GENOMIC META-ANALYSIS COMBINING MICROARRAY STUDIES WITH CONFOUNDING CLINICAL VARIABLES: APPLICATION TO DEPRESSION ANALYSIS

by

Xingbin Wang

MS, Shanghai Jiaotong University CHINA, 2005

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

2011

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Xingbin Wang

It was defended on

August 15 2011

and approved by

George C. Tseng, ScD, Associate Professor, Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

Etienne Sibille, Ph.D, Associate Professor, Department of Psychiatry

School of Medicine, University of Pittsburgh

Allan Sampson, Ph.D, Professor, Department of Statistics

School of Arts & Sciences, University of Pittsburgh

Yan Lin, Ph.D, Assistant Professor, Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

Dissertation Director: **George C. Tseng**, ScD, Associate Professor, Department of

Biostatistics

Graduate School of Public Health, University of Pittsburgh

GENOMIC META-ANALYSIS COMBINING MICROARRAY STUDIES WITH CONFOUNDING CLINICAL VARIABLES: APPLICATION TO DEPRESSION ANALYSIS

Xingbin Wang, PhD

University of Pittsburgh, 2011

Major depressive disorder (MDD) is a heterogeneous psychiatric illness with mostly uncharacterized pathology and is the fourth most common cause of disability according to the World Health Organization (WHO) and has a significant impact on public health in the United States. To understand the genetics of MDD, we aim to develop effective meta-analysis approaches to provide high-quality characterization of MDD related biomarkers and pathways with proper clinical variable adjustment. First, genomic meta-analysis in MDD faces multiple unique difficulties, such as weak expression signal of MDD, substantial clinical heterogeneity and small sample size. Given these obstacles, it is hard to identify consistent and robust biomarkers in an individual study. To achieve a more accurate and robust detection of differentially expressed (DE) genes and pathways associated with MDD, we proposed a statistical framework of meta-analysis for adjusting confounding variables (MetaACV). The result showed that more MDD related biomarkers and pathways were detected that greatly enhanced understanding of MDD neurobiology. Secondly, Meta-analysis has become popular in the biomedical research because it generally can increase statistical power and provide validated conclusions. However, its result is often biased due to the heterogeneity. Meta-regression has been a useful tool for exploring the source of heterogeneity among studies in a meta-analysis. In this dissertation, we will explore the use of meta-regression in microarray meta-analysis. To account for heterogeneities introduced by study-specific features such as sex, brain region and array platform in the meta-analysis of major depressive disorder

(MDD) microarray studies, we extended the random effects model (REM) for genomic meta-regression, combining eight MDD microarray studies. The result shows increased statistical power to detect gender-dependent and brain-region-dependent biomarkers that traditional meta-analysis methods cannot detect. The identified gender-dependent markers have provided new biological insights as to why females are more susceptible to MDD and the result may lead to novel therapeutic targets. Finally, we present an open-source R package called Meta-analysis for Differential Expression analysis (MetaDE) which provides 12 commonly used methods of meta-analysis. It is a friendly used software such that biologists implement meta-analysis in their research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This thesis is based upon studies conducted during August 2006 to August 2011 at the Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh. Many inspiring people have been involved in the work leading to my PhD thesis. I will like to acknowledge everyone for their contributions to the studies conducted during my time as a PhD student.

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Chien-Cheng Tseng (George), for his excellent and inspiring supervision of my work. He has been a great mentor and a great friend, without his instruction and unique support, this thesis would never have become a reality. Further I would like to thank my co-supervisor, Dr. Etienne Sibille, for his great and invaluable co-operation and help.

I thank all the members of my committee, Dr. Allan Sampson, Dr. Yan Lin for the many discussions and their critical comments on my thesis. I also thank all the members of my learning group, Jia Li, Zhibao Mi, Kui Shen, Shuya Lu, Sunghee OH, Chi Song(Chuck), Dongwan Kang and Lun-Ching Chang, for creating a nice studying and working environment in our office and their willingly participation in many interesting discussions.

Finally, I wish to express my greatest thanks to my family, my wife, Xinmei Zhu, who has been supporting me in many ways during the past years. She has been doing biomedical research for many years. It was her who introduced biostatistics to me several years ago, in that meanwhile, I was a mathematician. She let me know how important the statistical analysis is for medical research. Thereafter, I started to like biostatistics and finally chose the PhD program of biostatistics after I got my Master degree of Science. I have truly been fortunate, and I do appreciate her support during PhD study. I also thank my daughters, Ellin Wang and Raelyn Wang, who bring lots of joy to me.

1.0   INTRODUCTION

Major depressive disorder (MDD) is a heterogeneous psychiatric illness with mostly uncharacterized pathology, contributes to death by suicide, and is the fourth most common cause of disability according to the World Health Organization (WHO). To understand the genetics of MDD, gene expression analysis is a effective approach to identify the biomarkers associated with MDD. Differentially expressed (DE) gene detection is one of the most common analyses in microarray data, which are generally comprised of three components: (1) the gene expression data; (2) the outcome variable, such as disease status; and (3) patient-specific covariates, including treatment history and additional clinical and demographic information. The primary aim of many gene expression studies is to identify the DE genes by characterizing the relationship between the first two of these components, the gene expression and the disease outcome. Thus, in the literature, most psychiatric disease-related microarray studies of similar design did not carefully consider how these factors (the third component) influence the relationship between the gene expression and the disease status. Usually they either ignored the clinical variables or applied simple linear regression to include all variables in the model. Our results clearly show the limits to those two approaches. To our knowledge, this is the first study that systematically considers the critical elements in the data structure in order to obtain more accurate DE gene and pathway detection. In addition, due to the very weak expression signal of MDD, a substantial clinical heterogeneity and small sample size, it is hard to identify consistent and robust biomarkers in an individual study. In this dissertation, we aim to develop effective meta-analysis approaches to fill this gap and provide high-quality characterization of MDD related biomarkers and pathways with proper clinical variable adjustment.

This dissertation is organized as follows: in Chapter 1, the concept of MDD is first described in section 1.1; then, the statistics used in individual analysis, meta-analysis and pathway analysis methods are reviewed in sections 1.3,1.4 and 1.5, respectively. In chapter 2, we describe a statistical approach for meta-analysis to tackle weak signal expression profiles that have small sample size, case-control paired design and confounding covariates in each study. In Chapter 3, a meta-regression model with variable selection is described. In Chapter 4, the implementation and usage of the MetaDE package are described. Conclusions and future works are provided in Chapter 5.

1.1   MAJOR DEPRESSIVE DISORDER

Major depressive disorder (MDD) is a mental disorder characterized by an all-encompassing low mood accompanied by low self-esteem, and by loss of interest or pleasure in normally enjoyable activities. MDD is a disabling condition which adversely affects a person's family, work or school life, sleeping and eating habits, and general health[65]. It involves a minimum two-week continuous period of at least five of the following symptoms: lowered mood for the majority of the day, diminished pleasure in daily activities, weight loss or gain, sleep disturbance, agitation or lethargy, fatigue, feelings of worthlessness or helplessness, impaired thought or memory, and recurring thoughts of self-harm or death (DSM-IV 2000). Depression is a common human psychiatric disorder and the leading cause of disability in North America, afflicting an estimated 18% of the population with an approximate lifetime incidence of 12% in men and 20% in women [55]. Around 3.4% of people with major depression commit suicide, and up to 60% of people who committed suicide had depression or another mood disorder. The symptoms of depression are the greatest contributor to the global burden of disease [46] as calculated by total days lived with the disorder. It remains the fourth leading cause of worldwide disability, after accounting for higher mortality in other diseases. This ranking is expected to rise to second place by the year 2020, as current effective treatment for other diseases become more globally accessible.

Risk factors for depression: Major risk factors for depression include the sex of an individual, previous history of the illness, genetic predisposition/family history, and chronic or acute stress [30]. Some combination of these can prompt a depressive episode, but the requisite combination varies by individual. The threshold for depression is sensitive to social support, religiosity, age, and life stressors [14, 56, 57]. These environmental factors interact with the genetics of depression estimated at 33% heritance [30]. This is a lower heritability than bipolar disorder, or schizophrenia, which adds to the difficulty in teasing apart contributory factors. Depression itself is a risk factor for the disorder, as untreated depression is likely to reoccur [70]. This is particularly problematic as a significant percentage of patients (varying from placebo levels of 30%, up to 40% depending on the study) never meet the criteria for complete remission and will commonly endure increasingly lengthy bouts of depression [36, 58].

Complexity obscure the neuropathology of depression: Depression's continued toll on society is a function of multiple genetic and environmental susceptibilities that recruits a diverse cadre of further genetic factors to sustain the condition [6]. To date, most experiments have examined single aspects of the disease, but the complex causal factors in depression make it resistant to highly specific approaches. One immediate question is: Why not create sub-divisions of depression that have more homogeneous symptom groups that will be amenable to a pathology classification? However, clinical evidence does not strongly support this approach. In patients with repeated depressive episodes there is no correspondence of symptoms across episodes, preventing definitive clinical subdivisions that might have more consistent pathophysiology[74]. There is some evidence to suggest that classes of antidepressants have distinct response rates in different DSM-IV classifications of depression (atypical, psychotic, bipolar etc) [4]. However, a meta-analysis of over 100 antidepressant drugs trials found no difference in response rates as an interaction of drug class and putative subtype[17].

Table 1.1: Data description of eight MDD microarray studies

Study	Gender	Brain region	sample	Platform
MD1_ACC	M	ACC	32(16)	Affymetrix
MD3_ACC	F	ACC	44(22)	Illumina
C_MD2_ACC	F	ACC	18(9)	Illumina
C_MD2_ACC	M	ACC	26(13)	Affymetrix
MD1_AMY	M	AMY	28(14)	Illumina
MD3_AMY	F	AMY	42(21)	Illumina
C_MD2_DLPFC	F	DLPFC	28(14)	Affymetrix
C_MD2_DLPFC	M	DLPFC	32(16)	Affymetrix

1.2 DATA DESCRIPTION AND PROBLEMS ENCOUNTERED IN GENE EXPRESSION ANALYSIS

Data description In this dissertation, we will focus on 8 human studies listed in Table 1.1 obtained from Dr. Sibille's lab for meta-analysis. In most of the patient cohorts, MDD patients are matched to control patients by their demographics such as age, sex and race. MDD related clinical variables of the patients are available for most studies, including alcohol consumption (AH), history of taking anti-depressant drugs (AD), death by suicide (SC), pH level of brain tissues (pH), disease recurrence (RC) and postmortem interval (PMI). Each study has three study-level features that may need adjustment in the analysis: sex, brain region, and array platform.

Problems encountered in gene expression analysis: Detecting candidate markers in transcriptomic studies is often difficult in MDD studies: First, as described in section 1.1, MDD is thought to be a complex and heterogeneous disease, associated with multiple genetic, genomic, post-translational, and environmental factors. Furthermore, patients might have varying disease severity, with some having psychotic features as well as exposure to a variety of medications and dosage levels to control their illness. Secondly, the genetic disease effects

4

are potentially confounded by many covariates, which include (1) demographical variables such as age, sex and race; (2) clinical variables such as anti-depressant drug usage, death by suicide and alcohol dependence; (3) technical variables inherent in the use of post-mortem brain samples, such as the pH level of brain tissues, brain region and post-mortem interval (PMI). If the statistical models employed to identify differentially expressed genes fail to incorporate these sources of heterogeneity, not only can this reduce the statistical power, but also it will introduce sources of spurious signals to the gene detection. Finally, sample sizes for these studies are generally small (between 9-22 pairs of MDDs and controls) due to the limited availability of suitable brain specimens and the significant costs associated with their collection. These three features in MDD studies severely hamper accurate biomarker detection. In section 1.3, we listed several statistical methods often used in the literature for DE gene detection in individual analysis, such as paired or unpaired t-test or the simple linear regression model. The former approach undoubtedly ignored the effects from confounding covariates; the latter approach was not efficient or not even applicable when the number of covariates is large and the number of samples in each study is small. These methods have been shown to have low statistical power in each MDD study with weak signal expression profiles, small sample size, case-control paired design and confounding covariates.

1.3   EXISTING METHODS FOR DE GENE DETECTION IN SINGLE STUDY

Gene-expression microarrays hold tremendous promise for revealing the patterns of coordinately regulated genes. Because of the large volume and intrinsic variation of the data obtained in each microarray experiment, statistical methods has been used as a way to systematically extract biological information and to assess the associated uncertainty. SAM and LIMMA are popular methods for microarray. Methods we covered here are more naive versions.

1.3.1 T-TEST

The t test perhaps is the most popular method for detecting differentially expressed genes due to its simplicity and availability. The t statistic is defined as

$$T_g = \frac{\bar{Y}_D - \bar{Y}_C}{S\sqrt{\frac{1}{n_D} + \frac{1}{n_C}}}, \tag{1.1}$$

where $\bar{Y}_D$ and $\bar{Y}_C$ are the mean values of disease (MDD) and control groups; n_D and n_C are the number of replicates in disease and control groups. S is the pooled standard deviation, which is estimated by $S = \sqrt{\frac{(n_D-1)S_D^2 + (n_C-1)S_C^2}{n_D + n_C - 2}}$. Under normal assumption, T_g follows a central student's t distribution with degree of freedom $n_C + n_C - 2$ under null hypothesis if we assume that MDD and control group have the same variance and the experiment was not pair-designed.

1.3.2 Paired T-TEST

The matched groups design is another popular form in medicine research in which subjects from disease and control groups are matched on some demographic variables such as age, gender and race. In this situation, paired t-test is the conventionally used test, which is defined as,

$$T_g = \frac{\bar{Y}_D - \bar{Y}_C}{S\sqrt{\frac{1}{n}}}, \tag{1.2}$$

where S is the standard deviation of differences of each pair, which is estimated by $S = \sqrt{\frac{\sum_{i=1}^{n}[(Y_{Di}-Y_{Ci})-(\bar{Y}_D-\bar{Y}_C)]^2}{(n-1)}}$. T_g follows Student's t distribution with degree of freedom of $n - 1$ under assumptions that the paired differences are independent and identically normally distributed. In general, paired test has more power than unpaired test whenever the within-pairs covariance is positive. Note that an alternative to the paired Student's t-test when the population can not be assumed to be normally distributed is the Wilcoxon singed-rank test[102].

1.3.3   MODERATED T-TEST

The gene-specific t test is not affected by heterogeneity in variance across genes because it only uses information from one gene at a time. It may, however, have low power because the sample size is small. In addition, the variances estimated from each gene are not stable: for example, if the estimated variance for one gene is small, by chance, the T_g value can be large even when the corresponding mean difference is small. To account for gene-specific fluctuations, a moderated t statistics [27, 99] is defined as below,

$$T_g = \frac{\bar{Y}_D - \bar{Y}_C}{s_g + s_0},$$
(1.3)

where $\bar{Y}_D$ and $\bar{Y}_C$ are the mean values of expression for gene g in disease and control groups, respectively; s_g is the standard deviation of repeated expression measurements: $s_g = \sqrt{\frac{[(n_D-1)S_D^2+(n_C-1)S_C^2][\frac{1}{n_D}+\frac{1}{n_C}]}{n_D+n_C-2}}$; s_0 is a positive constant to minimize the variability among $s_g(1 \leq g \leq G)$. In SAM, a regression procedure was used to select the optimal value of s_0. For simplicity, s_0 was often selected as the median of s_g. With this modification, genes with small mean differences will not be selected as significant, and this removes the problem of stability mentioned above.

1.3.4   LINEAR REGRESSION MODEL

A simple linear regression model[67, 76] has been commonly used to detect DE genes to account for additional variability resulting from many confounding variables. (e.g., in MDD studies, Age, pH, PMI and RIN). The model is described as below:

$$Y_{gi} = \mu_g + \beta_{g0}X_{0i} + \sum_{l=1}^{L} \beta_{gl}X_{li} + \epsilon_{gi},$$
(1.4)

In the model, Y_{gi} was the gene expression value of gene $g(\leq g \leq G)$ and sample i . X_{0i} was the disease label that took value one if the sample was disease and zero if sample was a control. X_{li} represented values for potential confounding covariate l $(1 \leq l \leq L)$; 0-1 binary for categorical variables of two levels and numerical for continuous variables). Finally, ϵ_{gi} were independent random noises that followed a normal distribution with mean zero and

variance σ_g^2. Under this model, β_{g0} was the disease effect of gene g and was the parameter of major interest. To obtain a disease-associated biomarker candidate list in a single study analysis, likelihood ratio test (LRT) or wald test was used to assess the p-values of testing $H_0 : \beta_{g0} = 0$ (vs $H_A : \beta_{g0} \neq 0$).

