics with R Altuna Akalin Chapman & Hall/CRC, 2021, *International Statistical Review*, 89(2), 420-421.

1.1.8 MICROARRAY EXPERIMENT

Microarrays have played an important role in the recent biotechnological revolution. They have enabled researchers to simultaneously monitor the expression of thousands of genes and hence obtain snapshots of the state of a complete genome. This advancement has given rise to a new field of study in statistics and bioinformatics.

DNA microarray, also commonly known as DNA chip, is a technology that allows researchers to identify and quantify mRNA transcripts in cells. Researchers use DNA microarrays to quantify the expression levels of large numbers of genes simultaneously. The number of molecules of mRNA produced by the transcription of a specific gene can be used to approximate the gene's degree of expression. A microarray is made up of strands of polynucleotide (DNA) called probes that are connected to or produced at set spots on a solid surface. The mRNA is taken from the subject cells to begin a microarray experiment. The molecules are then tagged with fluorescent dye. These tagged transcripts are referred to as targets. After the samples have been processed, they are deposited on the array and permitted to hybridize in a hybridization chamber for some time. By hybridization, the tagged targets bind to probes on the array with which they share sufficient sequence complementarity. The array is then washed, removing any targets that did not hybridize. Spotted or cDNA microarrays and oligonucleotide chips are the two most common types of expression microarrays. Figure 4 depicts an example of a cDNA microarray experiment.

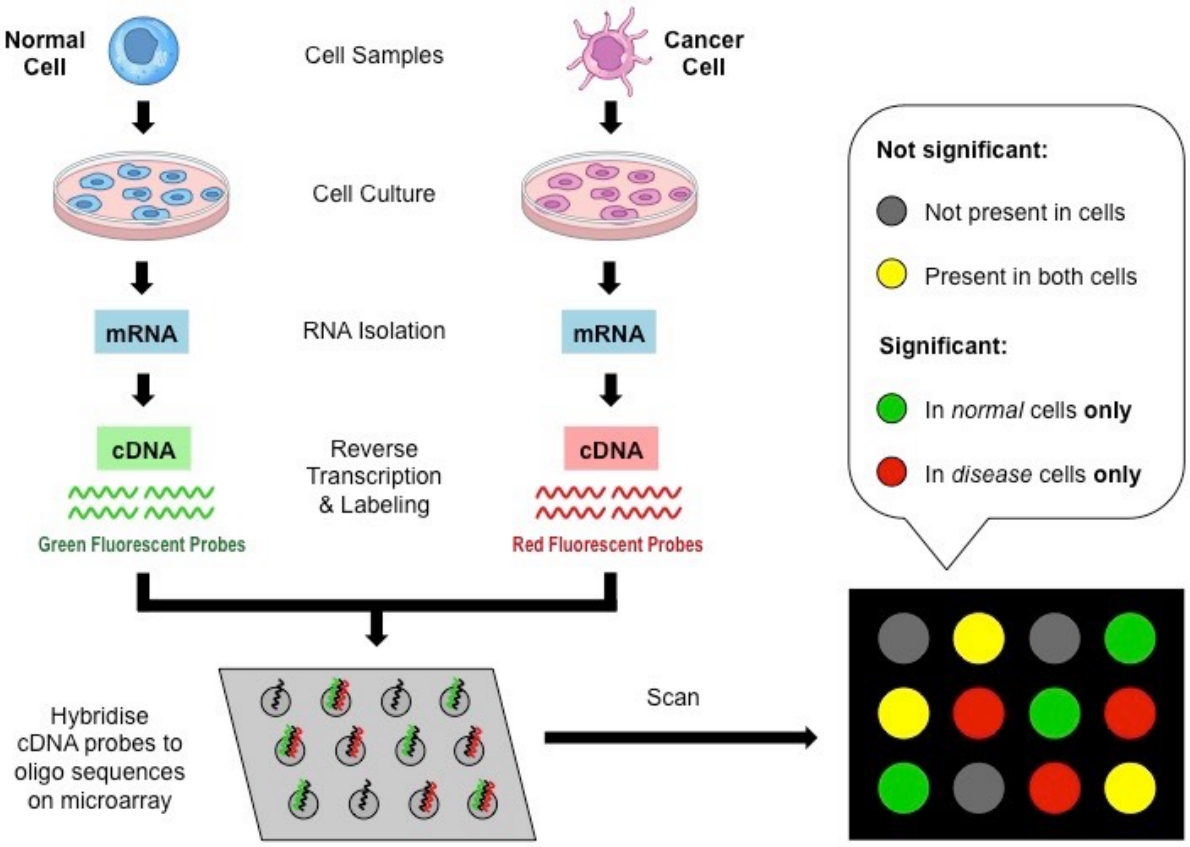

Fig. 4. A cDNA microarray experiment. Adapted from "cDNA and Microarrays" by the BioNinja (https://ib.bioninja.com.au/standard-level/topic-3-genetics/35-genetic-modification-and/cdna-and-microarrays.html).

At the end, this image will be analyzed to obtain information about gene expression levels: each spot on the array is identified, then its intensity is measured and compared to the background. This is known as image analysis. As with any statistical analysis, and particularly in image analysis, the quality of the data must be examined first. Once the quality of the data has been determined it is still necessary to carryout some preprocessing before the analysis. Microarray data preprocessing mainly consists of two steps: background correction and normalization. The next step is the differential expression analysis. The goal of differential expression analysis is to discover quantitative changes in expression levels between experimental groups. The analysis of gene expression data from a single study has

limitations, such as small sample sizes. The statistical significance established is substantially higher when multiple studies are taken into account at once than when only one study is taken into account. Therefore, it makes more sense to generalize the findings from a meta-analysis than from a single study since it incorporates many populations into the analysis and, as a consequence, takes into account the variations between various groups, which are most likely to respond in different ways.

1.2 META-ANALYSIS METHODS FOR GENOMIC DATA

Today, meta-analysis has become a key component in genomic research. A meta-analysis is a statistical method that integrate the summary results of multiple studies. Meta-analysis can be performed when there are multiple studies addressing the same overall hypothesis. It produces more robust results, which can assist researchers in better understanding the magnitude of an effect that generalized to a larger population. Meta-analysis has several benefits, including increased statistical power to detect an effect and improved precision and accuracy of effect estimates.

Meta-analysis methods can be broadly classified into two main categories. One where statistical significance measures such as p-values from individual studies are combined, and the other where statistical estimates such as effect sizes from individual studies are combined. Fisher's combined probability test [1] and Stouffer's Z-test [2] are the two widely used traditional p-value combination methods. Over the time, several weighted and generalized versions of these traditional p-value combination methods have also been discussed and illustrated [3–5]. Several methods combining effect sizes exist in the literature [6, 7]. There are advantages and disadvantages for both type of meta-analysis methods. Methods based on combining significance (p-values) are relatively flexible since they require minimal information and assumptions from the studies. On the other hand, combining effect sizes is statistically more powerful than combining p-values or z-scores but require more strict assumptions and they can not be easily extended to studies with multiple groups.

In our first project, we will focus on the p-value combining methods. Following is a brief description of some commonly used methods for combining significance (p-values). We consider K independent studies that have been carried out to detect a certain effect. Let θ_j denote the parameter that characterizes the effect of study j, $j = 1, ..., K$. The j^{th} study

is outlined to test the hypothesis $H_{0j} : \theta_j = 0$ against an alternative $H_{1j} : \theta_j \neq 0$ using the test statistic T_j. Assuming T_j follows a continuous distribution, the significance of a test is often defined as a p-value, that is, as $p_j = Pr(T_j > t_j|H_{0j})$. Note that when H_{0j} is true, p_j is assumed to be uniformly distributed.

Fisher's Method

The most popular method for combining p-values is the Fisher's combined probability test [1]. Fisher's method uses the product of p-values from the studies and convert it to chi-square scores using a $-2log$ transformation. Due to this reason it is also known as the inverse chi-square method. The Fisher's combined probability test statistic is defined as follows:

$$V_F = -\sum_{j=1}^{K} 2log(p_j)$$

When all the null hypotheses of the K tests are true, then V_F has a χ^2 distribution with $2K$ degrees of freedom, assuming that the K studies are independent. Fisher's method can be sensitive to smaller p-values.

Stouffer's Method

An alternative to Fisher's method is Stouffer's Z test [2], which is based on converting the p-values into Z-scores. Let $\phi(x) = exp(-x^2/2)/\sqrt{2\pi}$ denotes the standard normal probability density function and $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$ denotes the standard normal cumulative distribution function. Then the Stouffer's Z test statistic is defined as follows:

$$V_Z = \sum_{j=1}^{K} \Phi^{-1}(1 - p_j)$$

When all the null hypotheses of the K tests are true, V_Z has a normal distribution with mean zero and variance K, assuming that the K studies are independent.

Tippett's Minimum P-value Method

The Tippett's test statistic [8] is defined as:

$$V_T = Min_{j=1}^{K}\ p_j$$

whose null distribution, if the K studies are independent and continuous, is the $Beta(1, K)$. The Tippett's Minimum p-value test is closely related to the Bonferroni method [9] and is also known as the union-intersection method. This method is also sensitive to small p-values like Fisher's method but it is less powerful than Fisher's method especially when all studies are significant.

Most of these traditional meta-analysis methods aim at detecting signals in at least one study. Nowadays, researchers are more interested in detecting signals in a consistent pattern across multiple studies. This is especially true with meta-analysis of genomic studies. In that context some methods have been proposed recently. For example, Song and Tseng [10] proposed an order statistic of p-values (rth ordered p-value, rOP) across studies as the test statistic for detecting differentially expressed genes in a majority of studies. They also developed methods to estimate the parameter r for real-world applications. Later, Li and Ghosh [11] introduced a new class of meta-analysis methods based on summaries of weighted ordered p-values (WOP) with the goal of detecting significance in a majority of studies. They considered weighted versions of traditional methods such as Fisher's method and Stouffer's method, in which the weight for each p-value is determined based on its order among the studies.

In our second project, we will focus on meta-analysis techniques that combine effect size estimates from multiple studies addressing the same research question as an alternative to combining p-values. In meta-analysis, different weights are typically assigned to each study to reflect the relative contributions of individual studies to the total effect size. Many meta-analysis methods weight the studies based on precision. In the calculation of the effect size, more precise studies are given more weight. Either a fixed effects model or a random effects model can be used for this. Because fixed effects meta-analysis assumes that the genetic effects are constant across the studies, it differs from random effects meta-analysis. Compared to random effects models, fixed effects models offer much lower p-values and more precise confidence intervals for the estimates [12]. Below is a brief discussion of both the fixed effects and random effects models.

Fixed Effects Model

The fixed-effect model assumes that the true effect size for all studies is identical and that the only reason the effect size varies between studies is due to within-study estimation error.

Therefore, the weights in the fixed effect model are solely based on within-study variances. Inverse-variance weighting is frequently utilized for the fixed effects model. Suppose $\hat{\theta}_j$ and w_j denote the effect size estimate and its inverse of the estimated variance for the j^{th} study, $j = 1, 2, ..., K$. The weighted average of the effect size estimates can be computed as follows:

$$\hat{\theta}_F = \frac{\sum_{j=1}^{K} w_j \hat{\theta}_j}{\sum_{j=1}^{K} w_j}$$

The variance of the weighted average of the effect size estimate can be calculated as

$$var(\hat{\theta}_F) = \frac{1}{\sum_{j=1}^{K} w_j}$$

Random Effects Model

The goal of the random-effects model is to estimate the mean of a distribution of effects rather than one true effect. Those means usually assumed to follow a Normal distribution. The model estimates the variance of that Normal distribution and therefore the degree of between-study heterogeneity. Each study's weight incorporates the between-study variance of heterogeneity τ^2 [13] , which is expressed as

$$\tau^2 = (Q - (K-1))/\left(\sum_{j=1}^{K} w_j - \left(\frac{\sum_{j=1}^{K} w_j^2}{\sum_{j=1}^{K} w_j}\right)\right)$$

where k is the number of studies and $Q = \sum_{j=1}^{K} w_j(\hat{\theta}_j - \theta_F)^2$. Then the weight for the random effects model is computed as

$$w_j^R = 1/(\frac{1}{w_j} + \hat{\tau}^2)$$

In the second part of this dissertation, we will focus on the inverse variance method based on fixed effect model, which involves multiplying individual effect sizes by the inverse of their squared standard error. The widely used meta-analysis method called METAL [14] use this idea to test the hypothesis that the genes are differentially expressed in at least one study.

1.3 MULTIPLE HYPOTHESIS TESTING CORRECTION METHODS

Multiple hypothesis testing refers to testing several hypotheses simultaneously. This scenario is quite common in statistical research, especially in genomic studies where thousands of genes are tested simultaneously. The probability that some of the true null hypotheses

will be rejected can be large if our decisions about the individual hypotheses are based on the unadjusted marginal p-values. The concept of family wise error rate (FWER) was first introduced by Tukey in 1958. The FWER is defined as the probability of making at least one type I error in the family. Therefore, by assuring $FWER \leq \alpha$, the probability of making one or more type I errors in the family is controlled at level α. Tukey, Scheffé and Bonferroni methods are some of the classical solutions to control FWER.

Benjamini and Hochberg [15] proposed an alternative approach to control false discovery rate (FDR) that is more appropriate in scenarios, such as the detection of differentially expressed genes. FDR is the expected proportion of tests which are incorrectly called significant out of all the tests which are called significant. Considering M hypothesis tests, Table 1 lists all possible outcomes. Here, FDR is defined as $E(\frac{V}{R})$. The Benjamini-Hochberg method is a procedure which controls the FDR so that $FDR \leq \alpha$.

Table 1
Possible outcomes when simultaneously testing M null hypotheses.

	Not significant	Significant	Total
Null is true	U	V	M_0
Alternative is true	T	S	M_1
Total	M-R	R	M

1.4 OVERVIEW OF THE DISSERTATION

This dissertation is based on two different projects that are linked through the meta-analysis of genomic data. The first project focuses on developing a novel meta-analysis technique which aim to identify genes significant consistently in a majority of studies. The proposed method empirically adjusts the weighted ordered p-values from multiple studies before combining them. This proposed method is developed for situations when p-value

is the only available information for the genes in the studies considered for meta-analysis. A second project is developed for situations where the aim is to identify genes significant in at least one study. This method requires availability of effect size estimates and their standard errors from the studies which are then combined across the studies. Through various simulation studies, we showed that our proposed meta-analysis methods outperform the existing methods in terms of accurately identifying the truly significant set of genes. We applied our proposed methods on several real-life datasets (lung cancer, brain cancer and diabetes) which clearly demonstrate the utility of the proposed methods.

CHAPTER 2

EMPIRICALLY ADJUSTED WEIGHTED ORDERED P-VALUES METHOD

2.1 INTRODUCTION

One of the flexible approaches for meta-analysis methods is to directly combine the p-values from the studies. In that context, there are two popular traditional methods - Fisher's combined probability test [1] and the Stouffer's Z test [2]. These traditional methods assume that individual p-values which are coming from different studies are uniformly distributed under the null hypothesis of no differential expression in any study. However, in large-scale multiple testing problems, the theoretical and empirical null might differ, [16] hence, the distributional assumption of these traditional p-value combination methods can become problematic.

Recently, an empirically adjusted meta-analysis (EAMA) method [17] has been proposed to ensure that all the p-values from individual studies are uniformly distributed under the null hypothesis. In this study, they have applied an empirical adjustment to the raw p-values before combining them using Fisher's method and obtained better inferences compared to the Fisher's method. However, the traditional methods, including EAMA, aim to identify genes that have significant contributions to the outcome of interest in any of the studies. In recent years, researchers are often more interested in identifying those genes which are differentially expressed in majority of studies. To address this problem, in recent years several methods have been proposed [18–23]. For example, Song and Tseng [10] proposed the r^{th} ordered p-value (rOP) method that aims at testing the alternative hypothesis that there is signal in at least a given percentage of the studies. But, the rOP method considers information on only one ordered p-value for conclusion, which can lead to considerable loss of information for meta-analysis. Later, Li and Ghosh [11] proposed a weighted ordered p-values method (WOP) which combines all ordered p-values using Fisher's and Stouffer's statistics after weighting them based on their order, and tests whether the genes are differentially expressed

in majority of studies against the null hypothesis that they are not differentially expressed in any study. Based on the results obtained by Li and Ghosh [11], the WOP method appeared to be more robust than the rOP method as well as the traditional methods (Fisher's method and Stouffer's method). Under the WOP framework, although many classical p-value combination approaches can be expressed, Li and Ghosh focused on the Fisher's and Stouffer's methods. Hence it is still needed to ensure that the p-values are uniformly distributed under the null hypothesis.

We propose a meta-analysis method that implements WOP method but empirically adjusts the null distribution of the p-values to ensure that they are uniformly distributed under the null hypothesis. We considered the same problem of testing the alternative hypothesis that a gene is differentially expressed in at least half of the studies against the null hypothesis that it is not differentially expressed in any study. But instead of directly combining the weighted ordered p-values, we combined empirically adjusted p-values after weighting them based on their order. Our proposed method utilizes the Empirical Bayes method, proposed by Efron [16], for estimating the null distribution empirically. After appropriately combining empirically adjusted weighted ordered p-values across the studies using the WOP summary statistic, we computed multiple testing corrected p-values based on the numerical distribution with the aim of identifying a smaller proportion of interesting or significant genes that are differentially expressed in at least half of the studies.

Our proposed method has advantages over both the traditional p-value combination methods as well as the WOP method. Compared to the traditional p-value combination methods, the proposed method better focus on identifying genes that are differentially expressed in a majority of studies. On the other hand, our proposed method has much robust performance than the WOP method. The rest of this chapter is organized as follows. In section 2.2, we first discuss our hypothesis setting. We then propose our empirically adjusted weighted ordered p-values method. Next, we briefly describe the empirical estimation of null distribution using the Empirical Bayes method. In section 2.3, we present simulation results and the application of our method on three sets of micro-array data on lung cancer, brain cancer, and diabetes which clearly demonstrate the utility of the proposed method. We end this chapter with a discussion in section 2.4.

2.2 METHODS

2.2.1 HYPOTHESIS SETTING

Suppose there are K independent studies where each study consists of G genes. Let θ_{ij} denotes the underlying true effect size for the i^{th} gene in the j^{th} study where $\theta_{ij} = 0$ indicates that the i^{th} gene is not differentially expressed in the j^{th} study and $\theta_{ij} \neq 0$ indicates that the i^{th} gene is differentially expressed in the j^{th} study, $i = 1, 2 \ldots, G$: $j = 1, 2, \ldots, K$. The goal of our method is to detect genes that are differentially expressed in a majority of studies against the null hypothesis that they are not differentially expressed in any study. As a general rule, we target those genes that are differentially expressed in at least half of the studies. That is, for the i^{th} gene, the hypothesis setting for our meta-analysis method is

$$HS_m \colon \{H_0 \colon \sum_{j=1}^{K} I(\theta_{ij} \neq 0) = 0 \, vs \, H_1^m \colon \sum_{j=1}^{K} I(\theta_{ij} \neq 0) \geq m\}$$

where $m = \lceil K/2 \rceil$, i.e., m is the smallest integer that is not lower than $K/2$.

Note that, the hypothesis setting under our meta-analysis method can be generalized for any choice of m ranging from $\lceil K/2 \rceil + 1$ to K. Since the WOP meta-analysis method is mostly focused on testing HS_m for $m = \lceil K/2 \rceil$ for simplicity [11], we also focus on testing HS_m for $m = \lceil K/2 \rceil$ in this study.

2.2.2 PROPOSED EMPIRICALLY ADJUSTED WEIGHTED ORDERED P-VALUES METHOD

In this section, we describe our proposed meta-analysis method which empirically modifies the raw p-values from multiple studies and computes multiple testing corrected p-values after appropriately combining them across the studies. This empirical modification of the raw p-values will ensure that the p-values from all studies are uniformly distributed under the null hypotheses, so that the key assumption of the p-value combination methods such as the Fisher's [1] and the Stouffer's [2] methods is satisfied. Next, we provide the detailed steps of our proposed meta-analysis method:

- **Step 1**: For gene i in study j, we obtained the p-value p_{ij} for testing the null hypothesis $\theta_{ij} = 0$ against the alternative hypothesis $\theta_{ij} \neq 0$, $i = 1, 2 \ldots, G$: $j = 1, 2, \ldots, K$.

- **Step 2**: We considered the inverse z-transformation to get the corresponding z-scores as follows:

$$z_{ij} = \Phi^{-1}(p_{ij}), \qquad i = 1, 2 \ldots, G \colon j = 1, 2, \ldots, K$$

- **Step 3**: The z-scores in step 2 may not follow a standard normal distribution under the null hypotheses. Therefore, we estimated the parameters of the null distribution of the z-scores empirically assuming that the null distribution is normal with mean δ_0 (not necessarily 0) and standard deviation σ_0 (not necessarily 1) using an Empirical Bayes method as described in section 2.2.3. Let $\hat{\delta}_0$ and $\hat{\sigma}_0$ be the estimated mean and standard deviation of the null distribution. We modified the z-scores, obtained in step 2, using the estimated parameters as:

$$z_{ij}^{'} = \frac{z_{ij} - \hat{\delta}_0}{\hat{\sigma}_0}, \qquad i = 1, 2, \ldots, G \colon j = 1, 2, \ldots, K$$

  These modified z-scores $z_{ij}^{'}$ are expected to follow a standard normal distribution under the null hypotheses.

- **Step 4**: We converted the empirically adjusted z-scores into corresponding p-values as :

$$p_{ij}^{'} = \Phi(z_{ij}^{'}), \qquad i = 1, 2, \ldots, G \colon j = 1, 2, \ldots, K$$

- **Step 5**: For a gene i, we ordered the p-values over the K independent studies. Let $p_{i(j)}^{'}$ denote the j^{th} ordered p-value for gene i, $i = 1, 2, \ldots, G \colon j = 1, 2, \ldots, K$. We calculated the summary statistic as follows:

$$T_i = \sum_{j=1}^{K} w_j H(p_{i(j)}^{'})$$

  where $p_{(i(j))}^{'}$s denote the list of p-values, corresponding to the i^{th} gene, ordered over the K studies, $j = 1, 2, \ldots, K$. Here, w_j represents the weight corresponding to the j^{th} ordered p-value and $H(.)$ denotes the p-value combination method, $i = 1, 2, \ldots, G \colon j = 1, 2, \ldots, K$. In this study, we considered the same choices for the weights and p-value combination methods as considered by the WOP method [11].

Two different weighting schemes based on the binomial distribution are considered – binomial weighting and half-binomial weighting [11]. The weights are calculated based on the Binomial distribution so that p-values closer to the median are up-weighted and smallest/largest p-values are down-weighted. This makes sense because the behavior of the majority of studies are best captured by the p-values that are closer to the center of the distribution. In the binomial weighting scheme every p-value contribute to the combined summary statistic since all the weights are non-zero. Additionally, since the Fisher's method can be very sensitive to extremely small p-values, we considered the half-binomial weighting scheme where the effect of the smallest p-values are reduced on the combined summary statistic, by considering zero weights for them.

Thus, the binomial weighting scheme is defined as $w_j^b = f(j-1\colon K-1, 0.5), j = 1, 2, ..., K$, where $f(x\colon n, p)$ denotes the probability mass function of the binomial distribution $Bin(n, p)$ for $x = 0, 1, \ldots, n$. The half binomial weighting scheme is defined as $w_j^{hb} = w_j^b$ for $m \leq j \leq K$ and 0 for $j < m$. The function $H(.)$ in the WOP statistic depends on the choice of the p-value combination method. In particular, two popular p-value combination methods are considered – Fisher's method [1] where $H(p_{i(j)}) = -2log(p_{i(j)})$, and Stouffer's method [2] where $H(p_{i(j)}) = \Phi^{-1}(1 - p_{i(j)})$, $i = 1, 2\ldots, G\colon j = 1, 2, \ldots, K$.

