HYPOTHESIS SETTINGS AND METHODS FOR GENE EXPRESSION META-ANALYSIS

by

Chi Song

MS, Tsinghua University, Beijing, China, 2007

BS, Tsinghua University, Beijing, China, 2004

Submitted to the Graduate Faculty of
the Graduate School of Public Health in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Chi Song

It was defended on

April 16th 2012

and approved by

George C. Tseng, ScD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, PhD, Professor

Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Lisa Weissfeld, PhD, Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Xinghua Lu, MD, PhD, Associate Professor

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh

Dissertation Director: George C. Tseng, ScD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

HYPOTHESIS SETTINGS AND METHODS FOR GENE EXPRESSION META-ANALYSIS

Chi Song, PhD

University of Pittsburgh, 2012

With the advent of high-throughput technologies, biomedical research has been dramatically reshaped in the past two decades. Technologies such as microarrays are broadly utilized to study the relationship between genomic alterations and disease outcomes. However, genomic analyses are criticized for their low reproducibility and generalizability. Large-scale meta-analysis of multiple studies is a timely and important issue with great public health significance, because robust biomarkers can be found for complex human diseases such as major depression disorder using meta-analysis techniques. Accurate marker detection will improve the disease diagnosis, treatment selection and prognosis prediction.

In this dissertation, I first illustrate different hypothesis settings for two different types of biomarkers: biomarkers that are differentially expressed (DE) "in all" studies and biomarkers that are DE "in any" studies. Then I propose a robust setting HS_r to detect genes differentially expressed (DE) "in majority of" studies. For HS_r, I propose an order statistic of p-values (rth order p-value, rOP) across combined studies as the test statistic. I also explore statistical properties such as power and asymptotic behavior of rOP. The method is applied to three examples to demonstrate its robustness and sensitivity. I develop two methods to guide the selection of r.

The non-complementary property of HS_r causes anti-conservative inferences. To overcome this, I propose HS'_r as a complementary form of HS_r. For HS'_r, the major obstacle comes from the mixture nature of the null distribution. From a Bayesian point of view, I propose a semiparametric mixture model for the observed p-values in combined studies.

A Bayes factor is calculated based on the posterior distribution to substitute traditional hypothesis testing for HS'_r. I also develop an expectation-maximization (EM) algorithm to fit this model. Simulation results and real data analysis show improved specificity and sensitivity of this novel approach compared to traditional methods.

Beyond meta-analysis of single genes, I also propose a framework to integrate multiple biological networks. A conservative subnetwork in a subset of datasets can be identified using my approach.

In conclusion, I discuss various interesting questions in genomic meta-analysis in this dissertation. And I provide a series of statistical tools to address them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to offer my sincere gratitude to my advisor Dr. George C. Tseng for his encouragement, guidance and support. This dissertation would not have been possible without his continuous and selfless help. His support is not limited to research area, but also in my life. By sharing his wisdom and experience, he helped me establish interests in academic career.

I would like to thank my committee members, Drs. Eleanor Feingold, Lisa Weissfeld and Xinghua Lu for their constructive advices in research and accommodation in schedule.

I also thank my colleagues who made the group feel just like home.

Last but not least, thank my parents and my wife for their constant support without complaint.

1.0 INTRODUCTION

With the advances in high-throughput experimental technology in the past decade, the production of genomic data has become affordable and thus prevalent in biomedical research. Accumulation of experimental data in the public domain has grown rapidly, particularly of microarray data for gene expression analysis and single nucleotide polymorphism (SNP) genotyping data for genome-wide association studies (GWAS). For example, the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) from the National Center for Biotechnology Information (NCBI) and the Gene Expression Atlas (http://www.ebi.ac.uk/gxa/) from the European Bioinformatics Institute (EBI) are the two largest public depository websites for gene expression data and dbGaP (http://www.ncbi.nlm.nih.gov/gap) has the largest collection of genotype data.

Because individual studies usually contain a limited number of samples, and the reproducibility of genomic studies is relatively low, the generalizability of their conclusions is widely criticized. Therefore combining multiple studies to improve statistical power and provide validated conclusions has emerged as a common practice (see recent review papers Tseng et al., 2012; Begum et al., 2012). Such genomic meta-analysis is particularly useful in microarray analysis and GWAS.

In this dissertation, I will focus on the meta-analysis of microarray data. However, the methods developed could also be applied in other types of genomic data. In section 1.1, I briefly introduce microarray technology. And in section 1.2, traditional meta-analysis methods which have been extended to microarray data are introduced. Network analysis methods are discussed in section 1.3.

In chapter 2, existing meta-analysis methods that are commonly seen in genomics research are compared and categorized into two groups: those that pursue DE genes in "one or more" studies or "all" studies respectively. I further explore the underlying hypothesis settings (HS_A and HS_B) for these two groups of methods. Based on this observation, I propose a robust hypothesis setting which targets differentially expressed (DE) genes in the "majority of" studies (HS_r). To detect HS_r, I suggest the usage of the rth order p-value (rOP) as the test statistic and develop two methods to select parameter r. Statistical properties like power function and asymptotic behavior of rOP are investigated. This method is then applied to three real data examples and compared to other meta-analysis methods.

In chapter 3, I extend HS_r to a more realistic hypothesis setting, HS'_r. In HS'_r, the null hypothesis becomes a composite hypothesis which is the complementary event of the alternative hypothesis. However, because of the complexity of the null hypothesis, it is intractable to find an appropriate test statistic with a well-defined null distribution. Instead, to bypass the null hypothesis issue, I propose a Bayesian approach to address the meta-analysis problem from another angle. I infer the posterior probability of the alternative hypothesis being true. In this approach, I model the p-values for all the genes in all the studies using a semiparametric two-component mixture model. Gene effects and study effects are also taken into account. Both simulation and real data analysis show the advantage of this approach.

In chapter 4, meta-analysis of multiple network construction results is discussed. I propose a brand new framework to identify conservative subnetworks in combined studies. The method is also evaluated using simulation and real data analysis.

In chapter 5, the works in this dissertation are summarized.

1.1 GENE EXPRESSION TECHNOLOGY

Microarray technologies assay monitor the mRNA expression levels of tens of thousands of genes in each sample simultaneously. By analyzing the gene expression levels from a set of tissue samples, biomarkers that are related to the phenotype of interest or differentially expressed between specific groups of samples can be detected. The gene expression levels can also be used to detect important biological pathways using the correlations between the genes. In biomedical applications, microarrays can be an important diagnosis or prognosis tool for complex human diseases as well.

In most microarrays platforms, each sample is measured on a single array. First, the mRNAs are extracted from the sample tissue. Then polymerase chain reaction (PCR) technology is applied to retro-transcribe and amplify the mRNAs into cDNA libraries. The cDNAs are marked by fluorescent dyes. Then the cDNA libraries are incubated with the microarray. Since on the array, there are oligonucleotide probes designed to hybridize with the cDNA segments of each gene of interest, the cDNAs will be retained on the probes with matched sequences. Therefore, the fluorescent amount detected on each probe may reflect the expression level of the corresponding gene. For redundancy, generally multiple probes are designed for the same gene.

Currently, a number of commercial microarray platforms are provided by companies including Affymetrix, Agilent and Illumina. However, because every platform uses its own probe design and detailed technology, the signals detected by different platforms could be dramatically different. Even within the same platform, there are biases for a number of reasons including sample preparation, experimental protocol and batch effects. As a result, direct combination of multiple microarray studies requires careful normalization and tends to be error-prone.

To understand the mechanisms of certain diseases, it is particularly important to identify the set of genes that are DE between the diseased and normal tissues (or between the severe and non-severe patients). For DE gene detection, hypothesis testing is repeatedly applied to each of the genes. And because multiple hypothesis tests are performed, the problem of multiple comparisons should be addressed. False discovery rate (FDR) are generally controlled for microarray analysis by permutation test or by the procedure proposed by Benjamini and Hochberg [1995].

1.2 TRADITIONAL META-ANALYSIS METHODS

As microarray analysis becomes prevalent, the meta-analysis of multiple studies becomes commonplace and important. Many traditional methods have been extended and applied for microarray meta-analysis. Two major types of statistical procedures have been used: combining effect sizes and combining p-values. Generally, no method uniformly performs better than the others in all datasets for all biological goals, both from a theoretical point of view [Littell and Folks, 1971, 1973] and from empirical experiences.

In the methods combining effect sizes, the fixed effects model and random effects model are the most popular [Cooper et al., 2009]. These methods are usually more straightforward and powerful to directly synthesize information of the effect size estimates. However, they are only applicable to samples with two conditions so the effect sizes can be well-defined and combined.

Methods combining p-values provide better flexibility for various outcome conditions as long as p-values can be assessed for integration. Fisher's method is among the earliest p-value methods applied to microarray meta-analysis [Rhodes et al., 2002]. It adopts a sum score of log-transformed p-values to aggregate statistical significance across studies. Under the null hypothesis and assuming that studies are independent and the hypothesis testing procedure correctly fits the observed data, Fisher's statistic follows a chi-squared distribution. Other methods such as Stouffer's method [Stouffer et al., 1949], minP method [Tippett, 1931] and maxP method [Wilkinson, 1951] have also been widely used in microarray meta-analysis. It can be shown that these test statistics have simple analytical forms of null distributions and thus they are easy to apply to the genomic setting.

1.2.1 Fisher's method

For combining K p-values $(p_1, p_2, \ldots, p_K)$, the Fisher's statistic is

$$S^{Fisher} = -2 \sum_{k=1}^{K} \log p_k.$$

Under the null hypothesis that the effect size in all K studies equals zero ($\theta_1 = \theta_2 = \cdots = \theta_K = 0$), the K combined p-values independently follow $U(0,1)$. Therefore, it is easy to show that $-\log p_k \sim \text{EXP}(1)$. Accordingly, $S^{Fisher} \sim \chi^2(2K)$. It has been proven that Fisher's method is asymptotic Bahadur optimal (ABO) when it is assumed that the effect sizes are all the same in the alternative hypothesis [Littell and Folks, 1971, 1973].

1.2.2 Stouffer's method

Similar to Fisher's method, Stouffer's statistic is the summation of inverse Gaussian transformation of combined p-values.

$$S^{Stouffer} = \sum_{k=1}^{K} \Phi^{-1}(p_k)$$

It is easy to show that under null hypothesis, $S^{Stouffer} \sim N(0,k)$. The p-value of Stouffer's statistic can be calculated accordingly. Comparing to Fisher's method, Stouffer's method is more robust to outlying p-values. Thus it is increasingly used in biomedical researches where outliers are expected.

1.2.3 Minimum P-value

In the minimum p-value (minP) method, the test statistic is the smallest p-value of combined studies:

$$S^{minP} = \min\{p_k\}.$$

Under the null hypothesis, $S^{minP} \sim Beta(1,K)$. Because only the smallest p-value is considered in the test statistic, the minP method is sensitive to a single extreme p-value. Thus minP is used to detect genes that are differentially expressed in "one or more" studies. Similarly, in Fisher's method and Stouffer's method, only one extremely small p-value may result in a large test statistic and a significant result. Collectively, Fisher's method, Stouffer's method and the minP method are all designed to detect differential biomarkers in "any" of a set of studies.

1.2.4 Maximum P-value

Maximum P-value (maxP) uses a statistic that is the largest p-value of the combined studies:

$$S^{maxP} = \max\{p_k\}.$$

Under the null hypothesis, $S^{maxP} \sim Beta(K, 1)$. Although maxP looks similar to minP, the performance is very different. In maxP, because the test statistic is the largest p-value, it is robust to single small p-values. However, since maxP requires that all combined p-values are small to generate a significant result, it is sometimes too stringent to detect important biomarkers. Therefore, maxP is not commonly used in biomedical research because of its low statistical power.

1.3 NETWORK ANALYSIS

Understanding the roles of single genes is fundamental to investigating the mechanism of complex biological process. However, in biological systems, genes do not function independently. Different genes work together to carry out certain biological procedures. To address this question, networks are usually constructed for biological systems. The networks can be presented by graphs which are comprised of vertices and edges connecting them. The edges can be either directed or undirected in biological networks.

The networks that represent the interactions among molecules such as DNA, RNA and proteins are called molecular networks. Currently, multiple types of molecular networks are used to describe the interactions among genes and other components. The most commonly used networks include protein-protein interaction networks, metabolic networks, regulatory networks and RNA networks [Barabási et al., 2011]. Networks can be inferred by either experimental or computational methods. Among the available networks, regulatory networks are particularly interesting because genes involved in the same biological pathway are often co-regulated. It is known that genes are regulated by many molecular mechanisms including copy number variation, methylation, transcription factors and microRNAs. Complex diseases are sometimes found to be related to alteration of gene regulation networks. For example, in cancers, it is known that genes related to cell cycle and proliferation are often dysregulated. Therefore, constructing regulatory networks from gene expression data is important to understand the mechanisms of the disease.

Because in co-regulation networks, expression of co-regulated genes is also highly correlated, correlations between genes can be used to infer the co-regulation network. For example, in gene expression data, correlations between different genes can be calculated. The gene coexpression network can be constructed by thresholding the absolute correlations [Guilloux et al., 2010]. The expression levels of multiple genes can also be deemed as sampled from a Markov random field (MRF). Then methods such as graphical lasso [Friedman et al., 2008] can be applied to recover the MRF. Besides methods that use only the expression levels of mRNAs and their covariance structure, methods that accommodate other data types and prior informations have been proposed. For example, Huang et al. [2011] proposed mirConnX to incorporate both microRNA expression levels and prior knowledge of experimentally confirmed biological pathways to construct the gene regulation network.

In network construction, especially for ab initio methods, edges should be inferred between each pair of genes. The number of inferences increases dramatically as the number of genes increases. Therefore, large sample sizes are required to construct a stable network. However, in single studies, the sample sizes are usually limited. So the networks constructed are not satisfactory in terms of sensitivity and specificity. Meta-analysis methods can be generated to combine multiple studies to construct robust subnetworks.

2.0 RTH ORDER P-VALUE FOR ROBUST GENOMIC META-ANALYSIS

2.1 BACKGROUND

In microarray analysis, One commonly seen analysis is to detect differentially expressed (DE) genes when samples are collected with labels of two conditions (e.g. tumor recurrence versus non-recurrence), multiple conditions (e.g. multiple tumor subtypes), survival information or time series. In the literature, microarray meta-analysis usually refers to combining multiple studies of related hypothesis or conditions to better detect DE genes (also called candidate biomarkers).

According to section 1.2, many traditional meta-analysis methods have been applied to microarray meta-analysis. And there are two major categories of procedures: combining effect sizes and combining p-values. Because of the limitation of the combining effect sizes methods, they are not the focus of this dissertation. The null distributions and properties of commonly used methods that combine p-values are discussed in section 1.2. The assumptions and hypothesis settings behind these methods are, however, very different and have not been carefully considered in most microarray meta-analysis applications so far. In this chapter, I begin in Section 2.2.1 to elucidate the hypothesis settings and biological implications behind these methods. In many meta-analysis applications, detecting markers differentially expressed in all studies is more appealing. The requirement of DE in "all" studies, however, is too stringent when K is large and in light of the fact that experimental data are peppered with noisy measurements from probe design, sample collection, data generation and analysis. Thus, I describe in Section 2.2.1 a robust setting (called HS$_r$) that detects biomarkers differentially expressed in "majority of" studies (e.g. $> 60\%$ of the studies) and propose a robust order statistic, rth order p-value (rOP), for this hypothesis setting.

The remainder of this chapter is structured as follows. In Section 2.2.2, the rationale and algorithm of rOP is outlined, and methods for parameter estimation are described in section 2.2.3. Section 2.2.4 extends rOP with a one-sided test correction to avoid detection of DE genes with discordant fold change directions across studies. Section 2.3 demonstrates application of rOP to three examples in brain cancer, major depressive disorder (MDD) and diabetes and compares the result with other classical meta-analysis methods. I then further explore power calculation and asymptotic properties of rOP in section 2.4.1, and establish an unexpected but insightful connection of rOP with the traditionally undesirable vote counting method in section 2.4.2. Section 2.5 contains final conclusions and discussions.

2.2 RTH ORDER P-VALUE (ROP)

2.2.1 Hypothesis settings and motivation

I consider the situation when K transcriptomic studies are combined for meta-analysis and each study contains G genes for information integration. Denote by θ_{gk} the underlying true effect size of gene g and study k ($1 \leq g \leq G$, $1 \leq k \leq K$). For a given gene g, I follow the convention of Birnbaum [1954] and Li and Tseng [2011] to consider two complementary hypothesis settings, depending on the pursue of different types of target markers:

$$\mathrm{HS}_A: \quad \left\{ H_0 : \bigcap_k \{\theta_{gk} = 0\} \text{ versus } H_a^{(A)} : \bigcap_k \{\theta_{gk} \neq 0\} \right\}$$

$$\mathrm{HS}_B: \quad \left\{ H_0 : \bigcap_k \{\theta_{gk} = 0\} \text{ versus } H_a^{(B)} : \bigcup_k \{\theta_{gk} \neq 0\} \right\}$$

In HS_A, the targeted biomarkers are those differentially expressed in all studies (i.e. the alternative hypothesis is the intersection event that effect sizes of all K studies are non-zero), while HS_B pursues biomarkers differentially expressed in one or more studies (the alternative hypothesis is the union event instead of intersection in HS_A). Biologically speaking, HS_A is more stringent and more desirable to identify consistent biomarkers across all studies if the combined studies are homogeneous. HS_B, however, is useful when heterogeneity is expected.

For example, if studies analyzing different tissues are combined (e.g. study 1 uses epithelial tissues and study 2 uses blood samples), it is reasonable to identify tissue-specific biomarkers detected by HS$_B$. I note that HS$_B$ is identical to the classical union-intersection test (UIT) [Roy, 1953] but HS$_A$ is different from intersection-union test (IUT) [Berger, 1982, Berger and Hsu, 1996]. In IUT, the statistical hypothesis is in complementary form between null and alternative hypothesis $\{H_0 : \bigcup_k \{\theta_{gk} = 0\}$ versus $H_a : \bigcap_k \{\theta_{gk} \neq 0\}\}$. Solutions for IUT require more sophisticated mixture or Bayesian modeling to accommodate the composite null hypothesis and will be explored in chapter 3.