1.4   EXISTING MICROARRAY META-ANALYSIS METHODS

Many high-throughput genomic technologies have advanced dramatically in the past decade. Microarray experiment is one example that evolved into relative maturity with general consensus experimental protocol and data analysis strategy. Its extensive application in the biomedical field has led to an explosion of gene expression profiling studies publicly available. The noisy nature and small sample size in each dataset, however, often result in inconsistent biological conclusions. Consequently, meta-analysis methods for combining microarray studies have been widely applied to increase statistical power and provide validated conclusions. Four major categories of statistical methods have been used to combine microarray studies in differentially expressed (DE) gene detection: combining p-values, combining effect sizes, combining ranks and directly merge after normalization. In this dissertation, we mainly focused on the first two categories, one is to combine statistical significance (p-value)[52, 79, 80] from each individual study, and the other is to combine the effect sizes [15, 66]from each individual study. In general, among these microarray meta-analysis methods used in the literature, most methods have their pros and cons depending on the data structure and biological goal [47, 78]. Briefly, methods based on combining p-values are free of distribution assumptions and more powerful when the studies combined are heterogeneous, but do not support inferences about magnitudes and directions. On the other hand, methods based on combining effect sizes provide information about magnitudes and directions (e.g. down-regulated or up-regulated), but are more stringent on assumptions. In section 1.4.1, we described several representative methodologies for the first category. The representative methodologies for the second category were described in section 1.4.2. In these two sections, we consider K independent experiments have been performed to detect a certain

effect, θ_{gk} is the parameter that characterizes the condition (e.g. disease) effect in study $k, k = 1, 2, \cdots, K$ for gene $g, (1 \leq g \leq G)$. The kth experiment is concerned to test the hypothesis $H_{0gk} : \theta_{gk} = 0$ against an alternative $H_{1gk} : \theta_{gk} \neq 0$, and the p-value associated with the above test is denoted as p_{gk}.

1.4.1 METHODS COMBINING P-VALUES

1.4.1.1 Fisher's method(Fisher) Fisher's method (Fisher)[31, 32]is perhaps the most widely used combination procedure, which uses the product of p-values from tests in each study and transform it to chi-square scores using $-2 \log$ transformation.

$$V_g^{Fisher} = -2 \sum_{k=1}^{K} log(p_{gk}) \tag{1.5}$$

Under the null hypothesis, statistics V_g^{Fisher} follows a χ^2 distribution with $2K$ degrees of freedom. This method aggregates statistical significance from each study and generally has good detection power. It, however, can detect genes that are extremely significant (e.g. p=1E-20) in one study but not significant in the other four studies, a set of genes normally of less biological interests. See Li and Tseng [52] for more discussion.

1.4.1.2 Tippett's method(minP) This method is called minimum p-value (minP) method proposed by Tippett [82].

$$V_g^{minP} = \min_{1 \leq k \leq K} p_{gk} \tag{1.6}$$

Under the null hypothesis, V_g^{minP} has a $Beta$ distribution with degrees of freedom 1 and K. This method is also viewed as the union-intersection method. Say the rejection region for the test of H_{0gk} is $\{p_{gk} \leq \alpha\}$, where α is the overall significance level. Like Fisher's method, this method is also sensitive to very small p values in partial studies, but it is less powerful than Fisher's approach.

1.4.1.3 Wilkinson's Method(maxP) Maximum p-value(maxP) is a special case proposed by Wilkinson[103].

$$V_g^{maxP} = \max_{1 \leq k \leq K} p_{gk} \tag{1.7}$$

Under the null hypothesis, V_g^{maxP} has a *Beta* distribution with degrees of freedom K and 1. In contrast to Fisher's method, maxP detects genes that have small p-values in all studies but is usually less powerful than Fisher's method.

1.4.1.4 Generalized ordered statistics(rOp) maxP method is very conservative in that it requires all genes differentially expressed in all studies. A robust alternative is to apply the r-th ordered p-value (rOp). Let $p_{g(r)}$ denote the rth order statistic of K p-values, $p_{g1}, p_{g2}, \cdots, p_{gK}$.

$$V_g^{rop} = p_{g(r)} \tag{1.8}$$

Under the null hypothesis, V_g^{rop} has a *Beta* distribution with degrees of freedom r and $K - r + 1$. The r-th ordered p-value method (rOP) provides an alternative approach with robustness when large numbers of studies with potentially heterogeneous patient cohorts and variable quality are combined.

1.4.1.5 Stouffer's Method(Stouff) Stouffer's method is also called the inverse normal method proposed by Stouffer [90]. This procedure involves transforming each p-value to the corresponding normal score. and then taking the average. More specifically, define Z_k by $p_k = \Phi(Z_k)$, where $\Phi(x)$ is the standard normal cumulative distribution function. Then Stouffer's test statistic is defined as,

$$V^{Stuof} = \frac{\sum_{k=1}^{K} \Phi^{-1}(p_k)}{\sqrt{K}}, \tag{1.9}$$

Under null hypothesis, V^{Stuof} has the standard normal distribution. A weighted inverse normal method was generalized by Mosteller and Bush[69] to give different weights to each study according to their power. The weighted inverse normal test statistic is defined as

$$V^{W\_Stuof} = \frac{\sum_{k=1}^{K} w_k \Phi^{-1}(p_k)}{\sqrt{\sum_{k=1}^{K} w_k^2}}, \tag{1.10}$$

Under null hypothesis, V^{w_Stouff} also has the standard normal distribution. Whitlock [101] suggests that the weights can be chosen to be the inverse of squared standard error. He further shows weighted method is superior to the un-weighted version.

1.4.1.6 Adaptively weighted Fisher's Method(AW) Li and Tseng[52] elucidated two statistical hypothesis settings behind two separate biological goals in combining multiple array studies and developed an adaptively-weighted (AW) method. Genes that are differentially expressed in all studies were termed as HS_A type (hypothesis setting A) while genes differentially expressed in at least one study was called HS_B type. The adaptively-weighted statistic is defined as:

$$U_g(w_g) = -\sum_{k=1}^{K} w_{gk} \log(p_{gk}), \tag{1.11}$$

$$V_g^{AW} = \min_{w_g \in W} p_U(\mu_g(w_g)), \tag{1.12}$$

,where $w_g = (w_{g1}, w_{g2}, \cdots, w_{gK})$, and $\mu_g(w)$ is the observed statistic for $U_g(w)$, and $W = \{w | w_i \in \{0, 1\}\}$. Because the exact distribution of AW statistic can not be derived analytically, the p-value is usually calculated by permutation method. It has been shown that AW method has the power to identify DE genes considered significant in either partial or full data sets, and the resulting weight provides a natural categorization of the detected biomarkers for further biological investigation.

1.4.2 METHODS COMBINING EFFECT SIZES

The effect size (ES) reflects the magnitude of the disease effect or (more generally) the strength of association with clinical outcome and was widely used to combine information in meta-analysis. There are many different metrics that can be used to measure effect size,such as the r statistics(correlation coefficients)[81], d statistics[20, 44] and the odds ratio (OR)[35]. Here, we mainly focus on the d statistics proposed by Hegdes [44].Specifically,denote the gene expression value of gene $g(1 \leq g \leq G)$ in the disease (D) and control(C) groups of pair $i(1 \leq i \leq n_k)$ and study $k(1 \leq k \leq K)$ by X_{gki}^d and Y_{gki}^C, respectively. We assume that these studies are independent and that each of the X_{gki}^D and Y_{gki}^C is normally distributed. More

succinctly, $X_{gki}^D \sim \mathcal{N}(\mu_{gk}^d, \sigma_{gk}^2)$ and $Y_{gki}^C \sim \mathcal{N}(\mu_{gk}^c, \sigma_{gk}^2)$, $(1 \leq g \leq G, 1 \leq i \leq n_k, 1 \leq k \leq K)$. The effect size parameter δ_{gk} for gene g in kth study is defined as

$$\delta_{gk} = \frac{\mu_{gk}^d - \mu_{gk}^c}{\sigma_{gk}}, k = 1, 2, \cdots, K \tag{1.13}$$

To estimate the population effect size, the d statistic for standardized effect size measures is often used in the literature [15, 44]; however, it is a biased estimator of the population effect size(δ_{gk}), and underestimate when the sample size is relatively small. Thus, an unbiased estimator, d', is alternatively developed by multiplying a correction factor, $c(m) = \frac{\Gamma(m/2)}{\sqrt{m/2}\Gamma((m-1)/2)}$, in [66, 26], where $\Gamma(x)$ is the Gamma function and m is the degree of freedom of d statistics. Below, we show detailed formulation to estimate σ_{gk} from studies that are unpaired or paired design.

Computing d and d' from studies that are unpaired design: We can estimate the standardized mean difference (δ_{gk}) from studies that are unpaired design with two independent samples as:

$$d = \frac{\bar{Y}_D - \bar{Y}_D}{S_p} \tag{1.14}$$

where $\bar{Y}_D$ and $\bar{Y}_C$ are the sample means in the disease and control group, respectively. In the denominator, S_p is the pooled standard deviation across groups, $S_p = \sqrt{\frac{(n_D-1)S_D^2 - (n_C-1)S_C^2}{n_D+n_C-2}}$,where, S_D and S_C are the sample standard deviations in disease and control group, respectively. The estimator of the variance of d is given in [15, 44]

$$Var(d) = \frac{n_D n_C}{n_D + n_C} + \frac{d^2}{2(n_D + n_C)} \tag{1.15}$$

, which is an asymptotic estimator. Then, the exact form of the variance is provided by Hedges[43] and used by Marot [66], it can be shown that

$$Var(d) = \frac{m}{(m-2)\tilde{n}}[1 + \tilde{n}d^2] - \frac{d^2}{c^2(m)} \tag{1.16}$$

, where $\tilde{n} = \frac{n_D n_C}{n_D + n_C}$,and $m = n_D + n_C - 2$.

Correspondingly, the d' statistic and its variance is given by,

$$d' = c(m)d \tag{1.17}$$

$$Var(d') = c^2(m)Var(d). \tag{1.18}$$

Computing d and d' from studies that are paired design: While the studies are paired design with matched groups, the standardized mean difference (δ_{gk}) from studies can be estimated by :

$$d = \frac{\bar{Y}_D - \bar{Y}_D}{S_p}, \tag{1.19}$$

where $\bar{Y}_D$ and $\bar{Y}_C$ are the sample means in the disease and control group, respectively. In the denominator, S_p is the pooled standard deviation across groups, $S_p = \sqrt{S_D^2 + S_C^2 - 2S_D S_C r}$, where, S_D and S_C are the sample standard deviations in disease and control group, respectively, and r is the sample correlation coefficient, $r = \frac{\sum_{i=1}^n (Y_{iD} - \bar{Y}_D)(Y_{iC} - \bar{Y}_C)}{S_D S_C}$. The estimator of the variance of d is given in [9, 26]

$$Var(d) = \frac{2(1-r)}{n} + \frac{d^2}{2(n-1)} \tag{1.20}$$

, which is an asymptotic estimator. Then, the exact form of the variance is provided by Becker [9] and corrected by Morris [68], it can be shown that

$$Var(d) = \frac{2(1-r)}{n}(\frac{n-1}{n-3})[1 + \frac{nd^2}{2(1-r)}] - \frac{d^2}{c^2(n-1)} \tag{1.21}$$

, where n is the sample size in each group.

Correspondingly, the d' statistic and its variance is given by,

$$d' = c(m)d \tag{1.22}$$

$$Var(d') = c^2(m)Var(d). \tag{1.23}$$

13

1.4.2.1 Fixed Effects model(FEM) Fixed effects model is an often-used method of combining effect sizes when the studies to be combined are homogeneous, in which only within-study variability is considered. The assumption is that studies use identical methods, samples, and measurements; that they should produce identical results; and that differences are only due to within-study variation. The general model is given by

$$Y_{gk} = \mu_g + \alpha_{gk}. \tag{1.24}$$

Under the fixed-effect model we assume that there is one true effect size which underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error. Thus $Y_{gk} \sim N(\mu_g, \sigma_{gk}^2)$. The most efficient and unbiased estimator of 1 is the weighted average of estimates where the weight is determined by inverse of their standard errors. The estimate is

$$\hat{\mu}_g = \frac{\sum_{k=1}^{K} w_{gk} Y_{gk}}{\sum_{k=1}^{K} w_{gk}}, \tag{1.25}$$

where $w_{gk} = S_{gk}^{-2}$ and S_{gk}^2 s the estimated within-study variance in study k for gene g. The variance of $\hat{\mu}_g$ is then

$$Var(\hat{\mu}_g) = \frac{1}{\sum_{k=1}^{K} w_{gk}}. \tag{1.26}$$

So, a Z-score to test the null hypothesis that the common true effect μ_g is zero can be computed using

$$Z_g^{FEM} = \frac{\hat{\mu}_g}{\sqrt{Var(\hat{\mu}_g)}}. \tag{1.27}$$

which follows a standard normal distribution.

1.4.2.2 Random Effects model (REM) REM method is a popular method for combing effect sizes in meta-analysis, which makes the assumption that individual studies are estimating different treatment effects. Choi *et al*[15] were probably among the first authors to raise this issue of meta-analysis in the context of microarray data to find DE genes using this method, where the effect size is defined as the standardized mean difference $d = \frac{\bar{Y}_D - \bar{Y}_C}{S_p}$, where $\bar{Y}_D$ and $\bar{Y}_C$ represent the means of disease (MDD) and control groups, respectively, and S_p indicates an estimation the pooled variation. The corresponding model used was described as:

$$Y_{gk} = \mu_g + \alpha_{gk} + \eta_{gk}, \tag{1.28}$$

where Y_{gk} is the observed effect size in study k for gene g; the parameters α_{gk} and η_{gk} are the between-study and within-study errors, respectively. It assumeshin-study variances, respectively. Usually, the estimate of σ_{gk}^2 can be produced in each study k. The between-study variance can be estimated using a method of weighted moments (MM) estimator of τ_g^2, which can be derived from the heterogeneity statistic $Q_g = \sum_{k=1}^{K} w_{gk}(Y_{gk} - \hat{\mu}_g)^2$, where $\hat{\mu}_g = (\sum_{k=1}^{K} w_{gk}Y_{gk})/\sum_{k=1}^{K} w_{gk}$ is the feasible weighted least-squares estimator with weights $w_{gk} = s_{gk}^{-2}$, and s_{gk}^{-2} is the estimate of σ_{gk}^2. Then, the weighted unbiased MM estimator of τ_g^2 suggested by DerSimonian and Larird (DL)[22]: $\hat{\tau}_g^2 = \max\{0, \frac{Q-(K-1)}{s_1-(s_2/s_1)}\}$, where $w_{gk} = s_{gk}^{-2}$, and $s_r = w_{gk}^r (r = 1, 2)$, and K is the number of studies. Under the assumption that the gene expression levels were normally distributed, a z-score to test for DE genes was constructed as, $Z_g^{REM} = \frac{\hat{\mu}(\tau_g)}{\sqrt{Var(\hat{\mu}(\tau_g))}}$, which follows a normal distribution with zero mean and unit variance. The p-values of each gene could then be calculated and subsequent inferences could be made.

1.4.2.3 Fixed effects model versus Random effects model When we perform a meta-analysis using a fixed effects model or random effects model, one of first decisions we have to make is "Which model is more appropriate for current data?". The selection of a computational model should be based on our expectation about whether or not the studies share a common effect size and on our goals in performing the analysis. It makes sense to use the fixed effects model if we believe that all the studies included in the analysis are functionally identical. By contrast, when the data sets are accumulated from a series of studies

that had been performed by researchers operating independently, it would be unlikely that all the studies were functionally equivalent. Typically, the subjects or interventions in these studies would have differed in ways that would have impacted the results, and therefore we should not assume a common effect size. Therefore, in these cases the random effects model is more easily justified than the fixed-effect model. Therefore, a random effects model may be more general, in which both the random variation within the studies and the variation between the different studies is incorporated. However, more data are required for random effects models to achieve the same statistical power as fixed effects models. Testing how much heterogeneity there is is a way to determine whether the fixed effects model or random effects model is appropriate.Heterogeneity in meta-analysis refers to the variation in study outcomes between studies.

In practice, the question of which model is appropriate for given studies can be addressed by testing for the homogeneity of study effects. There are some general ways to assess heterogeneity in meta-analysis, but each has a liability for interpretation. In this dissertation, we focused on the one now widely-used chi-squared test (a Q-statistic) proposed by Cochran[8]. The Q statistic is defined as $Q_g = \sum_{k=1}^{K} w_{gk}(Y_{gk} - \hat{\mu}_g)^2$, where $\hat{\mu}_g = (\sum_{k=1}^{K} w_{gk} Y_{gk}) / \sum_{k=1}^{K} w_{gk}$ is the feasible weighted least-squares estimator with weights $w_{gk} = s_{gk}^{-2}$, and s_{gk}^{-2} is the estimate of σ_{gk}^2. Under the hypothesis of homogeneity, it follows a χ_{K-1}^2 distribution. A large observed value of the statistic Q relative to this distribution indicates rejection of the hypothesis of homogeneity, which therefore a random effect model is more appropriate. The previous method is based on gene by gene test. To further confirm the existence of the heterogeneities, we assume that the genes can be treated as independent samplings and the homogeneity can be explored over all the genes. The histogram of the observed Q values and quantile-quantile plots (Q-Q plot) of the observed versus expected values are used confirm the existence of the heterogeneity overall.