- **Step 6**: We obtained the p-value p^i, for gene i by comparing the statistic defined in step 5, to the numerical distribution by simulating $U(0, 1)$ random variables as described below:

  (i) We randomly generated p-values from $U(0, 1)$ distribution for all the G genes in the K studies. We repeated this data generation process B times. Let $p_{ij}^{(b)}$ denotes the p-value for the i^{th} gene in the j^{th} study in the b^{th} dataset, $i = 1, 2, .., G\colon j = 1.2, \ldots, K\colon b = 1, 2, \ldots, B$.

  (ii) We calculated the summary statistic $T_i^{(b)} = \sum_{j=1}^{K} w_j H(p_{i(j)}^{(b)})$, for gene i, using the simulated p-values obtained in (i), $i = 1, 2, .., G\colon b = 1, 2, \ldots, B$.

  (iii) For gene i, the p-value corresponding to the summary statistic T_i is computed as

$$p^i = \frac{\sum_{b=1}^{B} I\{T_i^{(b)} \geq T_i\}}{B}$$

An alternative way for obtaining the p-values of the WOP statistic is based on permutation analysis which requires the original data for each study. Since in the context of meta-analysis it is not always feasible to obtain the original data, we focus on the more practical solution by obtaining the p-values based on the numerical distribution of the summary statistic. Finally, the Benjamini-Hochberg method is applied to the p-values of the summary statistic to account for multiple testing [15].

After obtaining the list of differentially expressed genes using Benjamini-Hochberg p-value cutoff of 0.05, the performance of the proposed method is assessed using the following four performance measures:

- Sensitivity: Proportion of genes which are correctly identified as differentially expressed out of all the genes which are truly differentially expressed.
- Specificity: Proportion of genes which are correctly identified as non differentially expressed out of all the genes which are truly non differentially expressed.
- FDR: Proportion of genes which are incorrectly identified as differentially expressed that are actually differentially expressed in less than m studies.
- Type I error rate: Proportion of genes which are incorrectly identified as differentially expressed that are actually not differentially expressed in any of the studies.

2.2.3 EMPIRICAL ESTIMATION OF NULL DISTRIBUTION

Suppose the p-values corresponding to G genes in a study are denoted as $p_1, p_2, \ldots, p_G$. These p-values can be converted into z-scores as $z_i = \Phi^{-1}(p_i)$, $i = 1, 2 \ldots, G$. The null distribution of the z-scores is $N(0, 1)$ theoretically. However, in large-scale testing situations empirical and theoretical null might differ. The large-scale multiple testing situation enables us to estimate the null distribution of the z-scores. In this section, we will briefly discuss an empirical Bayes method, proposed by Efron [16], for estimating the null distribution empirically.

The z-scores, corresponding to the G genes, can be categorized into two groups – the "uninteresting" group if the z_i is obtained from the null distribution, and the "interesting" group if the z_i is obtained from the non-null distribution, $i = 1, 2 \ldots, G$. Let p_0 denotes

the prior probability of the z-scores belonging to the "uninteresting" group and $p_1 = 1 - p_0$ denotes the prior probability of the z-scores belonging to the "interesting" group. Suppose $f_0(z)$ and $f_1(z)$ be the densities of the z-scores in the "uninteresting" and the "interesting" groups respectively. The mixture density of the z-scores is defined as $f(z) = p_0 f_0(z) + p_1 f_1(z)$. Following Bayes theorem, the *a posteriori* probability of belonging to the "uninteresting" group given z is $Pr["uninteresting"|z] = \frac{p_0 f_0(z)}{f(z)}$.

The mixture density $f(z)$ is estimated using the Poisson regression method introduced by Lindsey as described in Efron and Tibshirani [24]. Here the range of z-scores is partitioned into k equal intervals, with k^{th} interval having midpoint x_k and containing count s_k of the total G z-scores. If the z-scores are independent then the counts s_k will follow a Poisson distribution where the expected value λ_k is approximately proportional to $f(x_k)$. These λ_ks are estimated using Lindsey's method which amounts to estimating $f(z)$.

The null density, f_0 is estimated from the central peak of the histogram of the z-scores. Assuming that f_0 is a normal distribution with mean δ_0 and standard deviation σ_0, for z-scores close to zero, we can write $log(f(z)) = -\frac{1}{2}(\frac{z-\delta_0}{\sigma_0})^2 + constant$. The parameters of f_0 are estimated as: $\delta_0 = argmax\{f(z)\}$ and $\sigma_0 = [-\frac{d^2}{dz^2} log f(z)]_{\delta_0}^{-\frac{1}{2}}$.However, the above estimate of σ_0 can be unstable [16]. Therefore, a smoothing step is applied where a quadratic curve $a_0 + a_1 x_k + a_2 x_k^2$ is fitted by ordinary least squares to the estimated $log(f(x_k))$ values, for x_k within 1.5 units of the maximum δ_0, yielding the final estimate of σ_0 as $[-2a_2]^{-\frac{1}{2}}$. Furthermore, assuming that the interesting z-scores always fall outside $\delta_0 \pm c\sigma_0$ one can obtain an unbiased estimate for p_0 as $\hat{p_0} = \frac{\hat{\pi}(c)}{G_0(c)}$, where $\hat{\pi}(c) = \#(z_i \in \delta_0 \pm c\sigma_0)/G$ and $G_0(c) = 2\Phi(c) - 1$.

This method of estimation of the parameters of null distribution is called the method of "central-matching". More details about this method can be found in Efron [16] and Efron [25].

2.3 RESULTS

2.3.1 SIMULATION STUDIES

We conducted simulation studies to evaluate the performance of our proposed method for accurate identification of significant genes in majority of studies. We simulated continuous gene expression datasets for multiple independent studies. Details of the data generation process are given below.

We considered 10 independent studies each involving continuous gene expression levels for 3000 genes, i.e., $K = 10$ and $G = 3000$. We considered two groups of subjects in each study where each group consists of 20 subjects, i.e., $n_1 = n_2 = 20$. We considered 50 genes as differentially expressed between the two subject groups in 1,2,. . . ,10 studies respectively (see Figure 5). That is, in total, 500 genes are differentially expressed between the subject groups in at least one study. Since our alternative hypothesis for a gene is that it is differentially expressed in at least five studies, we aim to identify only the 300 genes (10%) that are differentially expressed in at least five of the studies.

We generated the (log) expression level for the i^{th} gene, l^{th} subject in the k^{th} group for each study separately using the following model:

$$y_{ikl} = \mu + G_i + V_k + GV_{ik} + W_{ikl} + e_{ikl}$$

Here μ denotes the overall mean effect, G_i denotes the effect due to the i^{th} gene, V_k denotes the effect due to the k^{th} subject group, and GV_{ik} denotes the interaction effect between the i^{th} gene and the k^{th} subject group, W_{ikl} denotes the effect of a hidden variable and e_{ikl} denotes the error component corresponding to the i^{th} gene, l^{th} subject in the k^{th} group, $i = 1, 2. . . , G$: $k = 1, 2$: $l = 1, 2, . . . , n_k$.

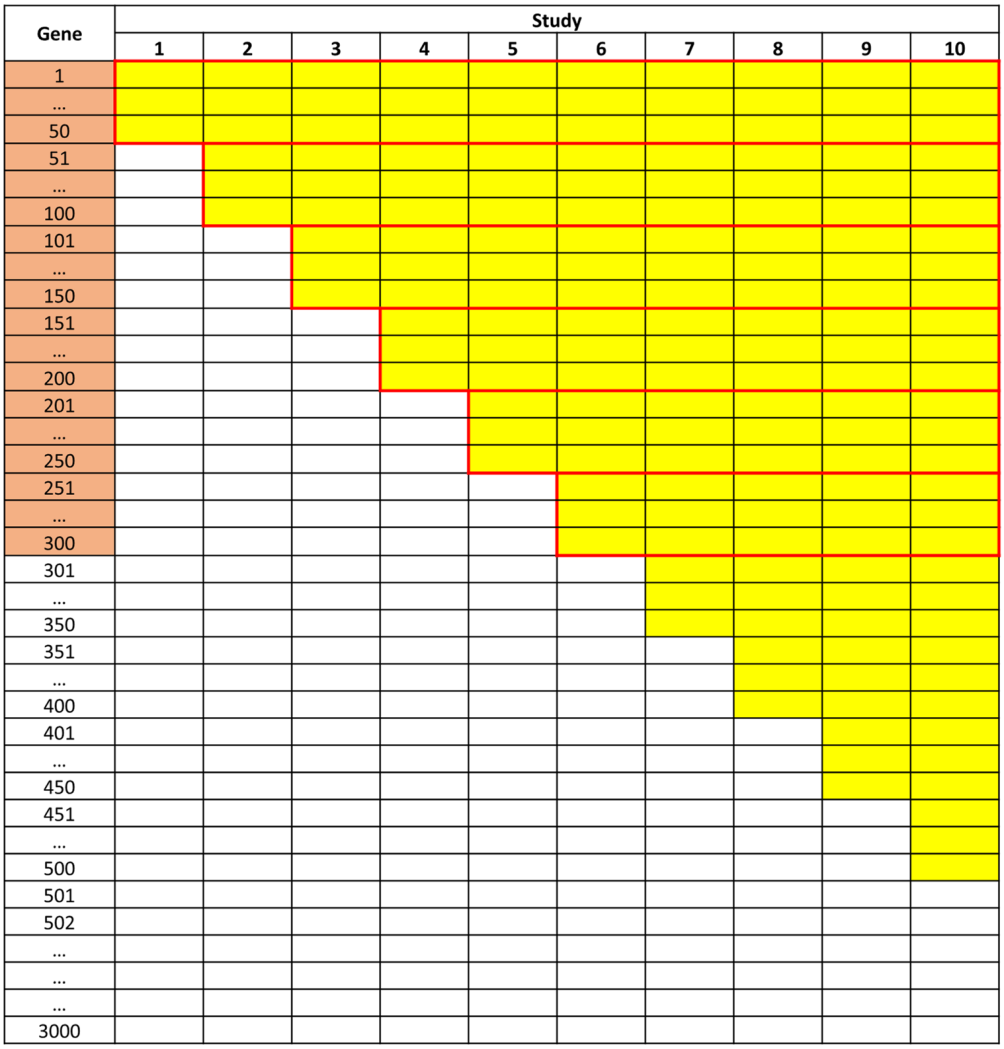

Fig. 5. Illustration of differentially expressed genes in 10 studies.

For our simulations, we considered μ, G_i, and V_k as zero for all i, k and l, for simplicity. Note that, we considered 50 genes to be differentially expressed between the two subject groups in 1,2,...,10 studies respectively. The differences in magnitudes of (log) expression values of these genes between the two groups are considered as eight, which are obtained through the generation of the interaction terms between the genes and the groups, $(GV)_{ik}$s, as follows:

For study j where $j = 1, \ldots, K$,

$(GV)_{i1} = -4, (GV)_{i2} = 4$ for $i = 1, \ldots, 25j$
$(GV)_{i1} = 4, (GV)_{i2} = -4$ for $i = 25j + 1, \ldots, 50j$
$(GV)_{i1} = (GV)_{i2} = 0$ for $i = 50j + 1, \ldots, G$

In our simulations, we assumed the presence of a hidden variable which acts as a confounder. The effect of the hidden confounder for the i^{th} gene, l^{th} subject in the k^{th} group was generated such that it varied over the two subject groups, different groups of genes as well as over different studies. We considered $W_{ikl} = u_{ikl} I(s_{ikl} = 1)$, where $s_{ikl} \sim Bernoulli(0.4)$ and u_{ikl} are generated depending on the gene, subject group and the study ID j as given below:

$$u_{ikl} = \begin{cases} N(-1 + j + \delta.I, 0.01^2), & \text{for } i = 1, \ldots, 25j \colon l = 1, \ldots, n_k. \\ N(2 + j + \delta.I, 0.01^2), & \text{for } i = 25j + 1, \ldots, 50j \colon l = 1, \ldots, n_k. \\ N(5 + j + \delta.I, 0.01^2), & \text{for } i = 50j + 1, \ldots, G \colon l = 1, \ldots, n_k. \end{cases}$$

where $i = 1, 2 \ldots, G \colon k = 1, 2 \colon j = 1, 2, \ldots, K$. Here, $I = 1$ for group 1 (i.e., for $k = 1$) and $I = 0$ for group 2 (i.e., for $k = 2$). The magnitude of the difference between the means of the distributions of u_{ikl} between the two subject groups is given by δ. In our simulations, we considered $\delta = 4$.

We introduced correlations among some of the genes through the generation of the error terms e_{ikl}s as described below:

We considered four groups of correlated genes given by $C_1 = \{1, 2, \ldots, 30\}$, $C_2 = \{121, 122, \ldots, 180\}$, $C_3 = \{1501, 1502, \ldots, 1560\}$ and $C_4 = \{2671, 2672, \ldots, 2730\}$. The error term e_{ikl} for the i^{th} gene, l^{th} subject in the k^{th} group , $i = 1, 2, \ldots, G \colon k = 1, 2 \colon l = 1, 2, \ldots, n_k$ is generated as:

$$e_{ikl} = \begin{cases} \frac{1}{\sqrt{2}} e^1_{ikl} + \frac{1}{\sqrt{2}} e^2_{ikl}, & \text{if } i \in \{C_1, C_2, C_3, C_4\}. \\ e^2_{ikl}, & \text{o.w.} \end{cases}$$

where e^1 are generated independently from $N(0, 1)$ in such a way that the values of e^1 are same for all the genes belonging to the same group, and e^2 are generated independently from $N(0, 2^2)$.

For each study, after generating the (log) gene expression values for all the subjects, we tested whether the genes are differentially expressed between the two subject groups using "limma" in Bioconductor [26] and stored the raw p-values. We then applied our proposed method, as described in section 2.2.2, and obtained the list of differentially expressed genes with a Benjamini-Hochberg adjusted p-value cutoff of 0.05. We obtained the type I error rate for our proposed method as well as evaluated its performance using sensitivity, specificity, and false discovery rate (FDR) based on 500 Monte-Carlo iterations. Because of the hypothesis setup, as described in section 2.2.1, false positives and type I error are not the same. Type I error is rejecting the null hypothesis for a gene that is not differentially expressed in any of the studies while a false positive is rejecting the null hypothesis for a gene that is differentially expressed in less than m studies [11]. We, additionally, compared the performance of our proposed method with the original WOP method (without any empirical adjustment) [11].

Table 2 summarizes the simulation results for our proposed method and the corresponding WOP method with the two choices for the p-value combination approach (Fisher [1] and Stouffer [2]) and the two weighting schemes (binomial and half-binomial [11]), as discussed in the methods section 2.2.2. The type I error rates for our proposed method are controlled at 0.05 but the WOP method has extremely high type I error rates in all settings. Our proposed method also has significantly lower FDR values compared to the WOP method. Although the proposed method has slightly lower sensitivity values, the specificity values are much higher compared to the WOP method. In general, the half binomial weighting scheme has slightly lower sensitivity and slightly higher specificity values compared to the binomial weighting scheme. The FDR values for both methods and the type I error rates for the WOP method are also slightly lower for half binomial weighting. The results did not vary significantly between the choices of the p-value combination approaches.

Table 2

Performances of the proposed method and the WOP method in presence of hidden confounder. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations. The proportion of differentially expressed genes between the two subject groups is 10%.

Weighting scheme	**P-value combination approach**	**Method**	**Type I error rate**	**Sensitivity**	**Specificity**	**FDR**
Binomial	Fisher	proposed	0.038	0.955	0.983	0.138
		WOP	0.530	0.999	0.767	0.677
	Stouffer	proposed	0.040	0.945	0.984	0.130
		WOP	0.518	0.997	0.784	0.661
Half-Binomial	Fisher	proposed	0.041	0.926	0.986	0.120
		WOP	0.489	0.992	0.815	0.627
	Stouffer	proposed	0.042	0.917	0.987	0.116
		WOP	0.483	0.990	0.822	0.618

We, additionally, varied the proportion of differentially expressed genes between the two subject groups, ranging from 5% to 20%. Figure 6 shows the performances of the proposed method as well as the WOP method with varying proportion of differentially expressed genes in presence of hidden confounder in the studies. Our proposed method has type I error rates controlled at 0.05 consistently in all settings. Although the type I error rates of the WOP method decreased with increase in proportion of differentially expressed genes in the studies, they are still unacceptably high (see Figure 6). The FDR values of our proposed method are also much lower than those of the WOP method in all settings, although the values for the WOP method decreased with increase in proportion of differentially expressed genes. Our proposed method has slightly lower sensitivity values compared to the WOP method but the values increased when the proportion of differentially expressed genes is increased. The specificity values of the proposed method are close to one in all settings but the WOP method has much lower specificity values which further decreased as the proportion of differentially expressed genes is increased.

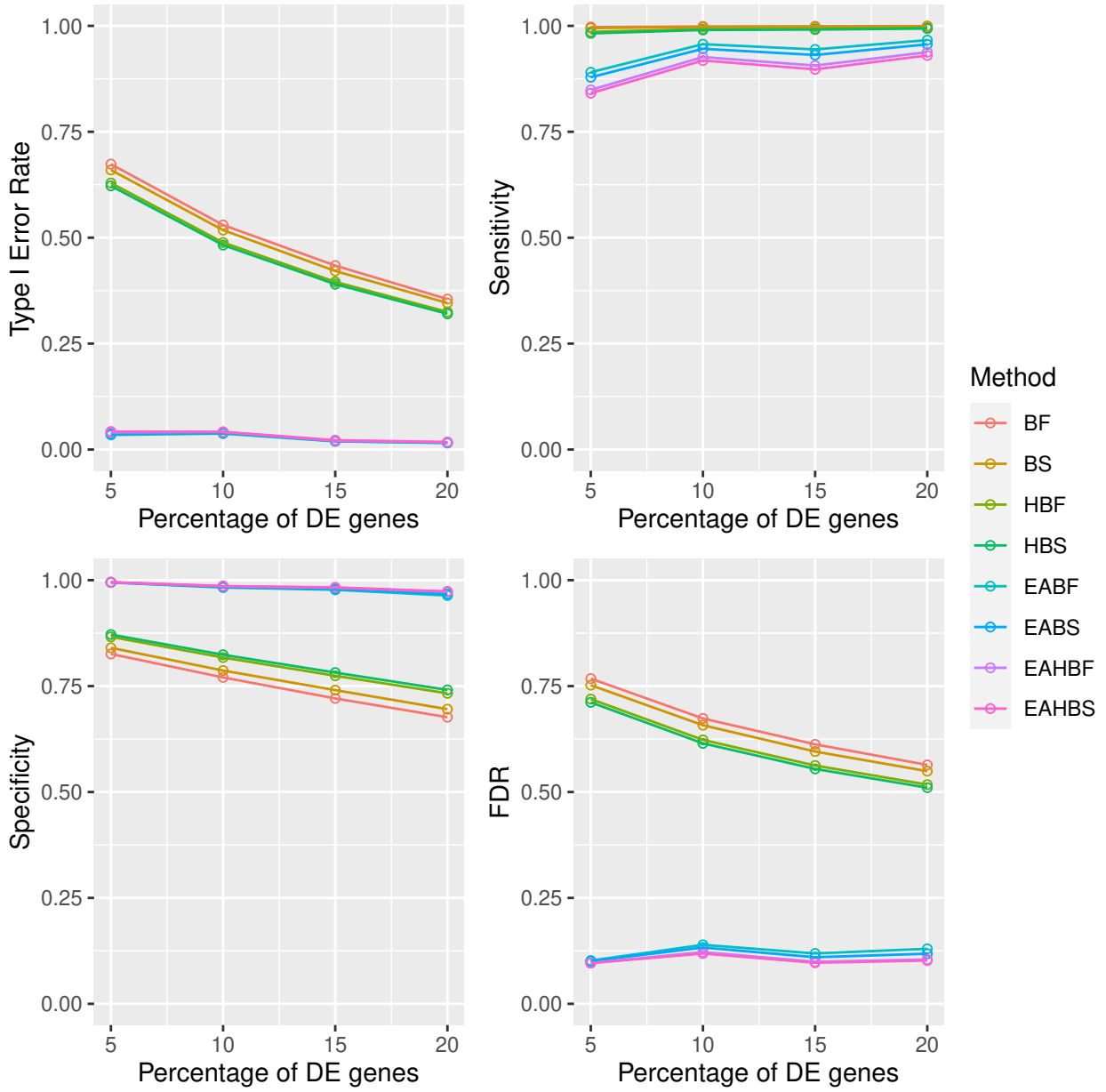

Fig. 6. Performances of the proposed method and the WOP method in presence of hidden confounder with varying proportion of differentially expressed genes between two subject groups. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations.(Continued on the following page.)

BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

We also considered some variations in our simulations. In particular, we considered a simulation scenario where we assumed the presence of a hidden variable that does not act as a confounder as well as a simulation scenario where there do not exist any effect of a hidden variable or confounder in the studies. We also looked at the effect of changing the number of studies. All these scenarios are described below.

a) Presence of a hidden variable that does not act as a confounder

In this simulation scenario, we assumed the presence of a hidden variable which affects the outcome but does not vary between the two subject groups. We generated the distribution of the hidden variable for the i^{th} gene, l^{th} subject in the k^{th} group, as $W_{ikl} = u_{ikl}I(s_{ikl} = 1)$, where $s_{ikl} \sim Bernoulli(0.4)$ and u_{ikl} are generated as given below:

$$u_{ikl} = N(-4 + j, 0.1^2) \qquad \text{for } i = 1, \ldots, G; k = 1, 2; l = 1, \ldots, n_k; j = 1, \ldots, K$$

We considered 10% of the genes as differentially expressed between the two subject groups in at least five studies as considered before. The differences in magnitudes of (log) expression values of these differentially expressed genes are considered as two. All the other terms in the model for simulation are generated in the same way as described previously.

Table 3 shows the results for our proposed method and the WOP method. In this simulation scenario, both the methods have controlled type I error rates under all settings. The FDR values of the proposed method are slightly lower compared to the WOP method. Both methods have very similar sensitivity and specificity values. The methods with half binomial weighting scheme have lower sensitivity as well as FDR values compared to those with binomial weighting scheme, similar to what we observed in presence of hidden confounder.

Table 3

Performances of the proposed method and the WOP method in presence of hidden variable that does not act as confounder. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations. The proportion of differentially expressed genes between the two subject groups is 10%.

Weighting scheme	**P-value combination approach**	**Method**	**Type I error rate**	**Sensitivity**	**Specificity**	**FDR**
Binomial	Fisher	proposed	0.044	0.913	0.986	0.123
		WOP	0.051	0.919	0.984	0.136
	Stouffer	proposed	0.043	0.884	0.988	0.110
		WOP	0.052	0.893	0.986	0.126
Half-Binomial	Fisher	proposed	0.043	0.839	0.990	0.095
		WOP	0.050	0.848	0.989	0.107
	Stouffer	proposed	0.043	0.824	0.991	0.092
		WOP	0.050	0.832	0.989	0.103

The performances of the two methods with varying proportion of differentially expressed genes, in the presence of hidden variable that does not act as confounder, are shown in Figure 7.