As discussed in [Tseng et al., 2012], most existing genomic meta-analysis methods target HS$_B$. Popular methods include classical Fisher's method [sum of minus log-transformed p-values; Fisher, 1925], Stouffer's method [sum of inverse-normal-transformed p-values; Stouffer et al., 1949], minP [minimum of combined p-values; Tippett, 1931] and a recently proposed adaptively weighted (AW) Fisher's method [Li and Tseng, 2011]. The random effects model targets a slight variation of HS$_A$, where the effect sizes in the alternative hypothesis are random effects drawn from a Gaussian distribution centered away from zero (but do not guarantee to be all non-zero). The maximum p-value method (maxP) is probably the only method available to specifically target on HS$_A$ so far. By taking the maximum of p-values from combined studies as the test statistic, the method requires that all p-values to be small for a gene to be detected. Assuming independence across studies and that the inferences to generate p-values in single studies are correctly specified, p-values (p_k as p-value of study k) are i.i.d. uniformly distributed in $[0, 1]$. Fisher's statistic ($S^{Fisher} = -2\sum \log p_k$) follows a chi-square distribution with degree of freedom $2K$ (i.e. $S^{Fisher} \sim \chi^2(2K)$) under null hypothesis H_0; Stouffer's statistic ($S^{Stouffer} = \sum \Phi^{-1}(p_k)$, where $\Phi^{-1}(\cdot)$ is the quantile of standard normal distribution) follows a normal distribution with variance K (i.e. $S^{Stouffer} \sim N(0, K)$); minP statistic ($S^{minP} = \min\{p_k\}$) follows Beta distribution with parameters 1 and K (i.e. $S^{minP} \sim Beta(1, K)$); and maxP statistic ($S^{maxP} = \max\{p_k\}$) follows Beta distribution with parameters K and 1 (i.e $S^{maxP} \sim Beta(K, 1)$).

The HS$_A$ hypothesis setting and maxP method is obviously too stringent in light of the generally noisy nature of microarray experiments. When K is large, HS$_A$ is not robust and inevitably detects too few genes. Instead of requiring differential expression in all studies, biologists may be more interested in, for example, "biomarkers that are differentially expressed in more than 70% of the combined studies." Denote by $\Theta_h = \left\{\sum_{k=1}^{K} I(\theta_{gk} \neq 0) = h\right\}$ the situation that exactly h out of K studies are differentially expressed. The new robust hypothesis setting becomes:

$$\text{HS}_r : \quad \left\{ H_0 : \bigcap_k \{\theta_{gk} = 0\} \text{ versus } H_a^{(r)} : \bigcup_{h=r}^{K} \Theta_h \right\},$$

where $r = \lceil p \cdot K \rceil$, $\lceil x \rceil$ is the smallest integer no less than x and p $(0 < p \leq 1)$ is the minimal percentage of studies required to be differentially expressed (e.g. $p = 50\%$ or 70%). I note that HS_A and HS_B are both special cases of the extended HS_r class (i.e. $\text{HS}_A = \text{HS}_K$ and $\text{HS}_B = \text{HS}_1$), but I will focus on large r (e.g. $p > 50\%$) in this chapter and view HS_r as a relaxed or robust form of HS_A.

In the literature, maxP has been used for HS_A and minP has been used for HS_B. An intuitive extension of these two methods for HS_r is to use the rth order p-value (rOP). Before I introduce the algorithm and properties of rOP, I consider below four hypothetical genes to compare Fisher, Stouffer, minP, maxP and rOP to illustrate the motivation of rOP. In the four example genes, gene A has marginally significant p-values ($p = 0.1$) in all five studies; gene B has strong p-value in study 1 ($p = 1e - 20$) but $p = 0.9$ in the other four studies; gene C is similar to Gene A but with much weaker statistical significance ($p = 0.25$ in all five studies); gene D differs from gene C in that studies 1-4 have small p-value ($p = 0.15$) but study 5 has large p-value ($p = 0.9$). Table 1 shows the resulting p-values from five meta-analysis methods that are derived from classical parametric inference in section 2.1. Comparing Fisher and minP in HS_B, minP is sensitive to a study that has very small p-value (e.g. gene B) while Fisher, as an evidence aggregation method, is more sensitive when all or most studies are marginally statistically significant (e.g. gene A). Stouffer behaves similarly to Fisher except that it is less sensitive to the extremely small p-value in gene B. When we turn our attention to HS_A, gene C and gene D cannot be detected by all three of Fisher, Stouffer and minP methods. Gene C can be detected by both maxP and rOP as expected ($p = 0.001$ and 0.015, respectively). For gene D, it cannot be identified by maxP method ($p = 0.59$) but can be detected by rOP at $r = 4$ ($p = 0.002$). Gene D gives a good motivating example that maxP may be too stringent when many studies are combined and rOP provides additional robustness when one or a small portion of studies are not significant. In genomic meta-analysis, genes similar to gene D are common due to noisy nature of high-throughput genomic experiments or when a low quality study is accidentally included in the meta-analysis. Although the types of desired markers (under HS_A, HS_B or HS_r) depend on the biological goal of a specific application, gene A, C and D are normally desirable marker candidates that researchers wish to detect in most situations while gene B is not (unless study specific markers are expected as mentioned in Section 2.1). This toy example motivates the development of a robust order statistic of rOP below.

Table 1: Four hypothetical genes to compare different mea-analysis methods and to illustrate the motivation of rOP (*: p-values smaller than 0.05)

	gene A	gene B	gene C	gene D
Study 1	0.1	1E-20	0.25	0.15
Study 2	0.1	0.9	0.25	0.15
Study 3	0.1	0.9	0.25	0.15
Study 4	0.1	0.9	0.25	0.15
Study 5	0.1	0.9	0.25	0.9
Fisher (HS_B)	0.01*	1E-15*	0.18	0.12
Stouffer (HS_B)	0.002*	0.03*	0.07	0.10
minP (HS_B)	0.41	5E-20*	0.76	0.56
maxP (HS_A)	1E-5*	0.59	0.001*	0.59
rOP ($r = 4$) (HS_r)	5E-4*	0.92	0.015*	0.002*

2.2.2 The rOP method

Below is the algorithm for rOP when the parameter r is fixed. For a given gene g, denote by $S_{g,r} = p_{g(r)}$ where $p_{g(r)}$ is the rth order statistic of p-values $\{p_{g1}, p_{g2}, \ldots, p_{gK}\}$. Under the null hypothesis H_0, $S_{g,r}$ follows a Beta distribution with shape parameters r and $K - r + 1$, assuming the model to generate p-value under the null is correctly specified and all studies are independent. To implement rOP, one may apply this null distribution to calculate p-values for all genes and perform a Benjamini-Hochberg correction [Benjamini and Hochberg, 1995] to control the false discovery rate. A more robust alternative to avoid the aforementioned assumptions is to perform permutation analysis as follows.

Step I. Study-wise p-value calculation before meta-analysis:

1. Considering two-group comparison in each study, compute the moderated t-statistics, t_{gk}, for gene g and study k [Efron et al., 2001, Tusher et al., 2001].
2. Randomly permute group labels in each study B times, and similarly calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \leq g \leq G$, $1 \leq k \leq K$, $1 \leq b \leq B$.
3. Estimate the p-value of t_{gk} as $p_{gk} = \left(\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(t_{g'k}^{(b)} \in R(t_{gk})\right)\right) / (B \cdot G)$, where $R(t_{gk})$ is the rejection region given the threshold t_{gk}. Similarly, given $t_{gk}^{(b)}$, compute its p-value as $p_{gk}^{(b)} = \left(\sum_{b'=1}^{B} \sum_{g'=1}^{G} I\left(t_{g'k}^{(b')} \in R(t_{gk}^{(b)})\right)\right) / (B \cdot G)$

Step II. Calculate rOP statistic:

Compute the rOP statistics: $S_{g,r} = p_{g(r)}$, where $p_{g(r)}$ is the rth order statistic of p-values $\{p_{g1}, p_{g2}, \ldots, p_{gK}\}$. Similarly, $S_{g,r}^{(b)} = p_{g(r)}^{(b)}$ is calculated as the rth order statistic of $\{p_{g1}^{(b)}, p_{g2}^{(b)}, \ldots, p_{gK}^{(b)}\}$.

Step III. Assess p-values and q-values:

1. The p-value of $S_{g,r}$ is calculated as $p(S_{g,r}) = \left(\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(S_{g',r}^{(b)} \leq S_{g,r}\right)\right) / (B \cdot G)$.
2. Estimate π_0, the proportion of null genes, as $\widehat{\pi}_0 = \left(\sum_{g=1}^{G} I\left(p(S_{g,r}) \in A\right)\right) / (G \cdot l(A))$ [Storey, 2002]. Normally I choose $A = [0.5, 1]$ and $l(A) = 0.5$.
3. Estimate the q-value for each gene as $q(S_{g,r}) = \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(S_{g',r}^{(b)} \leq S_{g,r}\right)}{B \cdot \sum_{g'=1}^{G} I\left(S_{g',r} \leq S_{g,r}\right)}$. The concluded DE gene list is $G^{rOP} = \{g : q(S_{g,r}) \leq 0.05\}$.

Remark 1. In step I, both statistics t_{gk} and rejection region $R(t_{gk})$ can be replaced, depending on the experimental design and hypothesis. For example, the F-statistic or Cox proportional hazard model can be used for multi-class or censored data design in each study.

Remark 2. Several forms of penalized or moderated t-statistics have been proposed and shown to outperform traditional t-statistics [Efron et al., 2001, Tusher et al., 2001, Smyth, 2004]. For my algorithm I recommend the penalized t-statistics used in Efron et al. [2001] and Tusher et al. [2001] because it is more robust to small variance estimations.

I note that both minP and maxP are special cases of rOP, but in this chapter I mainly consider properties of rOP as a robust form of maxP (specifically $K/2 \leq r \leq K$).

2.2.3 Selection of r in an application

The selection of r for rOP is obviously data-dependent. The purpose of selecting $r < K$ is to tolerate potentially outlying studies and noise in the data. This noise may come from experimental limitations (e.g. failure of probe design in certain studies, erroneous gene annotation or bias from experimental protocol) or heterogeneous patient cohorts in different studies. Another extreme case may come from inadequate inclusion of a low-quality study into the genomic meta-analysis. In this chapter, I use the empirical data across the entire genome to estimate the best r for a given application. I introduce two approaches for selecting r for rOP. The first is from the number of detected DE genes and the second is based on pathway analysis (a.k.a. gene set analysis) incorporating external biological knowledge.

2.2.3.1 Evaluation based on number of detected DE genes Under a reasonable wild guess, the number of detected DE genes should be maximized when the correct r is chosen in a genomic meta-analysis. Direct application of this intuition is, however, problematic and correction of bias is needed. When the studies combined have many DE genes, meta-analysis through rOP will detect genes when K is small even if the studies combined are totally irrelevant. For example, when $K = r = 2$ and 50% of genes in each study are DE genes, roughly 25% of the genes are DE genes from the meta-analysis, simply by chance. To eliminate this artificial trend, I apply a de-trend method by permutation similar to the GAP statistic [Tibshirani et al., 2001]. Using the original K studies, the number of DE genes detected by rOP using different r ($1 \leq r \leq K$) is first calculated as N_r (under certain false discovery rate threshold, e.g. FDR = 5%). I then randomly permute p-values in each study independently and re-calculate the number of DE genes as $N_r^{(b)}$ in the bth permutation. The permutation is repeated B times and the adjusted number of detected DE genes is defined as $N_r' = N_r - \sum_{b=1}^{B} N_r^{(b)}/B$. In other words, the adjusted number of DE genes is de-trended so that it is purely contributed by the consistent information among studies. The parameter r is selected so that N_r' is maximized or among the largest. I use $B = 100$ in this chapter.

Remark 3. Note that sometimes N_r' could be negative. This often happens when the signal in single study is strong and r is small. However, since I apply rOP for large K and r, the negative value is usually not an issue.

2.2.3.2 Evaluation based on gene set analysis Pathway analysis (a.k.a. gene set analysis) is a statistical tool to infer correlation of differential expression evidence in the data with pathway knowledge (usually sets of genes with known common biological function or interactions) from established databases. In this approach, I hypothesize that the best selection of r will produce a DE analysis result that generates the strongest statistical association in "important" (i.e. disease related) pathways. Normally, the "important" pathways related to a given application are not known. Instead, I adopt a novel selection procedure below. I perform pathway analysis using a large pathway database (e.g. GO, KEGG or BioCarta) and select pathways that are top ranking by aggregated committee decision of different r from rOP. The detailed algorithm is as follows:

Step I. Identification of related pathways: (committee decision by $[K/2] + 1 \leq r \leq K$)

1. Apply rOP method to combine studies and generate p-values for each gene. Run through different r, $[K/2] + 1 \leq r \leq K$.
2. For a given pathway m, apply Kolmogorov-Smirnov test to compare the p-values of genes in the pathway and those outside the pathway. The pathway enrichment p-values are generated as $p_{r,m}$. Its rank among all pathways in a given r is calculated as $R_{r,m} = \text{rank}_m(p_{r,m})$. Small ranks represent strong pathway enrichment for pathway m.
3. The sums of ranks of different r are calculated as $S_m = \sum_{r=[K/2]+1}^{K} R_{r,m}$. The top $U = 100$ pathways with the smallest S_m scores are selected and denoted as M. I treat M as the gold-standard disease-related pathway set.

Step II. Sequential testing of improved pathway enrichment significance:

1. I perform sequential hypothesis testing that starts from $r' = K$ since conceptually I would like to pick r as large as possible. I first perform Wilcoxon signed rank test to test difference of pathway enrichment significance for $r' = K$ and $r' = K - 1$. In other words, I perform two-sample test on paired vectors of $(p_{K,m}; m \in M)$ and $(p_{K-1,m}; m \in M)$ and record the p-value as $\tilde{p}_{K,K-1}$.
2. If the test is rejected (using conventional type I error 0.05), that means reducing from $r = K$ to $r = K - 1$ can generate DE gene list that produce more significant pathway enrichment in M. I will continue to reduce r' by one (i.e. $r' = K - 1$) and repeat the test between $(p_{r',m}; m \in M)$ and $(p_{r'-1,m}; m \in M)$. Similarly, the resulting p-values are recorded as $\tilde{p}_{r',r'-1}$. The procedure is repeated until the test from r' is not rejected. The final r' is selected for rOP. Note that for simplicity, I did not perform p-value correction for multiple comparison or sequentially dependent hypothesis testing here.

2.2.4 One-sided test correction to avoid discordant effect sizes

Methods combining effect sizes (e.g. random or fixed effects models) are suitable to combine studies with binary outcome, in which case the effect sizes are well-defined as the standardized mean difference or odds ratio. Methods combining p-values, however, have advantages to combine studies with non-binary outcomes (e.g. multi-class, continuous or censored data), in which case F-test, simple linear regression or Cox proportional hazard model can be used to generate p-values for integration. On the other hand, p-value combination methods usually combine two-sided p-values in binary outcome data. A gene may be found statistically significant with up-regulation in one study and down-regulation in another study. Such a confusing discordance although sometimes is reflection of biological truth, is often undesirable in most applications. Owen [2009] and Pearson [1934] applied a one-sided test form of Fisher's method to address the possible discordance issue. Two Fisher scores are first obtained from left and right one-sided p-values: $S^{Fisher;L} = -2\sum_{k=1}^{K}\log(\tilde{p}_k)$ and $S^{Fisher;R} = -2\sum_{k=1}^{K}\log(1-\tilde{p}_k)$, where $\tilde{p}_k$ is the left-sided p-value of study k. The one-sided corrected Fisher score is defined as $S^{Fisher;C} = \max\left(S^{Fisher;L}, S^{Fisher;R}\right)$. Below I similarly modify the rOP method for a one-sided corrected form. Denote by $S^{rOP;L} = \tilde{p}_{(r)}$, where $\tilde{p}_{(r)}$ is the rth order statistic of left one-sided p-values $\{\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_K\}$ from K studies. Similarly, $S^{rOP;R} = \tilde{q}_{(r)}$, where $\tilde{q}_{(r)}$ is the rth order statistic of right one-sided p-values $\{\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_K\} = \{1-\tilde{p}_1, 1-\tilde{p}_2, \ldots, 1-\tilde{p}_K\}$ from K studies. The test statistic is defined as $S^{rOP;C} = \min\left(S^{rOP;L}, S^{rOP;R}\right)$. Under the null hypothesis that the one-sided p-values $\{\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_K\}$ are independently and uniformly distributed in $[0,1]$. The null distribution of $S^{rOP;C}$ can be derived using integration by part. Equivalently, the null distribution could also be derived using the following property.

$$\begin{aligned} & \Pr\left(S^{rOP;C} \leq p|H_0\right) = \Pr\left(S^{rOP;L} \leq p \text{ or } S^{rOP;R} \leq p|H_0\right) \\ = \quad & \Pr\left(\sum_{k=1}^{K} I(\tilde{p}_k \leq p) \geq r \text{ or } \sum_{k=1}^{K} I(\tilde{p}_k \geq 1-p) \geq r \middle| H_0\right) \end{aligned}$$

Because $\sum_{k=1}^{K} I(\tilde{p}_k \leq p) \geq r$ and $\sum_{k=1}^{K} I(\tilde{p}_k \geq 1-p) \geq r$ are not mutually exclusive (except when $r \geq [K/2]+1$ and $p \leq 0.5$), the above probability should be calculated differently as follows.

1. For $r \geq [K/2]+1$

   a. If $p \leq 0.5$,
   $\Pr\left(S^{rOP;C} \leq p|H_0\right) = 2F(K-r; K, 1-p)$, where $F(K-r; K, 1-p) = \sum_{i=0}^{K-r} \binom{K}{i}(1-p)^i p^{K-i}$ is the Binomial CDF for having $K-r$ successes in K Bernoulli trails with success probability $1-p$.

   b. If $p > 0.5$,
   $\Pr\left(S^{rOP;C} \leq p|H_0\right) = 1 - \sum_{i=K-r+1}^{r-1} \sum_{j=K-r+1}^{K-i} \frac{K!}{i!j!(K-i-j)!}(1-p)^{i+j}(2p-1)^{K-i-j}$.

2. For $r \leq [K/2]$

   a. If $p \leq 0.5$,
   $\Pr\left(S^{rOP;C} \leq p|H_0\right) = 1 - \sum_{i=0}^{r-1} \sum_{j=0}^{r-1} \frac{K!}{i!j!(K-i-j)!} p^{i+j}(1-2p)^{K-i-j}$.

   b. If $p > 0.5$,
   $\Pr\left(S^{rOP;C} \leq p|H_0\right) = 1$.