1.5 PATHWAY ENRICHMENT ANALYSIS

In above sections, meta-analysis methods that combine gene expression information across studies were reviewed. Gene expression information can be also integrated within a study. Specifically, instead of studying each gene individually, we can also study a gene set. A gene set is a pre-defined set of genes that may have similar locations or functions or form a particular pathway. If genes in a gene set act in concert, this gene set may have important biological effects on the phenotype of concern [91]. Thus, it is important to test whether a set of genes is coherently associated with the phenotype of interest. This type of analysis is called gene set enrichment analysis or pathway enrichment analysis[73, 91, 96]. When gene sets are defined by biological pathways, the term gene set enrichment analysis and pathway enrichment analysis are interchangeable. The common gene set/pathway databases include KEGG, Biocarta, and the gene ontology (GO) databases [37, 54]. The molecular signatures database (MsigDB) [91] is a collection of gene sets (including KEGG, Biocarta and GO) that has five major categories, including C1: positional gene sets; C2: curated gene sets; C3: motif gene sets; C4: computational gene sets and C5: GO gene sets. In this dissertation, pathway enrichment analysis was mainly used to evaluate the findings in in individual analyses and meta-analyses.

In the following sections, we give a brief review of two most commonly used pathway enrichment methods. Fisher's exact test is described in Section 1.5.1, and Kolmogorov-Smirnov (KS) test is described in Section 1.5.2.

1.5.1 Fisher's Exact Test

The Fisher's exact test method has been widely used in pathway enrichment analysis as a result of its simplicity[12, 24, 25, 106, 108]. The purpose for Fisher's exact test in this study was to determine whether the ratio of DE genes in a gene set was higher than the ratio outside of the pathway. If the ratio was higher than would be expected by chance, the pathway was referred to as an enriched pathway. The first step in Fisher's exact test method was to identify DE genes, the number of DE genes both inside and outside of the pathway

Table 1.2: 2×2 Contingency Table for Pathway Enrichment Analysis

	In pathway	out of pathway	Total
DE	c	l-c	l
Non-DE	t-c	G-l-t+c	G-l
Total	t	G-t	G

was then counted as a 2×2 contingency Table (Table 1.2). The p-value for enrichment of a pathway was calculated by testing the independence of the 2×2 contingency Table using Fisher's exact test. The null and alternative hypothesis for the Fisher's exact test is: $H_0 : p_1 = p_2$ and $H_a : p_1 > p_2$, where p_1 and p_2 are the probability of DE genes inside and outside of the pathway. In the Fisher's exact test, suppose a total number of G genes in the genome were considered, among them t genes were in the pathway, l genes were contained in the biomarker list and c genes were common to the target pathway (gene set) and the biomarker list (shown in Table 1.2). The p-value of the pathway enrichment was calculated from a hypergeometric distribution by $p = \sum_{x=c}^{\min(l,t)} \binom{t}{x} \binom{G-t}{l-c} / \binom{G}{l}$.

1.5.2   Kolmogorov-Smirnov (KS) Test

The Kolmogorov-Smirnov test (KS test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test). The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples, so it was widely used in pathway enrichment analysis[91, 61]. Specifically, the p-values calculated from individual analyses or meta-analyises for assessing the DE genes are classified into two categories, in the pathways (P) and out of pathway (P^C). Let $p_{(1)}, p_{(2)}, \cdots, p_{(n)}$ and $\tilde{p}_{(1)}, \tilde{p}_{(2)}, \cdots, \tilde{p}_{(m)}$ denote the order statistics of the p-values in P and P^C, respectively. The corresponding empirical

distribution functions, $\hat{F}_P(x)$ and $\hat{F}_{PC}(x)$ for P and P^C can be defined as:

$$\hat{F}_P(x) = \begin{cases} 0, & \text{if } x < p_{(1)} \\ k/n, & p_{(k)} \leq x < p_{(k+1)} \text{ for } k = 1, 2, \cdots, n-1 \\ 1, & \text{if } x \geq p_{(n)} \end{cases}$$

and

$$\hat{F}_{PC}(x) = \begin{cases} 0, & \text{if } x < \tilde{p}_{(1)} \\ k/m, & \tilde{p}_{(k)} \leq x < \tilde{p}_{(k+1)} \text{ for } k = 1, 2, \cdots, m-1 \\ 1, & \text{if } x \geq \tilde{p}_{(m)} \end{cases}$$

Let F_P and F_{PC} denote the population distribution for P and P^C, respectively. The one-sided two sample KS test can be defined based on the formula:

$$T_{KS} = \sup_x (F_P(x) - F_{PC}(x)) \tag{1.29}$$

in which the null hypothesis and the alternative hypothesis are:

$$H_0: \quad F_P(x) = F_{PC}(x) \quad \text{for all } X \tag{1.30}$$

$$H_a: \quad F_P(x) \geq F_{FC}(x) \quad \text{for all } x \tag{1.31}$$

$$F_P(x) > F_{PC}(x) \quad \text{for some } x \tag{1.32}$$

Under the null hypothesis, the rejection region has the form of $T_{KS} > K_\alpha$ at level of α. Rejection of H_0 means that P is stochastically less than P^C (the CDF of P lies above and hence to the left of that for P^C). In other words, the p-values of genes in the pathway P are stochastically less than the p-values of genes outside of pathway P^C. This indicates that genes in the pathway P have a stronger association with phenotype than genes from outside of the pathway P^C; thus, the pathway is of interest. Small p-value associated with KS test indicates a good performance of the methods.

2.0 A SYSTEMATIC STATISTICAL APPROACH TO INTEGRATE WEAK-SIGNAL MICROARRAY STUDIES ADJUSTED FOR CONFOUNDING VARIABLES WITH APPLICATION TO MAJOR DEPRESSIVE DISORDER

2.1 MOTIVATIONS

Microarray technology enables researchers to examine the expression of thousands of genes in parallel. Differentially expressed (DE) gene detection is one of the most common analyses in microarray data. In such an analysis, genes differentially expressed under multiple conditions are detected and are used for generating further biological hypotheses, developing potential diagnostic tools, or investigating therapeutic targets. The extensive applications of microarray technology have led to an explosion of gene expression profiling studies publicly available. However, the noisy nature of microarray data, together with small sample size in each study, often results in inconsistent biological conclusions [28, 92, 107]. Therefore, meta-analysis, a set of statistical techniques to combine multiple studies under related research hypotheses, has been widely applied to microarray analysis to increase the reliability and robustness of results from individual studies. In the literature, three major categories of meta-analysis methods have been applied to genomic meta-analysis: combining effect sizes [15, 66] , combining p-values [52, 79, 80] and combining rank statistics [21, 48]. In general, different approaches have different underlying assumptions and pros and cons in the application [78]. Major depressive disorder is a heterogeneous illness with mostly uncharacterized pathology. Despite many gene expression studies of MDD [3, 53, 85, 88, 87] published, the biological mechanisms of MDD remain mostly uncharacterized [7]. Although biomarkers and pathways have been identified in specific studies, the findings are not consistently ob-

served from study to study. This variability may be due to several factors. Firstly, MDD is thought to be a complex and heterogeneous disease [72], associated with multiple genetic, genomic, post-translational, and environmental factors. Furthermore, patients might have varying disease severity, with some having psychotic features as well as exposure to a variety of medications and dosage levels to control their illness. Secondly, the genetic disease effects are potentially confounded by many covariates, which include (1) demographical variables such as age, gender and race; (2) clinical variables such as anti-depressant drug usage, suicide and alcohol consumption; (3) technical variables inherent in the use of post-mortem brain samples, such as the pH level of brain tissues, brain region and postmortem interval (PMI). If the statistical models employed to identify differentially expressed genes fail to incorporate these sources of heterogeneity, not only can this reduce the statistical power, but also it will introduce sources of spurious signals to the gene detection. Finally, sample sizes for these studies are generally small (between 10-25 pairs of MDDs and controls) due to the limited availability of suitable brain specimens and the significant costs associated with their collection. In this paper, we propose a statistical framework to tackle weak signal expression profiles that have small sample size, case-control paired design and confounding covariates in each study. We use a set of five major depressive disorder (MDD) expression profiles as an illustrative example. In the literature, most analyses of similar data structure either ignored the potentially confounding covariates by using paired or unpaired t-test [18, 51, 98] or applied simple linear regression model to incorporate all covariates [67, 76]. The former approach undoubtedly ignored effects from confounding covariates; the latter approach was not efficient or even not applicable when the number of covariates is large and the number of samples in each study is small. In this paper, we will propose a framework that uses a random intercept model (RIM) to account for the case-control paired design and confounding covariates in single study analysis. An improved RIM with gene-specific variable selection (namely RIM_minP or RIM_BIC to be introduced later) will be performed to accommodate the small sample size and relatively large number of covariates in individual studies. We will then apply and compare three popular meta-analysis methods: Fisher's method [31, 32], inverse variance weighted random effects model [15, 44], and maximum p-value method [50, 86, 103]. Our proposed framework is general and applicable in commonly

21

Table 2.1: Data description of five MDD microarray studies

Study	Gender	Brain region	sample	Platform
MD1_ACC	M	ACC	32(16)	Affymetrix
MD2_ACC	M	ACC	20(10)	Illumina
MD3_ACC	F	ACC	50(25)	Illumina
MD1_AMY	M	AMY	28(14)	Affymetrix
MD3_AMY	F	AMY	42(21)	Illumina

encountered microarray meta-analysis of complex genetic diseases. Simulations considering various correlation structures among disease state, gene expression and covariates will be performed to demonstrate the better performance of this framework. The application of combining five MDD microarray studies also show improved DE gene detection power and superior statistical significance of pathway detection using our proposed method.

2.2 MATERIALS

Description of motivating MDD data: This research is motivated from the meta-analysis of combining five MDD transcriptomic studies. Brain tissues of three patient cohorts (MD1, MD2 and MD3) obtained from different sources at different time were analyzed. For all three patient cohorts, tissues from the anterior cingulated cortex (ACC) brain region were analyzed by microarray experiments independently to generate three microarray studies: MD1_ACC, MD2_ACC and MD3_ACC. Similarly, tissues from the amygdala (AMY) brain region in MD1 and MD3 cohorts were analyzed to generate MD1_AMY and MD3_AMY. Details of the five patient cohorts and microarray studies are available in Table 2.1. In each patient cohort, MDD patients were matched to control patients by three demographic variables: age, sex and race. Three additional clinical variables (alcohol consumption, history of taking anti-depressant drugs and history of committing suicide) and two technical vari-

Table 2.2: Pearson correlation between covariates in three MDD cohorts (collinearity evaluation)

	Age	Alcohol	Antidep	Suicide	pH	PMI
Age	—	(-0.05, 0.34, 0)	(0.15, 0.14, 0.04)	(0.02, -0.26, 0)	(-0.12, -0.01, -0.04)	(-0.19, -0.17, 0.37)
Alcohol	(-0.05, 0.34, 0)	—	(-0.21, 0.63, 0.28)	(0.41, 0.15, 0.22)	(0.09, 0.22, -0.08)	(-0.02, -0.29, -0.04)
Antidep	(0.15, 0.14, 0.04)	(-0.21, 0.63, 0.28)	—	(0.31, 0.19,0.22)	(0.18, 0.36,-0.21)	(-0.13, -0.35, -0.18)
Suicide	(0.02, -0.26, 0)	(0.41, 0.15, 0.22)	(0.31, 0.19,0.22)	—	(0.19, -0.3, 0.06)	(-0.17, -0.38, -0.02
pH	(-0.12, -0.01, -0.04)	(0.09, 0.22, -0.08)	(0.18, 0.36,-0.21)	(0.19, -0.3, 0.06)	—	(0.41, -0.03, -0.03)
PMI	(-0.19, -0.17, 0.37)	(-0.02, -0.29, -0.04)	(-0.13, -0.35, -0.18)	(-0.17, -0.38, -0.02)	(0.41, -0.03, -0.03)	—

ables (pH level of brain tissues and post-mortem interval PMI) were also available for each patient. Among the covariates described above, six variables (age, alcohol, drug, suicide, pH and PMI) are considered potential confounders in the DE gene detection of MDD. These six covariates were not highly correlated in our analysis and thus the collinearity issue does not exist in the linear models below (see Table 2.2).

Data preprocessing, gene matching and gene filtering: Microarray images were scanned and summarized by manufacturers' defaults. Data from Affymetrix arrays were processed by RMA method and data from Illumina are processed by manufacturer's software for probe analysis. When samples in each study were processed in multiple batches, potential batch effects were evaluated and normalizations were performed to correct batch biases when necessary. Probes (or probe sets) were then matched to official gene symbols using Bioconductore package. When multiple probes (or probe sets) matched to an identical gene symbol, the probe that generated the best disease association (by paired t-test) was selected to match to the gene symbol. This selection may cause potential bias but can increase statistical power in such weak-signal data. After genes were matched across five studies, 16,715 unique gene symbols were available across all five studies and intensities were all log-transformed (base 2). Two sequential steps of gene filtering were then performed. In the first step, we filtered out genes with very low gene expression that were identified with small average expression values across majority of studies. Specifically, mean intensities of each gene across all samples in each study were calculated and the corresponding ranks were obtained. The sum of such ranks across five studies of each gene was calculated and genes with the highest 30% rank sum were considered un-expressed genes (i.e. small expression in-

tensities) and were filtered out. Similarly, in the second step, we filtered out non-informative (small variation) genes by replacing mean intensity in the first step with standard deviation. Genes with the lowest 40% rank sum of standard deviations were filtered out. Supplement Figure 1 shows the preprocessing diagram and the number of genes remained in each preprocessing step. Finally, 7,020 matched genes ($16715 \times 0.7 \times 0.6 = 7020$) in five studies were analyzed.

2.3 METHODS

2.3.1 Single study analysis for DE gene detection

Paired t-test and Wilcoxon signed rank test: As a comparison, paired t-test and Wilcoxon signed rank test were performed. These two methods took into the MDD and control paired design into consideration but ignored the confounding covariates.

Random intercept model (RIM) and fixed effects model (FEM): To account for paired design (MDD samples paired with corresponding controls) and existence of MDD related covariates, we applied a random intercept model (RIM). For a given gene g, we fit the model

$$Y_{gik} = \mu_g + \beta_{g0}X_{0ik} + \sum_{l=1}^{L} \beta_{gl}X_{lik} + \alpha_k + \epsilon_{gik}, \tag{2.1}$$

In the model, Y_{gik} was the gene expression value of gene $g(\leq g \leq G)$ and sample i ($i = 1$ for control and 2 for MDD) in pair $k(1 \leq k \leq K)$. X_{0ik} was the disease label that took value one if the sample was MDD and Zero if sample was a control. X_{lik} represented values for potential confounding covariate l ($1 \leq\leq 6$; 0-1 binary for alcohol, drug and suicide and numerical for age, pH and PMI). α_k was the random intercept from a normal distribution with mean zero and variance τ_g^2, which represented the deviation of averaged expression values in the kth pair from the average in the whole population. Finally, ϵ_{gik} were independent random noises that followed a normal distribution with mean zero and variance σ_g^2. Under this model, β_{g0}

24

was the disease effect of gene g and was the parameter of major interest. To obtain an MDD-associated biomarker candidate list in a single study analysis, likelihood ratio test (LRT) was used to assess the p-values of testing $H_0 : \beta_g 0 = 0$ (vs $H_A : \beta_g 0 \neq 0$). The p-values were then be corrected by Benjamini-Hochberg procedure [8] for multiple comparison.