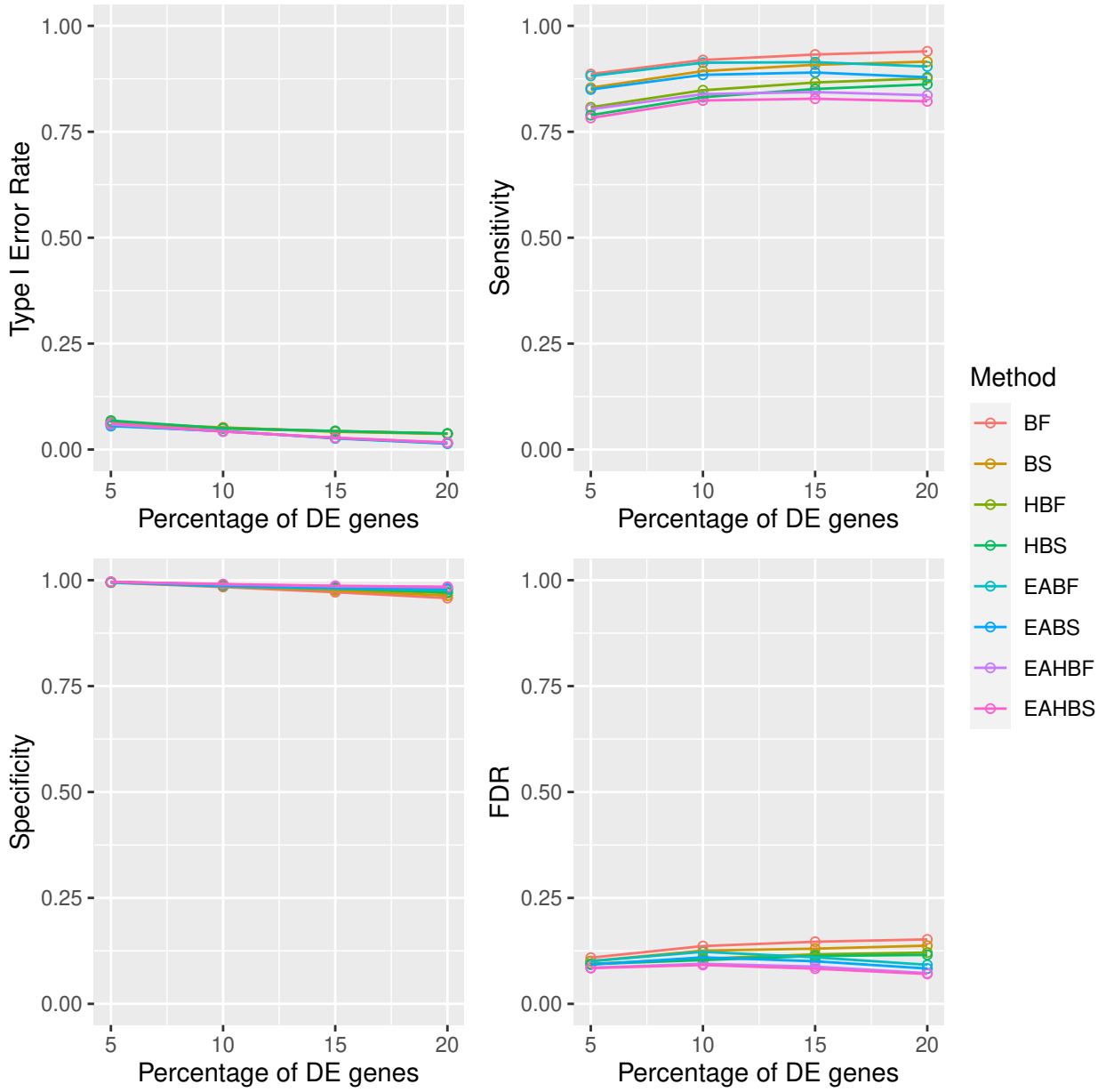

Fig. 7. Performances of the proposed method and the WOP method in presence of hidden variable that does not act as confounder with varying proportion of differentially expressed genes between two subject groups. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations.(Continued on the following page.)

BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

The type I error rates remained controlled at 0.05 for both methods consistently in all settings. Both the methods have similar FDR values for smaller proportion of differentially expressed genes, but the FDR values of the WOP method slightly increased with increase in the proportion of differentially expressed genes. The sensitivity values are very similar for both methods with half binomial weighting having slightly lower values than binomial weighting. Both methods have very similar specificity values.

b) No effect of any hidden variable or confounder

In this simulation scenario, we assumed that there is no effect of any hidden variable or confounder in the studies. Therefore, we set $W_{ikl} = 0$, for all i, k and l. Here also, we considered 10% of the genes as differentially expressed between the two subject groups in at least five studies and the differences in magnitudes of (log) expression values of these genes are considered as two. The random error term (e_{ikl}) are generated as before with e^1 drawn independently from $N(0, 0.5^2)$ and e^2 generated independently from $N(0, 5.5^2)$. All the other terms in the model for simulation are generated in the same way as described previously.

Table 4 shows the results for our proposed method as well as the WOP method. The type I error rates are controlled for both methods under all settings. The sensitivity values are slightly lower for the proposed method. FDR values are also slightly lower for the proposed

method compared to the WOP method. Both methods have very similar specificity values. The half binomial weighting scheme has lower sensitivity as well as FDR values compared to the binomial weighting scheme, consistent with what we observed in the previous simulation scenarios.

Table 4

Performances of the proposed method and the WOP method when there is no effect of any hidden variable or confounder. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations. The proportion of differentially expressed genes between the two subject groups is 10%.

Weighting scheme	**P-value combination approach**	**Method**	**Type I error rate**	**Sensitivity**	**Specificity**	**FDR**
Binomial	Fisher	proposed	0.037	0.905	0.984	0.137
		WOP	0.051	0.919	0.980	0.163
	Stouffer	proposed	0.038	0.882	0.986	0.122
		WOP	0.052	0.899	0.983	0.147
Half-Binomial	Fisher	proposed	0.036	0.837	0.990	0.100
		WOP	0.051	0.856	0.986	0.126
	Stouffer	proposed	0.038	0.824	0.990	0.098
		WOP	0.051	0.840	0.987	0.119

Figure 8 shows the simulation results for the two methods with varying proportion of differentially expressed genes when there is no effect of hidden variable or confounder in the studies. The performances of the methods are very similar to what we observed in the previous scenario in the presence of hidden variable that does not act as confounder.

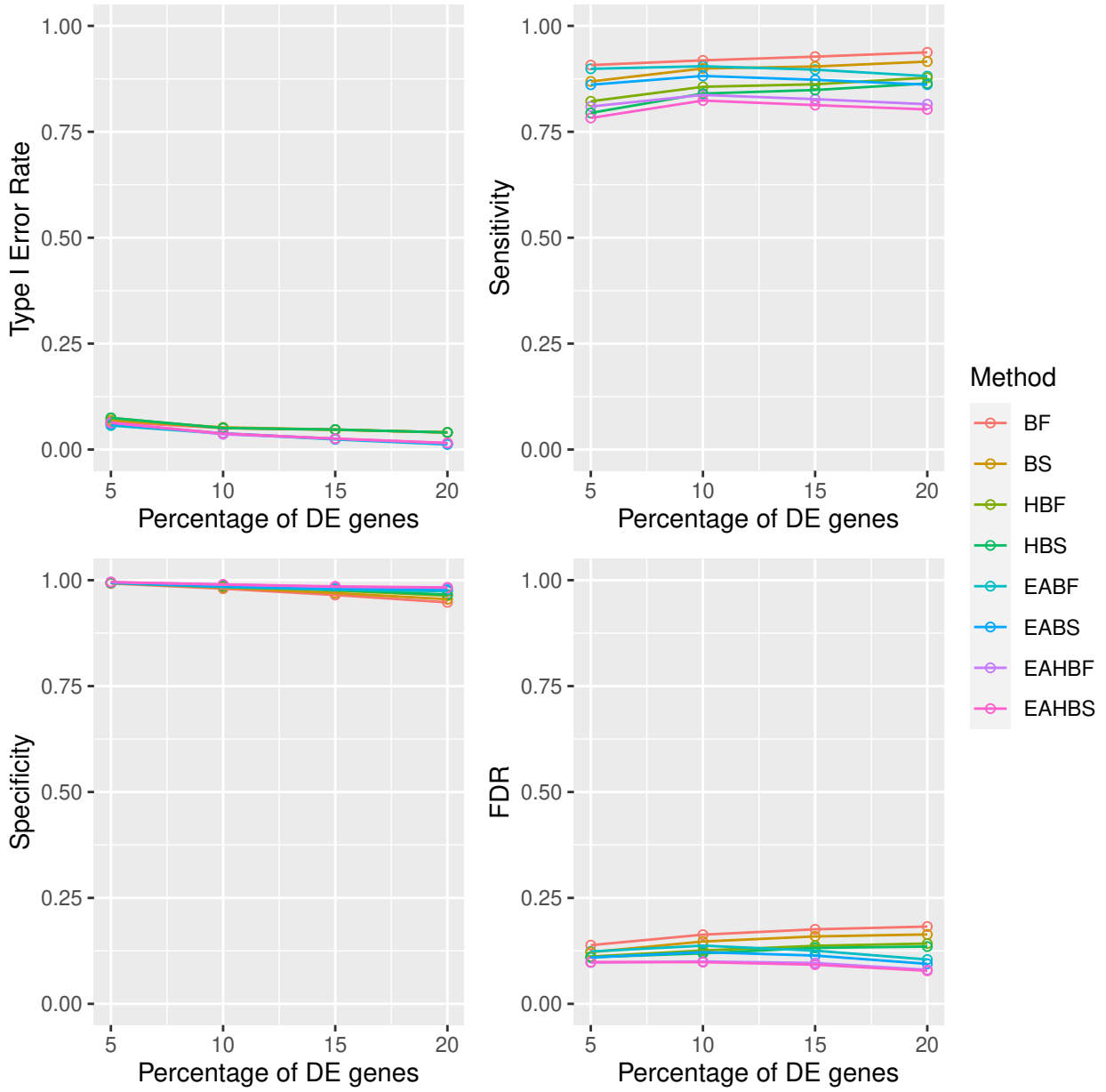

Fig. 8. Performances of the proposed method and the WOP method when there is no hidden variable or confounder with varying proportion of differentially expressed genes between two subject groups. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations. (Continued on the following page.)

BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

c) Effect of changing the number of studies

Here, we compared the performance of our proposed method with the WOP method for varying number of studies (5, 10 and 15). In this simulation scenario, we assumed presence of a hidden variable which acts as a confounder. We considered K independent studies each involving continuous gene expression levels for $G = 3000$ genes. Note that, we considered 10% of the genes are differentially expressed between two subject groups in at least $\lceil K/2 \rceil$ studies. The differences in magnitudes of (log) expression values of these genes between the two groups are considered as a function of number of studies (K), which are obtained through the generation of the interaction terms between the genes and the groups, $(GV)_{ik}$s, as follows:

For study j where $j = 1, \ldots, K$,

$(GV)_{i1} = -35/K, (GV)_{i2} = 35/K$ for $i = 1, \ldots, 25j$
$(GV)_{i1} = 35/K, (GV)_{i2} = -35/K$ for $i = 25j + 1, \ldots, 50j$
$(GV)_{i1} = (GV)_{i2} = 0$ for $i = 50j + 1, \ldots, G$

The effect of the hidden confounder for the i^{th} gene, l^{th} subject in the k^{th} group was generated such that it varied over the two subject groups, different groups of genes as well

as over different studies. We considered $W_{ikl} = u_{ikl}I(s_{ikl} = 1)$, where $s_{ikl} \sim Bernoulli(0.4)$ and u_{ikl} are generated depending on the gene, subject group and the study ID j as given below:

$$u_{ikl} = \begin{cases} N(-1 + j/2 + \delta.I, 0.01^2), & \text{for } i = 1, \ldots, 25j \colon l = 1, \ldots, n_k. \\ N(2 + j/2 + \delta.I, 0.01^2), & \text{for } i = 25j + 1, \ldots, 50j \colon l = 1, \ldots, n_k. \\ N(5 + j/2 + \delta.I, 0.01^2), & \text{for } i = 50j + 1, \ldots, G \colon l = 1, \ldots, n_k. \end{cases}$$

where $i = 1, 2 \ldots, G \colon k = 1, 2 \colon j = 1, 2, \ldots, K$. Here, $I = 1$ for group 1 (i.e., for $k = 1$) and $I = 0$ for group 2 (i.e., for $k = 2$). The magnitude of the difference between the means of the distributions of u_{ikl} between the two subject groups is given by δ. In our simulations, we considered $\delta = 4$. All the other terms in the model for simulation are generated in the same way as described previously.

Figure 9 summarize the simulation results for our proposed method and the corresponding WOP method with the two choices for the p-value combination approaches (Fisher and Stouffer) and the two weighting schemes (binomial and half-binomial) for varying number of studies (K). The type I error rates for our proposed method are controlled at 0.05 but the WOP method has extremely high type I error rates in all four settings and for all the choices of K. Our proposed method also has significantly lower FDR values compared to the WOP method in all four settings and for all the choices of K. When $K = 5$, sensitivity values are relatively lower for our proposed method. However, when $K = 10$ and 15, sensitivity values are very similar for both methods. The specificity values of the proposed method are close to one in all four settings and for all the choices of K but the WOP method has much lower specificity values in all four settings which further decreased as the number of studies is increased.

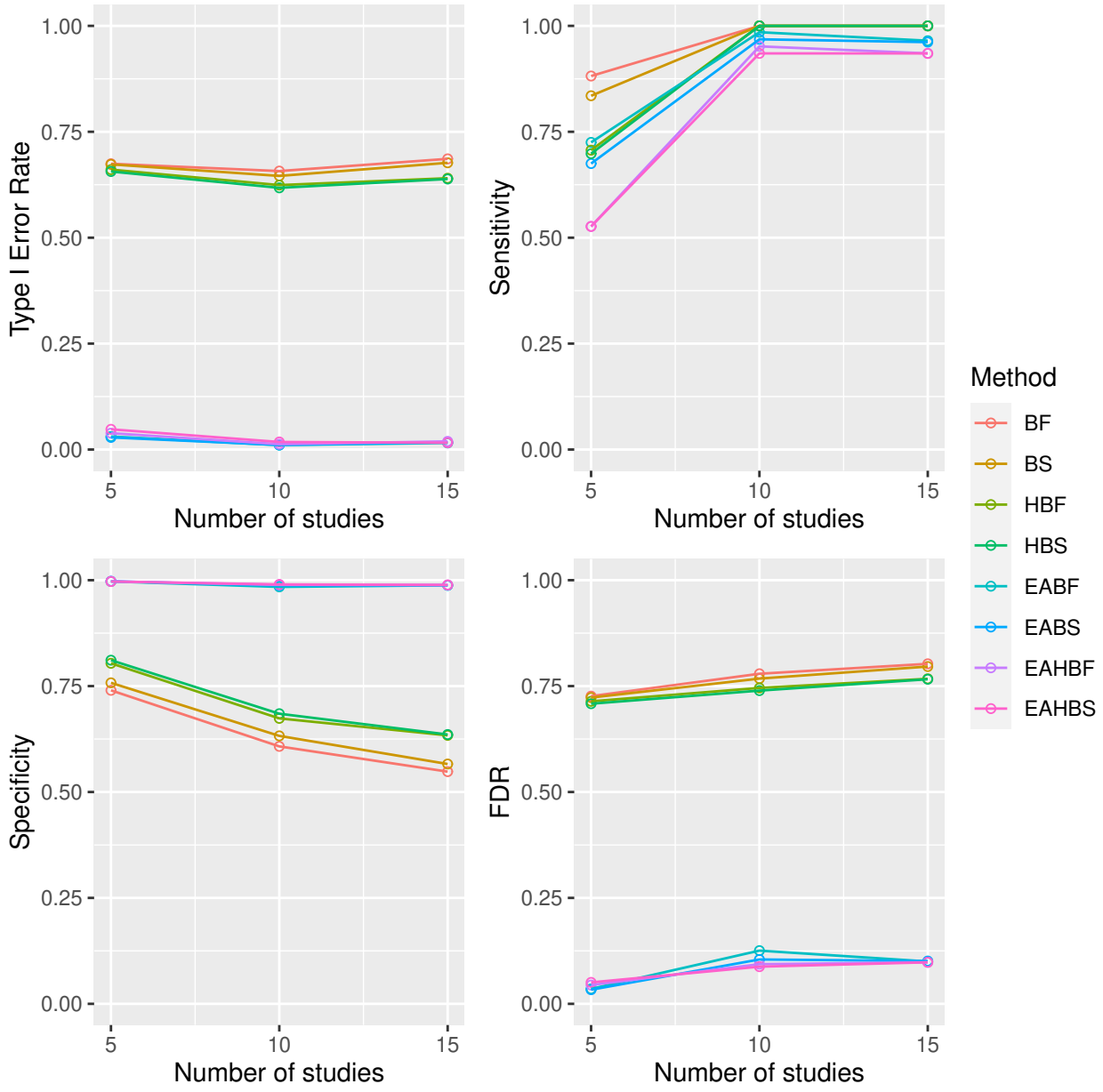

Fig. 9. Performances of the proposed method and the WOP method in presence of hidden confounder with varying number of studies. Type I error rate, sensitivity, specificity, and FDR values are obtained based on 500 Monte-Carlo iterations. (Continued on the following page.)

BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

2.3.2 AN APPLICATION TO LUNG CANCER STUDIES

We conducted meta-analysis using our proposed method on five lung cancer gene expression datasets. Details about these five studies can be found in [17, 27]. Each of the datasets contains normalized expression levels for 7,200 genes, and subjects with different types of lung cancer. We aimed to identify the genes that are differentially expressed between two lung cancer types - adenocarcinoma (AD) and squamous cell carcinoma (SQ) in at least three out of the five studies. Table 5 lists the detailed information on these five lung cancer studies. To obtain the p-values for the genes for each dataset, we tested for differential expression between AD and SQ subjects using "limma" [26]. We applied our proposed method, following the steps described in section 2.2.2, after empirically adjusting the p-values for the genes. The empirically estimated mean and the standard deviation of the original z-scores are – 0.83 and 1.99, respectively. That is, the empirically estimated null distribution of the original z-scores is much different from the theoretical null distribution. For comparison, we also applied the WOP method to identify the differentially expressed genes between the two lung cancer types in at least three studies.

Table 5

Detail information on lung cancer datasets.

Dataset	**Sample Size**	**Comparison**
Bhattacharjee	81	AD vs SQ
GSE11969	125	AD vs SQ
GSE29016	50	AD vs SQ
GSE30219	146	AD vs SQ
GSE43580	150	AD vs SQ

We considered a gene significant if the Benjamini-Hochberg adjusted p-value is less than 0.05. Table 6 summarizes the number of differentially expressed genes, identified by our proposed method as well as the WOP method, for the two choices of p-value combination approaches (Fisher [1] and Stouffer [2]) and the two weighting schemes (binomial and half binomial). The WOP method identified much higher number of significant genes compared to the proposed method at each combination of p-value combination approaches and weighting schemes. For the WOP method, the maximum number of significant genes (68.3%) is identified by the Fisher's p-value combination approach with binomial weighting scheme. Our proposed method with Stouffer's p-value combination approach and binomial weighting scheme identified the maximum number of significant genes, which is 1,474 (20.4%). In all settings, the WOP method identified more than 58% of the genes as significant which clearly indicates a possibility of high FDR. Consistent with our simulation results, the methods with binomial weighting scheme identified higher number of significant genes compared to the methods with half binomial weighting.

Table 6

The number of significant genes (percentage) identified by our proposed method and the WOP method with two choices of p-value combination approaches and two weighting schemes for the lung cancer data.

Weighting scheme	**P-value combination approach**	**Method**	**Number of significant genes (percentage)**
Binomial	Fisher	proposed	1406 (19.5%)
		WOP	4921 (68.3%)
	Stouffer	proposed	1474 (20.4%)
		WOP	4672 (64.9%)
Half-Binomial	Fisher	proposed	1317 (18.3%)
		WOP	4286 (59.5%)
	Stouffer	proposed	1371 (19.0%)
		WOP	4208 (58.4%)

Figure 10 shows the overlap between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and (b) Stouffer's p-value combination approaches for the lung cancer study. For the Fisher's p-value combination approach, there are 1,234 genes which are identified by both methods with both weighting schemes (see Figure 10(a)). Similarly, for the Stouffer's p-value combination approach, 1,312 genes are identified by both methods with both weighting schemes (see Figure 10(b)). At a specific combination of the p-value combination approach and weighting scheme, all the genes identified by our proposed method are also identified by the WOP method. In order to identify biological pathways associated with the gene lists identified by our proposed method with both p-value combination approaches and weighting schemes, we performed functional annotation clustering using Database for Annotation, Visualization and Integrated Discovery (DAVID) software [28]. Our proposed method identified several biologically relevant KEGG pathways including cell cycle, DNA replication, and p53 signaling pathway based on the Benjamini-Hochberg adjusted p-value cutoff of 0.05.

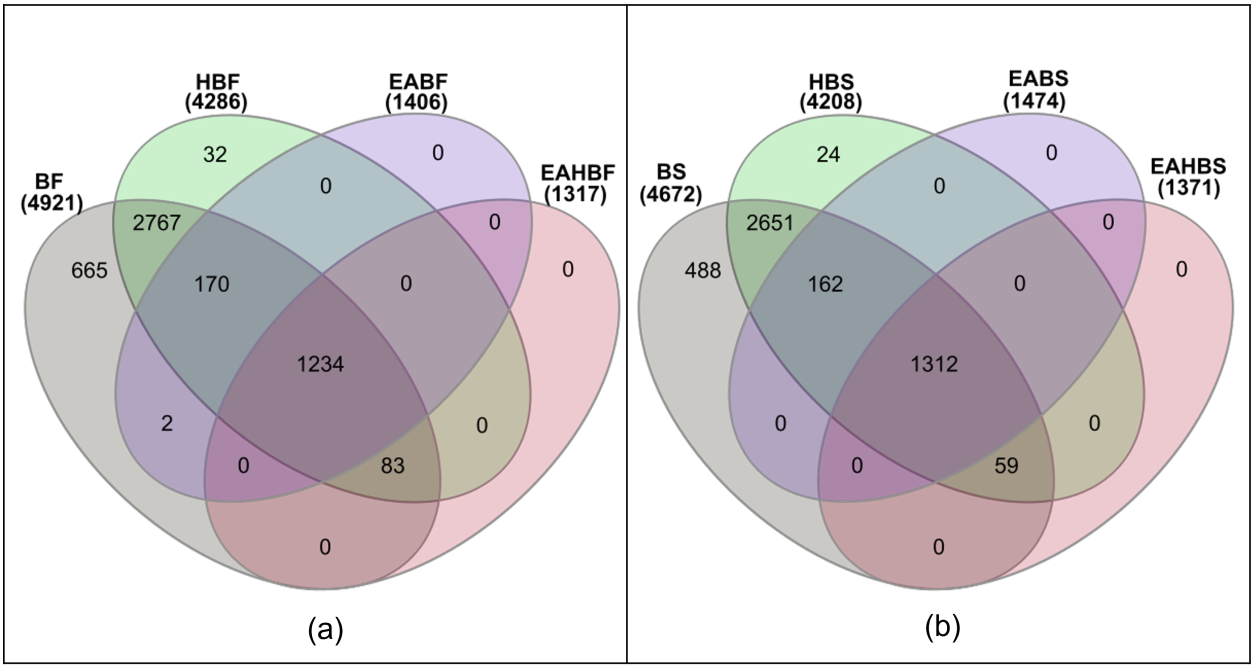

Fig. 10. Venn diagram showing the overlaps between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and (b) Stouffer's p-value combination approaches for the lung cancer study. BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

We further investigated some genes which are identified as significant by the WOP method but not by our proposed method. For the lung cancer study, 44 genes are not significant in any of the studies based on the p-values from the differential expression analysis in each study

but the WOP method with all choices of p-value combination approaches and weighting schemes have identified them as differentially expressed in at least three studies and the proposed method with any choice of the p-value combination approach and weighting scheme did not identify them as significant. For example, the genes with entrez IDs 11131 and 1723 are identified by the WOP method with both weighting schemes and p-value combination approaches, but not by our proposed method. Figure 11 and Figure 12 show the box plots of the expression levels of the genes for the two cancer types (AD and SQ) in each of the five studies. From the figures, we can see that the genes are not differentially expressed between the two cancer types in at least three studies. Further, based on the p-values from the differential expression analysis in each study, above mentioned genes are not significant in any of the studies. This suggests that these genes are unlikely to be differentially expressed between AD and SQ patients, and therefore, it is reasonable that our proposed method did not identify these genes.