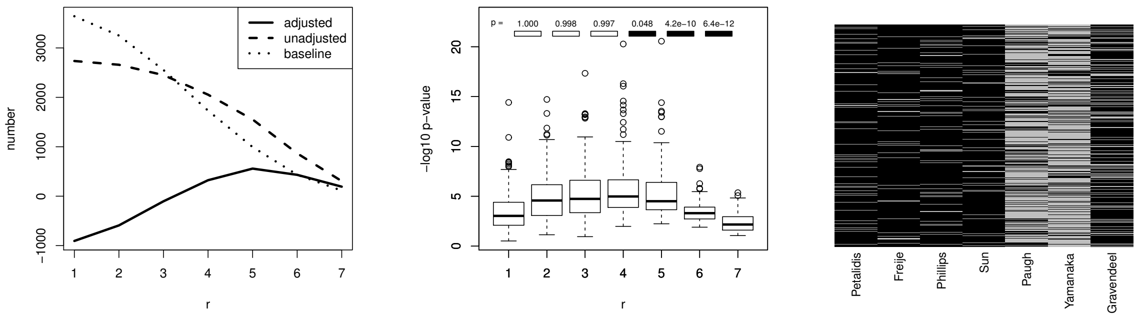
2.3 APPLICATIONS

In this section, I apply rOP as well as other meta-analysis methods to three microarray meta-analysis applications with different strengths of signal and different degrees of heterogeneity. Appendix table 4-6 lists detailed information on seven brain cancer studies, nine major depressive disorder (MDD) studies, and 16 diabetes studies for meta-analysis. I preprocess and normalize the data by standard procedures in each array platform. Probes are matched to the same gene symbols. When multiple probes (or probe sets) match to one gene symbol, the probe that contained the largest variability (i.e. inter-quartile range) was used to represent the gene. After gene matching and filtering, 6,005, 7,577 and 6,645 genes were remained in brain cancer, MDD and diabetes datasets, respectively. The brain cancer studies are collected from GEO database. The major depressive discarder (MDD) studies are obtained from Dr. Etienne Sibille's lab. A random intercept model is applied to each of the studies to get the p-values of single genes adjusted for potential confounders [Wang et al., 2012]. Preprocessed data of 16 diabetes studies described by Park et al. [2009] are obtained from the authors. For studies with multiple groups, I followed the procedure of Park et al. by taking the minimum p-value of all the pairwise comparisons and adjusted for multiple tests.

2.3.1 Application of rOP

I demonstrate the estimation of r for rOP using the two evaluation criteria based on the adjusted number of detected DE gene and gene set analysis in section 2.2.3. In the first dataset, two important subtypes of brain tumors - anaplastic astrocytoma (AA) and glioblastoma multiforme (GBM) - are compared in seven microarray studies. To estimate an adequate r for rOP application, I calculated the unadjusted number, baseline number from permutation and adjusted number of detected DE genes using $1 \leq r \leq 7$ under FDR=5% (Figure 1(a)). The result showed a peak at $r = 5$. For the second estimation method by pathway analysis, boxplots of $-\log_{10}(p)$ (p-values calculated from association of DE gene list with top pathways) versus r are plotted (Figure 1(b)). The sequential Wilcoxon signed rank tests showed that result from $r = 6$ is significantly more associated with pathways than that from $r = 7$ ($p = 6.4e - 12$) and similarly for $r = 5$ versus $r = 6$ ($p = 4.2e - 10$) and $r = 4$ versus $r = 5$ ($p = 0.048$). Combining results from Figures 1(a) and 1(b), I decided to choose $r = 5$ since I wanted to choose r as large as possible and the pathway analysis result of $r = 4$ in Figure 1(b) (although statistically significant) does not greatly improve over the result of $r = 5$. Figure 1(c) shows a heatmap of studies effective in rOP (when $r = 5$) for each detected DE gene (a total of 1,552 DE genes on the rows and seven studies on the columns). For example, if p-values for the seven studies are $(0.13, 0.11, 0.03, 0.001, 0.4, 0.7, 0.15)$, the test statistic for rOP is $S^{rOP} = 0.15$ and the five effective studies that contribute to rOP are indicated as $(1, 1, 1, 1, 0, 0, 1)$. In the heatmap, effective studies are indicated by black color and non-effective studies are in light gray. As shown in Figure 1(c), Paugh and Yamanaka are non-effective studies in almost all detected DE genes, suggesting that the two studies do not contribute to the meta-analysis and may potentially be problematic studies. This finding agrees with a recent quality control assessment result using the same seven studies [Kang et al., 2012]. In my application, AA and GBM patients are compared in all seven studies. I expect to detect biomarkers that have consistent fold change direction across studies and one-sided corrected rOP method is more preferable. Figure 2 shows plots similar to Figure 1 for one-sided corrected rOP. The result similarly concludes that $r = 5$ for the one-sided corrected rOP is the most suitable for this application.

For the second application, nine microarray studies used different areas of post-mortem brain tissues from major depressive disorder patients and control samples (Appendix table 5). Major depressive disorder is a complex genetic disease with largely unknown disease mechanism and regulatory networks. The post-mortem brain tissues usually result in weak signals which make meta-analysis an appealing approach. Figure 3 shows diagnostic plots to estimate r. In Figure 3(a), the maximizer of adjusted DE gene detection is at $r = 7$ ($r = 6$ or 8 are also good choices). For Figure 3(b), the statistical significance improved "from $r = 9$ to $r = 8$" ($p = 5.6e - 14$), "from $r = 8$ to $r = 7$" ($p = 8.7e - 7$) and "from $r = 7$ to $r = 6$" ($p = 0.045$). Combining the two results, I decided to choose $r = 7$ for the rOP method in this application. Figure 3(c) shows the heatmap of effective studies in rOP. No obvious problematic study is observed. The one-sided rOP is also applied (result not shown), good

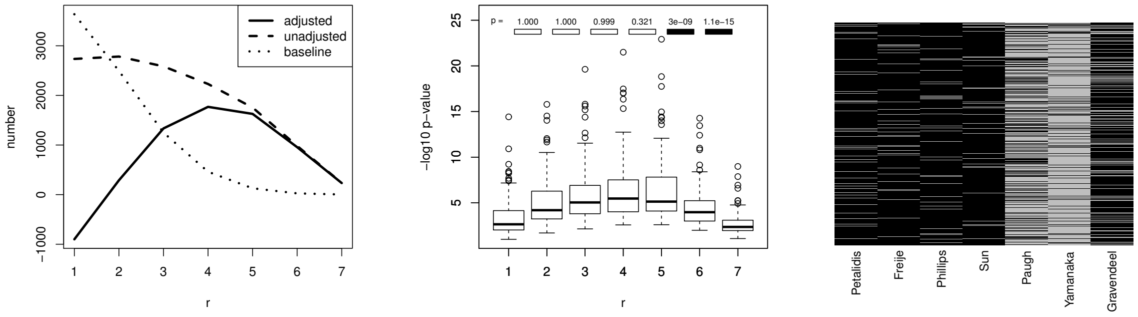


(a) Adjusted and unadjusted number of detected DE genes using different r

(b) Boxplot of $-\log(p)$ for the top 100 pathways using different r. P-values for sequential Wilcoxon signed rank tests are shown

(c) Heatmap shown effective studies of rOP in each genes. Effective studies are shown in black and non-effective ones are in light gray.

Figure 1: Results of brain cancer dataset



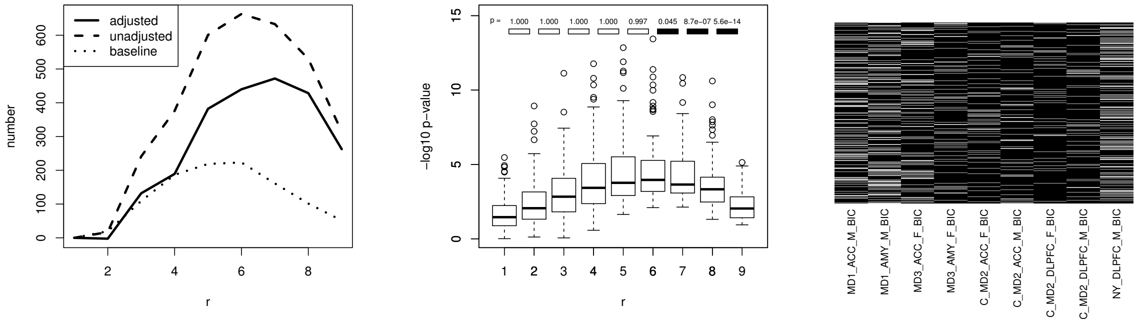
(a) Adjusted and unadjusted number of detected DE genes using different r

(b) Boxplot of $-\log(p)$ for the top 100 pathways using different r

(c) Heatmap shown effective studies of rOP in each genes.

Figure 2: A same figure of Figure 1 showing results of brain cancer dataset, except that one-sided corrected rOP method is used to focus on concordant fold change direction across studies.

selection of r appears between 5 and 7. $r = 7$ is chosen to make it comparable to two-sided rOP result.



(a) Adjusted and unadjusted number of detected DE genes using different r

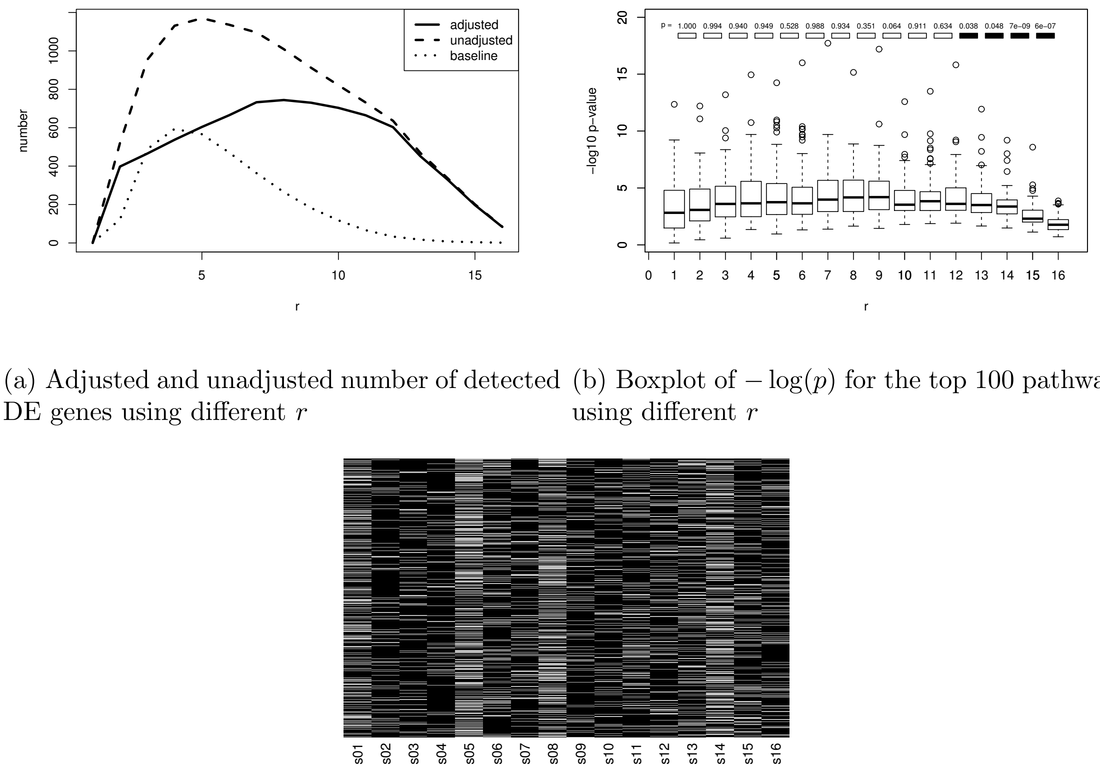
(b) Boxplot of $-\log(p)$ for the top 100 pathways using different r

(c) Heatmap shown effective studies of rOP in each genes

Figure 3: Results of MDD dataset

In the last application, 16 diabetes microarray studies are combined. These 16 studies are very heterogeneous in terms of the organisms, tissues and experimental design (Appendix table 6). Figure 4 shows diagnostic plots to estimate r. Although the number of studies and heterogeneity across datasets are relatively larger than previous two examples, I could still observe similar trends in Figure 4. Specifically, for Figure 4(a), it was shown that $r = 7 \sim 12$ detected higher adjusted number of DE genes. For pathway analysis, results from $r = 11 \sim 12$ are more associated with top pathways. As a result, I decided to use $r = 12$ in this application. It is noticeable that the r selection in these diabetes studies is relatively vague, compared to the other two examples. Figure 4(c) shows the heatmap of effective studies in rOP. Two to four studies appear to be problematic studies but the evidence is not as clear as the brain cancer example in figure 1(c).

I next explored the robustness of rOP by mixing a randomly chosen MDD study into seven brain cancer studies as an outlier and sensitivity analysis. The results in Figure 5 showed that $r = 5$ or 6 may be a good choice (Figure 5(a) and 5(b)). I used $r = 6$ in rOP for this application. Figure 5(c) interestingly shows that the mixed MDD study, together with Paugh and Yamanaka studies, are potentially problematic studies in the rOP meta-analysis.



(a) Adjusted and unadjusted number of detected DE genes using different r

(b) Boxplot of $-\log(p)$ for the top 100 pathways using different r

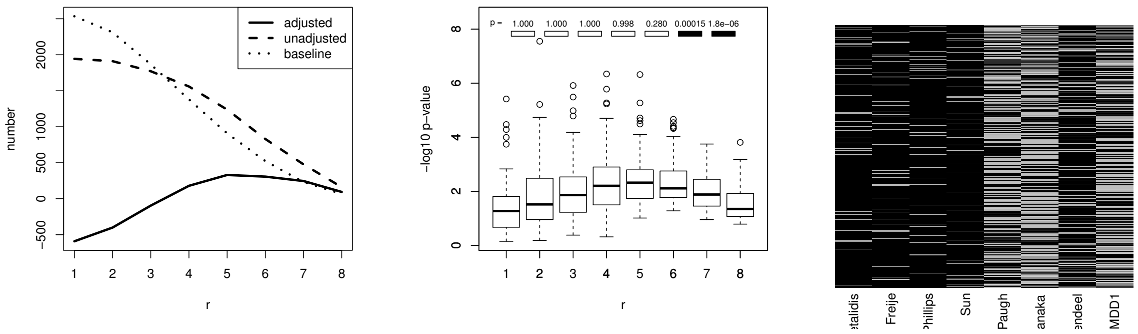
(c) Heatmap shown effective studies of rOP in each genes

Figure 4: Results of diabetes dataset

This result verifies my intuition that rOP is robust to an outlying study and the p-values of the outlying study minimally contribute to rOP statistic.

2.3.2 Comparison of rOP with other methods

I performed rOP using r's determined from section 2.3.1 in four applications (brain cancer, MDD, diabetes, and brain cancer + 1 random MDD) and compared to Fisher's method, Stouffer's method, minP and maxP. Two quantitative measures were used to compare the methods. The first measure compared the number of detected DE genes from each method as a surrogate of sensitivity (although the true list of DE genes is unknown and sensitivity



(a) Adjusted and unadjusted number of detected DE genes using different r

(b) Boxplot of $-\log(p)$ for the top 100 pathways using different r

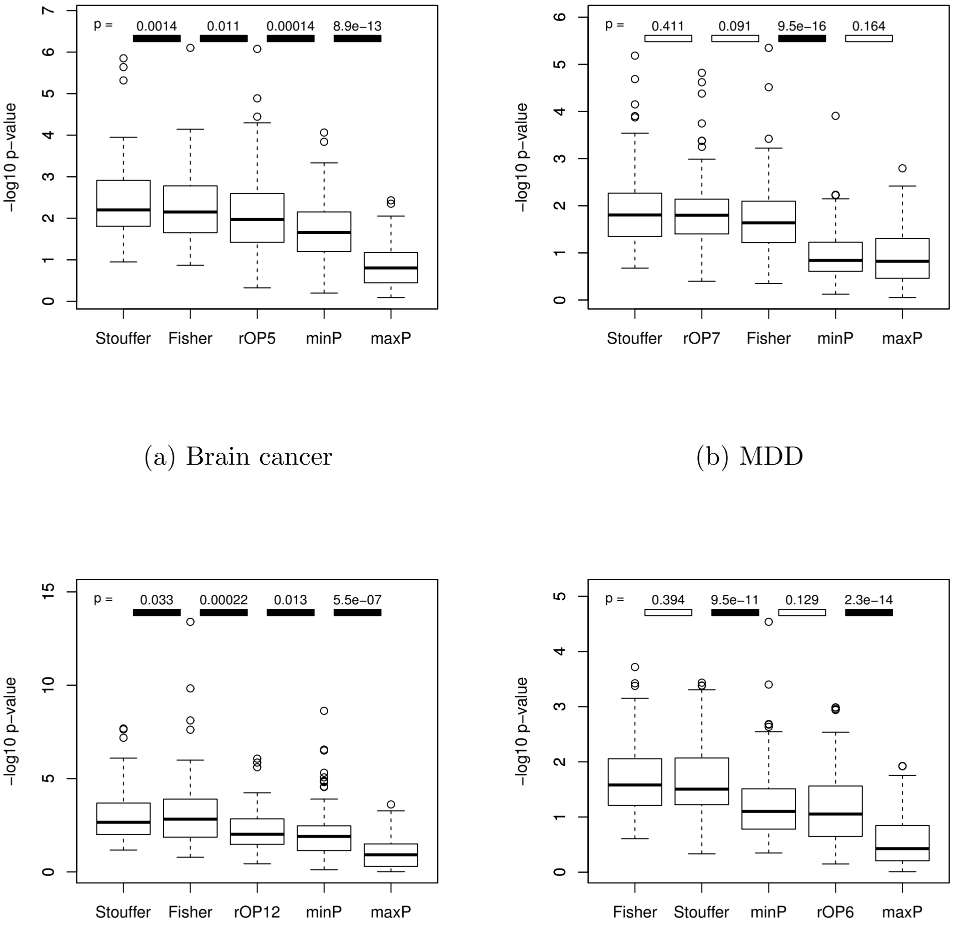
(c) Heatmap shown effective studies of rOP in each genes

Figure 5: Results of Brain Cancer and 1 random MDD dataset

cannot be calculated). The second approach was by pathway analysis; very similar to the method I introduced to select parameter r. However, in order to avoid bias in top pathway selection, single study analysis results were used as the committee to select disease related pathways. KEGG, BioCarta, Reactome and GO pathways were used in the pathway analysis. Wilcoxon signed rank test was then used to test if two methods performed similarly in detecting disease related pathways.

Table 2 shows the number of detected DE genes under FDR=5%. I can immediately observe that Fisher and Stouffer generally detect many more biomarkers because they target HS_B (genes differentially expressed in one or more studies). While minP sometimes has extremely low statistical power (in MDD and diabetes examples) because it requires at least one study with extremely small p-value to be detected. The stringent maxP method detected few numbers of DE genes. rOP detects many more genes than maxP. It identifies about 50 $\sim$ 65% fewer DE genes than the Fisher's and Stouffer's methods but guarantees that the gene list is differentially expressed in majority of studies. I also performed one-sided corrected rOP for comparison. The method detected similar number of DE genes from two-sided rOP, and majority of detected DE genes in two-sided and one-sided rOP were overlapped in the brain cancer example. The result shows that almost all DE genes detected by two-sided rOP had consistent fold change direction across studies. In MDD, one-sided rOP detected much fewer genes than two-sided methods. This implied that many genes related to MDD acted differently in different brain regions and in different cohorts.