Fixed effects model (FEM) below ignores the paired design while still considers the covariates in the model. It can be used when diseased and control samples are not paired.

$$Y_{gik} = \mu_g + \beta_{g0} X_{0ik} + s \sum_{l=1}^{L} \beta_{gl} X_{lik} + \epsilon_{gik} \tag{2.2}$$

RIM and FEM with variable selection: Although RIM model can effectively adjust for confounding covariates in DE gene detection, the small sample size (10-25 pairs) and relatively high number of potential confounders (6 covariates) can make the model inefficient and impractical. In this paper, we developed and evaluated two choices of variable selection procedures in the random intercept model (namely, RIM_BIC and RIM_minP). Specifically, all possible RIM models that included at most two (0, 1 or 2) clinical variables were computed and compared. In RIM_BIC, the model with the smallest Bayesian Information Criterion (BIC) [84] value was selected. For RIM_minP, we selected the model that yielded the smallest p-value associated with the likelihood ratio test for testing the disease effect $H_0 : \beta_{g0} = 0$. Conceptually, BIC selected the model with the best model fitting and prediction while minP focused on the model that gave the best statistical significance of the disease effect. This additional variable selection avoided to include more than 2 clinical variables in the model and allowed assessment of biomarkers affected by different sets of covariates in each gene (e.g. gene A is confounded by alcohol while gene B is confounded by drug), which biologically gave more appealing conclusions and interpretations. Similar to RIM model, likelihood ratio test were used to generate p-values of testing $H_0 : \beta_{g0} = 0$ in each gene for the selected model by BIC or minP. These attached p-value numbers were, however, not the true p-values for DE gene detection since they were biased from the variable selection procedure and the type I error control was voided. As a result, we performed a permutation test that randomly permuted the disease labels within each pair to generate a null distribution for p-value assessment. Figure 2.1 shows the simulated null distribution from permutation analysis. The skewed distribution deviating from uniform distribution between 0 and 1 showed the need
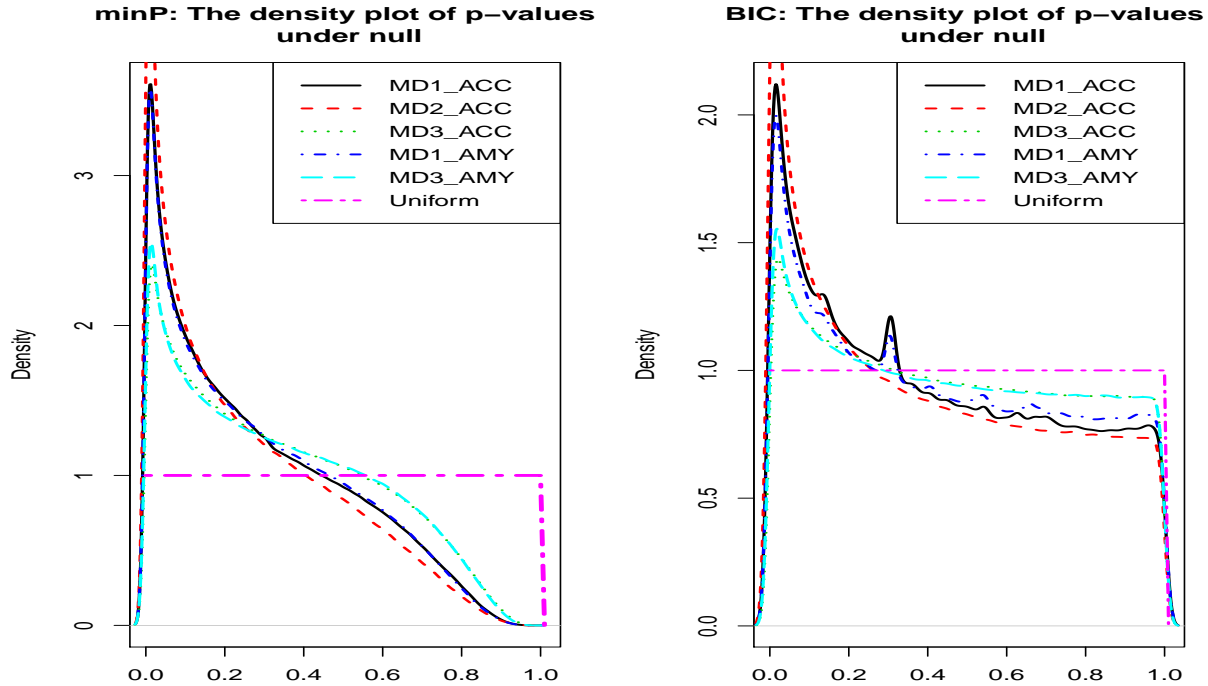
Figure 2.1: Simulated null distributions of disease effect p-value in the best model (left: RIM_minP; right: RIM_BIC) from permutation analysis in the five MDD studies. The result shows bias (deviation from uniform distribution) caused by variable selection.

of the permutation analysis for p-value correction. Subsequently, the resulting unbiased p-values after permutation correction were then corrected by Benjamini-Hochberg procedure for multiple comparison in each study for DE gene detection. Detailed algorithm of the permutation analysis is described in AppendixA. In contrast to RIM_minP and RIM_BIC, we denote by RIM_ALL the RIM model that includes all covariates without variable selection.

Testing significance of interaction terms of each covariate: In the literature, age as well as other covariates has been found to be confounders of the disease effect with significant interaction term in some important biomarkers. [34, 39] In other words, the disease effect on gene expression may be affected by age differently in older and in younger cohorts.

26

Table 2.3: The number of significant interaction terms between disease state and covariates in FEM model and RIM.

| | | FEM | | | | | RIM | | | | |
		MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY
	Age	0	0	0	0	0	54	1	0	0	0
FDR=0.05	pH	0	1	0	0	0	0	7	2	0	0
	PMI	0	0	0	0	0	0	2	0	0	0

To evaluate the overall impact of the interaction terms in each covariate, we performed the following simple linear model

$$Y_{gik} = \mu_g + \beta_{g0}X_{0ik} + \beta_{gl}X_{lik} + \gamma_{gl}X_{0ik}X_{lik} + \epsilon_{gik} \tag{2.3}$$

, and random intercept model

$$Y_{gik} = \mu_g + \beta_{g0}X_{0ik} + \beta_{gl}X_{lik} + \gamma_{gl}X_{0ik}X_{lik} + \alpha_k + \epsilon_{gik}, \tag{2.4}$$

where the notations were the same as in FEM model and RIM model with only one covariate l included and a corresponding interaction term involved. We performed likelihood ratio test for $H_0 : \gamma_{gl} = 0$ to test the statistical significance of the interaction term of gene g and covariate l. Table 2.3 summarizes the number of significant interaction terms in the genome of each covariate. The result shows that the interaction terms between each covariate (Age, pH or PMI) and MDD were not significant in most of the genes under false discovery rate FDR = 5% (Benjamini-Hochberg correction). As a result, we did not consider the interaction terms in our RIM models hereafter.

2.3.2   Meta-analysis for DE gene detection

Among the many microarray meta-analysis methods used in the literature, most methods have their pros and cons depending on the data structure and biological goal [47, 78]. In this paper, we compared three most popular methods described in sections 1.4.1.1 ,1.4.1.3,1.4.2.2, repectively, Fisher, maxP and IVW.

2.3.3   Pathway analysis

We applied Fisher's exact test to detect enriched pathways in detected DE gene lists from individual study analyses and three meta-analysis methods. Pathways were obtained from the MSigDB database [91] and Gene Ontology, in which we only considered pathways that included at least five genes. We evaluated C1-C4 in MSigDB and gene ontologies in "GOstats" package in Bioconductor. In the Fisher's exact test, suppose a total number of g genes in the genome were considered, among them t genes were in the pathway, l genes were contained in the biomarker list and c genes were common to the target pathway (gene set) and the biomarker list. The p-value of the pathway enrichment was calculated from a hypergeometric distribution by $p = \sum_{x=c}^{\min(l,t)} \binom{t}{x}\binom{g-t}{l-c}/\binom{g}{l}$. We evaluated p-values for each pathway independently and then corrected the p-values by Benjamini-Hochberg procedure for multiple comparison to generate q-values. As will be seen in Figure 6, the three meta-analysis methods detect different sets of DE genes. To avoid bias and as an attempt to retain advantages from all three meta-analysis methods, we develop a minimum p-value method (minP) for integrating results from Fisher, maxP and IVW. Specifically, a minP statistics is defined as $U_g^{minP} = \min(p_g^{Fishe}r, p_g^{maxP}, p_g^{IVW})$, where p_g^{Fisher}, p_g^{maxP} and p_g^{IVW} are p-values of gene g generated by each meta-analysis method. Under null hypothesis, U_g^{minP} follows a beta distribution with degrees of freedom 1 and 3. The resulting p-values are then adjusted by Benjamini-Hochberg procedure for q-values.

2.3.4 Post hoc analysis on the confounding variables after meta-analysis

An essential advantage of our gene-specific variable selection scheme is the possibility of post hoc analysis on the selected confounders across studies in a genome-wide scale. Three questions can be explored and answered: (1) Which variable(s) is the most or least frequently included in the model selection to confound with disease effect? (2) Are variables repeatedly selected across studies more frequently than by random (e.g. alcohol is selected in most or all studies in a given gene)? (3) Are the directions of effect sizes of a variable consistent across studies (e.g. patients who take alcohol have higher expression than non-alcohol in most studies for a given gene)? For the first question, we first generated a list of DE genes under a given FDR threshold and counted the frequency of each variable being selected in the gene list. The variables were ranked according to the frequencies in each study and a rank average of each variable was calculated across five studies. A small averaged rank of a given gene showed frequent appearance of the variable in the DE genes' models and was a frequent confounder. For question (2), we computed a pair-wise co-appearance score ($T1$) for a given gene set and assessed its statistical significance. For example, VGF in Table 4 had detected age effects in 2 studies, alcohol effects in 3 studies, anti-depressant effect in 1 study and suicide effect in 3 studies. By summing up co-appearing pairs of the five studies in each variable, we obtained a $T_{1,g}$ statistics of 7 ($C_2^2 + C_2^3 + 0 + C_2^3 = 7$) for g=VGF. Summing up all 10 genes, we obtained $T_1 = \sum_g T_{1,g} = 66$. Permutation test was then performed to assess the statistical significance of $T1$.

To answer question (3), we further computed rate of expression concordance among all co-appearance pairs. Specifically, we examined all co-appearing pairs that contributed to $T1$ and count the number of pairs that are concordant (up-regulation in both studies or down-regulation in both studies). The total aggregated score for pair-wise concordance was denoted as $T2$ and the ratio of concordance was $R = T1/T2$. In the example of Table 4, 45 out of 66 co-appearing pairs were concordant and $R = 0.68$. Similarly, permutation test was performed to assess the statistical significance of observed R scores. Detailed mathematical notation and permutation algorithm are outlined in Appendix B.

2.3.5 Evaluation and simulation

To evaluate performance of different models and methods in a real data analysis, we compared the number of detected DE genes and the statistical significance of important pathways. For the former criterion, we argue that with adequate modelling and multiple comparison correction, detecting more DE genes shows better statistical power of a method and should be a preferred method. There is, however, no rigorous proof to reason that detecting more DE genes guarantees better performance of a method, in terms of its type I error control and statistical power. Since the type I error and statistical power could not be evaluated in real data analysis, we performed extensive simulations to facilitate the evaluation. For a given gene, we considered three variables of a continuous vector of gene expression Y, a corresponding binary vector of disease state X and multiple vectors of potential binary confounding covariate Z. Figure 2.2 shows three correlation structures of interest among (X, Y, Z) that are systematically simulated. Scenario I demonstrated that both disease state X and confounding variables Z affect gene expression, a model we are most interested in this paper. Scenario II and III showed situations when confounding variables Z did not directly affect gene expression Y. In these latter two scenarios, including confounding variables Y in the model should not improve performance. The detailed simulation scheme and evaluation criteria are available in the Supplement Materials Part III. For each scenario, we simulated a data set with 1000 independent genes and 50 samples (25 diseased and 25 controls). Among the 1000 genes, 100 are true DE genes and 900 are non-DE genes. t-test, FEM_minP, FEM_BIC and FEM_ALL were applied to evaluate the effect of modelling confounding variables and variable selection in each correlation structure. We repeated the simulation 50 times. Type I error and power were calculated for each method in each data set and averaged over 50 repeated simulations.
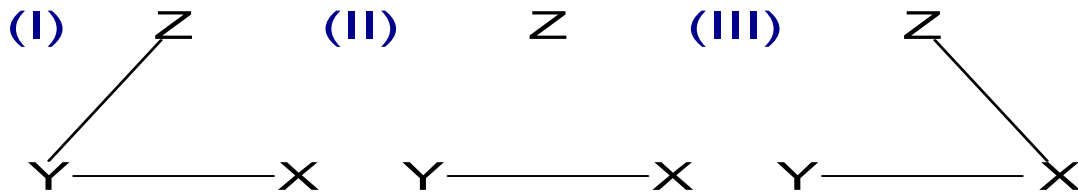
Figure 2.2: Three correlation structures of interest among disease variables X, gene expression variable Y and covariates Z that are used in the simulation. Scenario I: gene expression depends on both disease state and covariates. Scenario II: gene expression depends only on disease state. Scenario III: gene expression depends on disease state directly and depends on covariates indirectly through disease state.

2.4 RESULTS AND DISCUSSION

2.4.1 Recommended statistical framework

From the motivating MDD example, we proposed a statistical framework to consider potential confounding covariates, paired design and gene-specific variable selection in the meta-analysis modelling. Figure 2.3 shows a diagram of the framework. The framework consisted of four major steps: individual study analysis, meta-analysis, pathway analysis and post

hoc analysis. In the first "individual study analysis" step, collinearity of confounders was assessed and RIM_minP or FEM_minP method with variable selection was applied depending on paired or un-paired design. One or multiple meta-analysis methods were applied and compared in Step II. Pathway analysis was then performed on the detected DE gene list(s) to identify enriched pathways in Step III. Finally, post hoc analysis was performed to summarize importance of each confounding variables and to evaluate the consistency of disease effects and confounders' effects across studies. This framework is general and abstract that can be applied to any weak-signal data from a complex disease similar to the motivating MDD example.

2.4.2   Comparison of various methods in single study analysis

<u>Adjusting confounders and variable selection improve DE gene detection</u> For each single study analysis, we compared the number of detected DE genes under different p-value thresholds (p=0.001, 0.005, 0.01 and 0.05) from different methods. In Figure **??**, RIM_minP and RIM_BIC both detected more DE genes than RIM_ALL, showing the fact that variable selection helped to ignore irrelevant clinical variables when sample size was small. Among the two variable selection methods, RIM_minP detected more genes than RIM_BIC, supporting that the focus of RIM_minP to obtain the most significant disease effect outperformed RIM_BIC's focus for best model fitting in this example. Under p=0.005, RIM_minP detected (1.7 to 2.3) times of DE genes than RIM_BIC and (1.5 to 6) times than RIM_ALL in the five studies. The result suggested that RIM_minP is the most effective method in this data set to incorporate confounding variables in the model. In Figure 2.5, RIM_minP was further compared to paired t-test (PT) and Wilcoxon signed rank test (WT) and was found to detect more DE genes, showing the advantage of incorporating confounding covariates in the model. RIM_minP identified (0.8 to 3.9) times of DE genes than PT and (2.5 to 6.4) times than WT under p=0.005.

   <u>Paired design improves DE gene detection:</u> To evaluate the improvement of including paired design in the model, we compared RIM_minP and FEM_minP in Figure 2.6. We observed more powerful DE gene detection of RIM_minP compared to FEM_minP.
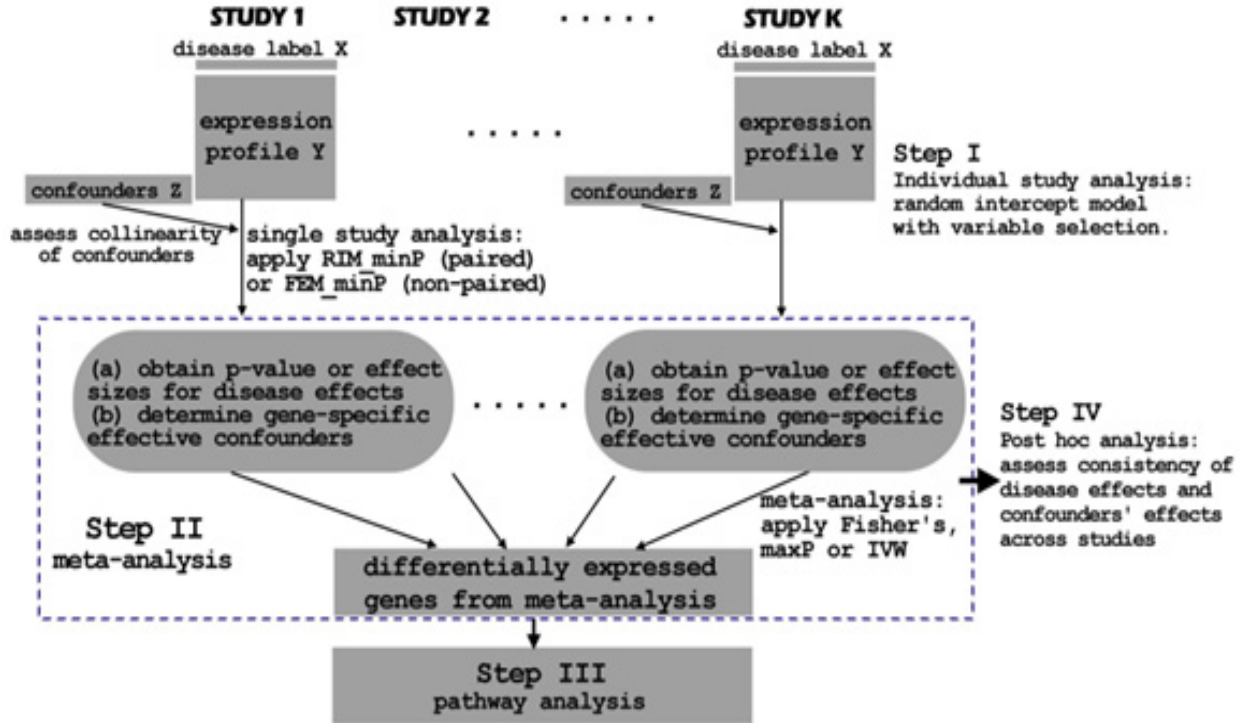
Figure 2.3: Three correlation structures of interest among disease variables X, gene expression variable Y and covariates Z that are used in the simulation. Scenario I: gene expression depends on both disease state and covariates. Scenario II: gene expression depends only on disease state. Scenario III: gene expression depends on disease state directly and depends on covariates indirectly through disease state.

RIM_minP detected more DE genes than FEM_minP in most studies except for MD1_AMY. The result showed that pairing cases to controls by age, race and sex usually helped increase statistical power.