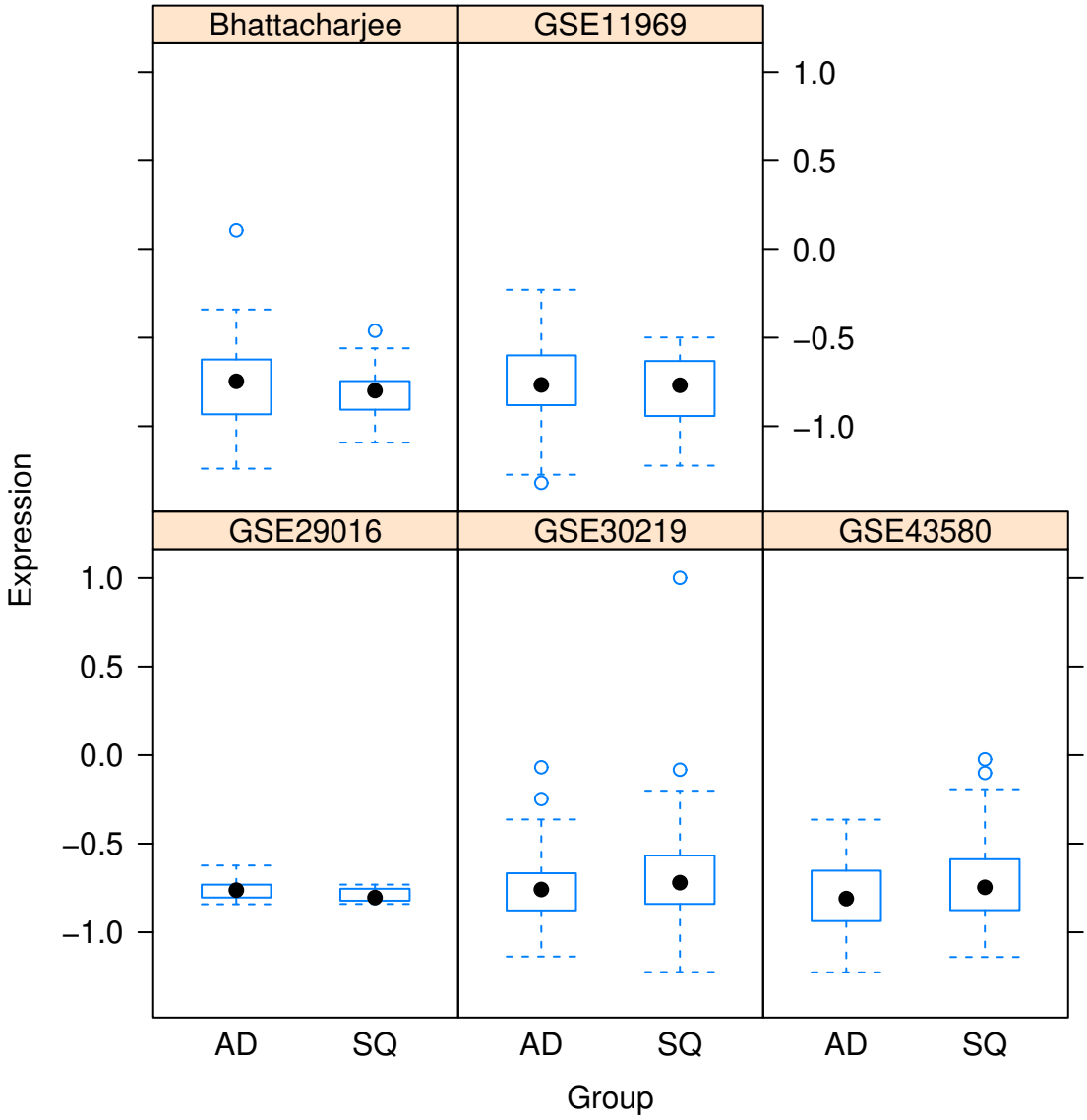

Fig. 11. The box plots of the gene with ID 11131 for the two cancer types in each of the five studies for lung cancer data.

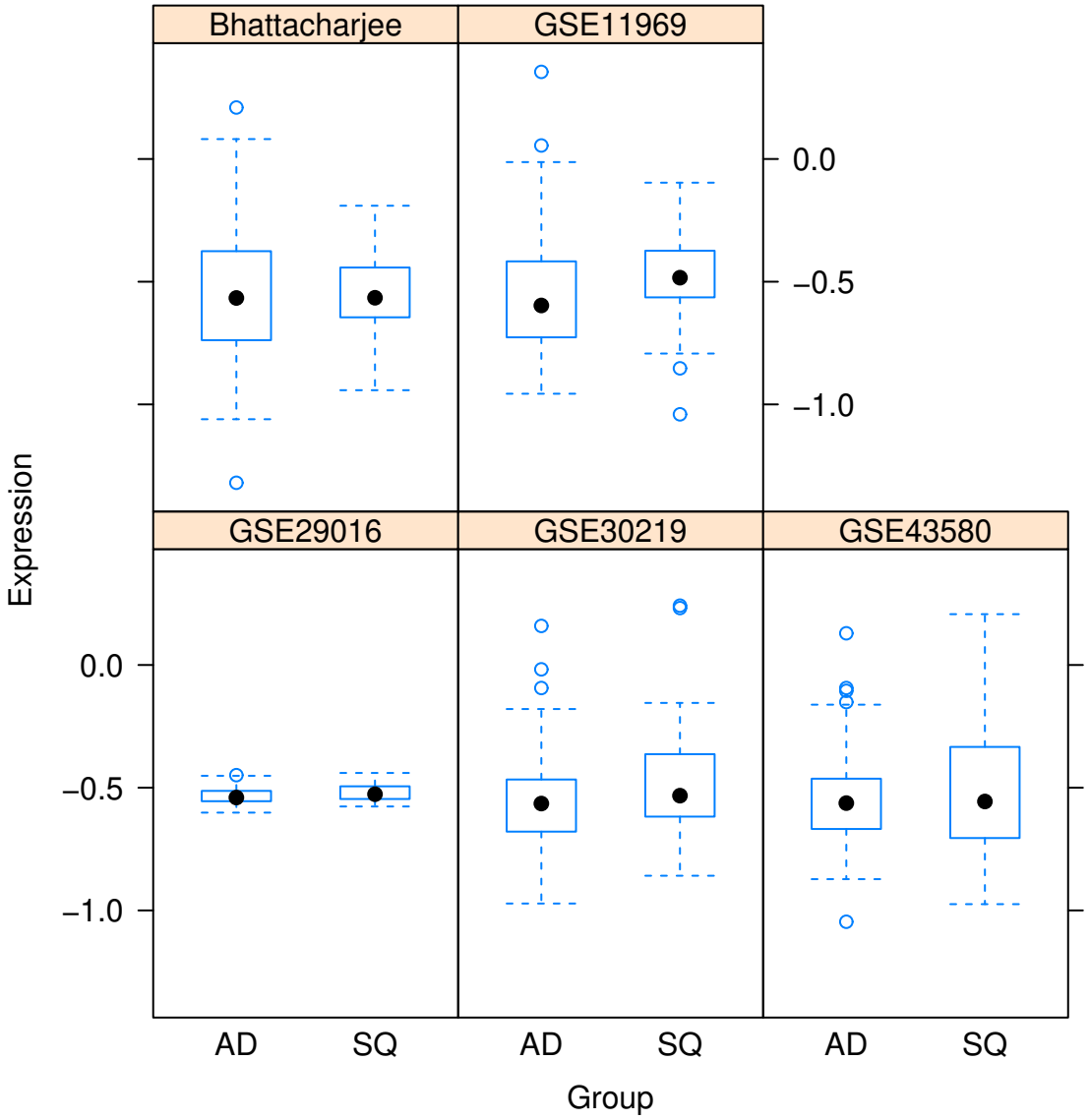

Fig. 12. The box plots of the gene with ID 1723 for the two cancer types in each of the five studies for lung cancer data.

2.3.3 FURTHER APPLICATIONS

To further compare the performance of our proposed method with the WOP method, we used two micro-array data sets. After the gene matching and filtering process, brain cancer and diabetes studies contain 5836 and 6645 genes, respectively. Table 7 and 10 provides the detailed information on seven brain cancer datasets and 16 diabetes datasets [10, 29]. The p-values for the genes for each dataset in brain cancer and diabetes studies were obtained from the supplementary material available in [10]. For the diabetes study with multiple groups, the p-values are obtained following the procedures as described in [29] i.e., by taking the minimum p-value of all the pairwise comparisons after adjusting for multiple testing.

Table 7

Detail information on brain cancer datasets.

Dataset	**Sample Size**	**Comparison**
GSE1993	65	AA vs GBM
GSE4412	85	AA vs GBM
GSE4271	100	AA vs GBM
GSE4290	180	AA vs GBM
GSE19578	53	AA vs GBM
GSE4381	29	AA vs GBM
GSE16011	284	AA vs GBM

For the brain cancer study, we identified the genes that are differentially expressed between two subtypes of brain tumors – anaplastic astrocytoma (AA) and glioblastoma (GBM) in at least four studies out of the seven studies. The estimated mean and the standard deviation of the empirical null distribution of the original z-scores are -0.36 and 1.42, respectively. That implies the empirically estimated null distribution of the original z-scores is different from the theoretical null distribution. We applied the WOP method and our proposed method to identify the differentially expressed genes between the two brain cancer

types. Table 8 shows the number of differentially expressed genes, identified by our proposed method and the WOP method, for the two choices of p-value combination approaches and two weighting schemes using the Benjamini-Hochberg adjusted p-value cutoff of 0.05. In all settings, the WOP method identified much higher number of significant genes ($> 30\%$) compared to the proposed method indicating possibility of high false discoveries. For both the methods, the maximum number of significant genes is identified by the Fisher's p-value combination approach with binomial weighting scheme. Consistent with our simulation results and previous data application, the methods with binomial weighting scheme identified higher number of significant genes compared to the methods with half binomial weighting.

Table 8

The number of significant genes (percentage) identified by our proposed method and the WOP method with two choices of p-value combination approaches and two weighting schemes for the brain cancer data.

Weighting scheme	**P-value combination approach**	**Method**	**Number of significant genes (percentage)**
Binomial	Fisher	proposed	635 (10.9%)
		WOP	2462 (42.2%)
	Stouffer	proposed	613 (10.5%)
		WOP	2260 (38.7%)
Half-Binomial	Fisher	proposed	518 (8.9%)
		WOP	1885 (32.3%)
	Stouffer	proposed	503 (8.6%)
		WOP	1796 (30.8%)

Figure 13 shows the overlap between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and

(b) Stouffer's p-value combination approaches for the brain cancer study. For the Fisher's p-value combination approach, there are 498 genes which are identified by both methods with both weighting schemes (see Figure 13(a)). Similarly, for the Stouffer's p-value combination approach, 482 genes are identified by both methods with both weighting schemes (see Figure 13(b)). We performed functional annotation clustering for the brain cancer study to identify pathways associated with the significant gene lists identified by our proposed method for all settings. Our proposed method identified important KEGG pathways for the brain cancer such as Focal adhesion, ECM-receptor interaction, Amoebiasis, PI3K-Akt signaling pathway, Pathways in cancer, p53 signaling pathway, Cell cycle, Shigellosis, Proteoglycans in cancer, TNF signaling pathway, Wnt signaling pathway and Protein processing in endoplasmic reticulum based on Benjamini-Hochberg adjusted p-value cutoff of 0.05.

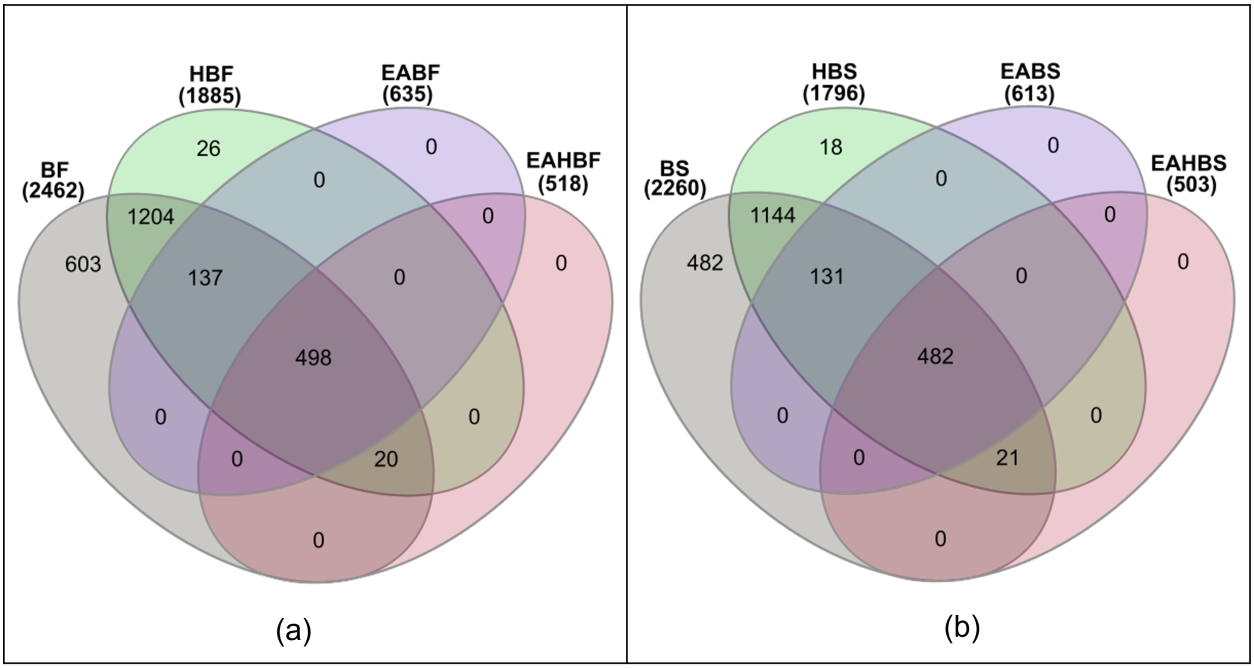

Fig. 13. Venn diagram showing the overlaps between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and (b) Stouffer's p-value combination approaches for the brain cancer study. BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

For the diabetes study, two datasets are excluded from the analysis because they are from human and we considered the remaining 14 mouse datasets. We tested the alternative hypothesis that the genes are differentially expressed in at least seven experiments out of 14

experiments. The estimated mean and the standard deviation of the empirically estimated null distribution of the original z-scores are -0.21 and 1.12, respectively. For this study also, we can see that the empirical null distribution of the original z-scores is different from the theoretical null distribution. Table 9 depicts the numbers of differentially expressed genes, identified by our proposed method and the WOP method, for all the settings using the Benjamini-Hochberg adjusted p-value cutoff of 0.05. Our proposed method has identified lesser number of differentially expressed genes compared to the WOP method in all scenarios, as we observed before with other studies.

Table 9

The number of significant genes (percentage) identified by our proposed method and the WOP method with two choices of p-value combination approaches and two weighting schemes for the diabetes data.

Weighting scheme	**P-value combination approach**	**Method**	**Number of significant genes (percentage)**
Binomial	Fisher	proposed	193 (2.9%)
		WOP	1131 (17.0%)
	Stouffer	proposed	204 (3.1%)
		WOP	1089 (16.4%)
Half-Binomial	Fisher	proposed	195 (2.9%)
		WOP	941 (14.2%)
	Stouffer	proposed	200 (3.0%)
		WOP	933 (14.0%)

Table 10

Detail information on diabetes datasets.

Dataset	Sample Size	Description
1	28	Brown preadipocyte IRS knockout profiling
2	14	Comparison of Low Fat and High Fat Diet on Mice of Two Genetic Backgrounds (B6 vs. 129) - Fat
3	16	Comparison of Low Fat and High Fat Diet on Mice of Backgrounds (B6 vs. 129) - Liver
4	17	Comparison of Low Fat and High Fat Diet on Mice of Backgrounds (B6 vs. 129) - Skeletal Muscle
5	18	Isolated adipocytes from normal and fat insulin receptor knockout (FIRKO) mice sorted into small and large cells
6	6	Liver - ob/ob mice
7	9	Mouse skeletal muscle - controls, streptozotocin diabetes and insulin treated
8	12	Human pancreatic islets from normal and Type 2 diabetic subjects
9	21	Transcription profiling of wild type and PGC-1alpha KO liver and skeletal muscle
10	12	Effect of PGC-1alpha and PGC-1beta on gene expression in myocytes and hepatocytes
11	57	Control Insulin Receptor (IR) and IRS-1 Single and Double Heterozygous (DH) Knockouts - Comparison of Age (6 weeks vs 6 months) and Genetic Background (B6 vs. 129) - Epididymal White Fat
12	55	Control Insulin Receptor (IR) and IRS-1 Single and Double Heterozygous (DH) Knockouts - Comparison of Age (6 weeks vs 6 months) and Genetic Background (B6 vs. 129) - Liver
13	52	Control Insulin Receptor (IR) and IRS-1 Single and Double Heterozygous (DH) Knockouts - Comparison of Age (6 weeks vs 6 months) and Genetic Background (B6 vs. 129) - Skeletal Muscle
14	12	Effect of insulin infusion on skeletal muscle
15	44	Skeletal Muscle - Muscle Insulin Receptor Knockout and Control Mice - Control, Streptozotocin Diabetic and Insulin Treated
16	54	Human skeletal muscle - type 2 diabetes - Swedish males

All datasets except 8 and 16 are used in the diabetes meta-analysis.

Figure 14 shows the overlap between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and (b) Stouffer's p-value combination approaches for the diabetes study. For the Fisher's p-value combination approach, there are 159 genes which are identified by both methods with both weighting schemes (see Figure 14(a)). Similarly, for the Stouffer's p-value combination approach, 175 genes are identified by both methods with both weighting schemes (see Figure 14(b)).

We performed pathway analysis using the significant gene lists identified by our proposed method for all settings. Our proposed method did not identify any significant KEGG pathway at Benjamini-Hochberg adjusted p-value cutoff of 0.05. However, our proposed method identified some significant GO term biological processes (BP) related to diabetes including response to lipopolysaccharide [30], response to estradiol [31] and response to glucocorticoid [32].

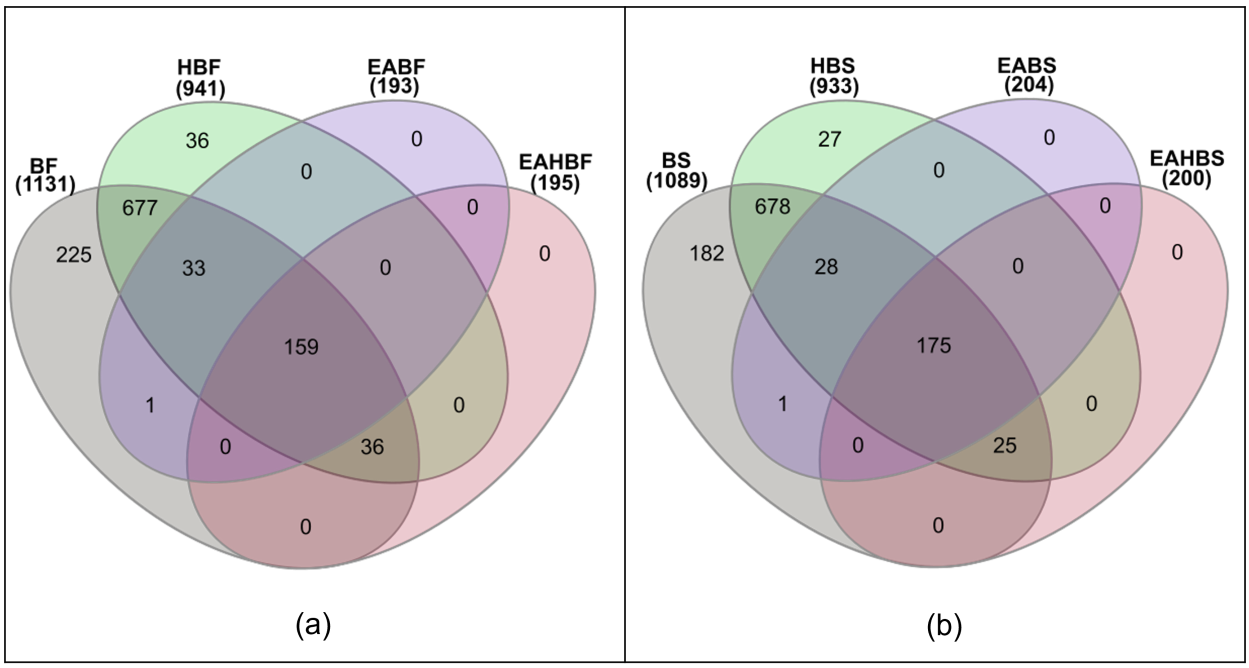

Fig. 14. Venn diagram showing the overlaps between the number of significant genes identified by the proposed method and the WOP method with both weighting schemes using the (a) Fisher's, and (b) Stouffer's p-value combination approaches for the diabetes study. BF: WOP method with binomial weighting scheme and Fisher's p-value combination approach; BS: WOP method with binomial weighting scheme and Stouffer's p-value combination approach; HBF: WOP method with half-binomial weighting scheme and Fisher's p-value combination approach; HBS: WOP method with half-binomial weighting scheme and Stouffer's p-value combination approach; EABF: proposed method with binomial weighting scheme and Fisher's p-value combination approach; EABS: proposed method with binomial weighting scheme and Stouffer's p-value combination approach; EAHBF: proposed method with half-binomial weighting scheme and Fisher's p-value combination approach; EAHBS: proposed method with half-binomial weighting scheme and Stouffer's p-value combination approach.

2.4 DISCUSSION

Meta-analysis is a popular method of integrating summary results from different sources to test a particular hypothesis. In this chapter, we mainly discussed the importance of developing new meta-analysis methods that focus on testing significance in a majority of studies in a consistent pattern. The WOP method is one of the important approaches that aims to identify genes significant in a majority of studies. However, the existing WOP method relies on some theoretical null distribution which can lead to incorrect statistical inference especially in presence of unobserved covariate effects. As a remedy to this problem, we have proposed a new meta-analysis method that empirically estimates the null distribution of the test statistic before combining them across the studies. Like the WOP method, our proposed meta-analysis method tests the same hypothesis with an aim to detect genes significant consistently in a majority of studies.

Several simulation studies are carried out under different scenarios to evaluate the performance of our proposed method. We considered sensitivity, specificity, false discovery rate and type I error rate as the performance measures. While generating the gene expression datasets for each study, we considered the scenarios that assumes no effect of any hidden variable or confounder, presence of a hidden variable which acts as a confounder as well as a scenario that assumes presence of a hidden variable that does not act as a confounder. We also introduced correlations among some of the genes through the generation of the error terms in the model.