Figure 6 shows results of biological association from pathway analysis that were similarly shown in 1(b). The result shows that Fisher and Stouffer seem to generate DE gene list more associated with biological pathways. The rOP method generally performs better than maxP and minP.



(a) Brain cancer

(b) MDD

(c) Diabetes

(d) Brain cancer and 1 random MDD

Figure 6: Comparison of methods by pathway analysis

Table 2: Number of DE gene detected by different methods under FDR=5%

	rOP		Fisher	Stouffer	minP	maxP
	Two-sided	One-sided				
Brain Cancer	1552 ($r = 5$)	1755 ($r = 5$)	3269	2774	2735	305
	overlap=1232					
MDD	633 ($r = 7$)	86 ($r = 7$)	1169	1463	0	314
	overlap=48					
Diabetes	636 ($r = 12$)		1698	1492	1	85
Brain + 1 MDD	830 ($r = 6$)		2359	2008	1943	161

2.4 STATISTICAL PROPERTIES OF ROP

2.4.1 Power calculation of rOP and asymptotic properties

When K studies are combined, suppose r_0 of the K studies have equal non-zero effect sizes and the rest of the $(K - r_0)$ studies have zero effect sizes. That is,

$$\begin{aligned} H_0 &: \quad \theta_1 = \cdots = \theta_K = 0 \\ H_1 &: \quad \theta_1 = \cdots = \theta_{r_0} = \theta \neq 0, \theta_{r_0+1} = \cdots = \theta_K = 0 \end{aligned}$$

Assume for single study, the power function given effect size θ is known as $\Pr(p_i \leq \alpha_0 | \theta)$. I will derive the statistical power of rOP under this simplified hypothesis setting when r_0 is given. Under H_0, the rejection threshold for rOP statistic is $\beta = B_\alpha(r, K - r + 1)$ (the α quantile of a beta distribution with shape parameters r and $K - r + 1$), where the significance level of the meta-analysis is set at α. The power of rejection threshold β under H_1 is $\Pr\left(p_{(r)} \leq \beta | H_1\right) = \Pr\left(\sum_{i=1}^{K} I(p_i \leq \beta) \geq r | H_1\right)$. By definition $\Pr(p_i \leq \beta | \theta_i = 0) = \beta$ and I further denote $\beta' = \Pr(p_i \leq \beta | \theta_i = \theta)$. The power calculation of interest is equivalent to finding the probabilities of having at least r successes in K independent Bernoulli trials, among which r_0 have success probabilities β' and $K - r_0$ have success probabilities β:

$$\Pr\left(p_{(r)} \leq \beta | H_1\right) = \sum_{i=r}^{K} \sum_{j=\max(0,i-K+r_0)}^{\min(i,r_0)} \binom{r_0}{j} \beta'^j (1-\beta')^{r_0-j} \binom{K-r_0}{i-j} \beta^{i-j} (1-\beta)^{K-r_0-i+j}$$

Below I demonstrate some asymptotic properties of rOP.

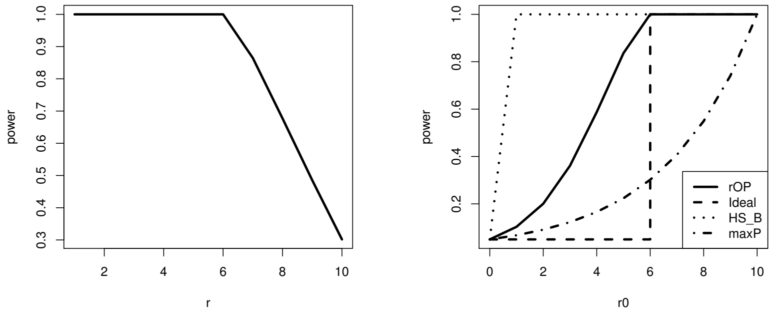
Theorem 2.4.1. *Assume r_0 is fixed. When effect size θ and K are fixed and the sample size of study k $N_k \to \infty$, $\Pr\left(p_{(r)} \leq \beta | H_1\right) \to 1$ if $r \leq r_0$. When $r > r_0$, $\Pr\left(p_{(r)} \leq \beta | H_1\right) \to c(r) < 1$ and $c(r)$ is a decreasing function in r.*

Proof. When $N_k \to \infty$, $\beta' \to 1$. The theorem easily follows from the power calculation formulae. □

Theorem 2.4.1 states that asymptotically if the parameter r in rOP is specified less or equal to r_0, the statistical power converges to 1 as intuitively expected. When specifying r greater than r_0, the statistical power is weakened with increasing r. Particularly, maxP will have weak power. In contrast to Theorem 2.4.1, for methods designed for HS_B (e.g. Fisher's method, Stouffer's method and minP), the power always converges to 1 if $N_k \to \infty$ and $r_0 > 0$. Figure 7(a) shows the power curve of rOP when $K = 10$, $r_0 = 6$ and $N_k \to \infty$.

Lemma 2.4.1. *Assume parameter r used in rOP is fixed. When effect size θ and K are fixed and the sample sizes $N_k \to \infty$, $\Pr\left(p_{(r)} \leq \beta | H_1\right) \to 1$ if $r_0 \geq r$. When $r_0 < r$, $\Pr\left(p_{(r)} \leq \beta | H_1\right) \to c(r_0) < 1$ and $c(r_0)$ is a increasing function in r_0.*

Lemma 2.4.1 takes a different angle from Theorem 2.4.1. When the parameter r used in rOP is fixed, it asymptotically has perfect power to detect all genes that are differentially expressed in r or more studies. It then does not have strong power to detect genes that are differentially expressed in less than r studies. Figure 7(b) shows a power curve of rOP for $K = 10$, $r = 6$ and $N_k \to \infty$ (solid line). I note that the dashed line ($f(r) = 0$ when $0 \leq r_0 < 6$ and $f(r) = 1$ when $6 \leq r_0 \leq 10$) is the hypothetical perfect method for HS_r (i.e. it detects all genes that are differentially expressed in r or more studies but does not detect any gene that are differentially expressed in less than r studies). Methods like Fisher, Stouffer and minP target on HS_B and their power is always 1 asymptotically when $r_0 > 0$. The maxP method has perfect asymptotic power when $r_0 = K = 10$ but has weak power when $r_0 < K$.



(a) $r_0 = 6$, $r = 1 \sim 10$

(b) $r = 6$, $r_0 = 0 \sim 10$

Figure 7: Power of rOP method when $N_k \to \infty$, $K = 10$

2.4.2 Connection with vote counting

Vote counting has been used in many meta-analysis applications due to its simplicity while it has been criticized as being problematic and statistically inefficient. Hedges and Olkin [1980] showed that the power of vote counting converges to 0 when many studies of moderate effect sizes are combined (see Theorem 2.4.2). We, however, surprisingly found that rOP has a close connection with vote counting and rOP can be viewed as a generalized vote counting with better statistical properties. There are many vote counting variations in the literature. One popular approach is to count the number of studies that have p-values smaller than α. I define this quantity as

$$r = f(\alpha) = \sum_{k=1}^{K} I\left\{p_k < \alpha\right\} \tag{2.1}$$

and define its related proportion as $\pi = E(r)/K$. The hypothesis testing used is

$$\begin{cases} H_0 : \pi = \pi_0 \\ H_A : \pi > \pi_0 \end{cases}$$

where $\pi_0 = 0.5$ is often used in applications. Under null hypothesis, $r \sim BIN(\alpha, K)$ and $\pi = \alpha$. The rejection region can be established. In the vote counting procedure, α and π_0 are two preset parameters and the inference is made on the test statistic r.

In the rOP method, I view equation (2.1) from another direction. I can easily show that if I solve equation (2.1) to obtain $\alpha = f^{-1}(r)$, the solution will be $\alpha \in \big[p_{(r)}, p_{(r+1)}\big)$ and one may choose $\alpha = p_{(r)}$ as the solution. In other words, rOP presets r as a given parameter and the inference is based on the test statistic $\alpha = p_{(r)}$.

The two theorems below show the criticized property of vote counting and show rOP does not have this issue.

Theorem 2.4.2. *When all K studies have equal effect sizes ($\theta_1 = \cdots = \theta_K = \theta \neq 0$) and the effect sizes are moderate (so that the single study power $\Pr(p_k < \alpha | \theta_k = \theta) < \pi_0$), the power of vote counting converges to 0 when $K \to \infty$.*

Proof. Denote by $\pi = \Pr(p_k < \alpha | \theta_k = \theta)$. Under the alternative hypothesis $\theta_1 = \cdots = \theta_K = \theta$, $r \sim BIN(\pi, K)$ and $r/K \to \pi$ as $K \to \infty$. Since $\pi < \pi_0$, $\Pr(\text{reject } H_0 | H_A) \to 0$. □

Theorem 2.4.3. *When all K studies have equal effect sizes ($\theta_1 = \cdots = \theta_K = \theta \neq 0$) and the effect sizes are moderate but informative (so that the single study power satisfies $\Pr(p_k < \alpha | \theta_k = \theta) > \alpha$, the power of rOP for r under significance level α converges to 1 when $K \to \infty$ and $r/K = c < 1$.*

Proof. Denote $\alpha_0 = B_\alpha(r, K - r + 1)$ (quantile of $Beta(r, K - r + 1)$). α_0 is the critical value for a single study. When $K \to \infty$, $Beta(r, K - r + 1)$ has mean converges to $r/K = c$ and variance converges to 0. As a result, $\alpha_0 \to c$ as $K \to \infty$. Assume $\Pr(p_k < \alpha_0 | \theta_k = \theta) = \alpha_0 + \epsilon$, and $\epsilon > 0$. Denote $m = \sum_{k=1}^{K} I(p_k < \alpha_0 | \theta_k = \theta)$. For $K \to \infty$, by the law of large numbers, $(m - r)/K = m/K - r/K \xrightarrow{p} \alpha_0 + \epsilon - c \xrightarrow{p} \epsilon > 0$. Therefore $\Pr(m \geq r) \to 1$ and the power of rOP $= \Pr(\text{reject } H_0 | H_A) \to 1$. □

2.5 DISCUSSION

In this chapter, I proposed a general class of order statistics of p-values, called r*th* order p-value (rOP), for genomic meta-analysis. The family of statistics includes traditional maximum p-value (maxP) and minimum p-value (minP) statistics that target on DE genes in "all studies" (HS_A) or "one or more studies" (HS_B). I extended HS_A to a robust form that detects DE genes in "majority of studies" (HS_r) and developed the rOP method for this purpose. The new robust hypothesis setting has an intuitive interpretation and is more adequate in genomic applications where unexpected noise is common in data. I developed the algorithm of rOP for microarray meta-analysis and proposed two methods to estimate r in applications. Under "two-class" comparisons, I proposed a one-sided-test corrected form of rOP to avoid detection of discordant expression change across studies (i.e. significant up-regulation in some studies but down-regulation in other studies). Finally, I performed power analysis and examined asymptotic properties of rOP to demonstrate appropriateness of rOP for HS_r over existing methods such as Fisher, Stouffer, minP and maxP. I further showed a surprising connection between vote counting and rOP and that rOP can be viewed as a generalized vote counting with better statistical property. Applications of rOP to three examples in brain cancer, major depressive disorder (MDD) and diabetes showed better performance of rOP over maxP in terms of detection power (number of detected markers) and biological association by pathway analysis.

There are two major limitations of rOP. Firstly, rOP is for HS_r but not the intersection-union test (IUT) setting (i.e. composite null hypothesis; see Section 2.2.1). Thus, it has weaker power to exclude markers that are differentially expressed in minor number (smaller than r) of studies since the null of HS_r is "differential expression in zero studies". One solution to improve it (which is addressed in chapter 3) is by Bayesian modeling of p-values with a family of beta distributions [Erickson et al., 2009]. Secondly, selection of r may not be conclusive from the two methods I proposed; especially the external pathway information may be prone to errors and may not be informative to the data. But since choosing slightly different r usually gives similar results, the problem is alleviated. I have tested a different approach by adaptively choosing the best gene-specific r that generates the best p-value.

The result is, however, not stable and the gene-specific parameter r is hard to interpret in applications.

Although many meta-analysis methods have been proposed and applied to microarray applications, it is still not clear which method enjoys better performance under what condition. Selection of an adequate (or best) method heavily depends on the biological goal (as the hypothesis settings illustrated in this chapter) and the data structure. In this chapter, I stated a robust hypothesis setting (HS_r) that is commonly targeted in biological applications (i.e. identify markers statistically significant in majority of studies) and developed an order statistic method (rOP) for the problem. The three applications covered “cleaner” data (brain cancer) to “noisier” data (complex genetics in the two diseases MDD and diabetes) and rOP performed well in all three examples. I expect that this order statistic methodology will find many future applications in genomic research and traditional univariate meta-analysis.

3.0 A SEMIPARAMETRIC MIXTURE MODEL APPROACH FOR GENOMIC META-ANALYSIS

3.1 BACKGROUND

In chapter 2, I proposed HS$_r$ as a robust form of HS$_A$. HS$_r$ targets detecting genes that are DE in "majority of" combined studies. And I proposed rOP to test HS$_r$. However in HS$_r$ the null hypothesis and alternative hypothesis are not complementary. I can extend the HS$_r$ to a robust form of intersection-union test (IUT, $H_0 : \bigcup_k \{\theta_{gk} = 0\}$ versus $H_a : \bigcap_k \{\theta_{gk} \neq 0\}$, where θ_{gk} denotes the effect size of gene g in study k) with complementary null and alternative hypotheses.

As discussed in section 2.2.1, IUT involves dealing with a composite null hypothesis and requires more sophisticated Bayesian modeling. Similarly, I can extend HS$_r$ and use a Bayesian modeling strategy to test it.

In the rest of this chapter, I first propose the new hypothesis setting in section 3.2. Then a novel semiparametric mixture model approach is proposed in section 3.3. The Bayesian interpretation of the model is explored. And an expectation-maximization algorithm (EM algorithm) is proposed to iteratively estimate the parameters of interest. Compared to the hypothesis testing framework of frequentist methods, a Bayes factor statistic is proposed based on the posterior distribution. In section 3.4 my model is evaluated by both simulation and real data analysis. The results are summarized, and potential advantages and limitations of this method are discussed in section 3.5.

3.2 MOTIVATION

In chapter 2, I introduced a robust meta-analysis hypothesis setting HS_r. It is easy to show that H_0 and $H_a^{(r)}$ are not complementary in HS_r - situations in which gene g is DE in fewer than r but more than 0 studies are in neither H_0 nor $H_a^{(r)}$. This is sometimes unfavorable because in reality, I could not conclude whether $H_a^{(r)}$ is true even though H_0 is rejected. To address this problem, I further extend HS_r to HS'_r as

$$\mathrm{HS}'_r : \quad \left\{ H_0^{(r)} : \bigcup_{h=0}^{r-1} \Theta_h \text{ versus } H_a^{(r)} : \bigcup_{h=r}^{K} \Theta_h \right\}$$

Notice that this hypothesis setting degenerates to IUT when $r = K$. However, HS'_r is a much more complicated problem than traditional hypothesis tests since the null distribution is contributed by both DE and non-DE genes. Traditional methods which combine p-values are anti-conservative for HS'_r because their null distributions are derived by assuming no DE gene exists. Although people realized that this problem could potentially be addressed by Bayesian approaches [Erickson et al., 2009] which borrow information from other genes within the same study, no established method has been proposed.

Within a single study k, the observed p-value x_{gk} for gene g could be thought of as sampled from two different distributions: the null distribution f_{0k} if gene g is not DE and another distribution f_{1k} if gene g is DE. Assuming the test scheme used in study k is performed correctly, f_{0k} would be exactly $U(0, 1)$. Because f_{1k} is a distribution on $[0, 1]$, it looks natural to model the p-value distribution as a mixture model of one uniform distribution and several beta distributions. Allison et al. [2002] used this mixture model to model the distribution of p-values in single studies. Pounds and Morris [2003] also applied a beta uniform mixture (BUM) model to model the p-value distribution and control the false discovery rate for a single study. However, these approaches have a drawback that the mixture proportion of the beta component can not be directly interpreted as the prior probability of a gene being DE because the distribution of DE genes is unknown and usually different from a beta distribution mixture even though the overall mixture may fit pretty well to the mixed distribution of both DE and non-DE genes. To extend this mixture model approach to the meta-analysis setting, I should also consider the correlations and differences between studies. Hereby, I propose a novel hierarchical Bayesian mixture model approach that can address all those issues mentioned.

3.3 HIERARCHICAL BAYESIAN MIXTURE MODEL

Similar to the mixture model approaches used in single studies, I can also assume that the observed p-values in each study k are from two distributions - f_{0k} and f_{1k}. By further assuming the statistical tests performed in single studies are correct, we know that f_{0k} are the same to uniform distribution ($U(0,1)$) for any k. However, as stated in section 3.1, f_{1k} can not be easily modeled using a mixture of beta distributions. Other parametric forms (e.g. polynomial) of f_{1k} are also difficult to pursue. And no parametric assumption is guaranteed to work without knowing the real distribution of the underlying effect sizes for all the genes and the particular experimental design for every study k. Therefore, I will model f_{1k} using non-parametric methods.

In order to calculate the posterior probability of H_0 being false, I need to assume the prior as well. Here I assume that the prior probability that gene g is DE in study k is p_{gk}. Intuitively, p_{gk} should depends on the study that it is from, and also the gene being tested. To account for the dependencies within the same genes and same studies, I assume that p_{gk} is determined by a logistic model with parameters β_g and w_k. Further distribution assumptions could be made on β_g and w_k. In this model, I assume that β_g and w_k are from two normal distributions.

Collectively, the observed p-values x_{gk} could be assumed to be generated from the following procedure, where y_{gk} denotes the underlying truth whether gene g is DE in study k.

Step I. Generate β_g and w_k from normal distributions:

$$\begin{aligned} \beta_g &\sim N(0, \sigma^2) \\ w_k &\sim N(0, \tau^2) \end{aligned}$$

Step II. Generate y_{gk} based on β_g and w_k:

$$\begin{aligned} y_{gk} &\sim Bin(1, p_{gk}) \\ \text{logit}(p_{gk}) &= \beta_g + w_k + c \end{aligned}$$

Step III. Generate x_{gk} given y_{gk} and f_{0k}, f_{1k}:

$$\begin{cases} x_{gk} \sim f_{0k}, \text{ if } y_{gk} = 0 \\ x_{gk} \sim f_{1k}, \text{ if } y_{gk} = 1 \end{cases}$$

This generative process could also be represented in graph (figure 8).