Conclusion In conclusion, incorporation of potential confounding covariates with variable selection and considering paired design in the model performed the best. We used RIM_minP hereafter for single study analyses and as the foundation of meta-analysis. In Table 1, the first five columns show the number of biomarkers detected by RIM_minP under different
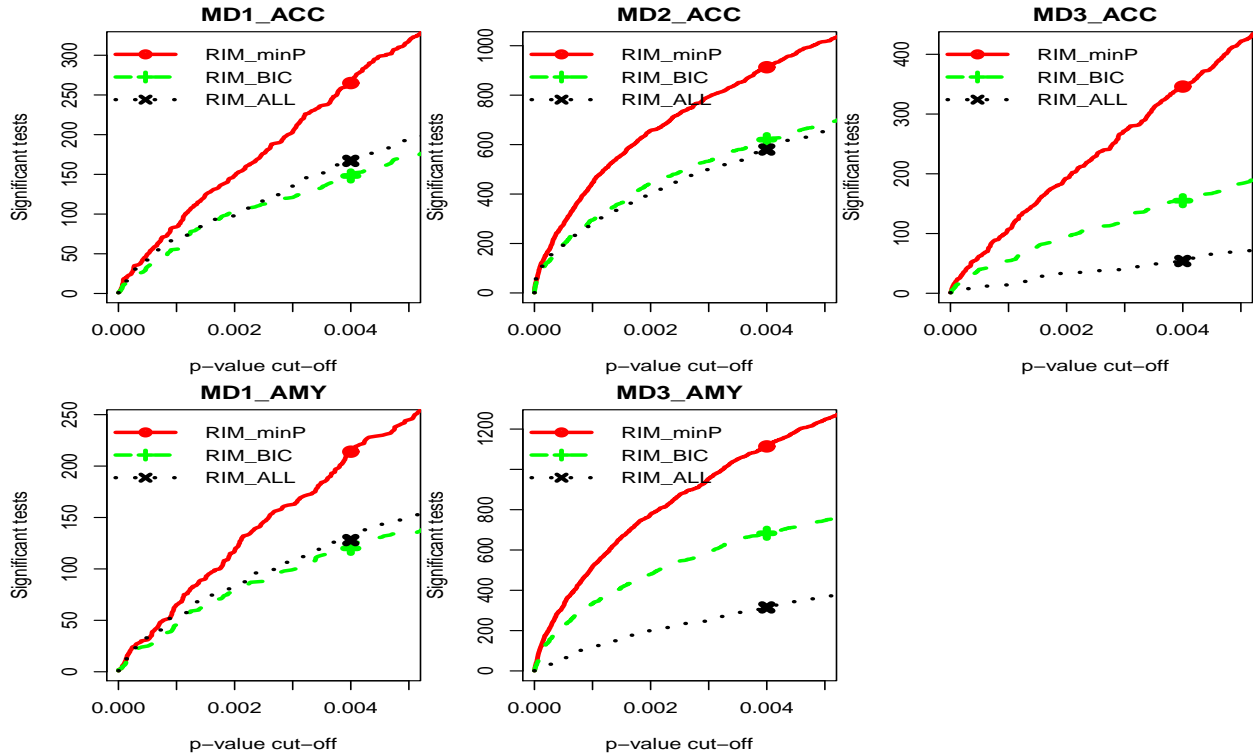
Figure 2.4: Comparison of RIM_minP, RIM_BIC and RIM_ALL in individual study analyses. The result showed that RIM_minP detected the largest number of DE genes among the three methods.

p-value and false discovery rate (FDR) thresholds. After multiple comparison correction by Benjamini-Hochberg procedure, only MD3_ACC detected one DE gene and all other four studies detected none DE gene under FDR=5%. This motivated us to perform meta-analysis below to increase the statistical power of DE gene detection.

2.4.3 Comparing three meta-analysis methods in combining all five studies

In the literature, many microarray meta-analysis methods have been proposed and compared [13, 47, 78]. As was discussed in the method section, different methods have different strength for detecting different types of differentially expressed genes. In Li et al [52], genes that are
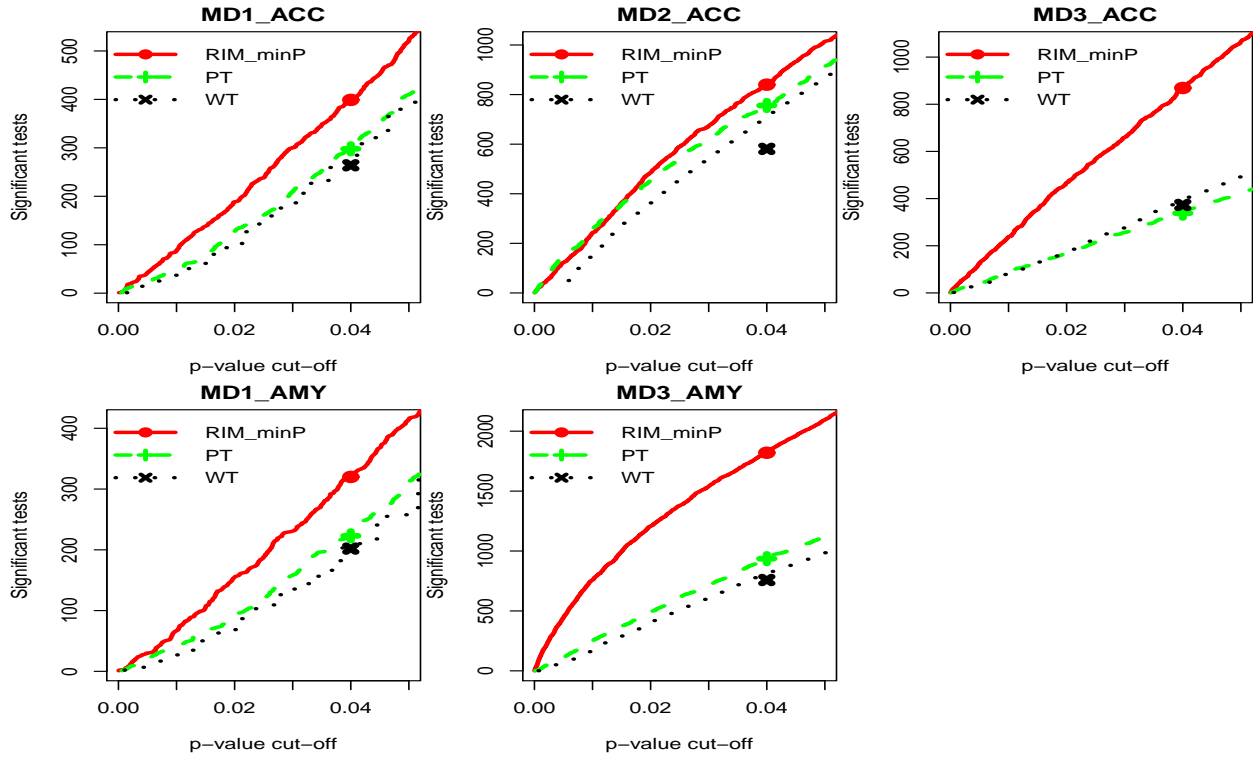
Figure 2.5: Comparison of RIM_minP, paired t-test (PT) and Wilcoxon signed-rank test (WT) in individual study analyses. The result showed that RIM_minP detected the largest number of DE genes among the three methods.

differentially expressed in all studies were termed as HS_A type (hypothesis setting A) while genes differentially expressed in at least one study was called HS_B type. Among the three methods compared in this paper, maxP and IVW were methods that detect HS_A type DE genes, while Fisher's method detected HS_B type DE genes. Table 2.4 shows the number of detected DE genes from five individual study analyses and from meta-analysis of five studies (Meta_3ACC+2AMY by Fisher, maxP and IVW) under different p-value and FDR threshold. A Venn diagram of DE gene lists detected by three meta-analysis methods under p=0.005 is shown in Figure F1. The result showed that the three meta-analysis methods detected different sets of DE genes, suggesting different algorithms and assumptions behind
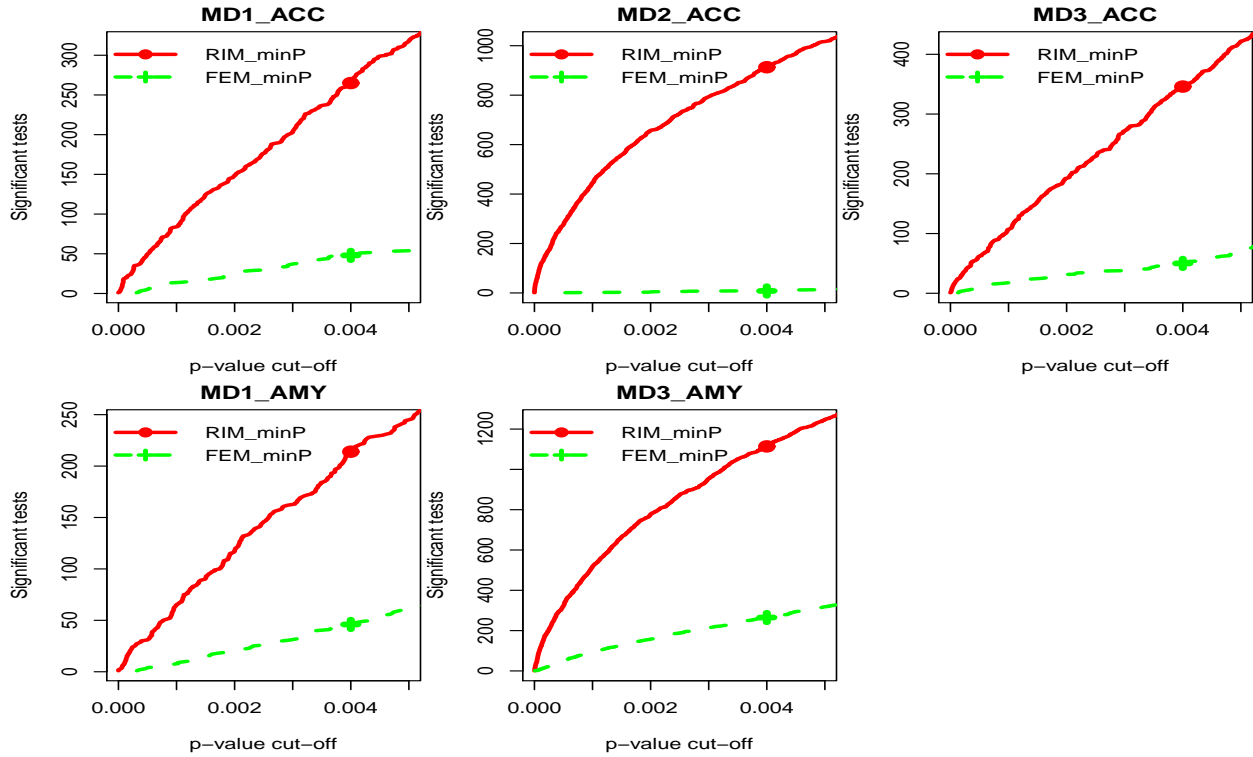
Figure 2.6: Comparison of RIM_minP and FEM_minP in individual study analyses. The result showed that RIM_minP usually detected more DE genes.

the methods. Figure 2.8 shows heamaps on genes detected by Fisher alone (A), maxP alone (B) or both (C). In Figure 2.8 A, majority of DE genes detected by Fisher but not by maxP were dominated by strong differential expression in one or two studies (many in MD3_AMY and some in MD2_ACC or MD3_ACC). Although Fisher's method has been popularly applied in the microarray meta-analysis literature, the result showed its weakness to be dominated by single strong signal studies that included potential false positives. On the other hand, maxP had better power to detect many genes with weak DE evidence in all studies (Figure 2.8B) that Fisher's method cannot detect. Conceptually, we were more interested in identifying genes differentially expressed across all studies through maxP or IVW although we still apply a unified minimum p-value method to integrate advantages of all three meta-analysis

Table 2.4: Results of individual study analyses and meta-analysis combining p-values calculated from RIM_minP

| | Individual studies | | | | | Meta-analysis | | | | | | | | |
| | | | | | | (3ACC) | | | (3ACC+2AMY) | | | (2AMY) | | |
	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	Fisher	maxP	IVW	Fisher	maxP	IVW	Fisher	maxP	IVW
p=0.001	5	25	29	3	118	73	123	246	220	255	425	64	50	340
p=0.005	42	122	123	30	448	304	371	572	658	664	569	283	185	828
FDR=0.05	0	0	0	0	0	8	86	106	574	605	552	0	0	143
FDR=0.1	0	0	1	0	882	149	534	812	1815	1909	616	33	0	996

methods. For a complete comparison, we also performed partial meta-analysis by combining three ACC studies (Meta_3ACC) and two AMY studies (Meta_2AMY). The results are outlined in Table 2.4.

To further evaluate the biological meaning of the detected DE genes by various methods, pathway analysis was performed. For a fair comparison, DE gene lists detected by various methods (individual study analyses or Fisher, maxP and IVW meta-analysis methods) under p=0.005 without multiple comparison adjustment were evaluated and Table 2 (See Appendix) shows the pathway analysis results. Comparing single study analysis and meta-analysis results in pathway analysis, the p-values of many important psychiatric related pathways were much smaller in the meta-analysis results than those from single study analysis, showing increased statistical power by meta-analysis in the functional analysis. For example, the gene set "ASTON_MAJOR_DEPRESSIVE_DISORDER_DN" obtained from a previous MDD study had none to marginal statistical significance in pathway analysis of each individual study. Meta-analysis by Fisher or maxP method generated high statistical significance from pathway analysis (p=4E-12 and 2E-9). Pathways with known or putative correlation with MDD were marked with asterisk in Table (in Appendix). Many of these insightful
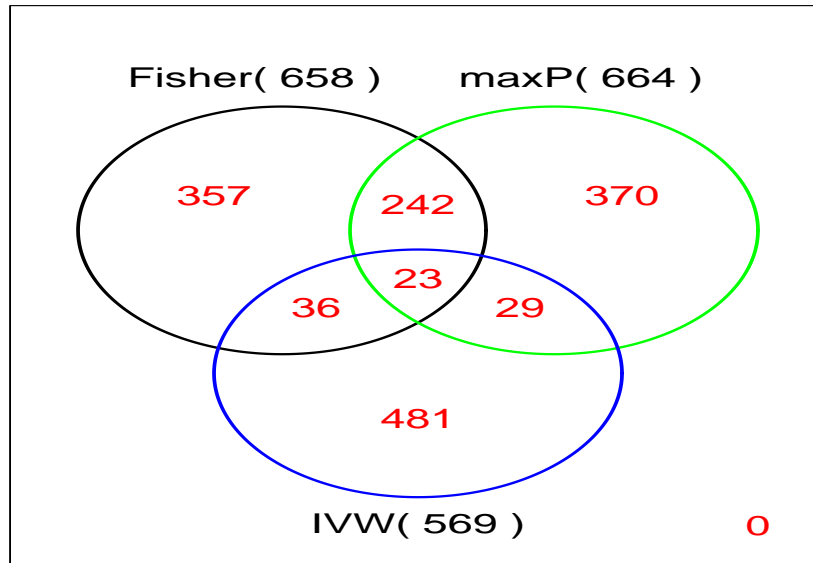
Figure 2.7: Venn diagram of DE gene lists obtained from Fisher, maxP and IVW under 0.005 p-value threshold.