The proposed method has shown significantly better performances than the WOP method especially in the presence of hidden confounder in the studies, a scenario very common in genomic studies. The type I error rates are controlled at 5% for our method while they are extremely high for the WOP method. The FDR values are also significantly lower for our proposed method compared to the WOP method. Our method has slightly lower sensitivity values but much higher specificity values compared to the WOP method. For the scenarios, that assumes no effect of any hidden variable or confounder and presence of a hidden variable which does not act as a confounder, the results showed that there is no significant difference between the WOP method and the proposed method. We showed the application of our method on three micro-array data sets: Lung cancer study, Brain cancer study and Diabetes study. For those real life data, we identified the genes that are differentially expressed in at least half of the experiments using the Benjamini-Hochberg adjusted p-value cutoff of

0.05. Further, we performed functional annotation clustering using DAVID with the genes that are identified as significant by our proposed method in order to identify the significant pathways. Future research can be done with different choices of the p-value combination approach and weighting scheme. Additionally, we can further extend our proposed method with hypothesis setting HS_m for choices of m ranging from $\lceil K/2 \rceil + 1$ to K.

CHAPTER 3

META-ANALYSIS METHOD COMBINING EFFECT SIZE ESTIMATES

3.1 INTRODUCTION

As discussed before in the introduction, there are two main approaches for the classical meta-analysis. The first approach combines p-values either using weights or without them, and the second approach combines effect size estimates in either fixed or random effects models. Both approaches have their own set of benefits and drawbacks as discussed in section 1.2. In this chapter, we focused on meta-analyzing multiple studies using effect size estimates to identify genes that are differentially expressed in at least one study.

Combining effect size estimates from studies that address the same research question is the primary alternative to combining p-values and/or z-scores. This can be accomplished using either a fixed effects model or a random effects model. A fixed-effects model assumes that there is a single underlying overall effect, whereas a random-effects model accounts for the possibility that each study represents its own population effect [33]. In this project we will focus on meta-analysis based on the fixed-effect model. Inverse-variance weighting is frequently utilized for the fixed effects model where the effect size estimates of the genes from each study are weighted by the inverse of their variances. Statistically, combining effect size estimates is more powerful than combining p-values or z-scores [13]. However, it necessitates that the outcome be assessed in each study using the exact same scale, unit, transformation, etc. For a outcome with well standardized measurements, this may be doable in a meta-analysis.

There are several software packages available that implement meta-analysis techniques for Genome-wide Association Studies (GWAS) and are discussed in [13]. METAL [14] is the most widely used software for meta-analysis. METAL employs two strategies: an effect-size based strategy that is weighted by the study-specific standard error and a weighted z-score

method based on sample size, p-value, and direction of effect in each study. Here we focus on the METAL method that is based on effect size estimates and their standard errors from independent studies. This method assumes that the overall z-score for a gene is standard normally distributed under the null hypothesis that the gene is not differentially expressed in any study. However, this assumption may be violated in large scale multiple testing problems [16]. It has been observed that even after using common multiple testing correction procedures, this assumption can still result in high number of false discoveries of significant genes [17]. To reduce the false discovery rate, it is critical to develop large-scale hypothesis testing meta-analysis methods based on empirically adjusted null distributions rather than relying on some theoretically assumed null distributions. In this chapter we proposed a new meta-analysis method that modifies the overall z-scores from METAL method by estimating the empirical null distribution parameters. For empirical estimation of the null distribution of z-scores, we considered the previously described Empirical Bayes approach [16].

We demonstrated, using a variety of simulated scenarios, that our proposed meta-analysis method with empirical adjustment has robust performance and outperforms the METAL method in terms of reducing false discoveries, particularly in the presence of hidden confounder variables. Furthermore, we showed the utility of the proposed meta-analysis approach by conducting a meta-analysis of lung cancer genomic studies. The rest of the chapter is organized as follows. In section 3.2, we first discuss our hypothesis setting for the meta-analysis. Then we provide a detailed description of our proposed meta-analysis method. Next, we briefly describe the methods for the empirical estimation of the null distribution of the z-scores. In section 3.3, we present various simulation results and the application of our method on lung cancer data. Our results clearly demonstrate the utility of our proposed method especially in presence of hidden confounder. We end this chapter with a discussion in section 3.4.

3.2 METHODS

3.2.1 HYPOTHESIS SETTING AND METAL META-ANALYSIS METHOD

Suppose there are K independent studies and G genes in each study. The goal of our meta-analysis method is to detect genes that are differentially expressed in at least one study. Therefore, we want to test the hypothesis that the true effect size of a gene is non-zero in at least one study against the null hypothesis that it is zero in all studies. That is, for a gene i, $i = 1, 2 \ldots, G$, the hypothesis setting for our meta-analysis method is

$$H_0 \colon \sum_{j=1}^{K} I(\theta_{ij} \neq 0) = 0 \, vs \, H_1 \colon \sum_{j=1}^{K} I(\theta_{ij} \neq 0) \geq 1$$

where θ_{ij} denotes the underlying true effect size for the i^{th} gene in the j^{th} study.

The widely used meta-analysis method called METAL [14], tests the above mentioned hypothesis to identify genes significantly differentially expressed in at least one study. For each gene, this method combines the estimates of the true effect size from multiple independent studies by weighting them with the inverse of their variances, under the assumption that the effect size estimates and their variances have same units across the studies. Suppose $\hat{\theta}_{ij}$ and SE_{ij} denote the effect size estimate and its standard error for gene i from study j, respectively, $i = 1, 2, \ldots, G \colon j = 1, 2, .., K$.

The effect size estimates, for each gene, are then combined across the studies in a weighted sum where the weights are proportional to the inverse of their variances. That is, for a gene i, the overall effect size estimate is defined as

$$\hat{\theta}_i = \frac{\sum_{j=1}^{K} \hat{\theta}_{ij} w_{ij}}{\sum_{j=1}^{K} w_{ij}}$$

where $w_{ij} = \frac{1}{SE_{ij}^2}, i = 1, 2, \ldots, G \colon j = 1, 2, .., K$. Then, for gene i, the overall z-score can be calculated as

$$z_i = \frac{\hat{\theta}_i}{SE_i}$$

where $SE_i = \sqrt{\frac{1}{\sum_{j=1}^{K} w_{ij}}}, i = 1, 2, \ldots, G\colon j = 1, 2, .., K$. Finally, an overall p-value for gene i, $i = 1, 2, \ldots, G$, is obtained as

$$p_i = 2(1 - \Phi(|z_i|))$$

3.2.2 PROPOSED META-ANALYSIS METHOD

In this section, we describe our proposed empirically adjusted meta-analysis method that modifies the overall z-scores from METAL meta-analysis method and computes multiple testing corrected p-values. This modification involves transforming the overall z-scores from METAL through an empirical correction of their null distribution. Following are the detailed steps of our proposed meta-analysis method.

- **Step 1**: Let $\hat{\theta}_{ij}$ and SE_{ij} denote the effect size estimate and its standard error for gene i from study j, respectively, $i = 1, 2, \ldots, G\colon j = 1, 2, .., K$. For gene i, we obtain the overall effect size estimate $\hat{\theta}_i$ as

  $$\hat{\theta}_i = \frac{\sum_{j=1}^{K} \hat{\theta}_{ij} w_{ij}}{\sum_{j=1}^{K} w_{ij}}$$

  where $w_{ij} = \frac{1}{SE_{ij}^2}, i = 1, 2, \ldots, G\colon j = 1, 2, .., K$.

- **Step 2**: For gene i, we obtain the overall z-score z_i as

  $$z_i = \frac{\hat{\theta}_i}{SE_i}$$

  where $SE_i = \sqrt{\frac{1}{\sum_{j=1}^{K} w_{ij}}}, i = 1, 2, \ldots, G\colon j = 1, 2, .., K$.

- **Step 3**: Since, these z-scores in step 2 may not follow $N(0, 1)$ under the null hypotheses, we empirically estimate the parameters of the null distribution of the z-scores. We consider the Empirical Bayes method [16], previously described, for estimating the parameters of the null distribution. Suppose $\hat{\mu}_{EB}$ and $\hat{\sigma}_{EB}$ are the Empirical Bayes method estimated mean and standard deviation of the null distribution of the z-scores. We modify the z-scores in step 2, using the estimated parameters as

  $$z_i^{EB} = \frac{z_i - \hat{\mu}_{EB}}{\hat{\sigma}_{EB}}, \qquad i = 1, 2, \ldots, G$$

These modified z-scores are expected to follow a standard normal distribution under the null hypotheses.

- **Step 4**: For gene i, we obtain the overall p-value as

$$p_i^{EB} = 2(1 - \Phi(|z_i^{EB}|)), \qquad i = 1, 2, \ldots, G$$

The final p-values are then corrected for multiple testing using the Benjamini-Hochberg (BH) method [15].

After obtaining the list of differentially expressed genes using Benjamini-Hochberg p-value cutoff of 0.05, the performance of the proposed method is assessed using the following three performance measures:

- Sensitivity: Proportion of genes which are correctly identified as differentially expressed out of all the genes which are truly differentially expressed.
- Specificity: Proportion of genes which are correctly identified as not differentially expressed out of all the genes which are truly not differentially expressed.
- FDR: Proportion of genes which are incorrectly identified as differentially expressed out of all the genes which are identified as differentially expressed.

The Empirical Bayes method for estimating the null distribution of the z-scores has been discussed before in section 2.2.3. The proposed method with Empirical Bayes adjustment to the z-scores will be referred to as Empirical Bayes adjusted method from now on. There is an alternative method of estimating the null distribution of the z-scores called BACON [34]. BACON considers a Bayesian framework for estimation and was previously introduced in the fixed effects meta-analysis in genomic studies [34]. We will compare the performance of our proposed (Empirical Bayes adjusted) method with the BACON adjusted method in our simulation studies and data applications. Estimation of the null distribution using Bayesian method (BACON) is briefly discussed below.

Bayesian method (BACON) for estimation of the null distribution parameters

Suppose the observed set of z-scores corresponding to G genes in a study are denoted as z_i, $i = 1, 2 \ldots, G$. The null distribution of the z-scores is supposed to be $N(0, 1)$ theoretically. However, in large-scale testing situations empirical and theoretical null might differ.

However, the large-scale multiple testing situation enables us to estimate the parameters of the null distribution of the z-scores. In this section, we will discuss a Bayesian method, known as BACON [34], proposed by van Iterson et al., for estimating the null distribution empirically.

BACON assumes that the observed set of z-scores can be modeled by a three-component normal mixture:

$$f(z\colon \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{k=1}^{3} p_k \ \phi(z\colon \mu_k, \sigma_k)$$

where the mixture proportions are constrained to sum to one, i.e., $\sum_{k=1}^{3} p_k = 1$. Here, $\phi(z\colon \mu_k, \sigma_k)$ represents the density of $N(\mu_k, \sigma_k^2)$, $k = 1, 2, 3$. Moreover, one of the components represents the empirical null distribution and other two components represent two separate non null distributions. This method applies Gibbs sampling algorithm to estimate the parameters of the mixture distribution. Conjugate prior distributions are assumed for means μ_k, variances σ_k^2, and proportions p_k as follows:

$$\mu_k|\sigma_k^2 \sim N\left(\lambda_k, \frac{\sigma_k^2}{\tau_k}\right)$$
$$\sigma_k^2 \sim Inverse\ Gamma(\alpha_k, \beta_k)$$
$$(p_1, p_2, p_3) \sim Dirichelet(\gamma_1, \gamma_2, \gamma_3)$$

We considered the same choices for hyper-priors $\lambda_k, \tau_k, \alpha_k, \beta_k, \gamma_k$, as suggested by van Iterson et al. [34]. Moreover, data-driven starting values are used to start the Gibbs sampling algorithm based on the median and median absolute deviation of the z-scores.

Given the z-scores z_i for $i = 1, 2 \ldots, G$, the Gibbs sampling algorithm iterate in the following steps:

(i) Generate the unobserved data x_{ik} from the multinomial distribution as follows:

$$x_{ik} \sim Multinomial(\tilde{\omega}_{ik})$$

where $\omega_{ik} = p_k \phi(z\colon \mu_k, \sigma_k)$ and $\tilde{\omega}_{ik}$ represents the normalized proportion such that $\sum_{k=1}^{3} \tilde{\omega}_{ik} = 1$

(ii) Calculate the following quantities:

$$\eta_k = \sum_{i=1}^{G} 1_{(x_{ik} \neq 0)}$$

$$s_k = \sum_{i=1}^{G} z_i 1_{(x_{ik} \neq 0)}$$

$$s_k^2 = \sum_{i=1}^{G} z_i^2 1_{(x_{ik} \neq 0)}$$

(iii) Generate samples from the posterior distributions as follows:

$$\mu_k | \sigma_k^2 \sim N\left(\frac{\lambda_k \tau_k + s_k}{\eta_k + \tau_k}, \frac{\sigma_k^2 + s_k}{\eta_k + \tau_k}\right)$$

$$\sigma_k^{-2} \sim Gamma\left(\alpha_k + \frac{1}{2}(\eta_k + 1), (\beta_k + \frac{1}{2}\tau_k(\mu_k - \lambda_k)^2 + \frac{1}{2}s_k^2)^{-1}\right)$$

$$p_k \sim Dirichelet(\gamma_k + \eta_k)$$

A burn-in period of 2000 iterations is considered.

3.3 RESULTS

3.3.1 SIMULATION STUDIES

We conducted simulation studies to evaluate the performance of our proposed method based on sensitivity, specificity and FDR. We simulated continuous gene expression datasets for multiple independent studies. Details of the data generation process are given below.

We considered 10 independent studies each involving continuous gene expression levels for 10000 genes, i.e., $K = 10$ and $G = 10000$. We considered two groups of subjects in each study where each group consists of 40 subjects, i.e., $n_1 = n_2 = 40$. We considered 1000 genes (i.e., 10%) as differentially expressed between the two subject groups in all 10 studies.

The (log) expression level for the i^{th} gene, l^{th} subject in the k^{th} group for each study is generated separately using the following model:

$$y_{ikl} = \mu + G_i + V_k + GV_{ik} + W_{ikl} + e_{ikl}$$

Here μ denotes the overall mean effect, G_i denotes the effect due to the i^{th} gene, V_k denotes the effect due to the k^{th} subject group, and GV_{ik} denotes the interaction effect between the i^{th} gene and the k^{th} subject group, W_{ikl} denotes the effect of a hidden variable and e_{ikl} denotes the error component corresponding to the i^{th} gene, l^{th} subject in the k^{th} group, $i = 1, 2 \ldots, G$: $k = 1, 2$: $l = 1, 2, \ldots, n_k$.

We considered μ, G_i, and V_k as zero for all i,k and l, for simplicity. The differences in magnitudes of (log) expression values of differentially expressed genes between the two groups are considered as four, which are obtained through the generation of the interaction terms between the genes and the groups, $(GV)_{ik}$s, as follows:

$(GV)_{i1} = -2, (GV)_{i2} = 2$ for $i = 1, \ldots, 500$
$(GV)_{i1} = 2, (GV)_{i2} = -2$ for $i = 501, \ldots, 1000$
$(GV)_{i1} = (GV)_{i2} = 0$ for $i = 1001, \ldots, 10000$

We introduced correlations among some of the genes through the generation of the error terms e_{ikl}s as described below:

We considered four groups of correlated genes given by $C_1 = \{1, 2, \ldots, 1000\}$, $C_2 = \{2001, 2002, \ldots, 3000\}$, $C_3 = \{6001, 6002, \ldots, 7000\}$ and $C_4 = \{8001, 8002, \ldots, 9000\}$. The error term e_{ikl} for the i^{th} gene, l^{th} subject in the k^{th} group , $i = 1, 2, \ldots, G$: $k = 1, 2$: $l = 1, 2, \ldots, n_k$ is generated as:

$$e_{ikl} = \begin{cases} \frac{1}{\sqrt{2}} e^1_{ikl} + \frac{1}{\sqrt{2}} e^2_{ikl}, & \text{if } i \in \{C_1, C_2, C_3, C_4\}. \\ e^2_{ikl}, & \text{o.w.} \end{cases}$$

where e^1 are generated independently from $N(0, 1)$ in such a way that the values of e^1 are same for all the genes belonging to the same group, and e^2 are generated independently from $N(0, 2^2)$.

In our simulations, we considered the following three settings.

Setting I

In this simulation setting, we assumed the presence of a hidden variable which acts as a confounder. The effect of the hidden confounder for the i^{th} gene, l^{th} subject in the k^{th} group is generated such that it varied over the two subject groups, different groups of genes as well as over different studies. We considered $W_{ikl} = u_{ikl} I(s_{ikl} = 1)$, where $s_{ikl} \sim Bernoulli(0.4)$ and u_{ikl} are generated depending on the gene, subject group and the study ID j as given below:

$$u_{ikl} = \begin{cases} N(-1 + j + \delta.I, 0.01^2), & \text{for } i = 1, \ldots, 500 \text{: } l = 1, \ldots, n_k. \\ N(2 + j + \delta.I, 0.01^2), & \text{for } i = 501, \ldots, 1000 \text{: } l = 1, \ldots, n_k. \\ N(5 + j + \delta.I, 0.01^2), & \text{for } i = 1001, \ldots, G \text{: } l = 1, \ldots, n_k. \end{cases}$$

where $i = 1, 2 \ldots, G$: $k = 1, 2$: $j = 1, 2, \ldots, K$. Here, $I = 1$ for group 1 (i.e., for $k = 1$) and $I = 0$ for group 2 (i.e., for $k = 2$). The magnitude of the difference between the means of the distributions of u_{ikl} between the two subject groups is given by δ. In our simulations, we considered $\delta = 2$.

For each study, after generating the (log) gene expression values for all the subjects, we tested whether the genes are differentially expressed between the two subject groups using "limma" in Bioconductor [26] and obtained the effect size estimates and their standard

errors. We then applied our proposed method (Empirical Bayes adjusted method), as described in the methods section, and obtained the list of differentially expressed genes with a Benjamini-Hochberg adjusted p-value cutoff of 0.05. We evaluated the performance of the proposed method using sensitivity, specificity, and false discovery rate (FDR) based on 500 monte-carlo iterations. We compared the performance of our proposed method with the BACON adjusted method as well as the METAL method (without any empirical adjustment) [14].

Table 11 summarizes the simulation results for our proposed method (Empirical Bayes adjusted method), BACON adjusted method and the METAL method without any empirical adjustment [14]. Both our proposed method and the BACON adjusted method have significantly lower FDR values compared to METAL which has extremely high FDR value. The sensitivity values are very similar for the all three methods whereas the specificity values are much higher for both the proposed method and the BACON adjusted method compared to METAL. In general, both Empirical Bayes and BACON adjusted methods have much better performances than METAL.

Table 11

Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden confounder. The proportion of differentially expressed genes between the two subject groups is 10%.

Method	**Sensitivity**	**Specificity**	**FDR**
METAL	1.000	0.597	0.784
Empirical Bayes adjusted	1.000	0.992	0.070
BACON adjusted	1.000	0.999	0.011

We, additionally, varied the proportion of differentially expressed genes between the

two subject groups, ranging from 5% to 20%. Figure 15 shows the performances of the proposed method, the BACON adjusted method, as well as the METAL method with varying proportion of differentially expressed genes in presence of hidden confounder in the studies. Although the FDR values of the METAL method decreased with increase in the proportion of differentially expressed genes in the studies, they remained unacceptably high in all scenarios. Both our proposed method and the BACON adjusted method had significantly lower FDR values compared to METAL. The sensitivity values are very similar for all the methods. Both the proposed method and the BACON adjusted method had significantly higher specificity values compared to METAL in all scenarios.

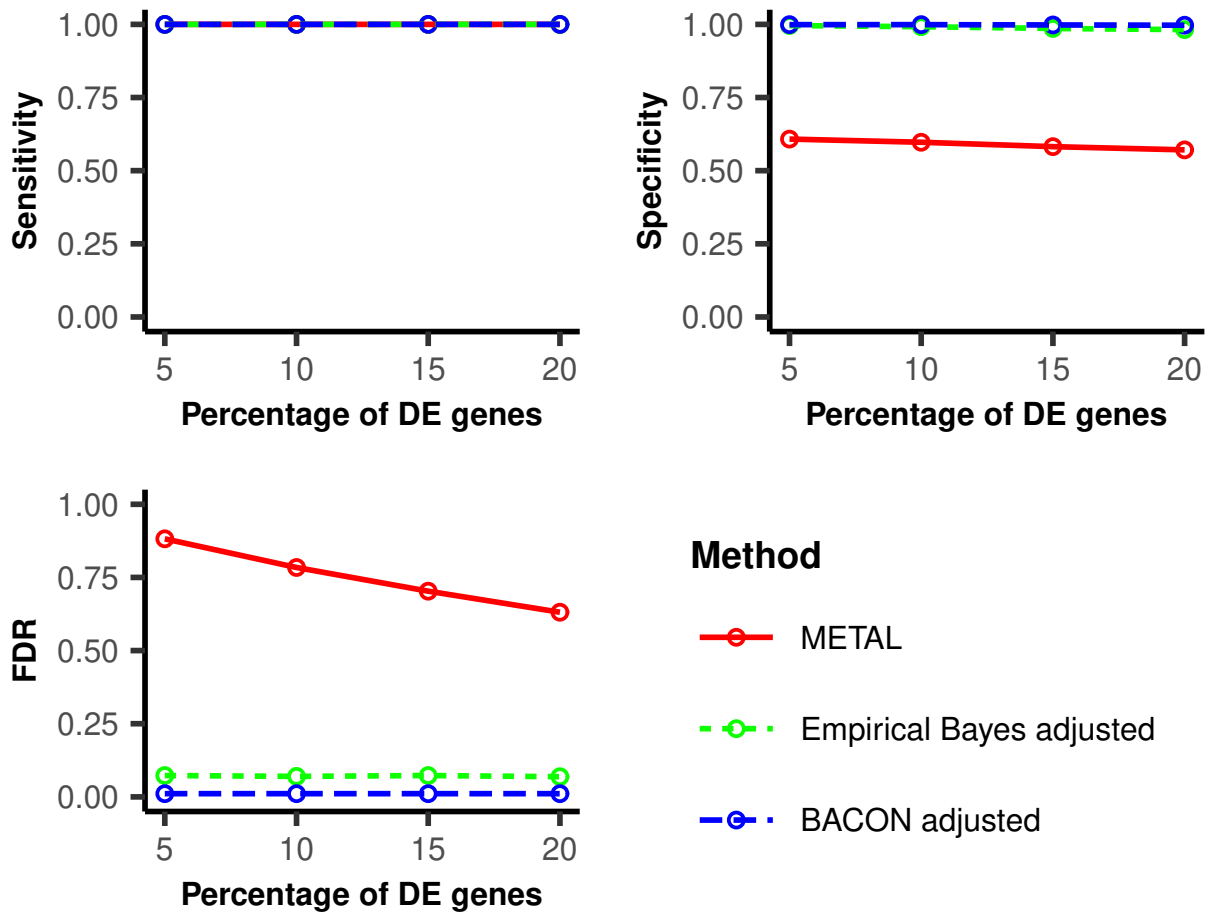

Fig. 15. Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden confounder with varying proportion of differentially expressed genes between two subject groups.

Setting II

In this simulation scenario, we assumed the presence of the hidden variable which affects the outcome but does not vary between the two subject groups. We generated this hidden variable which does not act as a confounder for the i^{th} gene, l^{th} subject in the k^{th} group, as $W_{ikl} = u_{ikl}I(s_{ikl} = 1)$, where $s_{ikl} \sim Bernoulli(0.4)$ and u_{ikl} are generated as given below:

$$u_{ikl} = N(2 + j, 0.1^2) \qquad \text{for } i = 1, \ldots, G; k = 1, 2; l = 1, \ldots, n_k; j = 1, \ldots, K$$

We considered 10% of the genes as differentially expressed between the two subject groups in all 10 studies as considered before. All the other terms in the model for simulation are generated in the same way as described previously.