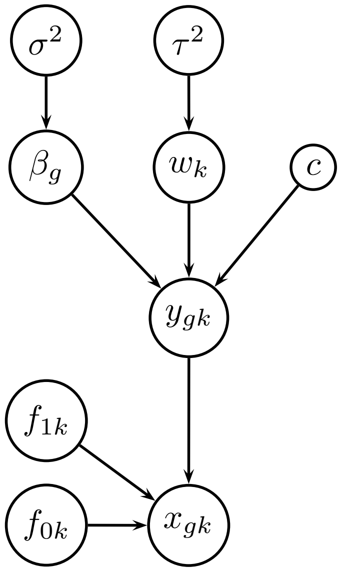


Figure 8: Bayesian generative model for x_{gk}

3.3.1 Mixture model with nonparamertic density estimation

Because of the difficulties of making parametric assumptions on f_{1k}, non-parametric approaches appear to be a more appealing method. Efron and Tibshirani [2002] used a spline smoother to model the empirical distribution of Wilcoxon rank sum statistics observed in a microarray study and utilized the known null distribution to estimate the mixing probability of DE and non-DE genes. The distribution of Wilcoxon rank sum statistic could be easily estimated because it is discrete. To estimate the continuous distribution of p-values, I can use the very similar approach but kernel density estimator instead of the spline smoother. Given a kernel function $K(\cdot)$ and the observed values $x_1, x_2, \ldots, x_n$, the kernel density using fixed bandwidth h could be estimated as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.1}$$

When applied to the p-value distribution, the density should be bounded in range $[0, 1]$. The kernel functions can be modified to reflect back into the range at the boundaries. In equation 3.1, $K\left(\frac{x-x_i}{h}\right)$ should be replaced by $\left\{K\left(\frac{x-x_i}{h}\right) + K\left(\frac{x+x_i}{h}\right) + K\left(\frac{2-x-x_i}{h}\right)\right\}$ accordingly [Silverman, 1986]. In practice, it is also desirable to make the bandwidth h variable when the distribution is too sharp. Usually h is adjusted adaptively such that $1/h$ is proportional to the local density [Terrell and Scott, 1992, Ghosh and Bandyopadhyay, 2006]. This variable kernel density estimation could be performed using R package `locfit` [Loader, 1999]. Kernel density can also be thought as an approximate maximum likelihood estimation (MLE), because it is the continuous approximation of histogram which is MLE for the multinomial distribution probabilities that the variable of interest falls into each bin.

3.3.2 Logistic model with ridge panelty

Based on the model specification, the distribution of y_{gk} can be modeled using the following model:

$$\begin{aligned} y_{gk} &\sim Bin(1, p_{gk}) \\ \text{logit}(p_{gk}) &= \beta_g + w_k + c \end{aligned}$$

where β_g is the main effect of gene g, w_k is the main effect of study k and c is the intercept term which reflects the overall probability for a gene being DE. However, since in general high-throughput genomic analysis there are usually thousands of genes analyzed, the total number of parameters (mainly β_g) will be too large, and the estimation of the parameters will be unstable. Also, though this parameterization makes good interpretability for the parameters, the design matrix is singular and has identifiable problem. So ridge penalty is introduced to regularize the likelihood function and stabilize the parameter estimations [Hoerl and Kennard, 1970, Le Cessie and Van Houwelingen, 1992]. The regularized loss function can be written as:

$$J(\boldsymbol{\beta}, \boldsymbol{w}, c) = -\log L(\boldsymbol{\beta}, \boldsymbol{w}, c|\boldsymbol{y}) + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{w}\|_2^2 \tag{3.2}$$

where λ_1 and λ_2 are the penalty parameters for gene and study effects respectively. And the parameters can be estimated by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}) = \arg\min J(\boldsymbol{\beta}, \boldsymbol{w}, c) \tag{3.3}$$

Remark 4. The ridge regression model can also be interpreted as the maximum a posteriori (MAP) estimation assuming that both β_g and w_k have normal prior distributions as

$$\begin{aligned} \beta_g &\sim N(0, \sigma^2) \\ w_k &\sim N(0, \tau^2) \end{aligned}$$

where $\sigma^2 = 1/(2\lambda_1)$ and $\tau^2 = 1/(2\lambda_2)$.

3.3.3 Model fitting using expectation-maximization algorithm

The full model can be fitted using iterative expectation-maximization (EM) algorithm. It is worth to mention that Dempster et al. [1977] had already extended EM-algorithm to MAP estimations in their original paper, although traditional EM-algorithm is mostly used on MLE estimations.

- *E-step:*

  Calculate the expectation of y_{gk} given $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{w}^{(s)}$, $c^{(s)}$ and $f_{1k}^{(s)}$ - the parameters estimated at round s:

$$E(y_{gk}) = \frac{\frac{\exp(\beta_g^{(s)}+w_k^{(s)}+c^{(s)})}{1+\exp(\beta_g^{(s)}+w_k^{(s)}+c^{(s)})} f_{1k}^{(s)}(x_{gk})}{\frac{1}{1+\exp(\beta_g^{(s)}+w_k^{(s)}+c^{(s)})} + \frac{\exp(\beta_g^{(s)}+w_k^{(s)}+c^{(s)})}{1+\exp(\beta_g^{(s)}+w_k^{(s)}+c^{(s)})} f_{1k}^{(s)}(x_{gk})} \tag{3.4}$$

- *M-step:*

  Update the parameter estimation given $E(y_{gk})$.

  - Update $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{w}^{(s+1)}$ and $c^{(s+1)}$: the MAP point estimations of these parameters can be obtained from equation 3.3 by replacing y_{gk} with $E(y_{gk})$. Optimization can be done using Newton-Raphson's method.
  - Update $f_{1k}^{(s+1)}$:

$$f_{1k}^{(s+1)}(x) = \frac{\sum_{g=1}^{G} E(y_{gk}) \left\{ K\left(\frac{x-x_g}{h_g(x)}\right) + K\left(\frac{x+x_g}{h_g(x)}\right) + K\left(\frac{2-x-x_g}{h_g(x)}\right) \right\}}{\sum_{g=1}^{G} E(y_{gk}) h_g(x)}$$

We can see that y_{gk}'s are sufficient statistics for both the logistic regression part and the mixture model part of the model. And in the M-step, all the parameter estimation used are either MAP or MLE (approximate). Therefore the EM algorithm will always converge in general.

3.3.4 Posterior inference

After model fitting, I can calculate the posterior distribution of y_{gk} using the same formula shown in equation 3.4 because $\Pr(y_{gk} = 1|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1) = E(y_{gk}|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$. To test HS'_r, I define $r_g = \sum_{k=1}^{K} y_{gk}$. The hypothesis testing then becomes deciding whether $r_g \geq r$. This question can be addressed by examining the posterior distribution of r_g. $\Pr(r_g|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$ can be calculated from the posterior distribution of y_{gk} using dynamic programming. Define $\gamma_g(i,j) = \Pr(\sum_{1\leq k\leq j} y_{gk} = i|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$, where $0 \leq j \leq K$. The probability of interest becomes $\Pr(r_g = r_0|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1) = \gamma_g(r_0, K)$. The dynamic programming can be performed as following:

Step I. Initialization: set $\gamma_g(0,0) = 1$, $\gamma_g(i,0) = 0$ and $\gamma_g(0,j) = \prod_{k=1}^{j}[1 - \Pr(y_{gk} = 1|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)]$, for all $1 \leq i \leq K$ and $1 \leq j \leq K$.

Step II. Extension: calculate $\gamma_g(i,j) = \gamma_g(i-1, j-1)\Pr(y_{gj} = 1|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1) + \gamma_g(i, j-1)[1 - \Pr(y_{gj} = 1|\boldsymbol{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)]$, for all for all $1 \leq i \leq K$ and $1 \leq j \leq K$.

To make a Bayesian decision on accepting $H_0^{(r)}$ or $H_a^{(r)}$, I further define the loss of making a type I error as l_1 and the loss of making a type II error as l_2. Then the decision should be made such that the posterior expected loss is minimized. The expected loss can be calculated using the following equations:

$$\begin{aligned}\rho_g(H_0^{(r)}) &= l_2 \sum_{i=r}^{K} \gamma_g(i, K) \\ \rho_g(H_a^{(r)}) &= l_1 \sum_{i=0}^{r-1} \gamma_g(i, K)\end{aligned}$$

I will accept $H_0^{(r)}$ if $\rho_g(H_0^{(r)}) < \rho_g(H_a^{(r)})$, and accept $H_a^{(r)}$ otherwise. This is equivalent to accepting $H_0^{(a)}$ if $\sum_{i=r}^{K} \gamma_g(i,K) / \sum_{i=0}^{r-1} \gamma_g(i,K) < l_1/l_2$. I can further denote $\phi_g = \sum_{i=r}^{K} \gamma_g(i,K) / \sum_{i=0}^{r-1} \gamma_g(i,K)$ and $\eta = l_1/l_2$. The decision rule becomes comparing ϕ_g to η. It can be shown easily that ϕ_g is the Bayes factor comparing these two hypothesis. And *eta* reflects the preference toward $H_0^{(r)}$: the larger η is, the more conservative the decision is. Generally, without particular preference toward any of the two competing hypotheses, I can choose $\eta = 1$ for convenience.

3.4 APPLICATIONS

3.4.1 Simulations

To evaluate this method, we simulated a dataset in the following way, such that there are 300 genes that are DE in r studies, for each r between 1 and K.

Step I. Set $G = 5000$ and $K = 10$.

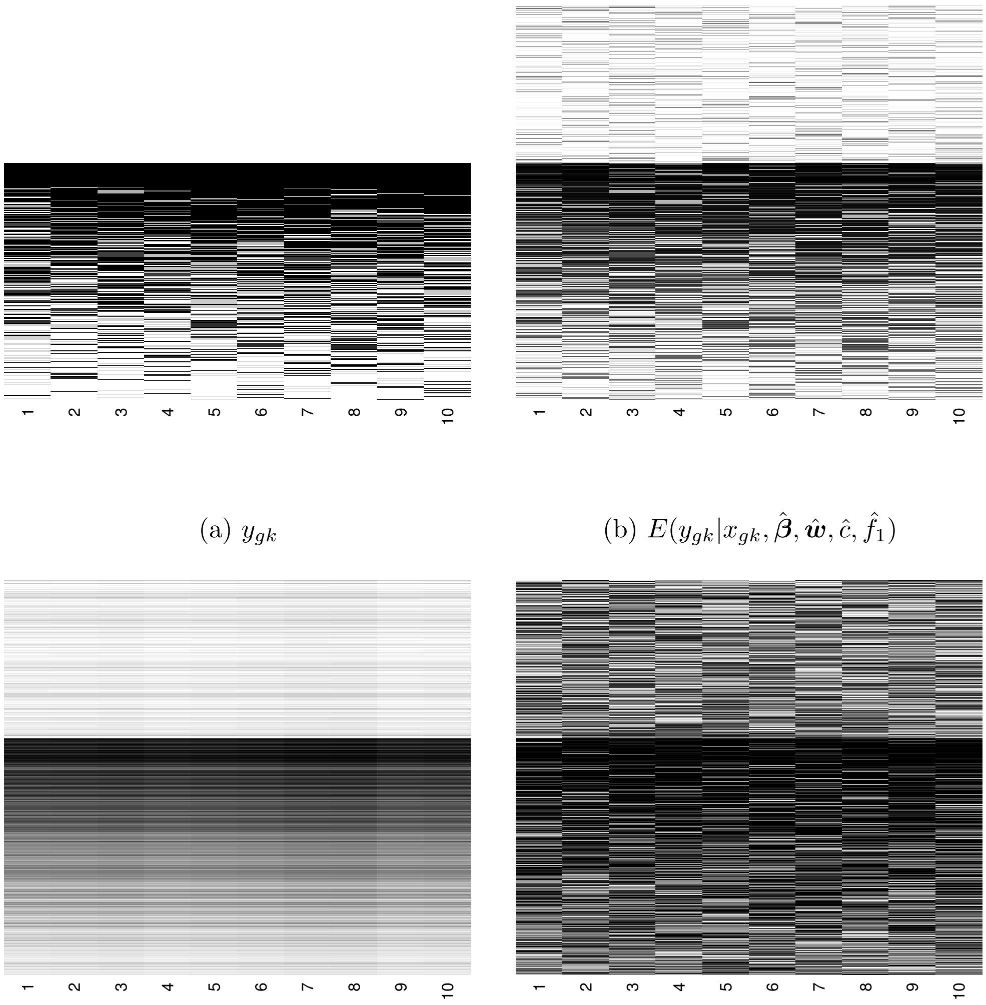
Step II. Given $1 \leq r \leq K$, for every vector $(y_{g1}, \ldots, y_{gK})$, where $300(r-1)+1 \leq g \leq 300r$, randomly sample r elements and set them to 1, and set others to 0.

Step III. Sample $\theta_{gk} \sim U(1,5)$. If $y_{gk} = 1$, sample $t_{gk} \sim T_{df=20}(\theta_{gk})$ and calculate $x_{gk} = 1 - F_{T,df=20}(t_{gk})$; otherwise, sample $x_{gk} \sim U(0,1)$.

Remark 5. Here I directly sampled the t-statistics and calculated their p-values instead of sampling from the raw microarray data. These two methods are equivalent and should give the same results. Note θ_{gk} is the noncentral parameter for the noncentral t-distribution and F_T is the CDF of the standard t-distribution.

In my analysis, I fixed the parameter at $\lambda_1 = 1$ and $\lambda_2 = 1$. This parameter setting roughly indicates that the odds ratio between the top 5% genes (studies) and bottom 5% genes (studies) is 10. The EM algorithm converges in 20 iterations (relative change of log-likelihood smaller than 1e-3).

Figure 9 shows the simulation results using heatmaps. Figure 9(a) is the underlying truth of y_{gk} (black indicates DE). Figure 9(b) shows the posterior distribution of y_{gk} given the parameters estimated using this model and the observation x_{gk}. Though not perfect, this looks similar to the patterns shown in figure 9(a). The total absolute loss is $\sum_{g,k} \|y_{gk} - E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)\| = 7528$ (histograms shown in figure 10). Figure 9(c) demonstrates, if x_{gk} is missing, the prior distribution of y_{gk} using the parameters estimated. I can see that the estimated prior distribution of y_{gk} is not too different from the true value, it is possible to use this prior to estimate the distribution of r_g, which can make the algorithm tolerant to missing values.



(a) y_{gk}

(b) $E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$

(c) $E(y_{gk}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$

(d) $1 - x_{gk}$

Figure 9: Heatmap results of simulation: each column is a study and each row is a gene.

After fitting the model using EM algorithm. For each $r \in \{1, \ldots, K\}$, I calculated ϕ_g for hypothesis setting HS'_r. For each gene g, the null hypothesis $H_0^{(r)}$ is accepted if $\phi_g < \eta$. The number of genes detected for each r and the corresponding FDR is shown in table 3 by setting $\eta = 1$. From the results, I can see that traditional combining p-value methods all lost control of FDR as r increases. Even the most conservative maxP method has FDR=62% when $r = 10$. However, my mixture model method has the FDR well-controlled except for

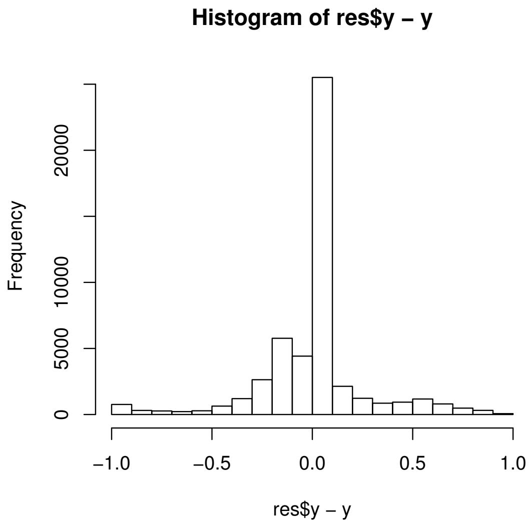


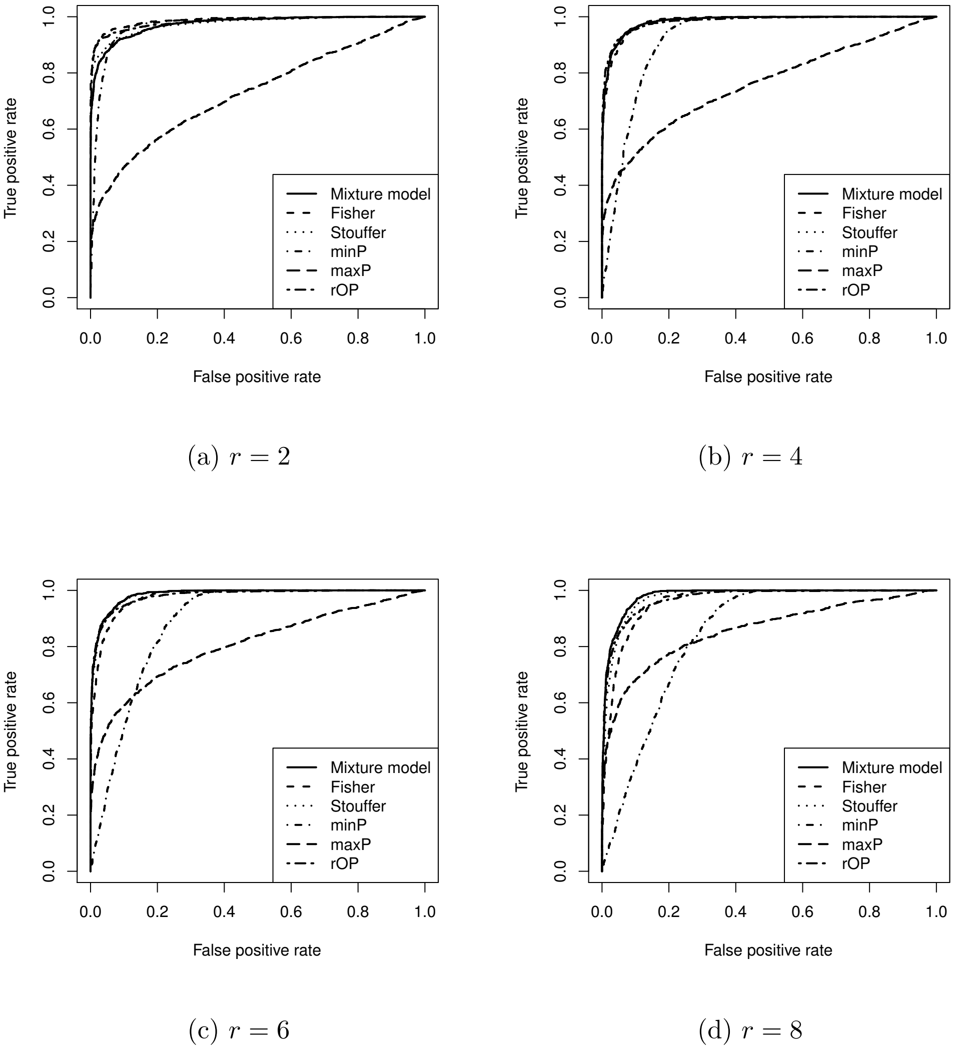
Figure 10: Histogram of residues of simulation data

$r = 1$ (FDR=28.6%). Since $\eta = 1$ is just an arbitrary cutoff, increasing η will give us better control of FDR.