pathways were, however, generated by split of the three methods, suggesting that Fisher, maxP and IVW may have their own characteristics and advantages to detect different pathways. To generate a more unbiased and unified result, we applied a minimum p-value (minP) method to integrate pathway analysis results from Fisher, maxP and IVW (details described in "Method" section). Table includes 87 pathways (in C2, C3 and gene ontology databases) detected by minP under a loose p-value threshold at 0.01 without multiple comparison. C1 and C4 databases in MsigDB did not generate any pathway with high statistical significance (3 out of 326 pathways in C1 and 9 out of 881 in C4 with p-value smaller than 0.01) and thus are excluded from the presentation. Many pathways listed in Appendix Table 1 were found related to signal transduction, neural development and neuropsychiatry, providing deep in-
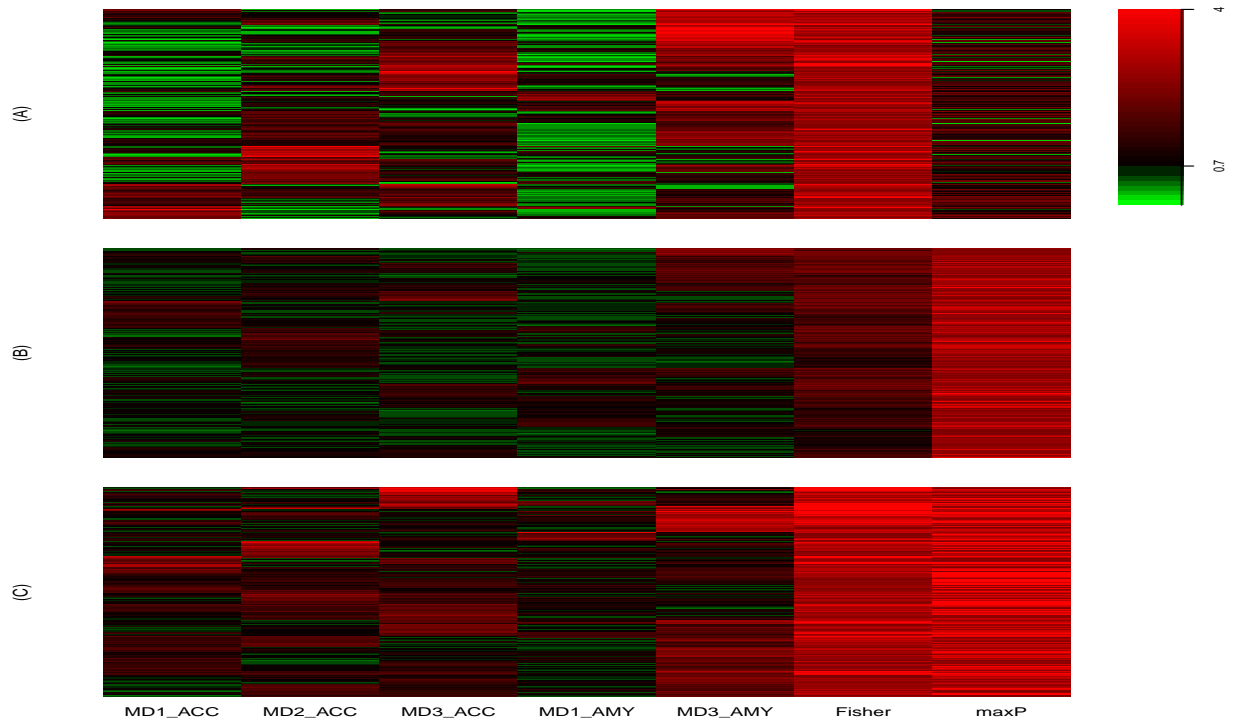
Figure 2.8: Heatmap of minus log10-transformed p-values obtained from all five studies and meta-analysis for detecting DE genes. Red indicates small p-values and green indicates large p-values. (A) DE genes detected by Fisher's method but not by maxP method; (B) DE genes detected by maxP but not by Fisher's method; (C) DE genes detected by both Fisher and maxP method.

sight to the underlying genetic mechanism of MDD. In C2 curated gene set category, the gene sets ("ASTON_MAJOR_DEPRESSIVE_DISORDER_DN") includes numerous down-regulated genes that are specific to oligodendrocytes [3], the major myelin-forming cell type in the brain, which have been shown to be reduced in numbers in depression [41]. The gene set ("BLALOCK_ALZHEIMERS_DISEASE_DN") highlights the potential biological interaction between the co-occurrence of de-pressive symptoms in Alzheimer disease patients. Other identified gene sets relate to biological functions that have been identified as causative, or at

least as risk factors for the development of depres-sion (i.e. interleukin-related function, "MA-HAJAN_RESPONSE_TO_IL1A_UP", "MARZEC_IL2_SIGNALING_UP"). Other gene sets potentially relate to biological functions involved in the therapeutic treatment of the illness (ie, "ROPERO_HDAC2_TARGETS ", "RODRIGUES_THYROID_CARCINOMA_DN". In C3 motif gene sets, six motifs "CCTGCTG,MIR-214", "CTTTGCA,MIR-527", "CAGGTG_V\$E12_Q6", "GAANYNYGACNY_UNKNOWN", "V\$SP1_Q2_01", "AAGCCAT,MIR-135A, MIR-135B", "V\$CDPCR3HD_01" for various miRNA and transcription factor targets had p-values smaller than 0.001. For gene ontologies, many pathways related to metabolism and signal transduction were identified. In particular, "cell communication", "nervous system de-velopment", "synaptic transmission" and "ensheathment of neurons" were insightful pathways related to MDD.

2.4.4   Distribution of covariate inclusion in the models of detected DE genes

To evaluate the impact of covariates on the gene expression values and degree of confounding with disease effect, especially among DE genes, we counted the number of appearances of covariates in the RIM_minP models for 664 DE genes detected by maxP method under 0.005 p-value threshold. We calculated the rank of each covariate in each study and computed rank averages of each co-variate to indicate relative degree of frequency that a covariate impacted gene expression and confounds with disease effect (see Table 2.5). PMI (appeared in 16-24% models of 664 DE genes) and pH (appeared 14-37%) consistently had high rank, indicating that they seldom confounded and influenced the disease effect estimate. Age (appeared 28-39%), alcohol (appeared 25-43%) and antide-pressant (appeared 17-48%) were three factors that consistently ranked among the most influential factors. Suicide ranked among the lowest in three studies (appeared 44-52%) but the highest in two studies (appeared 15-19%). The ranking of MD3_ACC and MD3_AMY was highly correlated (Spearman correlation=0.9) and the correlation between rankings of MD1_ACC and MD1_AMY was also high (Spearman correlation=0.71). The high within cohort correlations showed a cohort dependent structure and suggested that more studies may be needed to provide empirical evidence on the covariate impacts, particularly for the impact of antidepressant and suicide.

Table 2.5: Frequency of covariates appearing in RIM_minP models among 664 DE genes detected by maxP method under p-value threshold 0.005. Rank is shown in parentheses and rank average of each covariate is calculated to indicate relative degree of frequency that a covariate impacts gene expressions and confounds with disease effect

	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	rank average
Age	186 (4)	258 (2)	238 (2)	192 (3)	233 (3)	2.8
Alcohol	243 (3)	230 (3)	165 (4)	286 (1)	173 (4)	3
Antidep	291 (1)	115 (6)	202 (3)	277 (2)	320 (1.5)	2.7
pH	247 (2)	143 (4)	127 (6)	176 (4)	94 (6)	4.4
PMI	109 (5)	139 (5)	160 (5)	154 (5)	109 (5)	5
Suicide	97 (6)	295 (1)	345 (1)	126 (6)	320 (1.5)	3.1

To further explore effects of covariates, we identified a set of 10 genes that have been previously associated with MDD in the literature (see Appendix Figure F). Intuitively, we expected that a co-variate should be included in the model across studies more frequently than by random and effects of a covariate should have consistent differential expression direction across studies. We constructed two hypothesis testing using the co-appearing statistics T1 and concordant ratio statistics R described in Method section and performed the tests on the 10 MDD-related genes and on 664 DE genes detected by maxP under p=0.005 threshold. The result showed weak to marginal statistical significance of the first hypothesis (p=0.172 for the 10 MDD genes and p=0.05 for 664 DE genes), suggesting covariates were consistently selected across studies. For the second hypothesis, tests for both 10 MDD gene list and 664 DE gene list were statistically significant (p=0.019 and 0.002). The result demonstrated that covariates overall impacted gene expression changes consistently and confounded with disease effects among the two MDD-related candidate gene lists tested.

2.4.5 Simulation results

Simulation results are shown in Table 2.6. In Scenario I simulation, the effect of disease state X on gene expression Y was confounded by two out of ten clinical variables in Z. The result showed that t-test had low statistical power due to the confounders (power=0.679). FEM_ALL also had low power due to the inclusion of all ten clinical variables in the model (power=0.697). Both FEM models with variable selection perform well. FEM_BIC performed slightly better than FEM_minP (power=0.729 versus 0.746). The type I errors for all methods were close to the nominal 5% rate, showing adequacy of the models and statistical inference. For Scenario II, all clinical variables were independent from the gene expression. Not surprisingly, t-test performed the best with statistical power 0.938. FEM_minP and FEM_BIC both had similar high power at 0.929 and 0.925. FEM_ALL forced all variables in the model and obtained a statistical power at 0.85. From Scenario I and Scenario II simulation, FEM_BIC and FEM_minP performed well in both extreme cases, demonstrating its sensitivity and robustness. Scenario III examined a situation that variables Z impact gene expression Y through disease state X. Similar to Scenario II, t-test performed the best in this situation since Z is not confounded (power=0.938). Both FEM_BIC and FEM_minP had similar high power (power=0.925 and 0.916) but FEM_ALL again had low power (power=0.851). Overall, the simulation results confirmed our findings in MDD data analysis that variable selection by BIC and minP procedures had better sensitivity and robustness in DE gene detection.

2.5 DISCUSSION

In this paper, we described a statistical framework, namely MetaACV (Meta-analysis adjusted for confounding variables), to tackle weak signal expression profiles that have small sample size, case-control paired design and confounding covariates in each study. The results showed increased statistical power from confounding variable adjustment, paired design modelling and meta-analysis in this genomic setting and more profound biological findings

Table 2.6: Evaluation of t-test, FEM_minP, FEM_BIC and FEM_ALL methods by simulations. The Average of Type I errors, average of statistical powers, and average number of detected DE genes by each method are shown.

Scenario		Type (I) error (s.e)				Power (%) (s.e)				DE gene number (s.e)			
		t-test	FEM_minP	FEM_BIC	FEM_ALL	t-test	FEM_minP	FEM_BIC	FEM_ALL	t-test	FEM_minP	FEM_BIC	FEM_ALL
I	estimate	0.051	0.048	0.050	0.052	80.4	87.2	87.9	84.0	37.3	49.4	52.8	43.5
	s.e	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.004)	(0.004)	(0.005)	(1.228)	(1.063)	(1.029)	(1.162)
II	estimate	0.051	0.052	0.050	0.051	93.8	92.9	92.5	85.0	73.4	73.0	69.8	49.7
	s.e	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.005)	(0.846)	(0.915)	(0.956)	(1.368)
III	estimate	0.051	0.053	0.051	0.051	93.8	92.5	91.6	85.1	71.8	68.3	66.5	45.8
	s.e	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.004)	(0.004)	(0.005)	(0.928)	(0.940)	(0.876)	(1.047)

have been discovered in MDD neurobiology. Pathway analysis and post hoc analysis of variable selection revealed insightful biological conclusions. Simulations under three correlation structures were performed to verify improved performance of our proposed frame-work. In the literature, most psychiatric disease-related microarray studies of similar design either ignored the clinical variables or applied simple linear regression to include all variables in the model. Our results clearly show limits to those two approaches. To our knowledge, this is the first paper, which systematically considers the critical elements in the data structure in order to obtain more accurate DE gene and pathway detection. The framework is general and can be applied to microarray meta-analysis of other complex diseases with similar data structure. Specifically, this approach will be of great use in human post-mortem studies of the brain, where confounding factors are intrinsic (1) to the nature of the cohorts (demographic parameters), (2) to their method of collection (post-mortem interval) and (3) to the illness per se (clinical heterogeneity). Since dilution of expression signal is likely to occur in complex tissue such as the brain, DE genes often show small and weak effects, so reducing

the statistical interference of confounding factors is critical to detect disease effects. In the variable selection of the RIM model, we tested both BIC and minP approaches. The real data analysis showed that minP seemed to identify more DE genes and pathways in the MDD example while simulations showed similar performance and statistical power of the two methods. Another potential alternative is to apply popular regularization and shrinkage methods, such as Lasso or ridge regression, in the variable selection. A prohibitive down-side of such approaches is its extensive computation load for ge-nome-wide analysis, particularly in the estimation of the tuning parameter lambda. In our analysis, BIC and minP procedures limited to up to two covariates in the model balance well in biological inter-pretation and computation feasibility.

The goal of this study was to determine optimal analytical ap-proaches for complex datasets with multiple putative confounding variables. For this purpose, we focused on datasets produced by our group, in order to avoid additional confounding factors due to differences in laboratory protocols, brain bank collection, tissue treatment and sample handling. Now that we have established such analytical guidelines, the next step will be to increase the scope of meta-analyses by including additional datasets that are progressively made available in the literature. However, as expected, this also comes with added variability, which necessitates the development of complementary mathematical tools. For instance, we have designed a data-driven "meta-QC" quality control approach to rigorously assess the quality and potential load of confounding variables for any microarray datasets (Kang et al, paper in preparation). This preliminary quality control test is critical to assess whether the inclusion of additional datasets will increase the analytical power, or be detrimental to the meta-analysis, due to substantial quality control confounding load. Finally, as briefly elucidated in this report, mechanisms underlying neurological and neuropsychiatric disorders are likely to involve a distributed sets of brain regions linked in functional neural networks. The detection of molecular pathologies associated with those disorders will thus also critically depend on a priori hypotheses for converging or opposing effects in selected brain regions, for the presence (or not) of control brain regions. For instance, genetic risk factors may be hypothesized to similarly affect biological pathways across brain regions, while compensatory mechanisms leading to pathological dysfunction may display regional specificity, depending

on the respective activation or inhibition of different components of neural networks. Hence, the biological impact of the studies performed here will be investigated, validated and discussed more in-depth elsewhere. The studies combined in this paper have significant cohort features that may introduce significant heterogeneity. The five studies came from three distinct cohorts (MD1, MD2 and MD3), different sexes (male and female), array platforms (Affymetrix and Illumina) and brain regions (ACC and AMY). Future research is needed to further decipher such study-specific features. In this paper, we performed separated meta-analysis by brain region (Meta_3ACC and Meta_2AMY) for an initial comparison. A future direction is to collect more studies and apply meta-regression techniques to identify sex-specific or brain-region-specific genes in a unified meta-analysis.

3.0 META-REGRESSION MODELS TO DETECT BIOMARKERS CONFOUNDED BY STUDY-LEVEL COVARIATES IN MAJOR DEPRESSIVE DISORDER MICROARRAY DATA

3.1 INTRODUCTION

Meta-regression has grown in popularity in recent years, paralleling the increasing numbers of systematic and meta-analysis published medical literature. Traditional methods of meta-analysis attempt to combine results in order to obtain a single summarized effect size, such as the overall weighted effect size estimated by random effects model[44, 15]. The observed effect size in each study is an estimated , with some imprecision , of the true effect in that study. Usually, the statistical heterogeneity refers to the variation of effect sizes across studies, which cold result from clinical or methodological differences among the studies or could simply be because of chance[93]. In general, failure to consider the heterogeneity can cause bias in the results of a meta-analysis. Therefore, the potential scientific value of explorations of sources of heterogeneity has been emphasized in the past [9, 93, 94, 95],perhaps,this is the reason that meta-regression is now becoming a more widely used technique.

We conducted a systematic search from PubMed, SciSearch, Social SciSearch and AMEC (Allied and Complementary Medicine) using the search terms "meta-regression" in order to identify publications on meta-regression. The systematic review produced 205 publications relevant to meta-regression. We categorized the publications into two categories based on the primary focus of the article: The first category was the main meta-regression methods: fixed effects models (1 Publications [38]), random effects models (5 Publications [10, 11, 60, 45, 49]) and Bayesian or hierarchical models (2 publications[35, 89]); the second category was the application papers using the methods mentioned in the first category(197 publications). In

addition, there were two review paper about meta-regression [94, 5]. Among these papers related to meta-regression, we found that the the meta-regression was mainly used in clinical studies, and has never been used in genomic studies due to the availability of microarray studies. For example, it has been used to relate the standardised treatment response in RCTs to study-level variables (study duration, number of follow-up assessments, outpatients versus inpatients, per protocol analysis versus intention to treat analysis)[71], vaccine efficacy to geographical latitude[19], coronary risk benefit to serum cholesterol reduction [62]and properties of diagnostic tests to methodological quality of diagnostic accuracy studies[63]. In this paper, we will explore the use of meta-regression in gene expression analysis using 8 major depressive disorder (MDD) studies in which each study has three study-specific factors(sex,brain region and array platform). MDD is a heterogeneous illness with mostly uncharacterized pathology. Despite many gene expression studies of MDD [3, 53, 85, 88, 87], the biological mechanisms of MDD remain mostly uncharacterized [7]. MDD is thought to be a complex and heterogeneous disease [72], associated with multiple genetic, genomic, post-translational, and environmental factors. Furthermore, patients might have varying disease severity, with some having psychotic features as well as exposure to a variety of medications and dosage levels to control their illness. Secondly, the genetic disease effects are potentially confounded by many covariates, which include (1) demographical variables such as age, gender and race; (2) clinical variables such as anti-depressant drug usage, suicide and alcohol consumption; (3) technical variables inherent in the use of post-mortem brain samples, such as the pH level of brain tissues, brain region and postmortem interval (PMI). If the statistical models employed to identify differentially expressed genes fail to incorporate these sources of heterogeneity, not only can this reduce the statistical power, but also it will introduce sources of spurious signals to the gene detection. In our previous paper [100], we proposed a statistical framework to tackle weak signal expression profiles that have small sample size,case-control paired design and confounded with these sample-level covariates in each study. The results showed increased statistical power by incorporating the following considerations in the analysis: (1) inclusion of confounding clinical variables, (2) gene-specific variable selection, (3) random effects for paired design, and (4) meta-analysis. More MDD related biomarkers and pathways were detected that greatly enhanced understanding of MDD neurobiology.