Table 12 shows the results for our proposed method, BACON adjusted method, and the METAL method. In this simulation scenario, all the methods have controlled FDR values at 0.05. The sensitivity and specificity values are also very similar for all the methods.

Table 12

Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden variable that does not act as a confounder. The proportion of differentially expressed genes between the two subject groups is 10%.

Method	**Sensitivity**	**Specificity**	**FDR**
METAL	1.000	0.993	0.057
Empirical Bayes adjusted	1.000	0.994	0.049
BACON adjusted	1.000	0.994	0.051

The performances of the proposed method, BACON adjusted method, and METAL with varying proportion of differentially expressed genes, in the presence of hidden variable that does not act as confounder, are shown in Figure 16. The FDR values remained controlled at 0.05 for all methods consistently for all the considered proportions of significant genes. The sensitivity and specificity values are very similar for all methods.

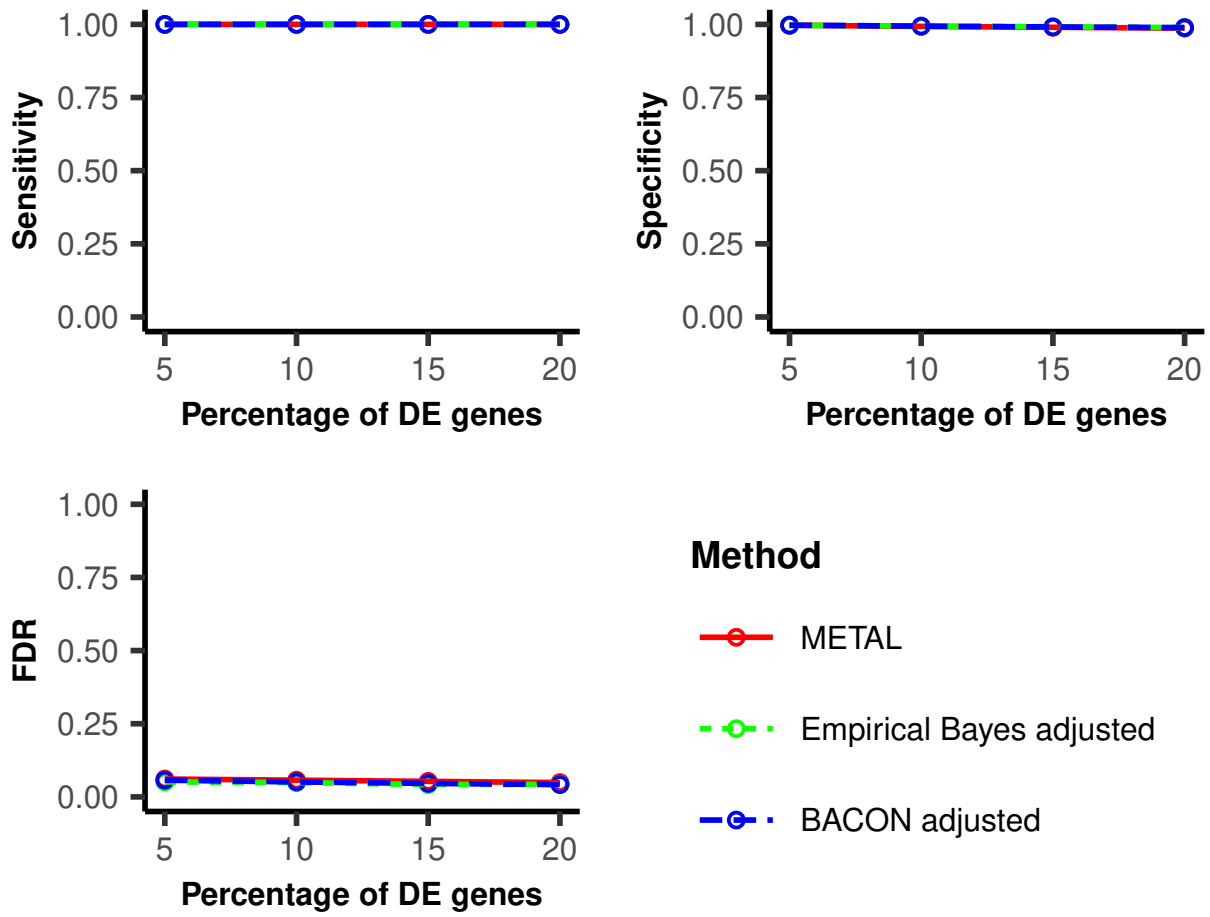

Fig. 16. Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden variable that does not act as a confounder with varying proportion of differentially expressed genes between two subject groups.

Setting III

In this simulation scenario, we assumed that there is no effect of any hidden variable or confounder in the studies. Therefore, we set $W_{ikl} = 0$, for all i, k and l. Here also, we considered 10% of the genes as differentially expressed between the two subject groups in all 10 studies. All the other terms in the model for simulation are generated in the same way as described previously.

Table 13 shows the results for our proposed method, BACON adjusted method, as well as the METAL method. Similar to setting II, all the methods have FDR values controlled at 0.05 with very similar sensitivity and specificity values.

Table 13

Performances of the proposed method, BACON adjusted method, and METAL when there is no effect of any hidden variable or confounder. The proportion of differentially expressed genes between the two subject groups is 10%.

Method	**Sensitivity**	**Specificity**	**FDR**
METAL	1.000	0.993	0.056
Empirical Bayes adjusted	1.000	0.995	0.045
BACON adjusted	1.000	0.995	0.047

Figure 17 shows the simulation results for the proposed method, BACON adjusted method, and METAL with varying proportion of differentially expressed genes when there is no effect of hidden variable or confounder in the studies. The performances of the methods are very similar to what we observed in the previous scenario in the presence of hidden variable that does not act as confounder.

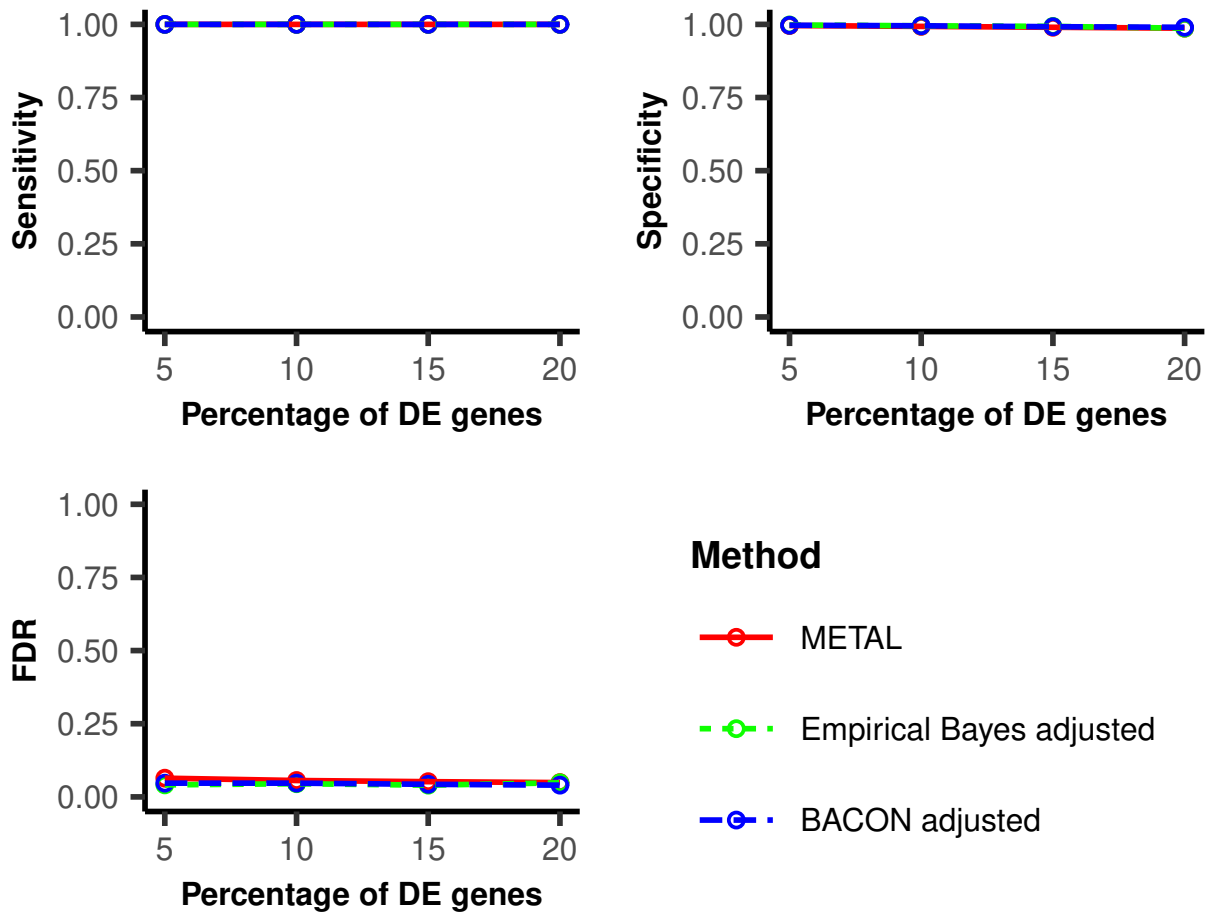

Fig. 17. Performances of the proposed method, BACON adjusted method, and METAL when there is no effect of any hidden variable or confounder with varying proportion of differentially expressed genes between two subject groups.

We also considered some variations in our simulations. In particular, we considered a simulation scenario where we evaluated the effect of changing the number of studies in the meta-analysis. We also considered simulation scenarios where we assumed that 10% of the genes were significant in lesser number of studies. In particular, we considered two choices for the number of studies, namely five and one, and evaluated the performances of our proposed method. Each of these scenarios is described below.

Effect of changing the number of studies

We compared the performance of our proposed method with BACON adjusted method, and METAL by varying the number of studies. In this simulation scenario, we assumed presence of a hidden variable which acts as a confounder. We considered K independent studies where $K = 5$, 10 and 15, each involving continuous gene expression levels for $G = 10000$ genes. Note that, we considered 10% of the genes as differentially expressed between two subject groups in all K studies. All the terms in the model for simulation are generated in the same way as described in setting I.

Figure 18 summarizes the simulation results for our proposed method, BACON adjusted method, and METAL method for varying number of studies (K) in presence of hidden variable which acts as a confounder. Both our proposed method and BACON adjusted method have significantly lower FDR values compared to the METAL method for all the choices of K. METAL has extremely high FDR values for all the choices of K and the values increase as the number of studies increases. The sensitivity values are very similar for all the methods while the specificity values of METAL remained much lower than the other two methods. The specificity values of METAL decreased with the increase in the number of studies.

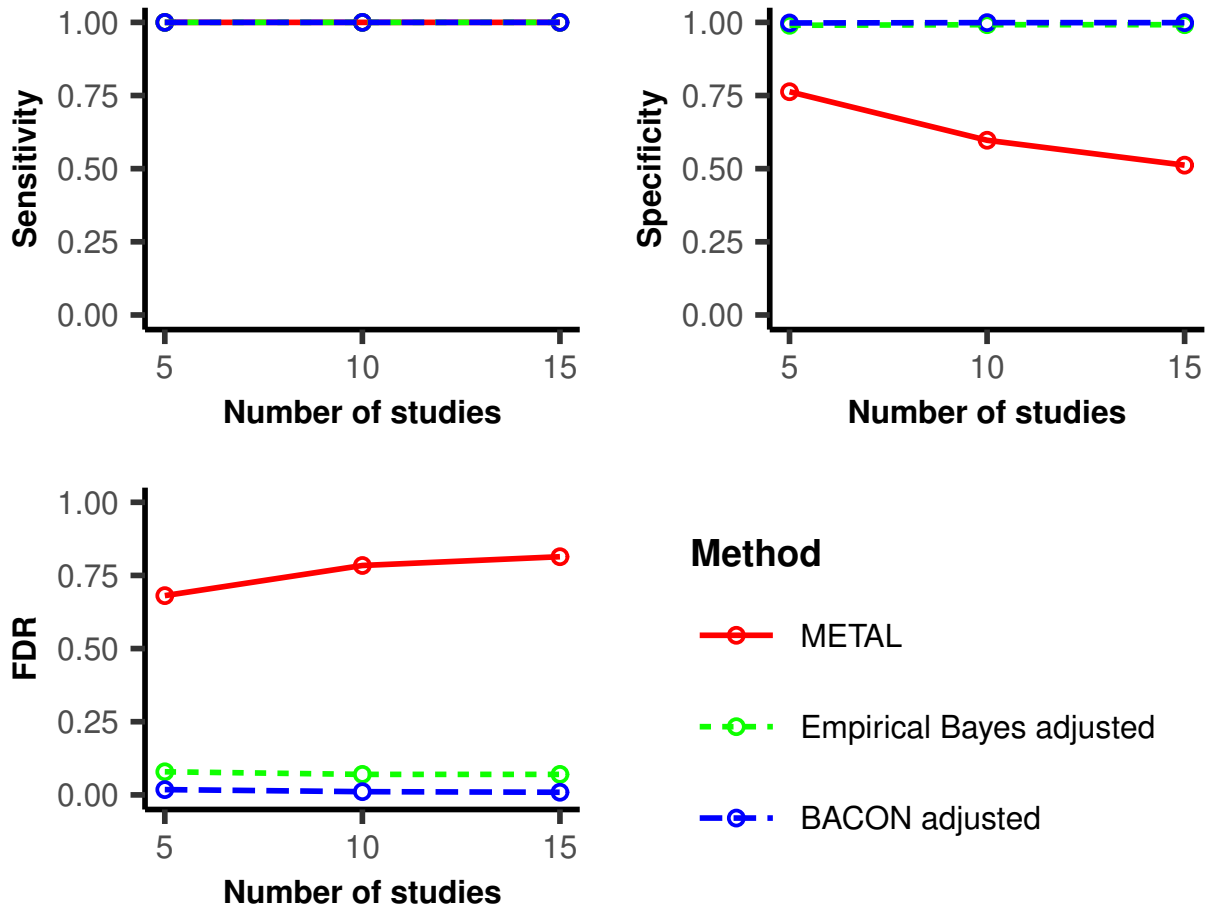

Fig. 18. Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden confounder with varying number of studies.

Further, we evaluated the performances of the proposed method, BACON adjusted method, and METAL with varying number of studies, in the presence of hidden variable that does not act as confounder. As shown in Figure 19, the FDR values remained controlled at 0.05 for all methods consistently for all the considered number of studies. The sensitivity and specificity values are very similar for all methods.

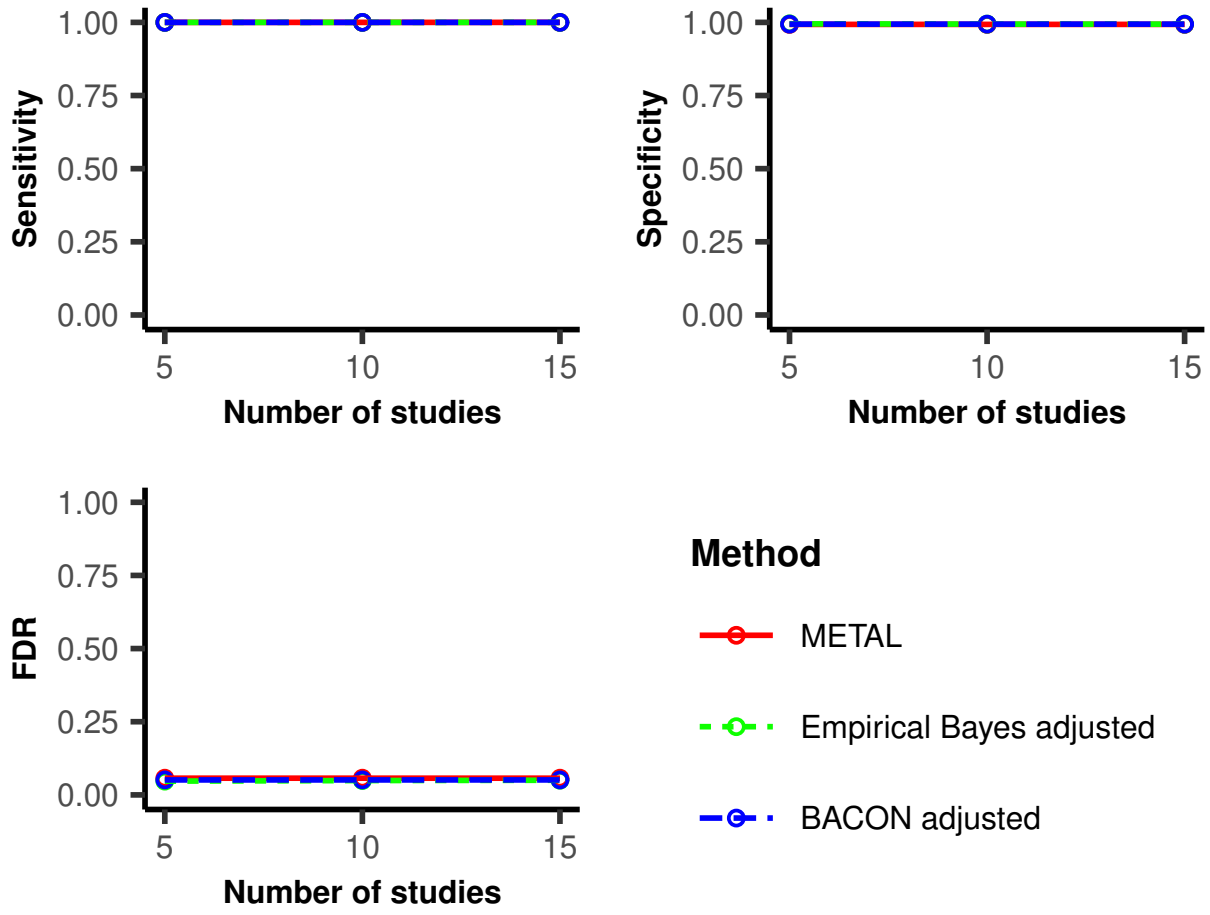

Fig. 19. Performances of the proposed method, BACON adjusted method, and METAL in presence of hidden variable that does not act as a confounder with varying number of studies.

Figure 20 shows the simulation results for the proposed method, BACON adjusted method, and METAL with varying number of studies when there is no effect of hidden variable or confounder in the studies. The performances of the methods are very similar to what we observed in the previous scenario in the presence of hidden variable that does not act as confounder.

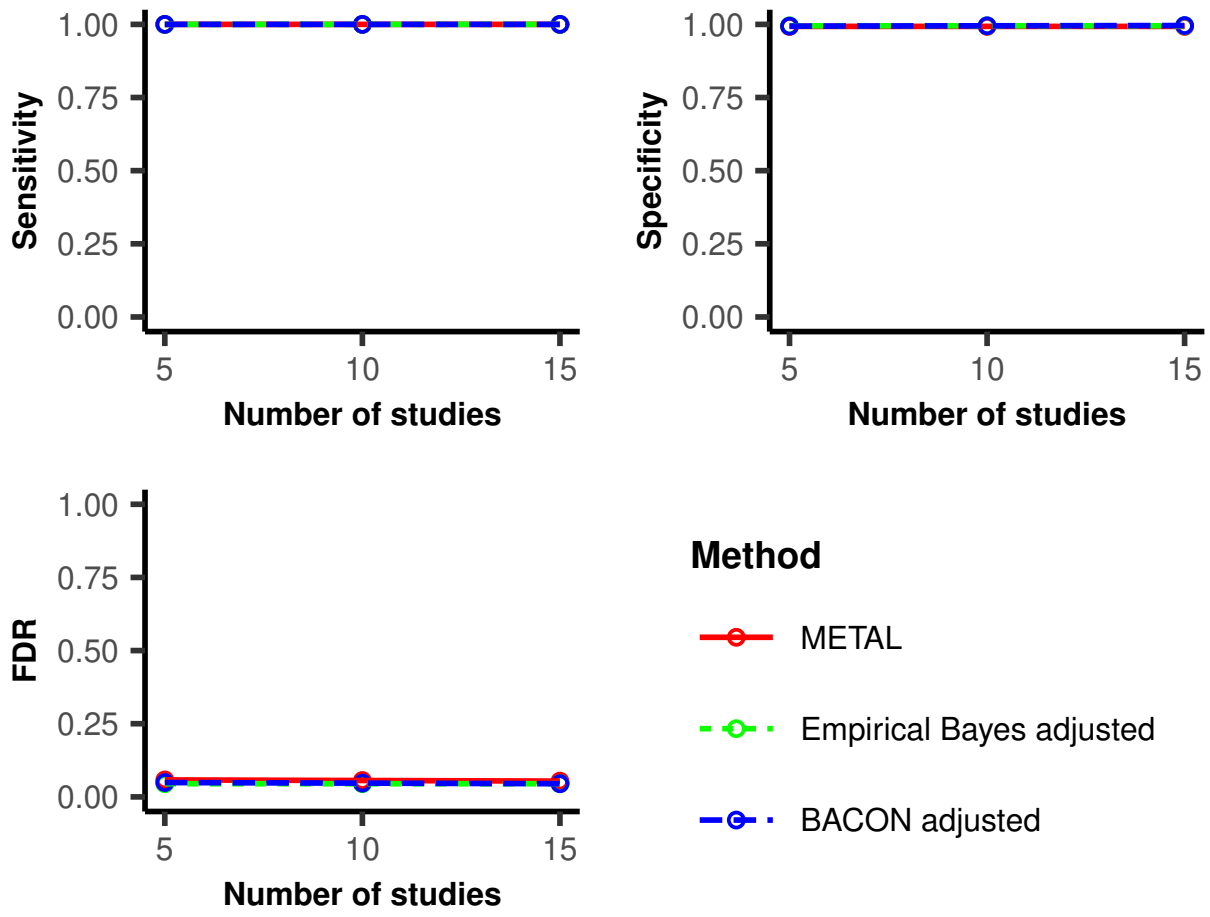

Fig. 20. Performances of the proposed method, BACON adjusted method, and METAL when there is no effect of any hidden variable or confounder with varying number of studies.

Effect of reducing the number of studies for the significant genes

In this scenario, we considered 10 independent studies each involving continuous gene expression levels for 10000 genes, i.e., $K = 10$ and $G = 10000$. We considered two groups of subjects in each study where each group consists of 40 subjects, i.e., $n_1 = n_2 = 40$. We considered 1000 genes as differentially expressed between the two subject groups but only in five of the ten studies. The results for the proposed method, BACON adjusted method, and METAL for all three simulation settings i.e., I, II and III are summarized in Table 14.

In setting I, i.e. in the presence of a hidden variable which acts as a confounder, METAL had extremely high FDR value, while it remained controlled for the other two methods. All three methods had very similar sensitivity values while METAL had much lower specificity values compared to the other two methods.

In setting II (in the presence of the hidden variable that does not act as a confounder) and setting III (no effect of any hidden variable or confounder), all three methods performed similarly, with very high sensitivity and specificity values and FDR values controlled at 5%.

Overall, the performances of the methods are very similar to what we observed in the previous scenario that considered 10% of the genes as differentially expressed in all 10 studies.

Table 14

Performances of the proposed method, BACON adjusted method, and METAL when 10% of the genes are significant in five studies for all three settings.

Setting	**Method**	**Sensitivity**	**Specificity**	**FDR**
Setting I	METAL	0.998	0.462	0.829
	Empirical Bayes adjusted	1.000	0.992	0.063
	BACON adjusted	1.000	0.999	0.009
Setting II	METAL	1.000	0.993	0.056
	Empirical Bayes adjusted	1.000	0.994	0.049
	BACON adjusted	1.000	0.994	0.054
Setting III	METAL	1.000	0.994	0.055
	Empirical Bayes adjusted	1.000	0.996	0.036
	BACON adjusted	1.000	0.994	0.050

We further reduced the number of studies for the differentially expressed genes. In particular, we assumed that 10% of the G=10000 genes are differentially expressed in only one of the ten studies. That is, we considered 1000 genes as differentially expressed between the two subject groups in only one study. As before, we considered 40 subjects in each group. The results for the proposed method, BACON adjusted method, and METAL for simulation settings I, II and III are summarized in Table 15.