Then I change η and plotted the ROC curve using the known y_{gk} for different HS'_r as $r \in \{1, \ldots, K\}$. The results are compared to Fisher's method, Stouffer's method, minP, maxP and rOP using ROC curves. Results for $r = 2, 4, 6, 8$ are shown in figure 11. In the ROC curves, my method is always among the best methods compared. Fisher's method and Stouffer's method perform well in the ROC curves. However, it is not easy to find appropriate cutoffs for these methods according to different HS'_r. Although anti-conservative, rOP is not a bad method because it performs well in the ROC curve and we can change r for different target hypothesis settings. As expected, minP performs well for HS'_r with small r and maxP performs in the opposite way. No matter how the ROC curves look, traditional methods all suffer from the difficulty of finding cutoffs because of the composite null hypothesis.

Table 3: Number of genes detected and FDR

Hypothesis Setting	Mixture Model	rOP	Fisher	Stouffer	minP	maxP
$r = 1$	4167	2752	2733	2479	2752	690
FDR(%)	28.6	2.2	2.1	2.3	2.2	1.9
$r = 2$	2927	2578	2733	2479	2752	690
FDR(%)	12.4	3.2	5.6	4.0	7.6	2.6
$r = 3$	2372	2456	2733	2479	2752	690
FDR(%)	6.7	7.5	13.4	8.6	15.9	3.0
$r = 4$	1957	2294	2733	2479	2752	690
FDR(%)	5.3	12.6	23.5	17.1	25.3	4.5
$r = 5$	1592	2090	2733	2479	2752	690
FDR(%)	5.0	17.3	34.2	27.9	35.3	7.0
$r = 6$	1225	1889	2733	2479	2752	690
FDR(%)	4.3	23.8	45.1	39.7	45.8	10.6
$r = 7$	876	1680	2733	2479	2752	690
FDR(%)	4.5	30.8	56.1	51.6	56.6	15.9
$r = 8$	557	1440	2733	2479	2752	690
FDR(%)	6.1	40.8	67.1	63.7	67.4	26.2
$r = 9$	260	1107	2733	2479	2752	690
FDR(%)	4.6	49.1	78.0	75.8	78.2	38.7
$r = 10$	77	690	2733	2479	2752	690
FDR(%)	10.4	62.0	89.0	87.9	89.1	62.0



(a) $r = 2$

(b) $r = 4$

(c) $r = 6$

(d) $r = 8$

Figure 11: ROC curves of meta-analysis methods for HS'_r with different r's

3.4.2 Real data analysis

I applied my method to the brain cancer datasets analyzed in chapter 2. I followed the same data preprocessing and analysis procedures described in the previous chapter to obtain the single study p-values. The result is shown in Figure 12. Figure 12(a) shows the heatmap of the posterior distribution of y_{gk} based on the fitted parameters and the observed p-values.

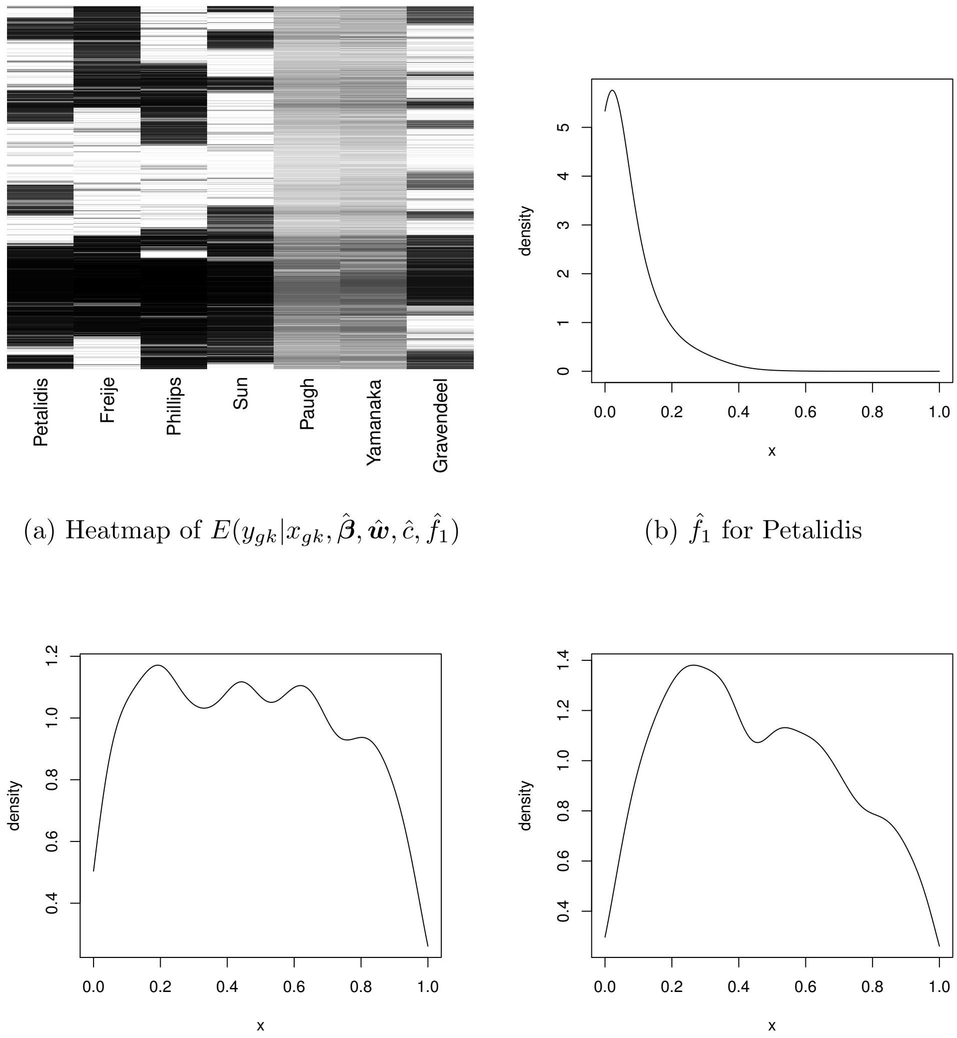
The genes are sorted to place the similar ones together. We can easily observe that the posterior distribution of y_{gk} is very different for the Paugh study and the Yamanaka study. By plotting the $\hat{f}_1$ of these two studies, I can confirm that these two studies have bad qualities. This is consistent with the findings of Kang et al. [2012] and chapter 2. Fixing $r = 4$, 2181 DE genes are detected using ϕ_g based on cutoff $\eta = 1$. Figure 13 shows the result using only the 5 studies with good quality. Figure 13(b) shows the histogram of $E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$. Most of the posterior probabilities are distributed close to either 0 or 1. This indicates that there are very limited ambiguities remained in the analysis result. 2710 genes are detected by setting $r = 3$.

3.5 DISCUSSION

In this chapter, I extended HS_r to a new hypothesis setting HS'_r with complementary null and alternative hypotheses. To model the p-value distribution, I proposed a semiparametric mixture model approach. I also proposed a EM algorithm to fit the model, and defined a Bayes factor test statistic ϕ_g to test HS'_r.

From the simulation result we can see that all of the traditional methods failed in FDR control when HS'_r is tested, because none of them is designed to work for the composite null hypothesis. My approach is the first meta-analysis method that address the composite null hypothesis problem in a genomic setting. Using my mixture model approach, even though an arbitrary cutoff $\eta = 1$ is used, the FDR is well controlled. Moreover, η could be chosen according to the cost of making type I and type II error, which would be helpful in making decisions in real life. Another advantage of my method is that I did not make any assumption for the p-value distribution for DE genes in single studies. Instead, kernel density estimation is applied to estimate the p-value distribution for DE genes from the data. In real data analysis, I also showed that $\hat{f}_1$ could be used for quality control purposes. Studies with noisy $\hat{f}_1$ are potentially identifiable as low quality and could be excluded from the analysis.

However, there are two limitations of this method. First, the selection of the ridge penalty parameters λ_1 and λ_2 is arbitrary. Although this could be addressed by setting hyper-



(a) Heatmap of $E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$

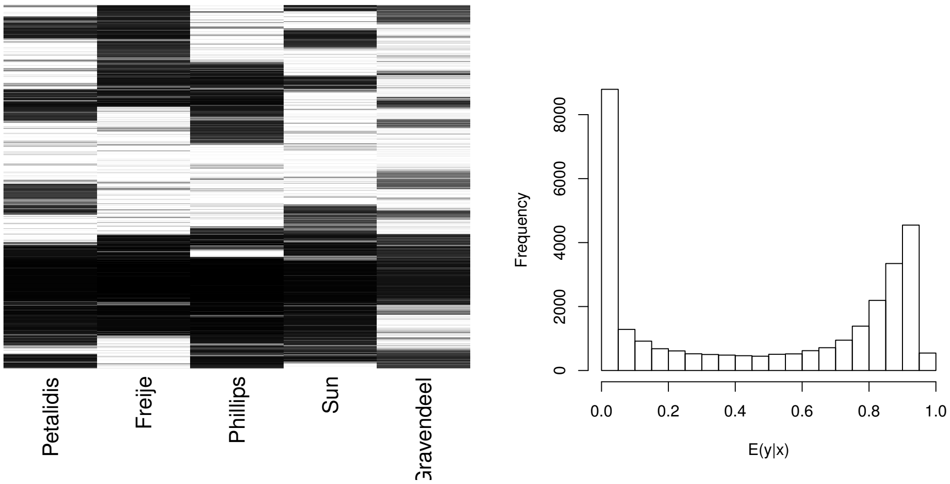
(b) $\hat{f}_1$ for Petalidis

(c) $\hat{f}_1$ for Paugh

(d) $\hat{f}_1$ for Yamanaka

Figure 12: Brain cancer result using all studies

parameters for σ^2 and τ^2 and using Markov Chain Monte Carlo (MCMC) to fit the model, the computation would be much more expensive than the EM-algorithm. And because λ_1 and λ_2 are corresponding to the variability of the log-odds in prior probabilities p_{gk}, it is appropriate to set them according to the assumptions. The larger the λ's, the more independent the genes (studies) are assumed. The second limitation is that this method could not be applied to single studies whose null distribution of p-values are unknown. For example, for studies



(a) Heatmap of $E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$ (b) Histogram of $E(y_{gk}|x_{gk}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}, \hat{c}, \hat{f}_1)$

Figure 13: Brain cancer result using only 5 studies

using one-sided test, the null hypothesis is composite and the null distribution is no longer $U(0, 1)$. To address this problem, parametric distribution of p-values could be used instead. And more than two mixing components could be assumed as well.

Extensions could be made to this model. As illustrated in section 3.4.1, using the estimated prior distribution of y_{gk}, I can allow missing values in the meta-analysis. Also, I can allow the prior distribution of β_g and w_k be correlated. For example, I can extend this method to the meta-analysis of continuous regions (e.g. SNP or copy number data), by assuming an auto-regression correlation structure for β_g. This is equivalent to replacing the penalty term in equation 3.2 by $\boldsymbol{\beta}'\Lambda_1\boldsymbol{\beta} + \boldsymbol{w}'\Lambda_2\boldsymbol{w}$. Here Λ_1 and Λ_2 are the precision matrices for the prior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{w}$. Because I can assume simple conditional dependency structures (e.g. $AR(1)$) among $\boldsymbol{\beta}$ and $\boldsymbol{w}$, the precision matrix and the second-order derivative matrix for $J(\cdot)$ would both be sparse, thus the computation should still be affordable.

4.0 META-NETWORK ANALYSIS

4.1 BACKGROUND

In chapter 2 and chapter 3, I developed two meta-analysis methods for genomic studies. Both these methods target differentially expressed genes by combining multiple studies. As stated in section 1.3, however, genes in biological systems do not function independently. Coregulation networks help regulate the expression levels of the genes that are involved in the same biological process. In systems biology, it is particularly important to discover the networks among the genes, in order to understand how the genes cooperate with each other to carry out certain biological functions.

Various methods have been developed to construct the network. For example, the pairwise correlations between single genes can be used to infer the coregulation network [Guilloux et al., 2010]. By assuming the gene expression levels are sampled from a Markov random field, graphical lasso can be applied to infer the sparse inverse of the covariance matrix, which is equivalent to the structure of the Markov network [Friedman et al., 2008]. And Huang et al. [2011] developed a tool named mirConnX by integrating both prior knowledge from experiments and microRNA data. In this chapter, I propose a meta-analysis framework to integrate network results from single studies to construct robust subnetworks in subsets of the combined studies.

In the remainder of this chapter, I introduce the idea of detecting conservative subnetworks in a subset of studies in section 4.2. In section 4.3 I define a likelihood based target function and propose a search algorithm to optimize it. My algorithm is applied in both simulated datasets and real data in section 4.4. The advantages and potential pitfalls of this algorithm are summarized, and possible future work is discussed in section 4.5.

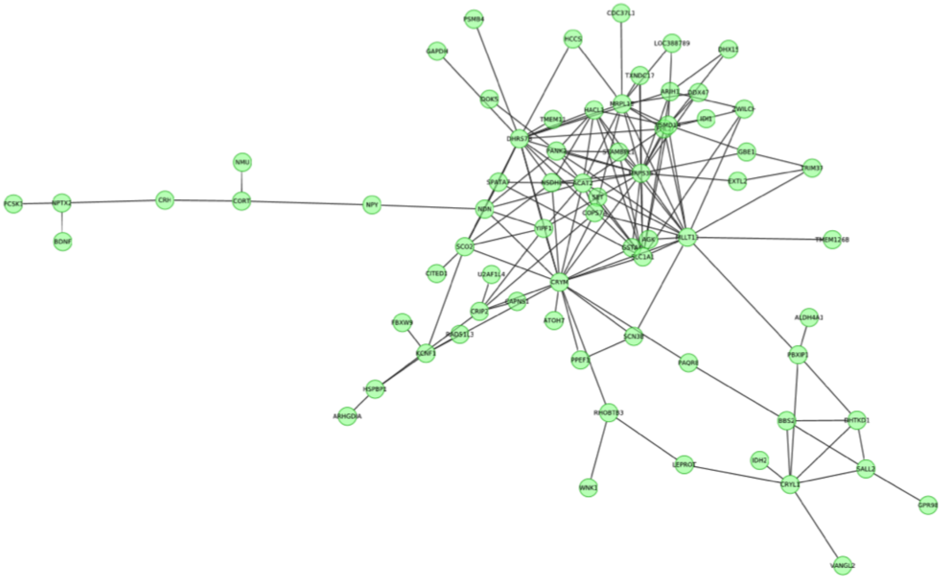


Figure 14: An example of gene coregulation network

4.2 MOTIVATION

Structure learning is a difficult problem in machine learning. Because in genomic studies usually thousands of genes are involved, which yield numerous possible connections in a network structure. Learning of this structure will suffer from issues of false positives and false negatives. Especially when the sample size of a single study is relatively small, the sensitivity and specificity of the result will be fairly low. Therefore, meta-analysis methods can be applied to the results of networks generated from multiple single studies in order to improve the network construction.

Biologically, networks of important biological functions are usually stably conserved across studies. By identification of subnetworks that are conservative in a majority of single studies, we expect to discover networks that are important for certain phenotypes. However, there is no statistically rigorous method available to combine network results across studies. I propose a brand new likelihood based framework to construct subnetworks from multiple studies.

4.3 METHODS

Suppose that we have obtained network results in $|K|$ studies, and each of the studies has $|G|$ genes involved. For each single study k, denote $t_{ij}^{(k)}$ as the edge between gene i and j in study k: if $t_{ij}^{(k)} = 1$, then i and j are connected in study k; otherwise, i and j are not connect in k. The goal of my method is to find a subnetwork $H \subseteq G$ that is conserved in a subset of studies $L \subseteq K$. In this chapter, we only consider networks that are undirected graphs.

4.3.1 Likelihood based score

The identification of H and L is nontrivial. I assume that given H, $t_{ij}^{(k)}$ is sampled from a Bernoulli trial with success probability p_{ij} if $k \in L$, otherwise $t_{ij}^{(k)}$ is sample from a Bernoulli trial with success probability p_{null}. Then I can calculate the likelihood of $P_L(H) = \{p_{ij}; i, j \in H\}$, p_{null} and L given H and the subnetwork $T(H)$ defined on H.

$$\begin{aligned} l(P_L(H), p_{null}, L; T(H), H) = &\sum_{k \in L} \left[\sum_{i,j \in H} \left(t_{ij}^{(k)} \log p_{ij} + (1 - t_{ij}^{(k)}) \log(1 - p_{ij}) \right) \right] \\ &+ \sum_{k \in K \setminus L} \left[\left(\sum_{i,j \in H} t_{ij}^{(k)} \right) \log p_{null} + \left(\binom{|H|}{2} - \sum_{i,j \in H} t_{ij}^{(k)} \right) \log(1 - p_{null}) \right] \end{aligned} \tag{4.1}$$

Suppose that the subset of studies that are conserved are known as $L = L_0$. The MLE of $P_L(H)$ and p_{null} can be obtained:

$$\widehat{p_{ij}} = \frac{\sum_{k \in L_0} t_{ij}^{(k)}}{|L_0|} \tag{4.2}$$

$$\widehat{p_{null}} = \frac{\sum_{k \in K \setminus L_0} \sum_{i,j \in H} t_{ij}^{(k)}}{\binom{|H|}{2} |K \setminus L_0|} \tag{4.3}$$

The likelihood function 4.1 can be maximized by plugging in the MLE of p_{ij} and p_{null} in 4.2 and 4.3. Then the best selection of L is obtained by optimizing the likelihood function $l(\widehat{P(L)}, \widehat{p_{null}(L)}, L; T(H), H)$. In order to capture well connected networks instead of sparse ones, I introduce a penalty term into the likelihood function. The resulting score function is

$$G(L) = l(\widehat{P(L)}, \widehat{p_{null}(L)}, L; T(H), H) + \lambda_L \sum_{i,j \in H} \widehat{p_{ij}}$$

Then the best selection of L with given network structure H is

$$L^* = \arg\max_{L \subseteq K} G(L)$$

4.3.2 Decomposition of the target function

Notice that the target function can be decomposed into a summation of sub-scores on each of the edges that belong to H.

$$\begin{aligned}
G(L) =& l(\widehat{P(L)}, \widehat{p_{null}(L)}, L; T(H), H) + \lambda_L \sum_{i,j \in H} \widehat{p_{ij}} \\
=& \sum_{i,j \in H} \left[\sum_{k \in L} t_{ij}^{(k)} \log \widehat{p_{ij}} + \left(|L| - \sum_{k \in L} t_{ij}^{(k)} \right) \log(1 - \widehat{p_{ij}}) \right] \\
&+ \sum_{i,j \in H} \left[\sum_{k \in K \setminus L} t_{ij}^{(k)} \log \widehat{p_{null}} + \left(|K \setminus L| - \sum_{k \in K \setminus L} t_{ij}^{(k)} \right) \log(1 - \widehat{p_{null}}) \right] \\
&+ \sum_{i,j \in H} \lambda_L \widehat{p_{ij}} \\
=& \sum_{i,j \in H} E_{ij}
\end{aligned}$$

Then the target function could be decomposed into scores on each edge. And we can select λ_L such that around 0.1% of the edges have positive scores.