The studies combined in that paper have significant cohort features that may introduce significant heterogeneity. The five studies came from three distinct cohorts (MD1, MD2 and MD3), different sexes (male and female), array platforms (Affymetrix and Illumina) and brain regions (ACC and AMY). In previous paper, we only had five studies, so we just performed sub-group meta-analysis by brain region (Meta3_ACC and Meta2_AMY) for initial comparison and did not decipher this study-specific features. After collecting another three MDD studies,in this paper, which has an appealing property to identify confounded study-level covariates and obtain a more accurate disease effect size after adjustment. An improved MetaRG with variable selection (namely MetaRG_BIC to be introduced later) will be performed to accommodate the small number of studies and relatively large number of study-level variables, especially, some variables may be multi-class variables. The result shows additional statistical power to detect gender-dependent and brain-region-dependent biomarkers that traditional meta-analysis methods, such as random effects model and Fisher's method, cannot detect.

3.2 METHODS

3.2.1 Description of motivating MDD data

Description of motivating MDD data This research is motivated from the meta-analysis of combining eight MDD transcriptomic studies. Brain tissues of four patient cohorts (MD1, MD2_M, MD2_F, and MD3) obtained from different sources at different time were analyzed. For all three patient cohorts, tissues from the anterior cingulated cortex (ACC) brain region were analyzed by microarray experiments independently to generate four microarray studies: MD1_ACC, C_MD2_ACC_F, C_MD2_ACC_M and MD3_ACC. Tissues from the amygdala (AMY) brain region in MD1 and MD3 cohorts were analyzed to generate MD1_AMY and MD3_AMY. Similarly, tissues from the dorsolateral prefrontal cortex (DLPFC) brain region in MD2 cohorts were analyzed to generate C_MD2_DLPFC_F, C_MD2_DLPFC_M. Details of the five patient cohorts and microarray studies are available in Table 1.1. Within each patient

cohort, MDD patients were matched to control patients by three demographic variables: age, sex and race. Three additional clinical variables (alcohol consumption, history of taking anti-depressant drugs and history of committing suicide) and two technical variables (pH level of brain tissues and post-mortem interval PMI) are also available for each patient. Six variables (age, alcohol-consumption, anti-depressant drug, suicide, pH and PMI) are considered as sample-level covariates. A random intercept model with Bayesian Information Criteria (BIC) variable selection has been proposed previously to account for these sample-level factors in single study analysis. For each of the eight microarray studies, three study-level factors (sex, brain region and array platform) are available and will be considered in the proposed meta-regression approach in this paper.

3.2.2 Data preprocessing, gene matching and gene filtering

Microarray images are scanned and summarized by manufacturers' defaults. Data from Affymetrix arrays are processed by RMA method and data from Illumina are processed by manufacturer's software for probe analysis. When samples in each study are processed in multiple batches, potential batch effects are evaluated and normalizations are performed to correct batch biases when necessary. Probes (or probe sets) are then matched to official gene symbols using Bioconductore package. After genes are matched across nine studies, 11840 unique gene symbols are available across all five studies. Two sequential steps of gene filtering are then performed. In the first step, we filter out genes with very low gene expression that are identified with small average expression values across majority of studies. Specifically, mean intensities of each gene across all samples in each study are calculated and the corresponding ranks are obtained. The sum of such ranks across nine studies of each gene is calculated and genes with the lowest 20% rank sum are considered un-expressed genes and are filtered out. Similarly, in the second step, we filter out non-informative (small variation) genes by replacing mean intensity in the first step with standard deviation. Genes with the lowest 20% rank sum of standard deviations are filtered out. Finally, $7,577 = 11840 \times (1-0.2) \times (1-0.2)$ matched genes in nine studies are analyzed.

3.2.3 Single study analysis incorporation sample-level variables

To account for paired design (MDD samples paired with corresponding controls) and sample-level covariates, we applied a random intercept model (RIM). For a given gene g, we fit the model

$$Y_{gik} = \mu_g + \beta_{g0}X_{0ik} + \sum_{l=1}^{L} \beta_{gl}X_{lik} + \alpha_k + \epsilon_{gik}, \tag{3.1}$$

where Y_{gik} was the gene expression value of gene $g(1 \leq g \leq G)$ and sample i (i=1,2 representing control and MDD, respectively) in pair $k(1 \leq k \leq K)$. X_{0ik} was the MDD indicator that took value one if the sample was MDD and zero if the sample was a control. X_{lik} represented values for clinical variable l (e.g. 0-1 binary for alcohol consumption or numerical for pH values in brain) and α_k was the random intercept from a normal distribution with mean zero and variance τ_g^2, which represented the deviation of averaged expression values in the kth pair from the average in the whole population. Finally, ϵ_{gik} were independent random noises that followed a normal distribution with mean zero and variance σ_g^2. Under this model, β_{g0} was the disease effect of gene g and was the parameter of major interest. To obtain an MDD-associated biomarker candidate list in a single study analysis, likelihood ratio test (LRT) was used to assess the p-values of testing $H_0 : \beta_{g0} = 0(vsH_A : \beta_{g0} \neq 0)$. Although RIM model can effectively adjust for confounding varables, the small sample size (9-22 pairs) and relatively high number of potential confounders (6 covariates) can make the model inefficient and impractical. In previous paper, we performed further variable selection to generate an optimal random intercept model(RIM_BIC), where the "optimal" referred to the model with the smallest Bayesian Information Criterion (BIC)[84]. Specifically, all possible RIM models that included at most two (0, 1 or 2) clinical variables were computed and compared. The model with the smallest BIC value was selected. This additional variable selection avoided to include more than 2 clinical variables in the model and allowed assessment of biomarkers affected by different sets of covariates in each gene (e.g. gene A is confounded by alcohol while gene B is confounded by drug), which biologically gave more appealing conclusions and interpretations. Similar to RIM model, likelihood ratio test were used to generate p-values

of testing $H_0 : \beta_{g0} = 0$ in each gene for the selected model. These attached p-value numbers were, however, not the true p-values for DE gene detection since they were biased from the variable selection procedure and the type I error control was voided. As a result, we performed a permutation test that randomly permuted the disease labels within each pair to generate a null distribution for p-value assessment. Subsequently, the resulting unbiased p-values after permutation correction were then corrected by Benjamini-Hochberg procedure for multiple comparisons in each study for DE gene detection. Based on the optimal model selected, the standardized effect size for each gene was defined as the coefficient of MDD divided by its standard error (i.e. $\hat{\beta}_{g0}/s_{g0}$ in the RIM_BIC model) from single study analysis; s_{g0} represented the estimated standard error of $\hat{\beta}_{g0}$.

3.2.4 Meta-analysis and Meta-regression

<u>Fixed effects model</u>: Fixed effects models is one of often used methods of combining effect sizes when the studies to be combined are homogeneous, in which only within-study variability is considered. The assumption is that studies use identical methods, samples, and measurements; that they should produce identical results; and that differences are only due to within-study variation. The general model is given by

$$Y_{gk} = \mu_g + \alpha_{gk} \tag{3.2}$$

Under the fixed-effect model we assume that there is one true effect size which underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error. Thus $Y_{gk} \sim N(\mu_g, \sigma_{gk}^2)$. The most efficient and unbiased estimator of 1 is the weighted average of estimates where the weights is determined by inverse of their standard errors. The estimate is

$$\hat{\mu}_g = \frac{\sum_{k=1}^{K} w_{gk} Y_{gk}}{\sum_{k=1}^{K} w_{gk}}, \tag{3.3}$$

where $w_{gk} = S_{gk}^{-2}$ and S_{gk}^2 s the estimated within-study variance in study k for gene g. The variance of $\hat{\mu}_g$ is then

$$Var(\hat{\mu}_g) = \frac{1}{\sum_{k=1}^{K} w_{gk}}. \tag{3.4}$$

51

So, a Z-score to test the null hypothesis that the common true effect μ_g is zero can be computed using

$$Z_g^{FEM} = \frac{\hat{\mu}_g}{\sqrt{Var(\hat{\mu}_g)}}. \tag{3.5}$$

which follows a standard normal distribution.

Random effect model (REM) is a popular method for combining of effect sizes in meta-analysis. Choi et al [15] was probably among the first to raise the issue of meta-analysis in the context of microarray data to find DE genes using this method, where the effect size is defined as the standardized mean difference $d = \frac{Y_D - Y_C}{S_p}$, where Y_D and Y_C represent the means of disease (MDD) and control groups, respectively and S_p indicates an estimation of the pooled variation. The corresponding model used was described as:

$$Y_{gk} = \mu_g + \alpha_{gk} + \eta_{gk} \tag{3.6}$$

, where Y_{gk} is the observed effect size in study k for gene g; the parameters α_{gk} and η_{gk} are the between-study and within-study errors, respectively. It assume that $\alpha_{gk} \sim N(0, \tau_g^2)$ and $\eta_{gk} \sim N(0, \sigma_{gk}^2)$, and τ_{gk}^2 and σ_{gk}^2 are the between-study and within-study variance. Usually, the estimate of σ_{gk}^2 can be produced in each study k. The between-study variance can be estimated using a method of weighted moments (MM) estimator of τ_{gk}^2, which can be derived from the heterogeneity statistic $Q_g = \sum_{k=1}^{K} w_{gk}(Y_{gk} - \hat{\mu}_g)^2$, where $\mu = (\sum_{k=1}^{K} w_{gk}Y_{gk})/\sum_{k=1}^{K} w_{gk}$ is the feasible weighted least-squares estimator with weights $w_{gk} = 1/S_{gk}^2$. Then, the weighted unbiased MM of τ_g^2 suggested by DerSimonian and Larird (DL) [22]: $\tau_g^2 = \max\{0, (Q_g - (K-1))/(S_1 - (S_2/S_1))\}$, where $w_{gk} = S_{gk}^{-2}$, and $S_r = w_{gk}^r$, and K is the number of studies. The average weighted effect size was estimated as $\mu(\tau_g) = \frac{\sum v_{gk}Y_{gk}}{\sum v_{gk}}$ and $Var(\mu(\tau_g) = 1/v_{gk}$, where $v_{gk} = 1/(\hat{\tau}_g^2 + S_{gk}^2)$. Under the assumption that the gene expression levels are normally distributed, a z-score to test for DE genes is constructed as, $Z_g = \frac{\mu(\tau_g)}{\sqrt{Var(\mu(\tau_g))}}$, which follows a normal distribution with zero mean and unit variance.

Random effects meta-regression model with one study-level Knapp and Hartung [60] proposed the following random effects meta-regression model with s single covariate:

$$Y_{gk} \sim N(\mu_g + \beta_g x_k, \tau_g^2 + \sigma_{gk}^2), k = 1, 2, \cdots, K \tag{3.7}$$

Where the parameter σ^2_{gk} stands for the within-study variance, and τ^2_g is the between-study variance. Like in REM, every study k produces an estimate of σ^2_{gk} denoted S^2_{gk}. The between-study variance can be estimated using a method of weighted moments (MM) estimator of τ^2_g, which can be derived from the heterogeneity statistic $\tilde{Q}_g = \sum_{k=1}^{K} w_{gk}(Y_{gk} - \hat{\mu}_g - \hat{\beta}_g x_k)^2$, where $\hat{\mu}_g$ and $\hat{\beta}_g$ are the feasible weighted least-squares estimator weights $w_{gk} = 1/S^2_{gk}$. Then, the weighted unbiased MM of τ^2_g is given in its truncated form as: $\tau^2_g = \max\{0, \frac{\tilde{Q}_g - (K-2)}{F(w_{gk}, x)}\}$ with $F(w_{gk}, x) = \sum w_{gk} - \frac{\sum w^2_{gk} \sum w_{gk} x^2_k - 2 \sum w^2_{gk} x_k + \sum w_{gk} \sum w^2_{gk} x^2_k}{\sum w_{gk} \sum w_{gk} x^2_k - (\sum w_{gk} x_k)^2}$ When there is no covariate, the above estimator reduces to the DL estimator.

Random effects Meta-regression model with multiple study-level covariates: Model 3.7 can be straightforwardly extended to the case where there are multiple study-specific covariates. The general random effect meta-regression model is

$$Y_{gk} = \mu_g + \sum_{l=1}^{L} \beta_{gl} X_{kl} + \alpha_{gk} + \eta_{gk}, \quad k = 1, 2, \cdots, K \tag{3.8}$$

, which can be written as a matrix form $Y_g \sim N(X\beta_g, \tau^2_g I_K + \Delta)$, where $Y_g = (Y_{g1}, Y_{g2}, \cdots, Y_{gK})'$ is a $K \times 1$ vector including the observed effect sizes from K studies; $X = (X_0, X_1, \cdots, X_L)$ is the $K \times (L+1)$-dimensional predictor matrix with rank $r(X) = r < K - 1$ and $X_0 = (1, 1, \cdots, 1)'$ and $X_l = (X_{1l}, X_{2l}, \cdots, X_{Kl})'$ $l = 1, 2, \cdots, L$; $\beta_g = (\mu_g, \beta_{g1}, \beta_{g2}, \cdots, \beta_{gL})'$ is the unknown parameter vector of the fixed effects; τ^2_g stands for the between-study variance; I_K is a $K \times K$ dimensional identity matrix; Δ is a $K \times K$ dimensional diagonal diagonal matrix with entries $\sigma^2_{gk}, k = 1, 2, \cdots, K$, that is, Δ contains the within-study variances. In this setting, the method of moment estimator of τ^2_{gk} is given by[60]: $\hat{\tau}^2_g = \frac{Q - (K-r)}{F(X, \Delta^{-1})}$, where $Q = Y'P'\Delta^{-1}PY_g$ with $P = (I_K - X(X'\Delta^{-1}X)^{-1}X'\Delta^{-1})$, and $F(X, \Delta^{-1}) = tr(\Delta^{-1} - tr((X'\Delta^{-1}X)^{-1}X'\Delta^{-2}X)$, in which $tr(X)$ denotes the trace of matrix X.

Meta-regression with variable selection : Exploring sources of heterogeneity my result in false positive conclusions through 'data dredging'[95]. Unlike meta-analysis in clinical or epidemiological research where up to hundreds of studies may be available for meta-regression model, only a small number (e.g. 5-15) of studies are available in a common microarray meta-analysis. When the number of study-level variables that potentially contribute to heterogeneity becomes large (greater than 2-3 studies), the regression model is not applicable. It is, however, reasonable to assume that only very small number (e.g. 0-1) of variables

contribute to the expression heterogeneity in each gene and the contributing variables are gene-specific. This leads to the variable selection approach we adopt in this paper. Specifically, all possible meta-regression models that include at most one (0 or1) study-specific variables are computed and compared.

$$MetaRG(w_g) : Y_{gk} = \mu_g + \sum_{l=1}^{L} w_l \beta_{gl} X_{lk} + \alpha_{gk} + \eta_{gk}, \qquad (3.9)$$

where w_l is the weight assigned to the lth study-level variable and $w_g = (w_{g1}, w_{g2}, \cdots, w_{gL})$, which belongs to $W = \{(w_1, w_2, \cdots, w_L)|w_l \in 0, 1 and \sum w_l \le 1\}$. We denote by $BIC(w_g)$ as the Bayesian Information Criterion value associated with meta-regression model $MetaRG(w_g)$. The adaptive weight w_g^* is defined as: $w_g^* = (w_{g1}^*, w_{g2}^*, \cdots, w_{gL}^*) = argmin_{(w \in W)} BIC(w)$, which serves as a convenient basis for gene categorization in follow-up biological interpretations and explorations. Based on the selected model $MetaRG(w_g^*)$, likelihood ratio test is applied to test $H_0 : \mu_g = 0 \quad and \quad \{\beta_{gl} = 0 \quad if \quad w_{gl}^* = 1\}$ versus $H_A : \mu_g \neq 0 \quad or \quad \{\beta_{gl} \neq 0 \quad if \quad w_{gl}^* = 1\}$. to derive the p-value of gene g. This added variable selection avoids including more than one study-specific variable in the model and allows assessment of biomarkers related to different study-specific variable (e.g. gene A might be related gender while gene B might be related to brain region or platform), which biologically gives a more appealing conclusion and interpretation. These attached p-value numbers are, however, not the real p-values for DE gene detection since they are biased from the variable selection procedure. As a result, we perform a permutation test that randomly permutes the disease labels within each pair to generate a null distribution effect sizes, then we repeat above variable selection procedure, and calculate the p-values, which server as the null distribution of p-values. Subsequently, the resulting unbiased p-values can then be corrected by Benjamini-Hochberg procedure for multiple comparisons for DE gene detection. Detailed algorithm of the permutation analysis is described in Appendix A.

3.2.5 Post hoc analysis on study-level variables after meta-regression

An essential advantage of our gene-specific mete-regression with variable selection scheme is the possibility of post hoc analysis on the selected study-specific variables in a genome-wide

scale. Two questions can be explored and answered: (1) Which variable(s) is the most or least frequently included in the model selection to influence disease effect? (2) Are variables repeatedly selected across genes more frequently than by random (e.g. sex is selected in most or all genes)? For the first question, we first generated a list of DE genes under a given FDR threshold and counted the frequency of each variable being selected in the gene list. Higher frequency showed that the variable more frequently influence the disease effects in different genes. For the second question, permutation test was performed to assess the statistical significance of the observed frequency. Specifically, suppose that we identified N DE genes among which we observed variable X was selected N_0 times. To generate the null distribution for hypothesis testing, we randomly chose N genes from the entire genome and count the frequency of variable X appearing in models associated with these genes as N_b. By repeating B times, the p-value of of the observed frequency is $p(X) = \frac{\sum_{b=1}^{B} I(N_b \geq N_0)}{B}$. We use B=10000 in this paper.