In setting I, i.e. in the presence of a hidden variable which acts as a confounder, the METAL method has extremely high FDR value while both the proposed method and BACON adjusted method have much lower FDR values. Although, the sensitivity values of all the three methods are low, the Empirical Bayes adjusted proposed method has slightly higher value compared to others. METAL also has significantly lower specificity value compared to the other methods. Overall, the Empirical Bayes adjusted proposed method performs better than the other two methods in this scenario with higher sensitivity and specificity values and FDR controlled at 5%. The BACON adjusted method is a bit conservative with low sensitivity and FDR values compared to the Empirical Bayes adjusted proposed method.

In setting II (presence of the hidden variable that does not act as a confounder), all the three methods have similar performances, although METAL had slightly lower sensitivity value. In this setting, the FDR values of all three methods are slightly higher than 0.05. In setting III (no effect of any hidden variable/confounder), the performances of all the three methods are very similar. However, the Empirical Bayes adjusted proposed method has slightly higher sensitivity and FDR values compared to the other two methods.

Table 15

Performances of the proposed method, BACON adjusted method, and METAL when 10% of the genes are significant in only one study for all three settings.

Setting	Method	Sensitivity	Specificity	FDR
Setting I	METAL	0.519	0.540	0.889
	Empirical Bayes adjusted	0.619	0.996	0.050
	BACON adjusted	0.517	1.000	0.006
Setting II	METAL	0.799	0.995	0.057
	Empirical Bayes adjusted	0.815	0.992	0.077
	BACON adjusted	0.812	0.994	0.066
Setting III	METAL	0.727	0.995	0.056
	Empirical Bayes adjusted	0.758	0.992	0.080
	BACON adjusted	0.736	0.995	0.060

3.3.2 AN APPLICATION TO LUNG CANCER STUDIES

We conducted meta-analysis using our proposed method on five lung cancer gene expression datasets. Details about these five studies are discussed in section 2.3.2. We aimed to identify the genes that are differentially expressed between two lung cancer types - adenocarcinoma (AD) and squamous cell carcinoma (SQ) in at least one study. To obtain the effect size estimates and their standard errors for the genes for each dataset, we tested for differential expression between AD and SQ subjects using "limma" [26]. We then applied our proposed method, following the steps described in methods section, using Empirical Bayes adjustment. For comparison, we also applied the BACON adjusted method as well as the METAL method to identify the differentially expressed genes between the two lung cancer types in at least one study.

The Empirical Bayes approach estimated mean and the standard deviation of the overall z-scores are – 0.88 and 4.77, respectively, whereas those estimated from the BACON method are – 0.20 and 5.86, respectively. Figure 21 shows the histogram of the original overall z-scores for the genes obtained from METAL. We also superimposed the density curves of the

theoretical null distribution as well as the empirical null distributions estimated from the Empirical Bayes and BACON methods.

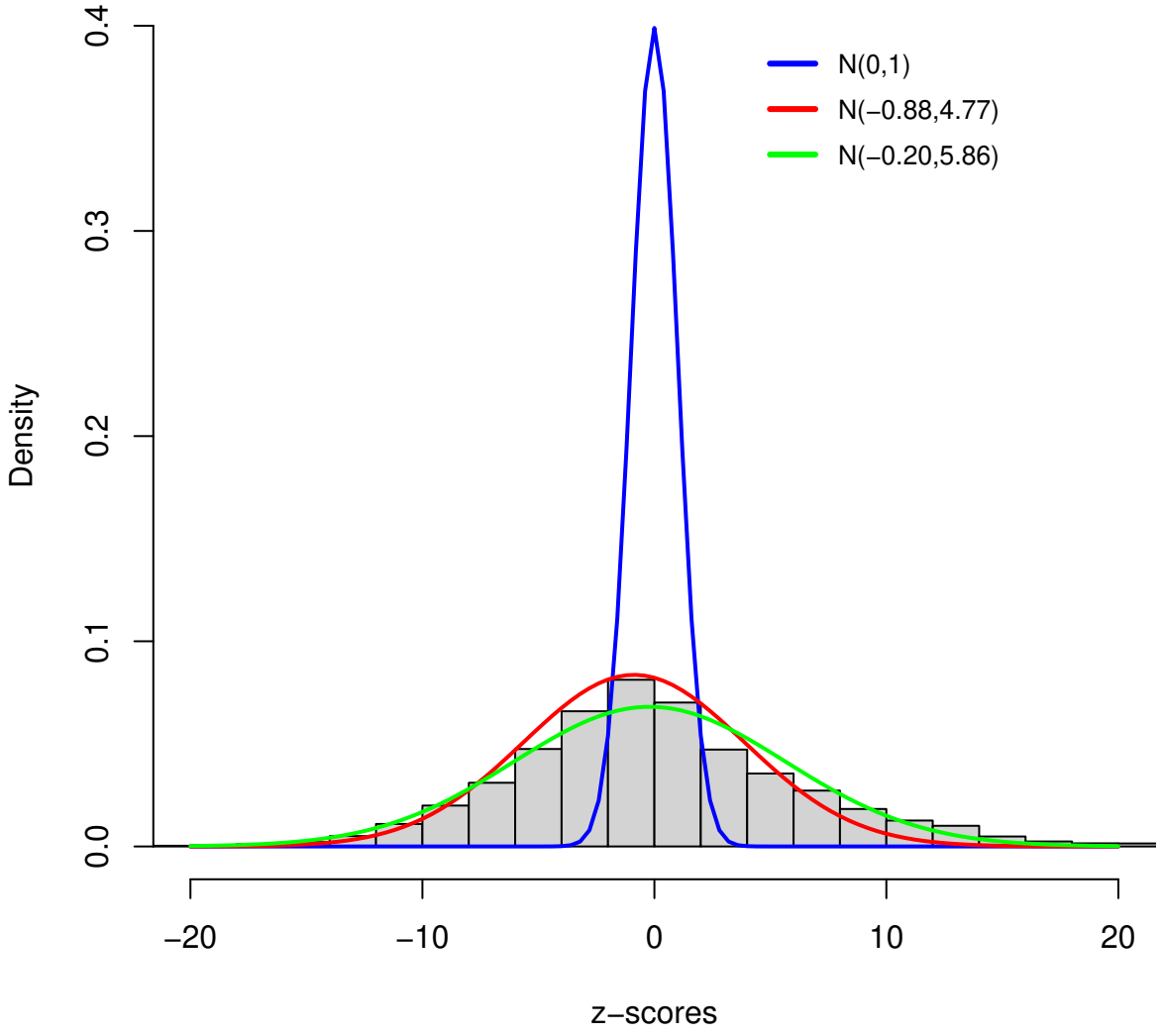

Fig. 21. Histogram of the overall z-scores from METAL along with the empirically estimated null distributions using Empirical Bayes method (red line) and BACON method (green line).

As evident from the figure, the theoretical null distribution i.e., $N(0,1)$ is much deviated from the histogram of the original z-scores. The empirically estimated null distributions using the Empirical Bayes and BACON methods are much closer to the histogram of the original z-scores. However, as evident from the figure, the Empirical Bayes estimated null distribution is more closer to the histogram compared to the BACON adjusted method.

Table 16 shows the number of differentially expressed genes, identified by our proposed method, BACON adjusted method, and METAL, significant at Benjamini-Hochberg adjusted p-value cutoff of 0.05. METAL identified significantly higher number of differentially expressed genes ($> 65\%$) compared to the other two methods, indicating a possibility of high false discoveries. The Empirical Bayes adjusted method identified about 3% of the genes as differentially expressed while the BACON adjusted method identified less than 1% of the genes as differentially expressed. Further, all the genes identified by the BACON adjusted method are also identified by the Empirical Bayes adjusted method. The BACON adjusted method might be a bit conservative with lower sensitivity and/or FDR values, as we observed in setting I of Table 15, where the genes are differentially expressed in a smaller number of studies.

Table 16

The number of significant genes (percentage) identified by our proposed method, BACON adjusted method, and METAL for the lung cancer data.

Method	**Number of significant genes (percentage)**
METAL	4901 (68.1%)
Empirical Bayes adjusted	205 (2.8%)
BACON adjusted	52 (0.7%)

In order to identify biological pathways associated with the gene lists identified by the our proposed method, we performed functional annotation clustering using Database for Annotation, Visualization and Integrated Discovery (DAVID) software [28]. Our proposed method with Empirical Bayes adjustment identified several biological processes related to lung cancer, including keratinocyte differentiation, epidermis development, intermediate filament organization, cell-cell adhesion, epithelial cell differentiation, keratinization, cell adhesion, negative regulation of endopeptidase activity, peptide cross-linking and hair follicle morphogenesis based on the Benjamini-Hochberg adjusted p-value cutoff of 0.05.

3.4 DISCUSSION

In this chapter, we proposed a new meta-analysis method that focus on testing for significance of a gene in at least one of the studies. The METAL method [14], based on effect size estimates and their standard errors from independent studies, is a widely used meta-analysis method which aims to test significance of a gene in at least one study. But METAL relies on a theoretical null distribution which can be different from the true null distribution of test statistics. Hence, we proposed a robust meta-analysis method that empirically modifies the overall z-scores from METAL method by estimating the parameters of the empirical null distribution. We considered an Empirical Bayes method [16], previously described in section 2.2.3, for empirical estimation of the null distribution of z-scores. For comparison, we considered an alternative approach for empirical estimation of the null distribution, called BACON [34] as described in section 3.2.2.

Simulation studies are carried out under different settings to evaluate the performance of our proposed method using sensitivity, specificity and false discovery rate as the performance measures. We mainly considered three different simulation settings that assumed no effect of any hidden variable or confounder in the studies, presence of a hidden variable which acts as a confounder and which does not act as a confounder while generating the gene expression values. Additionally, we considered some variations in our simulations. In particular, we considered simulation scenarios where we evaluated the effect of changing the number of studies in the meta-analysis. We also considered simulation scenarios where we assumed that the genes were significant in different number of studies.

For the simulation setting, that assumed presence of a hidden confounder variable, both

our proposed method and BACON adjusted method have shown significantly better performances than the METAL method. The FDR values are significantly lower for both the proposed method and BACON adjusted method compared to METAL in all the considered simulation scenarios. The sensitivity values are very similar for the all three methods whereas the specificity values are much higher for both the proposed method and BACON adjusted method compared to METAL in all the considered simulation scenarios. For the simulation settings, that assumed no effect of any hidden variable or confounder and presence of a hidden variable which does not act as a confounder, the results showed that there is no significant difference between the performances of METAL and the other two methods in all the considered simulation scenarios. In the simulation scenario, where the genes are differentially expressed in fewer number of studies, we observed differences in the performances of our proposed method and the BACON adjusted method. The BACON adjusted method is a bit conservative with lower sensitivity and FDR compared to the proposed method. We observed similar results in the lung cancer meta-analysis where the BACON adjusted method identified very few genes (0.7%) as significant. Our proposed method identified about 3% of the genes as significant which include all the genes identified by the BACON adjusted method. On the other hand, METAL identified extremely large number of genes as significant ($\sim 68\%$) indicating possibility of gross false discoveries.

CHAPTER 4

CONCLUSIONS AND FUTURE WORK

Recent advances in high-throughput technologies have made it possible to analyze thousands of genes at once. One can quickly obtain the summary results of several genomic studies, each of which includes the significance testing results of thousands of genes, thanks to the availability of numerous public databases. Meta-analysis is a popular method for combining the summary results of hypothesis testing from multiple studies. It is widely used in variety of scientific fields, including genomic research and psychological research. Traditional meta-analysis methods need to be modified when combining summary results of large-scale simultaneous hypotheses testing across multiple studies as they rely on a theoretical null distribution which may differ from the true null distribution especially in the presence of hidden confounder effects.

We discussed some recent advances in estimation of null distributions of test statistics empirically and proposed methods for incorporating such empirically estimated null distributions in meta-analysis of large scale genomic experiments in this dissertation. In Chapter 2, we focused on identifying genes that are differentially expressed in a majority of studies when p-value is the only source of information that is consistently available for all the genes in all of the studies taken into account for meta-analysis. We demonstrated that, in large-scale simultaneous testing of thousands of genes, the existing p-value combination methods of meta-analysis can experience increased type-I error rates and significant false discoveries, even after multiplicity corrections. As a remedy, we proposed a reliable meta-analysis technique that modifies each individual p-value empirically before combining them across the studies. We considered an Empirical Bayes method to estimate the parameters of the null distribution of the test statistics. Various simulated scenarios showed that our proposed empirically modified meta-analysis performed better than the existing method. In Chapter 3, we focused on identifying genes that are differentially expressed in at least one study. The METAL method combining effect size estimates and their standard errors from independent studies is one of the important approaches that test for significance of a gene in at least one study. We emphasized the shortcomings of this technique for significance testing in large scale hypothesis testing problems. We proposed a new meta-analysis method

that empirically modifies the overall z-scores from METAL by estimating the parameters of the empirical null distribution. We considered the Empirical Bayes approach for empirically estimating the null distribution of the z-scores. Particularly in the presence of hidden confounders, the proposed method (Empirical Bayes adjusted) has demonstrated significantly improved performances than the existing methods. We have shown the effectiveness of the proposed method in finding significant and biologically relevant genes using real genomic datasets.

Future research can be done to explore an alternative way to empirically estimate the null distribution of the p-values directly instead of relying on the empirical estimation of the null distribution of the test statistics. The idea is to fit a Beta distribution to the p-values for the genes and then convert the Beta distribution into Uniform distribution using the CDF of the particular Beta distribution.

REFERENCES

[1] R. A. Fisher, Statistical methods for research workers, Oliver & Boyd, Edinburgh, Scotland, 1925.

[2] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, R. M. Williams Jr, The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1 (1949).

[3] F. Mosteller, R. R. Bush, Selected quantitative techniques, Addison-Wesley, 1954.

[4] T. Lipták, On the combination of independent tests, Magyar Tud Akad Mat Kutato Int Kozl 3 (1958) 171–197.

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

[6] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, A. M. Chinnaiyan, Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, Cancer research 62 (2002) 4427–4433.

[7] H. Choi, R. Shen, A. M. Chinnaiyan, D. Ghosh, A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments, BMC bioinformatics 8 (2007) 1–20.

[8] L. H. C. Tippett, The methods of statistics. an introduction mainly for workers in the biological sciences., The Methods of Statistics. An introduction mainly for workers in the biological sciences. (1931).

[9] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8 (1936) 3–62.

[10] C. Song, G. C. Tseng, Hypothesis setting and order statistic for robust genomic meta-analysis, The annals of applied statistics 8 (2014) 777.

[11] Y. Li, D. Ghosh, Meta-analysis based on weighted ordered p-values for genomic data with heterogeneity, BMC bioinformatics 15 (2014) 1–12.

[12] J. P. Higgins, S. G. Thompson, J. J. Deeks, D. G. Altman, Measuring inconsistency in meta-analyses, Bmj 327 (2003) 557–560.

[13] F. Begum, D. Ghosh, G. C. Tseng, E. Feingold, Comprehensive literature review and statistical considerations for gwas meta-analysis, Nucleic acids research 40 (2012) 3777–3784.

[14] C. J. Willer, Y. Li, G. R. Abecasis, Metal: fast and efficient meta-analysis of genomewide association scans, Bioinformatics 26 (2010) 2190–2191.

[15] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal statistical society: series B (Methodological) 57 (1995) 289–300.

[16] B. Efron, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, Journal of the American Statistical Association 99 (2004) 96–104.

[17] S. Sikdar, S. Datta, S. Datta, Eama: Empirically adjusted meta-analysis for large-scale simultaneous hypothesis testing in genomic experiments, PloS one 12 (2017) e0187287.

[18] B. Wilkinson, A statistical consideration in psychological research., Psychological bulletin 48 (1951) 156.

[19] D. Phillips, D. Ghosh, Testing the disjunction hypothesis using voronoi diagrams with applications to genetics, The Annals of Applied Statistics (2014) 801–823.

[20] Y. Benjamini, R. Heller, Screening for partial conjunction hypotheses, Biometrics 64 (2008) 1215–1222.

[21] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, J. Chory, Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis, Bioinformatics 22 (2006) 2825–2827.

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

[23] I. Olkin, H. Saner, Approximations for trimmed fisher procedures in research synthesis, Statistical Methods in Medical Research 10 (2001) 267–276.

[24] B. Efron, R. Tibshirani, Using specially designed exponential families for density estimation, The Annals of Statistics 24 (1996) 2431–2461.

[25] B. Efron, Size, power and false discovery rates, The Annals of Statistics 35 (2007) 1351–1377.

[26] M. E. Ritchie, B. Phipson, D. I. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for rna-sequencing and microarray studies, Nucleic acids research 43 (2015) e47–e47.

[27] J. J. Hughey, A. J. Butte, Robust meta-analysis of gene expression using the elastic net, Nucleic acids research 43 (2015) e79–e79.

[28] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, David: database for annotation, visualization, and integrated discovery, Genome biology 4 (2003) 1–11.

[29] P. J. Park, S. W. Kong, T. Tebaldi, W. R. Lai, S. Kasif, I. S. Kohane, Integration of heterogeneous expression data sets extends the role of the retinol pathway in diabetes and insulin resistance, Bioinformatics 25 (2009) 3121–3127.

[30] L. Khondkaryan, S. Margaryan, D. Poghosyan, G. Manukyan, Impaired inflammatory response to lps in type 2 diabetes mellitus, International journal of inflammation 2018 (2018).

[31] F. Mauvais-Jarvis, Is estradiol a biomarker of type 2 diabetes risk in postmenopausal women?, Diabetes 66 (2017) 568–570.

[32] K. T. Bauerle, C. Harris, Glucocorticoids and diabetes, Missouri medicine 113 (2016) 378.

[33] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein, A basic introduction to fixed-effect and random-effects models for meta-analysis, Research synthesis methods 1 (2010) 97–111.

[34] M. van Iterson, E. W. van Zwet, B. T. Heijmans, Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution, Genome biology 18 (2017) 1–13.