4.3.3 Search algorithm

Given $L = L_0$, I define a subnetwork as optimal if and only if adding one more edge will not improve the score of the target function. Because in section 4.3.2, I have proven that the target function can be decomposed into the summation of edge-specific scores, I can easily see that the optimal subnetwork can be found by searching through all the connected components which are connected only by edges with positive E_{ij}'s. The connected component with largest the $\sum_{i,j \in H} E_{ij}$ would be selected as the optimal subnetwork given L_0. In order to find subnetworks that are conservative in majority of the studies, I use the restriction $|L| \geq |K|/2$. The searching algorithm is shown as below.

STEP I. For each $L \subseteq K$ and $|L| \geq |K|/2$, find λ_L such that around 0.1% of the E_{ij}'s are positive. Calculated all the E_{ij}'s.

STEP II. Find all connected components that are connected by only edges with positive E_{ij}'s.

STEP III. Record the connected component H with the largest $\sum_{i,j \in H} E_{ij}$ and its corresponding L. H is a conservative subnetwork in studies L.

STEP IV. Remove H from G and repeat from Step I, till no more significant subnetwork can be identified.

4.4 APPLICATION

4.4.1 Simulation

In order to evaluate my algorithm, I simulated data using the following procedure.

STEP I. Sample $t_{ij}^{true} \sim Bern(p_{net})$ for any $i, j \in H$. t_{ij}^{true} indicates whether gene i and j are connected in the true network.

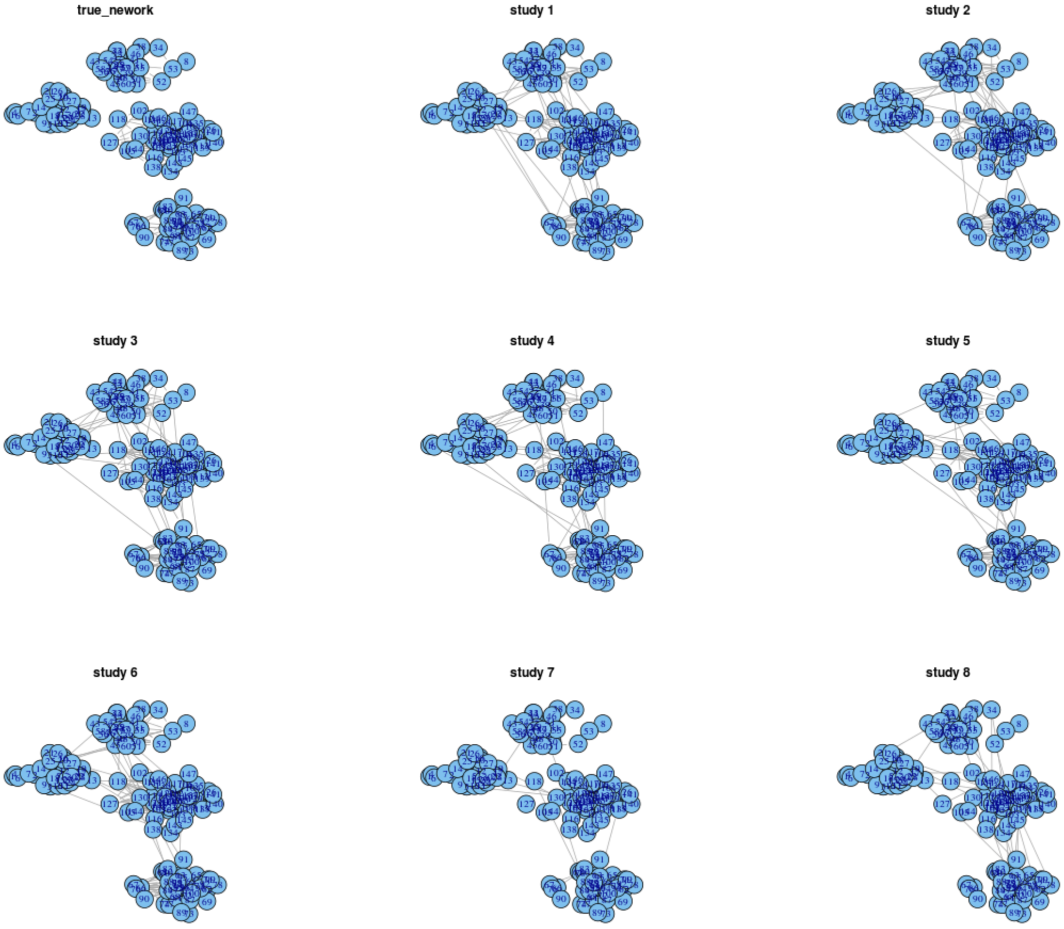
STEP II. For $i, j \in H$, sample $p_{ij} \sim Beta(s_1, 1)$ if $t_{ij}^{true} = 1$, and sample $p_{ij} \sim Beta(1, s_2)$ if $t_{ij}^{true} = 0$.

STEP III. For $k \in L$ and $i, j \in H$ sample $t_{ij}^{(k)} \sim Bern(p_{ij})$. Otherwise, sample $t_{ij}^{(k)} \sim p_{null}$ if $k \notin L$ or $i, j \notin H$.

In my simulation, I used parameters $|G| = 1000$, $|H| = 150$, $K = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $L = \{1, 2, 3, 4, 5, 6\}$, $p_{net} = 0.2$, $p_{null} = 0.002$, $s_1 = 5$ and $s_2 = 9$. Figure 15 shows my simulated true network and the networks in the 8 single studies for the 150 genes (genes in H). The 150 genes fall into 4 disconnected components in the true network. And in each single study, there are numbers of false positive and false negative edges.

Figure 16 compares the true network and the meta-network identified by my algorithm. We can see that the meta-network is closer to the true network comparing to any other single study networks, which shows better performance of the meta-networks compared to single studies.

In figure 17, the top 4 subnetworks (modules) identified by my meta-network algorithm are compared to the true subnetworks. The edge sensitivity and specificity are calculated for each module. For module 1, the sensitivity is 98.0% and the specificity is 98.7%. For module 2, the sensitivity is 90.6% and the specificity is 99.7%. For module 3, the sensitivity is 93.5% and the specificity is 99.7%. For module 4, the sensitivity is 91.7% and the specificity is 99.7%. For all these 4 modules, the true L is recovered with high accuracy. This result


true_nework

study 1

study 2

study 3

study 4

study 5

study 6

study 7

study 8

Figure 15: Simulated network

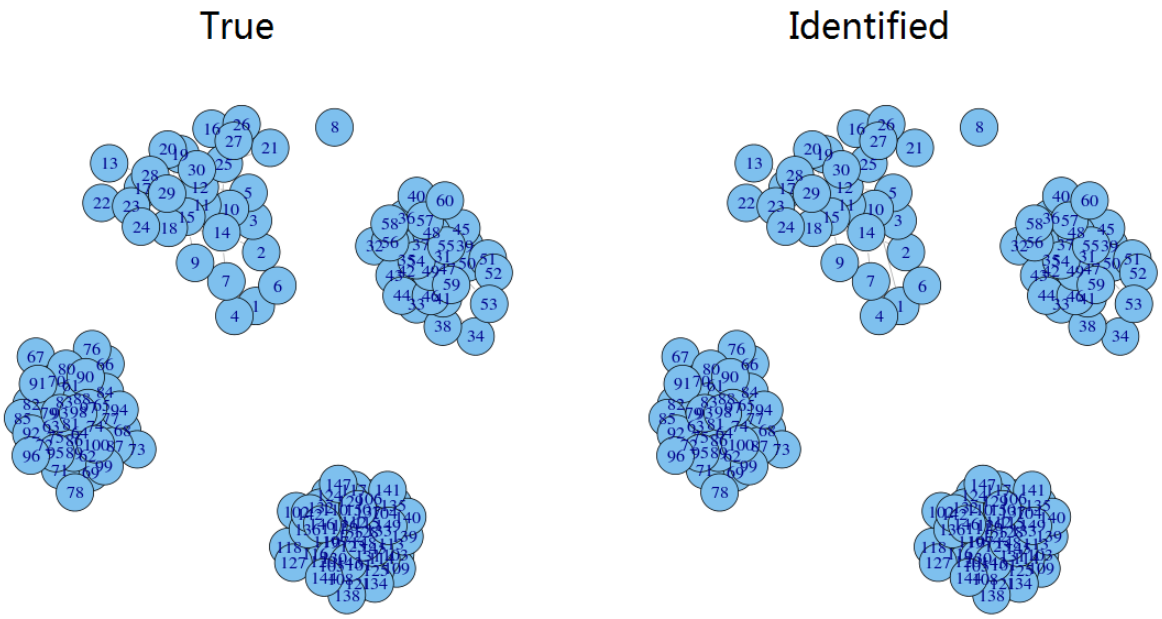


Figure 16: True and identified meta-network

indicates that my meta-network algorithm has good sensitivity and specificity for meta-network detection.

4.4.2 Real data analysis

I applied my method to the MDD dataset described in chapter 2 and obtained 520 DE genes detected by Wang et al. [2012]. Correlation between each pair of genes i and j is calculated in every single study. The single study networks are constructed by thresholding the correlations such that only 1% of the edges with the highest absolute correlations are kept.

After applying my algorithm, I have identified 3 large modules. The meta-network and the corresponding subnetwork structures in the 8 MDD studies for the 3 modules are shown in figure 18, 19 and 20. Comparing to the simulation result, the real data result is noisier. But I can still observe conserved substructures in each of the modules.

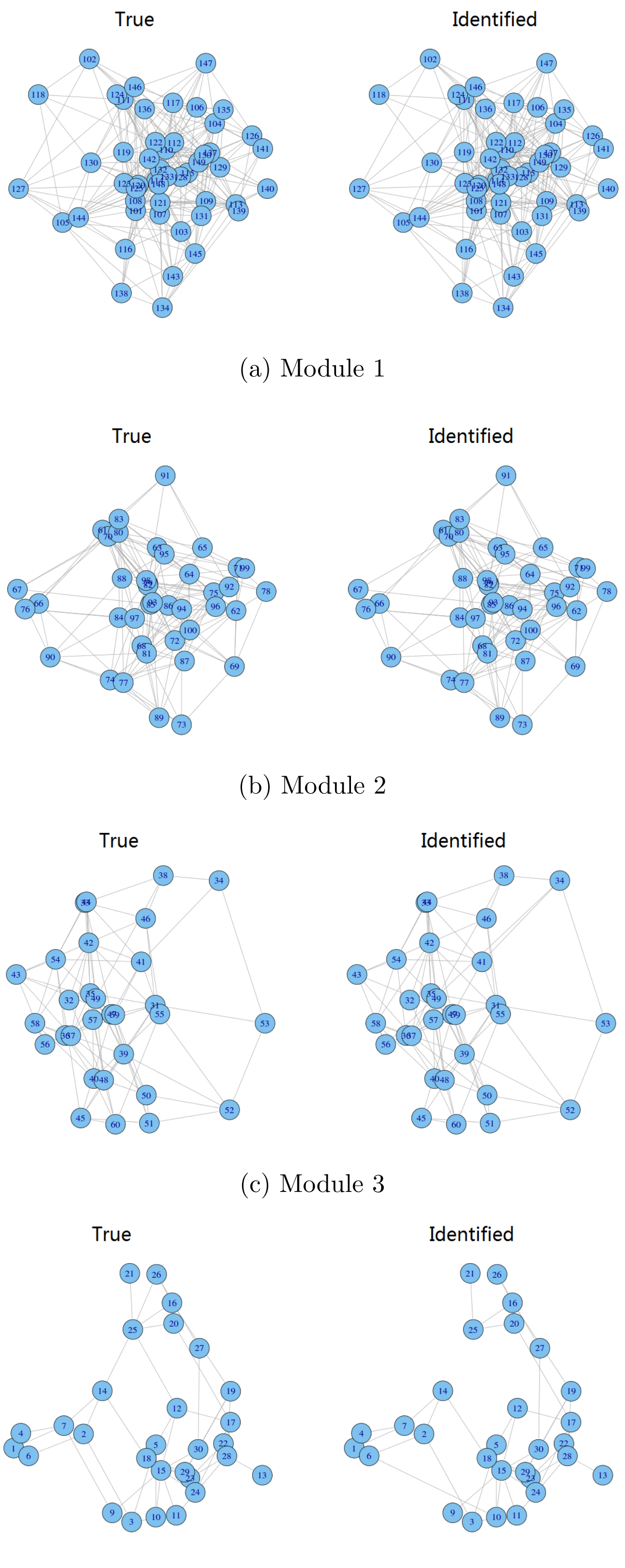
4.5 DISCUSSION

As the development of system biology, network analysis becomes a fundamental tool to understand how different genes cooperates with each other to carry out certain biological functions. However, because single studies often contain limited number of samples and the structure learning problem is very complex, meta-analysis methods that combines network results are demanding. Generally, in network analysis, the substructures are more important than pairwise relationship between genes. Therefore, traditional univariate meta-analysis methods are difficult to be generalized to network analysis. Existing studies that combine multiple network results are all relatively ad hoc.

In this chapter, I proposed a likelihood based method and defined a target function. Then the meta-analysis could be converted to an optimization problem. My framework provides the first systematic solution to the problem of network meta-analysis. In simulation analysis, I have shown that my method could recover the true network efficiently. Both the sensitivity and specificity are very impressive. In real data analysis, my method could still be used to find some conservative subnetworks, although the networks found are less conservative comparing to simulation. The results demonstrate that my algorithm provides a robust framework to combine multiple network analysis results.

However, there are potential pitfalls for my method. First, in this method, I treat the connections between genes as independent. In reality, however, some connections may tend to be correlated with others. For example, suppose gene A regulates gene B and C through a mediator D. When D is missing, A is connected to neither B or C; when D is present, A tend to be correlated with both B and C. Then the connections A-B and A-C are positively correlated. Second, scores of the target function may not be comparable between subnetworks of different sizes. Subnetworks with more edges may have larger scores. Third, I select subnetworks until no more significant ones could be found. The termination rule for the searching algorithm now depends on the size of the discovered subnetwork, which is ad hoc. And currently no formal statistical test is performed for the selected modules. Fourth, the performance of my algorithm depends on the network construction method used for single studies. Using a reliable method in single studies is important for the framework.

In the future, several modifications and extensions can be applied. First, I can use a permutation test to assign p-values for the subnetworks detected. Because the computation is expensive for my algorithm, I will also try to make the algorithm more efficient and use parallel computing techniques. Second, I will further investigate into the parameter selection issue for λ_L. If necessary, the target function will also be modified to make the scores comparable between subnetworks of different sizes. Third, different network construction methods for single studies will be evaluated. By applying different network construction methods, I can understand how the method selection affects the performance of the framework.