APPENDIX A

R SCRIPT FOR CHAPTER 2

The following R code is used to implement the Empirically Adjusted Weighted Ordered P-value Method in the presence of hidden variable that act as a confounder.

```
library(limma)
library(locfdr)

N=500
sensitivity<−matrix(NA,nrow=N,ncol=8)
specificity<−matrix(NA,nrow=N,ncol=8)
FDR<−matrix(NA,nrow=N,ncol=8)
type1error<−matrix(NA,nrow=N,ncol=8)

for(l in 1: N) {

  ############data generation############
  numgene<−3000
  m=10
  delta=4
  geneID<−1:numgene
  Ssize<−20
  Control_results<−matrix(NA,nrow=numgene,ncol=Ssize)
  Trt_results<−matrix(NA,nrow=numgene,ncol=Ssize)
  pval.res<−matrix(NA,nrow=numgene,ncol=m)

  for(j in 1:m){

    for(i in 1:Ssize){

      ####Control group####
```

```
gv1<−rep(c(−4,4,0),c(25∗j,25∗j,numgene −(50∗j)))

s_i1k<−rbinom(numgene,1,0.4)
datac<−data.frame(ID=geneID,s_i1k=s_i1k,gv1=gv1)

genec_1<−datac[1:(25∗j),]
genec_2<−datac[((25∗j)+1):(50∗j),]
genec_3<−datac[((50∗j)+1): numgene,]

w_11k<−ifelse(genec_1$s_i1k==1,rnorm(1,(−1+j),0.01),0)
w_21k<−ifelse(genec_2$s_i1k==1,rnorm(1,(2+j),0.01),0)
w_31k<−ifelse(genec_3$s_i1k==1,rnorm(1,(5+j),0.01),0)

datac$w_i1k<−c(w_11k,w_21k,w_31k)

e11<−c(rep(rnorm(1,0,1),30),rep(0,90),rep(rnorm(1,0,1),60),
      rep(0,1320),rep(rnorm(1,0,1),60),rep(0,1110),
      rep(rnorm(1,0,1),60),rep(0,270))
e12<−rnorm(3000,0,2)
erdat1<−cbind(e11,e12)
ec11<−((1/sqrt(2))∗(erdat1[1:30,1]+erdat1[1:30,2]))
ec12<−erdat1[31:120,2]
ec13<−((1/sqrt(2))∗(erdat1[121:180,1]+erdat1[121:180,2]))
ec14<−erdat1[181:1500,2]
ec15<−((1/sqrt(2))∗(erdat1[1501:1560,1]+erdat1[1501:1560,2]))
ec16<−erdat1[1561:2670,2]
ec17<−((1/sqrt(2))∗(erdat1[2671:2730,1]+erdat1[2671:2730,2]))
ec18<−erdat1[2731:3000,2]
ei1k<−c(ec11,ec12,ec13,ec14,ec15,ec16,ec17,ec18)

y_i1k<−datac$gv1+datac$w_i1k+ei1k

Control_results[,i]<−y_i1k

####Treatment group####
```

```
  gv2<−rep(c(4,−4,0),c(25∗j,25∗j, numgene −(50∗j)))

  s_i2k<−rbinom(numgene,1,0.4)
  datat<−data.frame(ID=geneID,s_i2k=s_i2k,gv2=gv2)

  genet_1<−datat[1:(25∗j),]
  genet_2<−datat[((25∗j)+1):(50∗j),]
  genet_3<−datat[((50∗j)+1): numgene,]

  w_12k<−ifelse(genet_1$s_i2k==1,rnorm(1,(−1+j+delta),0.01),0)
  w_22k<−ifelse(genet_2$s_i2k==1,rnorm(1,(2+j+delta),0.01),0)
  w_32k<−ifelse(genet_3$s_i2k==1,rnorm(1,(5+j+delta),0.01),0)

  datat$w_i2k<−c(w_12k,w_22k,w_32k)

  e21<−c(rep(rnorm(1,0,1),30),rep(0,90),rep(rnorm(1,0,1),60),
       rep(0,1320),rep(rnorm(1,0,1),60),rep(0,1110),
       rep(rnorm(1,0,1),60),rep(0,270))
  e22<−rnorm(3000,0,2)
  erdat2<−cbind(e21,e22)
  ec21<−((1/sqrt(2))∗(erdat2[1:30,1]+erdat2[1:30,2]))
  ec22<−erdat2[31:120,2]
  ec23<−((1/sqrt(2))∗(erdat2[121:180,1]+erdat2[121:180,2]))
  ec24<−erdat2[181:1500,2]
  ec25<−((1/sqrt(2))∗(erdat2[1501:1560,1]+erdat2[1501:1560,2]))
  ec26<−erdat2[1561:2670,2]
  ec27<−((1/sqrt(2))∗(erdat2[2671:2730,1]+erdat2[2671:2730,2]))
  ec28<−erdat2[2731:3000,2]
  ei2k<−c(ec21,ec22,ec23,ec24,ec25,ec26,ec27,ec28)


  y_i2k<−datat$gv2+datat$w_i2k+ei2k

  Trt_results[,i]<−y_i2k

}
```

```
  mj<−cbind(Control_results,Trt_results)
  colnames(mj)<−rep(c("C","T"),each=Ssize)

  f<−factor(colnames(mj))
  design <− model.matrix(~f)
  fit <− eBayes(lmFit(mj,design))
  pres<−fit$p.value[,2]
  pval.res[,j]<−pres

}


z<−qnorm(as.vector(pval.res))
w<−locfdr(z,plot=0)
mu<−w$fp0["cmest","delta"]
sigma<−w$fp0["cmest","sigma"]

z.pval<−qnorm(pval.res)
empadjz<−(z.pval−mu)/sigma
empadj.p<−pnorm(empadjz)


############Orderd P−Values for emp_adj_wop############
res<−data.frame(empadj.p)
rownames(res)<−paste("g_",1:numgene)
colnames(res)<−paste("e_",1:m)
res.list <− split(res, seq(nrow(res)))

sorted.res.list<−lapply(res.list,sort)

############Orderd P−Values for wop####################
res1<−data.frame(pval.res)
rownames(res1)<−paste("g_",1:numgene)
colnames(res1)<−paste("e_",1:m)
```

```
res.list1 <- split(res1, seq(nrow(res1)))

sorted.res.list1<-lapply(res.list1,sort)

###############Test statistics##########################
m=10
wb<-dbinom(0:(m-1),m-1,0.5)
k=5
whb<-c(0,0,0,0,dbinom((k-1):(m-1),m-1,0.5))

Fisher_Binom=function(x) -2*wb*log(x)
Stouffer_Binom=function(x) wb*qnorm(as.numeric(x),lower.tail=FALSE)
Fisher_HalfBinom=function(x) -2*whb*log(x)
Stouffer_HalfBinom=function(x) whb*qnorm(as.numeric(x),lower.tail=FALSE)

################Calculating test statistic for EAWOP
    ################

resFB<-lapply(sorted.res.list,Fisher_Binom)
resSB<-lapply(sorted.res.list,Stouffer_Binom)
resFHB<-lapply(sorted.res.list,Fisher_HalfBinom)
resSHB<-lapply(sorted.res.list,Stouffer_HalfBinom)



eawopFB_test_stat<-sapply(resFB,sum)
eawopSB_test_stat<-sapply(resSB,sum)
eawopFHB_test_stat<-sapply(resFHB,sum)
eawopSHB_test_stat<-sapply(resSHB,sum)

################Calculating test statistic WOP
    ####################

res1FB<-lapply(sorted.res.list1,Fisher_Binom)
res1SB<-lapply(sorted.res.list1,Stouffer_Binom)
res1FHB<-lapply(sorted.res.list1,Fisher_HalfBinom)
res1SHB<-lapply(sorted.res.list1,Stouffer_HalfBinom)
```

```
wopFB_test_stat<−sapply(res1FB,sum)
wopSB_test_stat<−sapply(res1SB,sum)
wopFHB_test_stat<−sapply(res1FHB,sum)
wopSHB_test_stat<−sapply(res1SHB,sum)


##########################calculating p values
    ##########################
eawopFB_results<−matrix(NA,nrow=numgene,ncol=1000)
eawopSB_results<−matrix(NA,nrow=numgene,ncol=1000)
eawopFHB_results<−matrix(NA,nrow=numgene,ncol=1000)
eawopSHB_results<−matrix(NA,nrow=numgene,ncol=1000)


wopFB_results<−matrix(NA,nrow=numgene,ncol=1000)
wopSB_results<−matrix(NA,nrow=numgene,ncol=1000)
wopFHB_results<−matrix(NA,nrow=numgene,ncol=1000)
wopSHB_results<−matrix(NA,nrow=numgene,ncol=1000)


for(k in 1:1000) {

  pval<−matrix(runif(numgene∗10),nrow=numgene,ncol=10)
  rownames(pval)<−paste("g_",1:numgene)
  pval.list <− split(pval, seq(nrow(pval)))
  sorted.pval.list<−lapply(pval.list,sort)

  ##########################################
  stat_res1<−lapply(sorted.pval.list,Fisher_Binom)
  stat_wop1<−sapply(stat_res1,sum)

  ind1<−ifelse(stat_wop1 > eawopFB_test_stat , 1 , 0)
  eawopFB_results[,k]<−ind1
  ind11<−ifelse(stat_wop1 > wopFB_test_stat , 1 , 0)
```

```
  wopFB_results[,k]<−ind11
  ##########################################
  stat_res2<−lapply(sorted.pval.list,Stouffer_Binom)
  stat_wop2<−sapply(stat_res2,sum)

  ind2<−ifelse(stat_wop2 > eawopSB_test_stat , 1 , 0)
  eawopSB_results[,k]<−ind2
  ind22<−ifelse(stat_wop2 > wopSB_test_stat , 1 , 0)
  wopSB_results[,k]<−ind22
  ##########################################
  stat_res3<−lapply(sorted.pval.list,Fisher_HalfBinom)
  stat_wop3<−sapply(stat_res3,sum)

  ind3<−ifelse(stat_wop3 > eawopFHB_test_stat , 1 , 0)
  eawopFHB_results[,k]<−ind3
  ind33<−ifelse(stat_wop3 > wopFHB_test_stat , 1 , 0)
  wopFHB_results[,k]<−ind33
  ##########################################
  stat_res4<−lapply(sorted.pval.list,Stouffer_HalfBinom)
  stat_wop4<−sapply(stat_res4,sum)

  ind4<−ifelse(stat_wop4 > eawopSHB_test_stat , 1 , 0)
  eawopSHB_results[,k]<−ind4
  ind44<−ifelse(stat_wop4 > wopSHB_test_stat , 1 , 0)
  wopSHB_results[,k]<−ind44


}

eawop_pval_FB<−apply(eawopFB_results,1,sum)/1000
eawop_pval_SB<−apply(eawopSB_results,1,sum)/1000
eawop_pval_FHB<−apply(eawopFHB_results,1,sum)/1000
eawop_pval_SHB<−apply(eawopSHB_results,1,sum)/1000

wop_pval_FB<−apply(wopFB_results,1,sum)/1000
wop_pval_SB<−apply(wopSB_results,1,sum)/1000
```

```
wop_pval_FHB<−apply(wopFHB_results,1,sum)/1000
wop_pval_SHB<−apply(wopSHB_results,1,sum)/1000


############apply the Benjamini−Hochberg##################

pval_results<−data.frame(null=c(rep(0,500),rep(1,2500)),
truth=c(rep(1,300),rep(0,2700)),eawop_pvalFB=eawop_pval_FB,
eawop_pvalSB=eawop_pval_SB,eawop_pvalFHB=eawop_pval_FHB,
eawop_pvalSHB=eawop_pval_SHB,wop_pvalFB=wop_pval_FB,
wop_pvalSB=wop_pval_SB,wop_pvalFHB=wop_pval_FHB,
wop_pvalSHB=wop_pval_SHB)
rownames(pval_results)<−paste("g_",1:numgene)

pval_results$adj.eawop_pval_FB<−p.adjust(pval_results$eawop_pvalFB,
method = "BH")
pval_results$pred.eawopFB<−ifelse(pval_results$adj.eawop_pval_FB < 0.05,
1 , 0)

pval_results$adj.eawop_pval_SB<−p.adjust(pval_results$eawop_pvalSB,
method = "BH")
pval_results$pred.eawopSB<−ifelse(pval_results$adj.eawop_pval_SB < 0.05,
1 , 0)

pval_results$adj.eawop_pval_FHB<−p.adjust(pval_results$eawop_pvalFHB,
method = "BH")
pval_results$pred.eawopFHB<−ifelse(pval_results$adj.eawop_pval_FHB < 0.05,
1 , 0)

pval_results$adj.eawop_pval_SHB<−p.adjust(pval_results$eawop_pvalSHB,
method = "BH")
pval_results$pred.eawopSHB<−ifelse(pval_results$adj.eawop_pval_SHB < 0.05,
1 , 0)

pval_results$adj.wop_pval_FB<−p.adjust(pval_results$wop_pvalFB,
method = "BH")
```

```
pval_results$pred.wopFB<−ifelse(pval_results$adj.wop_pval_FB < 0.05,
1 , 0)

pval_results$adj.wop_pval_SB<−p.adjust(pval_results$wop_pvalSB,
method = "BH")
pval_results$pred.wopSB<−ifelse(pval_results$adj.wop_pval_SB < 0.05,
1 , 0)

pval_results$adj.wop_pval_FHB<−p.adjust(pval_results$wop_pvalFHB,
method = "BH")
pval_results$pred.wopFHB<−ifelse(pval_results$adj.wop_pval_FHB < 0.05,
1 , 0)

pval_results$adj.wop_pval_SHB<−p.adjust(pval_results$wop_pvalSHB,
method = "BH")
pval_results$pred.wopSHB<−ifelse(pval_results$adj.wop_pval_SHB < 0.05,
1 , 0)

########################calculating performance
    ########################

EAFBsens<−sum(pval_results$pred.eawopFB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
EAFBspeci<−sum(pval_results$pred.eawopFB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
EAFBfdr<−sum(pval_results$pred.eawopFB[pval_results$truth==0]==1)/
sum(pval_results$pred.eawopFB==1)
EAFBtype1error<−sum(pval_results$pred.eawopFB[pval_results$null==1]==1)/
sum(pval_results$pred.eawopFB==1)

EASBsens<−sum(pval_results$pred.eawopSB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
EASBspeci<−sum(pval_results$pred.eawopSB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
EASBfdr<−sum(pval_results$pred.eawopSB[pval_results$truth==0]==1)/
sum(pval_results$pred.eawopSB==1)
```

```
EASBtype1error<−sum(pval_results$pred.eawopSB[pval_results$null==1]==1)/
sum(pval_results$pred.eawopSB==1)

EAFHBsens<−sum(pval_results$pred.eawopFHB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
EAFHBspeci<−sum(pval_results$pred.eawopFHB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
EAFHBfdr<−sum(pval_results$pred.eawopFHB[pval_results$truth==0]==1)/
sum(pval_results$pred.eawopFHB==1)
EAFHBtype1error<−sum(pval_results$pred.eawopFHB[pval_results$null==1]==1)/
sum(pval_results$pred.eawopFHB==1)

EASHBsens<−sum(pval_results$pred.eawopSHB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
EASHBspeci<−sum(pval_results$pred.eawopSHB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
EASHBfdr<−sum(pval_results$pred.eawopSHB[pval_results$truth==0]==1)/
sum(pval_results$pred.eawopSHB==1)
EASHBtype1error<−sum(pval_results$pred.eawopSHB[pval_results$null==1]==1)/
sum(pval_results$pred.eawopSHB==1)

WOPFBsens<−sum(pval_results$pred.wopFB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
WOPFBspeci<−sum(pval_results$pred.wopFB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
WOPFBfdr<−sum(pval_results$pred.wopFB[pval_results$truth==0]==1)/
sum(pval_results$pred.wopFB==1)
WOPFBtype1error<−sum(pval_results$pred.wopFB[pval_results$null==1]==1)/
sum(pval_results$pred.wopFB==1)

WOPSBsens<−sum(pval_results$pred.wopSB[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
WOPSBspeci<−sum(pval_results$pred.wopSB[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
WOPSBfdr<−sum(pval_results$pred.wopSB[pval_results$truth==0]==1)/
sum(pval_results$pred.wopSB==1)
```

```
  WOPSBtype1error<−sum(pval_results$pred.wopSB[pval_results$null==1]==1)/
  sum(pval_results$pred.wopSB==1)

  WOPFHBsens<−sum(pval_results$pred.wopFHB[pval_results$truth==1]==1)/
  sum(pval_results$truth==1)
  WOPFHBspeci<−sum(pval_results$pred.wopFHB[pval_results$truth==0]==0)/
  sum(pval_results$truth==0)
  WOPFHBfdr<−sum(pval_results$pred.wopFHB[pval_results$truth==0]==1)/
  sum(pval_results$pred.wopFHB==1)
  WOPFHBtype1error<−sum(pval_results$pred.wopFHB[pval_results$null==1]==1)/
  sum(pval_results$pred.wopFHB==1)

  WOPSHBsens<−sum(pval_results$pred.wopSHB[pval_results$truth==1]==1)/
  sum(pval_results$truth==1)
  WOPSHBspeci<−sum(pval_results$pred.wopSHB[pval_results$truth==0]==0)/
  sum(pval_results$truth==0)
  WOPSHBfdr<−sum(pval_results$pred.wopSHB[pval_results$truth==0]==1)/
  sum(pval_results$pred.wopSHB==1)
  WOPSHBtype1error<−sum(pval_results$pred.wopSHB[pval_results$null==1]==1)/
  sum(pval_results$pred.wopSHB==1)

  sensitivity[l,]<−c(EAFBsens,EASBsens,EAFHBsens,EASHBsens,
  WOPFBsens,WOPSBsens,WOPFHBsens,WOPSHBsens)
  specificity[l,]<−c(EAFBspeci,EASBspeci,EAFHBspeci,EASHBspeci,
  WOPFBspeci,WOPSBspeci,WOPFHBspeci,WOPSHBspeci)
  FDR[l,]<−c(EAFBfdr,EASBfdr,EAFHBfdr,EASHBfdr,WOPFBfdr,
  WOPSBfdr,WOPFHBfdr,WOPSHBfdr)
  type1error[l,]<−c(EAFBtype1error,EASBtype1error,EAFHBtype1error,
  EASHBtype1error,WOPFBtype1error,WOPSBtype1error,WOPFHBtype1error,
  WOPSHBtype1error)

}

colnames(sensitivity)<−c("sens_EAFB","sens_EASB","sens_EAFHB",
"sens_EASHB","sens_WOPFB","sens_WOPSB","sens_WOPFHB","sens_WOPSHB")
colnames(specificity)<−c("speci_EAFB","speci_EASB","speci_EAFHB",
```

```
"speci_EASHB","speci_WOPFB","speci_WOPSB","speci_WOPFHB","speci_WOPSHB")
colnames(FDR)<−c("fdr_EAFB","fdr_EASB","fdr_EAFHB","fdr_EASHB",
"fdr_WOPFB","fdr_WOPSB","fdr_WOPFHB","fdr_WOPSHB")
colnames(type1error)<−c("type1error_EAFB","type1error_EASB",
"type1error_EAFHB","type1error_EASHB","type1error_WOPFB",
"type1error_WOPSB","type1error_WOPFHB","type1error_WOPSHB")

(avg_sensitivity<−apply(sensitivity,2,mean))
(avg_specificity<−apply(specificity,2,mean))
(avg_FDR<−apply(FDR,2,mean))
(avg_type1error<−apply(type1error,2,mean))
```

APPENDIX B

R SCRIPT FOR CHAPTER 3

The following R code is used to implement Empirical Bayes Adjusted and Bacon Adjusted methods in the presence of hidden variable that act as a confounder.

```
library(limma)
library(locfdr)
library(bacon)

N=10 # replicates
g<−10000 # no of genes
sg<−1000 # no of sig genes
m=10 # no of experiments
Ssize<−40 #sample size in each group
geneID<−1:g
delta<−2

Control_results<−matrix(NA,nrow=g,ncol=Ssize)
Trt_results<−matrix(NA,nrow=g,ncol=Ssize)

z<−matrix(NA,nrow=g,ncol=m)
we<−matrix(NA,nrow=g,ncol=m)
b<−matrix(NA,nrow=g,ncol=m)

originalhypo<−matrix(NA,nrow=N,ncol=9)

for(l in 1: N) {

  for(j in 1:m){

    for(i in 1:Ssize){
```

```
####Control group####

gv1<−rep(c(−2,2,0),c(sg/2,sg/2,g −(sg)))
s_i1k<−rbinom(g,1,0.4)
datac<−data.frame(ID=geneID,s_i1k=s_i1k,gv1=gv1)

genec_1<−datac[1:(sg/2),]
genec_2<−datac[((sg/2)+1):(sg),]
genec_3<−datac[((sg)+1): g,]

w_11k<−ifelse(genec_1$s_i1k==1,rnorm(1,(−1+j),0.01),0)
w_21k<−ifelse(genec_2$s_i1k==1,rnorm(1,(2+j),0.01),0)
w_31k<−ifelse(genec_3$s_i1k==1,rnorm(1,(5+j),0.01),0)

datac$w_i1k<−c(w_11k,w_21k,w_31k)

e11<−c(rep(rnorm(1,0,1),1000),rep(0,1000),
        rep(rnorm(1,0,1),1000),rep(0,3000),
        rep(rnorm(1,0,1),1000),rep(0,1000),
        rep(rnorm(1,0,1),1000),rep(0,1000))
e12<−rnorm(10000,0,2)
erdat1<−cbind(e11,e12)
ec11<−((1/sqrt(2))∗(erdat1[1:1000,1]+erdat1[1:1000,2]))
ec12<−erdat1[1001:2000,2]
ec13<−((1/sqrt(2))∗(erdat1[2001:3000,1]+erdat1[2001:3000,2]))
ec14<−erdat1[3001:6000,2]
ec15<−((1/sqrt(2))∗(erdat1[6001:7000,1]+erdat1[6001:7000,2]))
ec16<−erdat1[7001:8000,2]
ec17<−((1/sqrt(2))∗(erdat1[8001:9000,1]+erdat1[8001:9000,2]))
ec18<−erdat1[9001:10000,2]
ei1k<−c(ec11,ec12,ec13,ec14,ec15,ec16,ec17,ec18)

y_i1k<−datac$gv1+datac$w_i1k+ei1k

Control_results[,i]<−y_i1k
```

```
####Treatment group####

gv2<−rep(c(2,−2,0),c(sg/2,sg/2, g −(sg)))

s_i2k<−rbinom(g,1,0.4)
datat<−data.frame(ID=geneID,s_i2k=s_i2k,gv2=gv2)

genet_1<−datat[1:(sg/2),]
genet_2<−datat[((sg/2)+1):(sg),]
genet_3<−datat[((sg)+1): g,]

w_12k<−ifelse(genet_1$s_i2k==1,rnorm(1,(−1+j+delta),0.01),0)
w_22k<−ifelse(genet_2$s_i2k==1,rnorm(1,(2+j+delta),0.01),0)
w_32k<−ifelse(genet_3$s_i2k==1,rnorm(1,(5+j+delta),0.01),0)

datat$w_i2k<−c(w_12k,w_22k,w_32k)

e21<−c(rep(rnorm(1,0,1),1000),rep(0,1000),
      rep(rnorm(1,0,1),1000),rep(0,3000),
      rep(rnorm(1,0,1),1000),rep(0,1000),
      rep(rnorm(1,0,1),1000),rep(0,1000))
e22<−rnorm(10000,0,2)
erdat2<−cbind(e21,e22)
ec21<−((1/sqrt(2))∗(erdat2[1:1000,1]+erdat2[1:1000,2]))
ec22<−erdat2[1001:2000,2]
ec23<−((1/sqrt(2))∗(erdat2[2001:3000,1]+erdat2[2001:3000,2]))
ec24<−erdat2[3001:6000,2]
ec25<−((1/sqrt(2))∗(erdat2[6001:7000,1]+erdat2[6001:7000,2]))
ec26<−erdat2[7001:8000,2]
ec27<−((1/sqrt(2))∗(erdat2[8001:9000,1]+erdat2[8001:9000,2]))
ec28<−erdat2[9001:10000,2]
ei2k<−c(ec21,ec22,ec23,ec24,ec25,ec26,ec27,ec28)

y_i2k<−datat$gv2+datat$w_i2k+ei2k
```

```
    Trt_results[,i]<−y_i2k

  }

  s<−cbind(Control_results,Trt_results)

  colnames(s)<−rep(c("C","T"),each=Ssize)

  f<−factor(colnames(s))
  design <− model.matrix(~f)
  fit <− lmFit(s,design)
  beta<−fit$coefficients[,2]
  se<−(fit$stdev.unscaled∗fit$sigma)[,2]

  z[,j]<−beta/se
  we[,j]<−1/seˆ2
  b[,j]<−beta

}

##############P−value####################

cbeta<−apply(b∗we,1,sum)/apply(we,1,sum)
zval<−cbeta/apply(we,1,function(x) sqrt(1/sum(x)))
pval<−2 ∗ pnorm( abs(zval), lower.tail=FALSE)

###########emp.adj Bayes#################

zres2<−as.vector(zval)
wtry2<−try(locfdr(zres2,plot=0,nulltype = 2),silent=TRUE)
if('try−error' %in% class(wtry2)) next
else w2<−wtry2
mu2<−w2$fp0["cmest","delta"]
sigma2<−w2$fp0["cmest","sigma"]

ezval2<−(zval−mu2)/sigma2
```

```
epval2<−2 ∗ pnorm( abs(ezval2), lower.tail=FALSE)

##########bacon pval ################
bc<−bacon(zres2)
#fit(bc)
#posteriors(bc)
w0<−estimates(bc)
mu0<−w0[1,4]
sigma0<−w0[1,7]

ezval0<−(zval−mu0)/sigma0
epval0<−2 ∗ pnorm( abs(ezval0), lower.tail=FALSE)

##############performance original hypo
    ######################
pval_results<−data.frame(truth=c(rep(1,sg),rep(0,(g−sg))),
                pval=pval,epval2=epval2,epval0=epval0)
pval_results$adj_pval<−p.adjust(pval_results$pval, method = "BH")
pval_results$adj_epval2<−p.adjust(pval_results$epval2, method = "BH")
pval_results$adj_epval0<−p.adjust(pval_results$epval0, method = "BH")

pval_results$pred<−ifelse(pval_results$adj_pval < 0.05 , 1 , 0)
pval_results$epred2<−ifelse(pval_results$adj_epval2 < 0.05 , 1 , 0)
pval_results$epred0<−ifelse(pval_results$adj_epval0 < 0.05 , 1 , 0)

osens<−sum(pval_results$pred[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
ospeci<−sum(pval_results$pred[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
ofdr<−sum(pval_results$pred[pval_results$truth==0]==1)/
sum(pval_results$pred==1)

oesens2<−sum(pval_results$epred2[pval_results$truth==1]==1)/
sum(pval_results$truth==1)
oespeci2<−sum(pval_results$epred2[pval_results$truth==0]==0)/
sum(pval_results$truth==0)
```

```
  oefdr2<−sum(pval_results$epred2[pval_results$truth==0]==1)/
  sum(pval_results$epred2==1)

  oesens0<−sum(pval_results$epred0[pval_results$truth==1]==1)/
  sum(pval_results$truth==1)
  oespeci0<−sum(pval_results$epred0[pval_results$truth==0]==0)/
  sum(pval_results$truth==0)
  oefdr0<−sum(pval_results$epred0[pval_results$truth==0]==1)/
  sum(pval_results$epred0==1)

  originalhypo[l,]<−c(osens,ospeci,ofdr,oesens2,
  oespeci2,oefdr2,oesens0,oespeci0,oefdr0)

}

colnames(originalhypo)<−c("original_sens","original_speci","original_fdr",
"EAsens_Bayes","EAspeci_Bayes","EAfdr_Bayes",
"EAsens_Bacon","EAspeci_Bacon","EAfdr_Bacon")

(avg_original<−apply(originalhypo,2,mean,na.rm=TRUE))
```

VITA

Wimarsha Thathsarani Jayanetti
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529

Education

Ph.D. Old Dominion Universtiy, Norfok, VA. (Anticipated December 2022)
Major: Computational Applied Mathematics (Statistics).

M.S. University of Texas at El Paso, El Paso, TX. (August 2018)
Major: Statistics.

B.S. University of Colombo, Sri Lanka. (April 2014)
Major: Statistics.

Experience

Graduate Research Assistant, Old Dominion University Research Foundation, Norfolk, VA, (05/2022-12/2022).

Graduate Teaching Assistant, Old Dominion University, Norfolk, VA, (08/2018 - 05/2022).

Graduate Teaching Assistant, University of Texas at El Paso, El Paso, TX, (08/2016-08/2018).

Lecturer, Sri Lanka Institute of Information Technology, Sri Lanka, (01/2016-07/2016)

Assistant Lecturer, University of Colombo, Sri Lanka, (02/2014-11/2015)

Typeset using LaTeX.