(a) Module 1

(b) Module 2

(c) Module 3

(d) Module 4

Figure 17: Identified modules

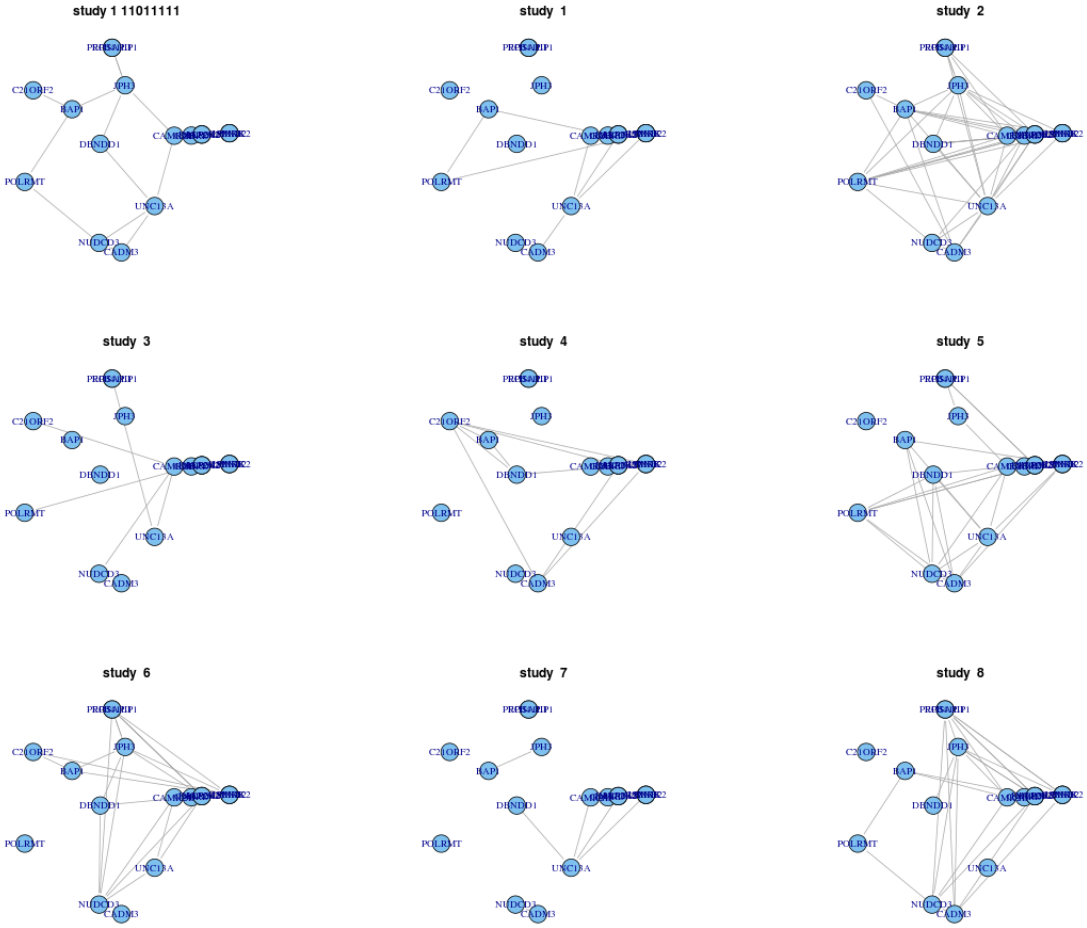


Figure 18: Module 1 in MDD data

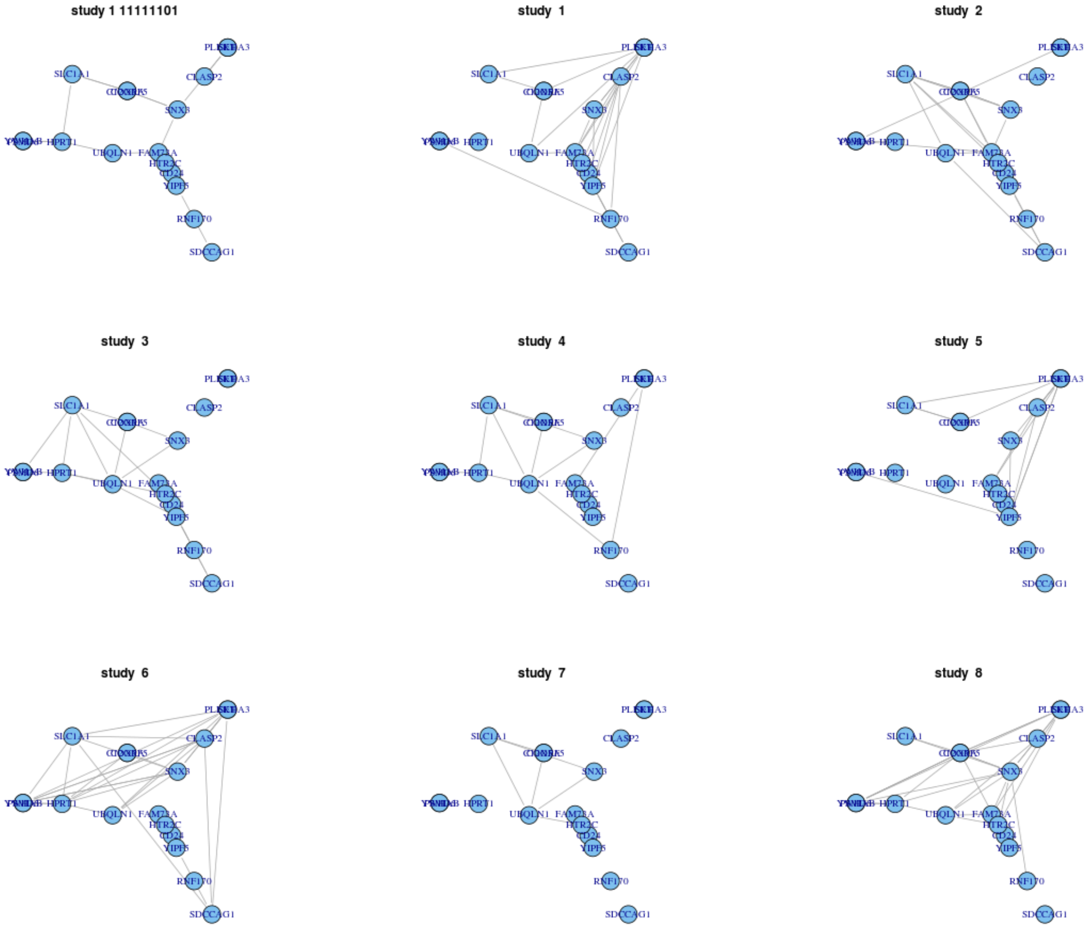
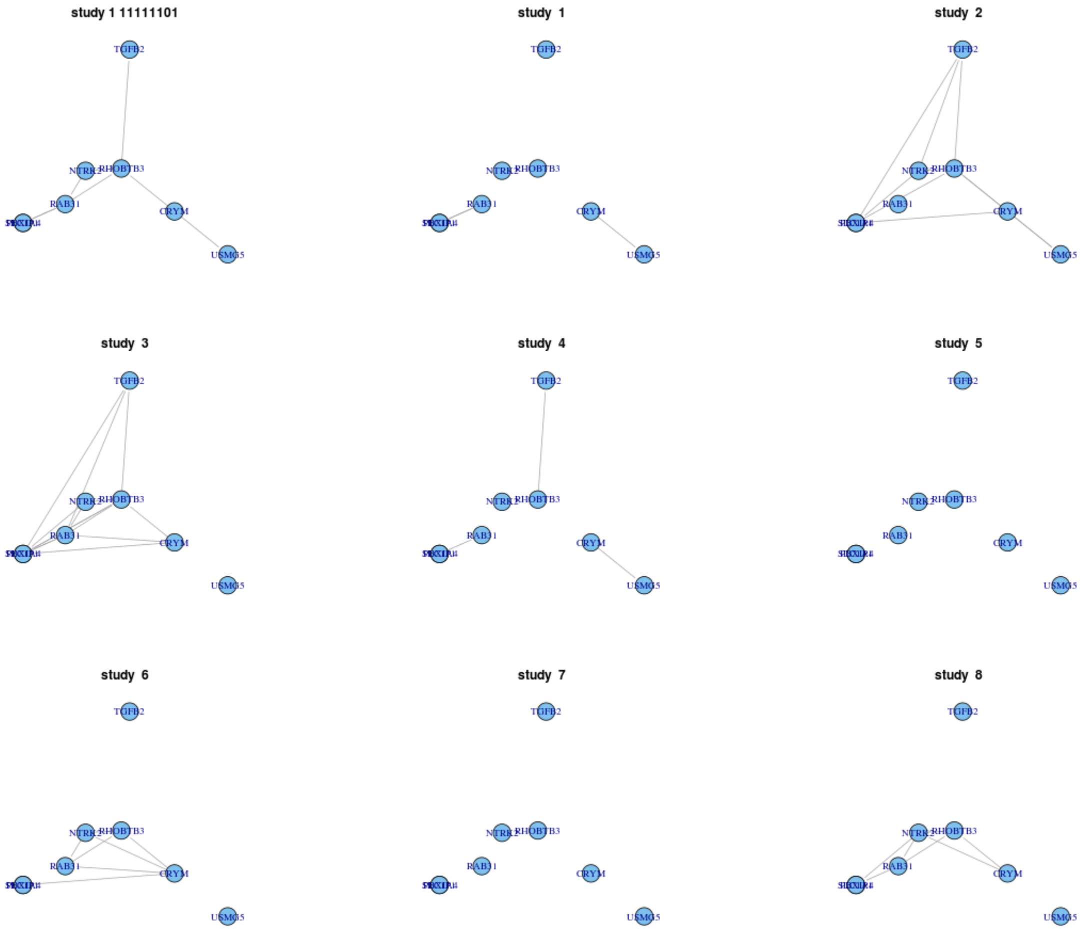


Figure 19: Module 2 in MDD data


study 1 11111101

study 1

study 2

study 3

study 4

study 5

study 6

study 7

study 8

Figure 20: Module 3 in MDD data

5.0 CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

As high-throughput technology thrives in biomedical research, meta-analysis methods that combine multiple genomic studies to generate statistically powerful and robust conclusions become timely and demanding. In this dissertation, I developed a series of tools for gene expression meta-analysis.

In chapter 2, I compared two hypothesis settings defined by Li and Tseng [2011]. Genes detected by these two hypothesis setting would be either DE "in all" studies or "in one or more" studies. Then current meta-analysis methods were compared and categorized based on their target hypotheses. However, the two complimentary settings are either too sensitive or too stringent. Therefore following the discussion of Li and Tseng, I proposed a robust hypothesis setting HS_r that pursues biomarkers DE "in majority" of combined studies. Biologically, genes that are DE in most of the combined studies are usually more meaningful because they may play important roles in disease development or could serve as potential drug targets. To test HS_r, I proposed to use the rth order p-value (rOP) as the test statistic to combine multiple studies. Two methods were proposed for the selection of r: one is based on the adjusted number of detected DE genes; the other is based on available pathway information. The statistical power of rOP was calculated analytically. I also found that rOP is closely related to the vote counting method, but does not share its asymptotically powerless property. I evaluated rOP by real data analysis. The results demonstrated that my selection of r is reliable, since the two methods agreed and suggested r's that are greater than half of the combined studies in all of the applications. Moreover, by adding one MDD dataset into the brain cancer studies, I showed that rOP is robust to outliers: even though a few studies might be mistakenly included in the meta-analysis, rOP method is still able to get the similar conclusion.

However, despite the advantages of rOP, there is an intrinsic limitation associated with hypothesis settings HS_A and HS_r. In HS_A and HS_r, the null and alternative hypotheses are not complementary. Therefore, the alternative hypothesis is not automatically accepted when the null hypothesis is rejected. To address this issue, in chapter 3 I proposed HS'_r as the complementary form of HS_r. Because in HS'_r, the null hypothesis is modified to a composite hypothesis with no known null distribution, hypothesis testing based methods no longer work. And traditional hypothesis testing based methods such as Fisher's method and Stouffer's method would be anti-conservative under HS'_r. To overcome this problem, I proposed a semiparametric mixture model approach to model the generation process of p-values in all the combined studies. In the generative model, a logistic prior is used to account for the effects of different genes and different studies. Then the p-values are assumed to be generated from either $U(0,1)$ or an unknown distribution depending on whether the gene is DE or not. Instead of assuming any parametric form, I used kernel density for the unknown distribution to make the model flexible. An EM algorithm is proposed for the model fitting, such that the computation is affordable in large genomic settings. I also proposed a Bayes factor to substitute the traditional hypothesis testing procedure. My new method is compared to others by simulation. The result indicates that my new approach performs as well as methods like Stouffer's method and Fisher's method in terms of ROC curves, while my method is much better than any existing methods in terms of FDR control since all other methods failed to control FDR under HS'_r. Then I applied my method in real data analysis. From the results, we saw that my method could also be used for quality control purposes. Additionally, this method could be easily extended. For example, the model could be generalized for datasets with missing p-values. Moreover, different logistic priors could be assumed to account for different types of data.

In addition to meta-analysis methods for single biomarker detection, I discussed the meta-analysis of biological networks in chapter 4. I proposed a likelihood based score function to evaluate the conservative subnetworks in a subset of the combined studies. Thus the meta-analysis of multiple network results is converted to an optimization problem. To further simplify the computation, I decomposed the target function into summation of edge specific scores. I evaluated my algorithm in both simulation and real data analysis. From simulation, I demonstrated that my method could successfully identify the true conserved network with high sensitivity and specificity. Conserved subnetworks could also be identified in real data analysis, although the result was noisier than in the simulation. This is the first method that provides a systematic framework to combine multiple network results. And this method will be further developed and extended in the future. For example, I will try to assign p-values for each identified subnetworks by permutation testing. Different network construction methods will be applied for single studies. If necessary, the target function could also be modified in the future. Moreover, I would like to extend this method to identify differential subnetworks between two cohorts (such as male vs female or white vs black).

In summary, I have developed a series of tools for genomic meta-analysis. My works attempt to provide robust meta-analysis solutions for different questions raised in genomic studies.

5.2 FUTURE WORK

In the future, I intend to further extend my current work in several ways, including methodology, software and application.

First, I will extend my methodologies into the vertical integration of multiple data types. Currently, the studies I combined have only a single data type. As high-throughput technologies improve, it becomes affordable to measure multiple types of data on the same set of samples or individuals. For example, the cancer genome atlas (TCGA) project measures gene expression, copy number, methylation, micro RNA expression, SNP and other kinds of data on hundreds of cancer patients with different types of cancer. Therefore, it is important for us to extend my methods to accommodate multiple data types. Especially, I will attempt to extend the mixture model approach to jointly model multiple data types simultaneously.

Second, I will polish my R code and make it easy to use. R packages will be developed for the methods discussed in this dissertation. And the packages will be submitted to the comprehensive R archive network (CRAN, `http://cran.r-project.org`). To facilitate biologists who may not be R users, I will develop Java based softwares with graphical user interfaces (GUI), such that the methods could be easily applied in real data analysis.

Third, the methods developed could be applied to different datasets to draw interesting biological conclusions. For example, I will apply my method in the major depression disorder studies to find interesting biomarkers that affects the disease. I can also apply my method in TCGA datasets to find important genes that could be potential drug targets for cancer treatment.

APPENDIX

DATA DESCRIPTION

Table 4: Seven brain cancer studies

Tissue	Author	Year	Platform	Sample size	Comparison	Source
Brain	Petalidis	2008	HG-U133A	65	AA vs GBM	GSE1993
Brain	Freije	2006	HG-U133A,B	85	AA vs GBM	GSE4412
Brain	Phillips	2006	HG-U133A,B	100	AA vs GBM	GSE4271
Brain	Sun	2006	HG-U133_Plus_2	180	AA vs GBM	GSE4290
Brain	Paugh	2010	HG-U133_Plus_2	53	AA vs GBM	GSE19578
Brain	Yamanaka	2006	Agilent	29	AA vs GBM	GSE4381
Brain	Gravendeel	2009	HG-U133_Plus_2	284	AA vs GBM	GSE16011

6,005 genes remained in the combined dataset after gene matching (AA: Grade 3 Anaplastic astrocytoma; GBM: Grade 4 Glioblastoma multiforme)

Table 5: Nine MDD studies

Study name	Gender	Brain region	Sample size	Platform
MD1_ACC	Male	ACC	32 (16 pairs)	Affymetrix
MD3_ACC	Female	ACC	44 (22 pairs)	Illumina
C_MD2_ACC_F	Female	ACC	18 (9 pairs)	Affymetrix
C_MD2_ACC_M	Male	ACC	26 (13 pairs)	Affymetrix
MD1_AMY	Male	AMY	28 (14 pairs)	Affymetrix
MD3_AMY	Female	AMY	42 (21 pairs)	Illumina
C_MD2_DLPFC_F	Female	DLPFC	28 (14 pairs)	Affymetrix
C_MD2_DLPFC_M	Male	DLPFC	32 (16 pairs)	Affymetrix
NY_DLPFC_M	Male	DLPFC	26 (13 pairs)	Affymetrix

7,577 genes remained in the combined dataset after gene matching.

Table 6: 16 diabetes studies

Study	Organism	Platform	Description
1	Mouse	MG-U74Av2	Brown preadipocyte IRS knockout profiling
2	Mouse	MG-U74Av2	Comparison of Low Fat and High Fat Diet on Mice of Two Genetic Backgrounds (B6 vs. 129) - Fat
3	Mouse	MG-U74Av2	Comparison of Low Fat and High Fat Diet on Mice of Two Genetic Backgrounds (B6 vs. 129) - Liver
4	Mouse	MG-U74Av2	Comparison of Low Fat and High Fat Diet on Mice of Two Genetic Backgrounds (B6 vs. 129) - Skeletal Muscle
5	Mouse	MG-U74Av2	Isolated adipocytes from normal and fat insulin receptor knockout (FIRKO) mice sorted into small and large cells
6	Mouse	MG-U74Av2	Liver - ob/ob mice
7	Mouse	MG-U74Av2	Mouse skeletal muscle - controls, streptozotocin diabetes and insulin treated
8	Human	HG-U133A,B	Human pancreatic islets from normal and Type 2 diabetic subjects
9	Mouse	MG-U74Av2	Transcription profiling of wild type and PGC-1alpha KO liver and skeletal muscle
10	Mouse	MG-U74Av2	Effect of PGC-1alpha and PGC-1beta on gene expression in myocytes and hepatocytes
11	Mouse	MG-U74Av2	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	Mouse	MG-U74Av2	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	Mouse	MG-U74Av2	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	Mouse	MG-U74Av2	Effect of insulin infusion on skeletal muscle
15	Mouse	MG-U74Av2	Skeletal Muscle - Muscle Insulin Receptor Knockout and Control Mice - Control, Streptozotocin Diabetic and Insulin Treated
16	Human	HG-U133A	Human skeletal muscle - type 2 diabetes - Swedish males

6,645 genes remained in the combined dataset after gene matching.

BIBLIOGRAPHY

D.B. Allison, G.L. Gadbury, M. Heo, J.R. Fernández, C.K. Lee, T.A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20, 2002.

A.L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

F. Begum, D. Ghosh, G.C. Tseng, and E. Feingold. Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic Acids Research*, 2012.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

R.L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, pages 295–300, 1982.

R.L. Berger and J.C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, pages 283–302, 1996.

A. Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, pages 559–574, 1954.

H.M. Cooper, L.V. Hedges, and J.C. Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation Publications, 2009.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

S. Erickson, K. Kim, and D.B. Allison. *Meta-analysis and combining information in genetics and genomics*, chapter 6, page 90. Chapman & Hall/CRC, 2009.

R.A. Fisher. *Statistical methods for research workers.* Edinburgh, 1925.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

A.K. Ghosh and S. Bandyopadhyay. Adaptive smoothing in kernel discriminant analysis. *Nonparametric Statistics*, 18(2):181–197, 2006.

JP Guilloux, C. Gaiteri, and E. Sibille. Network analysis of positional candidate genes of schizophrenia highlights... more than... myelin-related pathways. *Molecular psychiatry*, 15(8):786–788, 2010.

L.V. Hedges and I. Olkin. Vote-counting methods in research synthesis. *Psychological Bulletin*, 88(2):359, 1980.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

G.T. Huang, C. Athanassiou, and P.V. Benos. mirconnx: condition-specific mrna-microrna network integrator. *Nucleic acids research*, 39(suppl 2):W416–W423, 2011.

D.D. Kang, E. Sibille, N. Kaminski, and G.C. Tseng. Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Research*, 40(2): e15–e15, 2012.

S. Le Cessie and JC Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

J. Li and G.C. Tseng. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5 (2A):994–1019, 2011.

R.C. Littell and J.L. Folks. Asymptotic optimality of fisher's method of combining independent tests. *Journal of the American Statistical Association*, pages 802–806, 1971.

R.C. Littell and J.L. Folks. Asymptotic optimality of fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, pages 193–194, 1973.

C. Loader. *Local regression and likelihood.* Springer Verlag, 1999.

A.B. Owen. Karl pearson's meta-analysis revisited. *The Annals of Statistics*, 37(6B):3867–3892, 2009.

P.J. Park, S.W. Kong, T. Tebaldi, W.R. Lai, S. Kasif, and I.S. Kohane. Integration of heterogeneous expression data sets extends the role of the retinol pathway in diabetes and insulin resistance. *Bioinformatics*, 25(23):3121, 2009.

K. Pearson. On a new method of determining "goodness of fit.". *Biometrika*, 26(4):425, 1934.

S. Pounds and S.W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236, 2003.

D.R. Rhodes, T.R. Barrette, M.A. Rubin, D. Ghosh, and A.M. Chinnaiyan. Meta-analysis of microarrays. *Cancer research*, 62(15):4427, 2002.

SN Roy. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, pages 220–238, 1953.

B.W. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1): 3, 2004.

Chi Song and George C. Tseng. Order statistic for robust genomic meta-analysis. *Annals of Applied Statistics*, under review.

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

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

G.R. Terrell and D.W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

L.H.C. Tippett. *The Methods of Statistics.* London: Williams Norgate Ltd., 1931.

G.C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, 2012.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116, 2001.

X. Wang, Y. Lin, C. Song, E. Sibille, and G.C. Tseng. Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder. *BMC bioinformatics*, 13(1):52, 2012.

B. Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156, 1951.