3.2.6 Evaluation

To evaluate performance of different models and methods in a real data analysis, we compared the number of detected DE genes among a specified set of MDD-related genes and performed gene enrichment analysis using the same set of MDD-related genes, where the MDD-related genes set was defined by web-tool (Gene Prospector: http://hugenavigator.net) by inputting the search term, major depressive disorder. The search review produced 297 genes reported with MDD among which 147 genes were included in our study. In addition, we identified a set of 9 genes ("SST" , "VGF", "TAC1", "MBP", "MOBP", "RTN4", "QPRT", "DGCR2", "EPHB6") that have been previously associated with MDD in the literature, but not included in the data base of Gene Prospector. For gene enrichment analysis, we performed Kolmogorove-Smirnov (KS) test, which was widely used in gene enrichment analysis [61, 91] because it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. Specifically, the p-values calculated from individual analyses or meta-analyises for assessing the DE genes are classified into two categories, in the specified MDD-related gene set (P) and out of pathway (P^C). Let $p_{(1)}, p_{(2)}, \cdots, p_{(n)}$

and $\tilde{p}_{(1)}, \tilde{p}_{(2)}, \cdots, \tilde{p}_{(m)}$ denote the order statistics of the p-values in P and P^C, respectively. The corresponding empirical distribution functions, $\hat{F}_P(x)$ and $\hat{F}_{PC}(x)$ for P and P^C can be defined as:

$$
\hat{F}_P(x) = \begin{cases} 0, & if \ x < p_{(1)} \\ \frac{k}{n}, if \ p_{(k)} \leq x < p_{(k+1)} & k = 2, 3, \cdots, n-1 \\ 1, & if \ x \geq p_{(n)} \end{cases} \tag{3.10}
$$

and

$$
\hat{F}_{PC}(x) = \begin{cases} 0, & if \ x < -\tilde{p}_{(1)} \\ \frac{k}{m}, if \ \tilde{p}_{(k)} \leq x < \tilde{p}_{(k+1)} & k = 2, 3, \cdots, m-1 \\ 1, & if \ x \geq \tilde{p}_{(n)} \end{cases} \tag{3.11}
$$

Let F_P and F_{PC} denote the population distribution for P and P^C, respectively. The one-sided two sample KS test can be defined based on the formula:

$$
T_{KS} = \max_x [F_p(x) - F_{PC}(x)], \tag{3.12}
$$

where the null hypothesis and the alternative hypothesis are :

$$
H_0 : F_P(x) \quad = F_{PC}(x) \ for \ all \ x \tag{3.13}
$$
$$
H_a : F_P(x) \quad \geq F_{PC}(x) for \ all \ x \tag{3.14}
$$
$$
\& F_P(x) \quad > F_{PC}(x) for \ some \ x \tag{3.15}
$$

Under the null hypothesis, the rejection region has the form of $T_{KS} > C_\alpha$ at level of α. Rejection of H_0 means that P is stochastically less than P^C (the CDF of P lies above and hence to the left of that for P^C). In another words, the p-values of genes in the genes set P are stochastically less than the p-values of genes outside of gene set PC. This indicates that genes in the pathway P have a stronger association with MDD than genes from outside of the gene set P^C. Small p-value associated with KS test indicates a good performance of the methods.

3.3 RESULTS

3.3.1 Fixed effect model and Random effect model

We applied the chi-squared test using a Q-statistic proposed by Cochran [18] to assess homogeneity of the studies. Under the hypothesis of homogeneity, if follows a χ^2_{K-1} distribution. A large deviation of observed Q statistic relative to the null distribution indicates rejection of the hypothesis of homogeneity, which therefore a random effect model is more appropriate. Although this test is known to have low statistical power [42], the result shows that the heterogeneity exits among 3594 genes after controlling FDR at 5%. The previous method is based on gene by gene test. To further confirm the existence of the heterogeneities , we assume that the genes can be treated as independent samplings and the homogeneity can be explored over all the genes. The histogram of the observed Q values and quantile-quantile plot (Q-Q plot) of the observed versus expected values are shown in Figure 3.1. The sample mean and variance of Q values are 17 and 114, respectively, which are much larger than that of the expected mean and variance, 7 and 14 and argue overall heterogeneity of the studies in meta-analysis.

3.3.2 comparing individual analysis and meta-analysis

Among the many microarray meta-analysis methods used in the literature, most methods have their pros and cons depending on the data structure and biological goal [47, 78]. In this paper, although we mainly focused on comparing REM and MetaRG, we also included the results of Fisher's method since it has been popularly applied in the microarray meta-analysis literature. Figure 3.2 showed the DE number plots of both meta-analysis results and compared with individual study analyses under various FDR thresholds. The exact DE numbers were listed in Table 3.1, and the result shows that individual study results had very weak signal and meta-analysis improved the statistical power and provided validated conclusions.

The histogram of Q value   QQ Plot
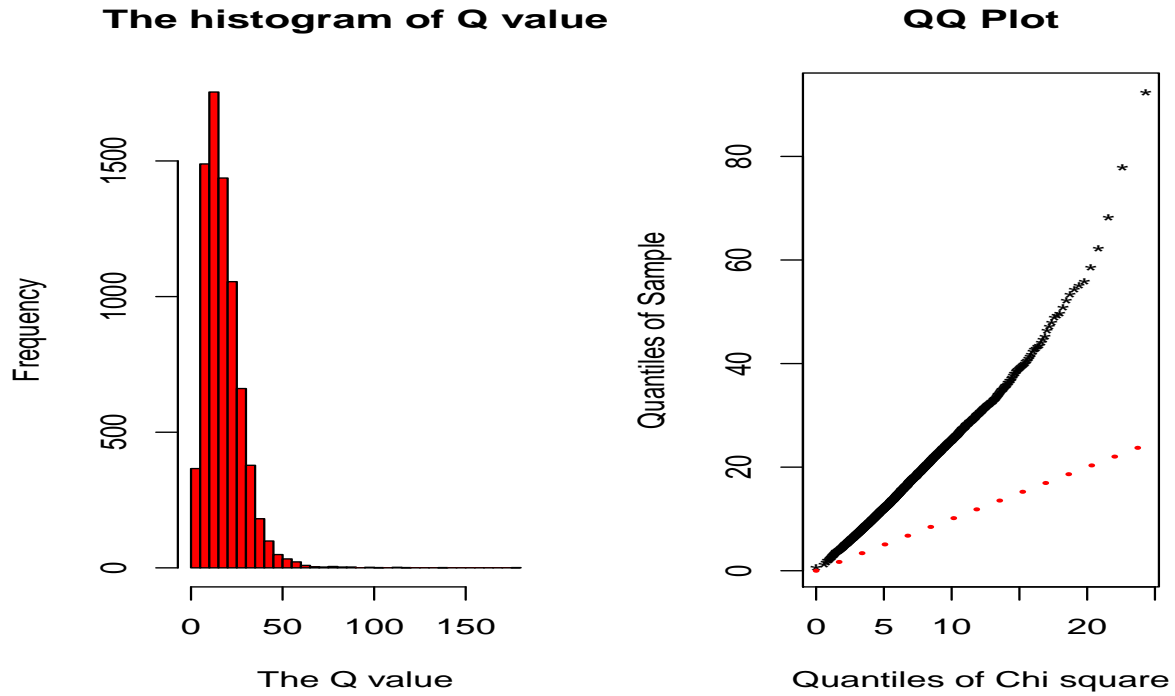
Figure 3.1: Gene by gene testing for the homogeneity of study effects. Overall test results are shown by the histogram of the observed Q values and the plot of the observed versus expected Q quantiles for the 8 MDD studies

3.3.3   Comparing REM and MetaRG

REM versus MetaRG: MetaRG detects both consistent markers that REM can detect and also markers confounded by study-level variables that REM cannot detect. MetaRG detects more markers than REM in general. A Venn diagram of DE gene lists detected by REM and MetaRG methods under FDR=1% is shown in Figure 3(a). The result showed that the three meta-analysis methods detected different sets of DE genes. Furthermore, in Figure 4(b), we drew the density plot of q-values calculated by MetaRG method for those 73 DE genes detected by REM only and the density plot of q-values calculated by REM method for those 175 DE genes detected by MetaRG only, the result shew that almost all genes detected

Table 3.1: Results of individual study analyses and meta-analysis

	C_MD2_DLPFC_M	MD1_ACC_M	C_MD2_ACC_M	MD1_AMY_F	MD2_MD2_DLPFC_F	MD3_ACC_F	C_MD2_ACC_F	MD3_AMY_F	Fisher	REM(IVW)	MetaRG
p=0.001	30	12	28	5	79	26	22	113	394	116	139
p=0.005	186	50	144	42	292	131	107	391	997	241	339
FDR=0.05	0	0	0	0	4	0	0	0	1307	103	124
FDR=0.1	0	0	0	0	97	1	0	406	2407	167	269

by REM can be detected by MetaRG if the FDR threshold is slightly relaxed. But many genes detected by MetaRG cannot be detected by REM. For example, Figure 2 shows the forest plot of an example gene SST (somatostatin; a gene known to affect neurotransmission in the central nervous system). When using conventional random effect (REM) model, the p-value is marginally significant (p=0.01) and the gene cannot be detected after multiple comparison. The forest plot shows a clear pattern that female MDD patients generally have down-regulation in SST while males have only slight down-regulation. Another example, Figure 3 shows the forest plot of an example gene ELP3 (elongation protein 3 homolog; a gene known to regulate the maturation of projection neurons). When using conventional random effect (REM) model, the p-value is marginally significant (p=0.02) and the gene cannot be detected after multiple comparison. The forest plot shows a clear pattern that this gene generally have down-regulation in AMY brain region while have slight up-regulation in ACC and DLPFC brain region.
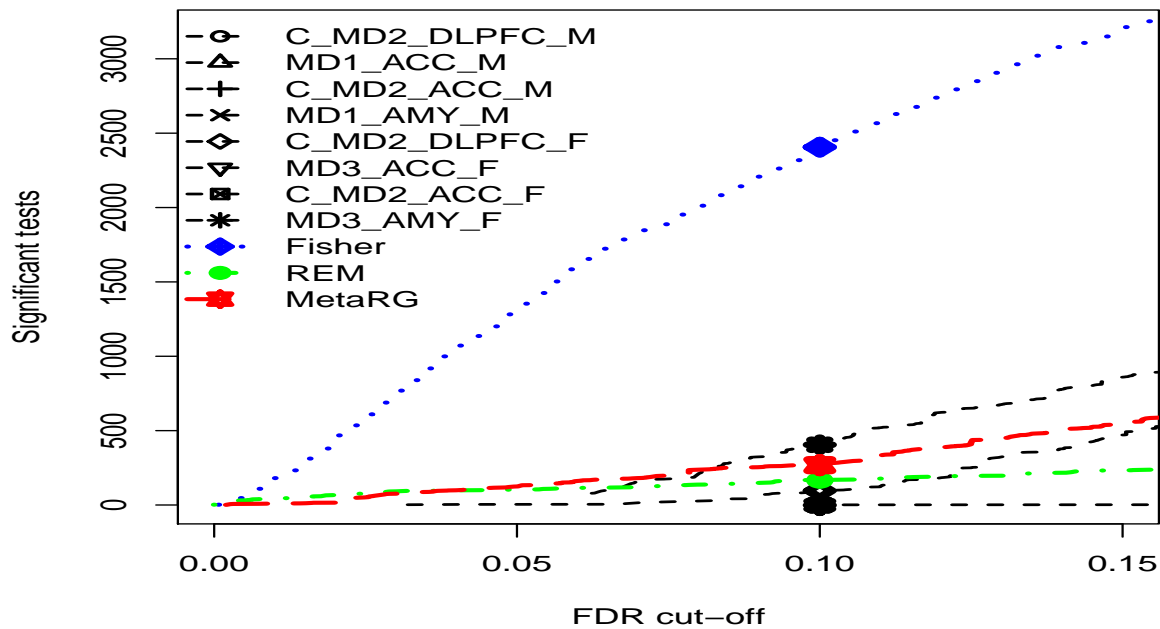
Figure 3.2: The DE number plot of both meta-analysis and individual analyses under various FDR thresholds.

Fisher: the most sensitive but can be tricky. Powerful for virtually all kinds of potential markers but may also contain many false positives. Can detect: (1) most of genes detected by REM (2) many of the genes detected by meta-regression (3) genes confounded unknown factors that can not be explicitly identified by meta-regression.

3.3.4 Frequencies of study-level covariates confounded with disease effect

MetaRG not only detects more biomarkers than REM by including biomarkers confounded by study-level covariates, but also has the advantage of showing the overall impact (frequency) of a study-level covariate confounded with the disease effect in the genome. One of MetaRG's potential abilities is to work out whether particular characteristics of studies are related to
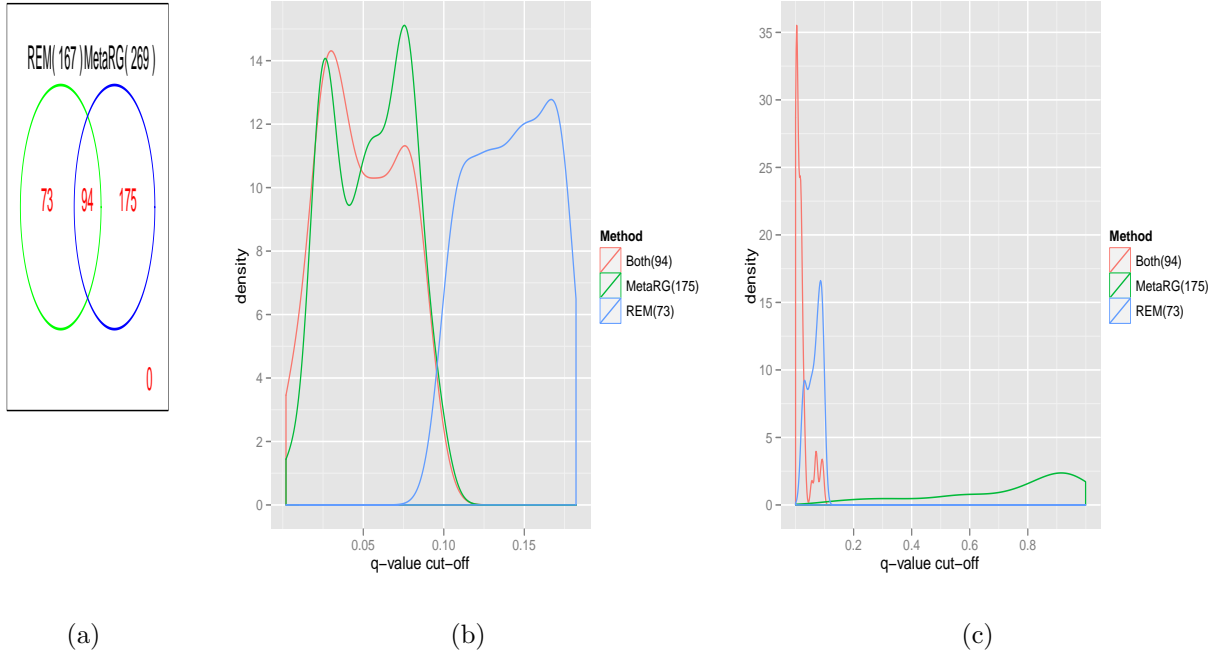
Figure 3.3: (a) the venn diagram of DE gene lists detected by REM and MetaRG at FDR 1%. (b) the density plot of q-values calculated from MetaRG for three categorical sets of DE genes.(c) The density plot of q-values calculated from REM for three categorical sets of DE genes

the effect sizes or not. To evaluate the impact of each characteristic of studies on the disease effect, especially among DE genes, we counted the numbers of appearances of study-specific variables in the meta-regression models for DE genes detected by MetaRG method under various p-value or FDR thresholds (see Table3.1). For example, under FDR=0.1 threshold, among which, 146 were sex-dependent markers, 37 were brain-region-dependent, and 39 were array-platform dependent. The p-values for testing whether sex, brain region and array platform are selected more frequently than by random are 0, 0.04 and 0.2, respectively, which indicates that sex is more frequently to influence the MDD effect, especially, among those

Gene: SST			
Study	Effect size	Weights(%)
C_MD2_DLPFC_M	0.8661	12.4
MD1_ACC_M	−1.0486	12.6
C_MD2_ACC_M	−1.2539	12.1
MD1_AMY_M	−0.1815	12.5
C_MD2_DLPFC_F	−3.5856	12.6
MD3_ACC_F	−1.3245	12.8
C_MD2_ACC_F	−1.9945	12.3
MD3_AMY_F	−5.1846	12.7
REM	−1.7249	0.4658
	p−value	0.01149

Gene: ELP3			
Study	Effect size	Weights(%)
C_MD2_DLPFC_M	2.5711	12.5
MD1_ACC_M	1.7345	12.7
C_MD2_ACC_M	1.8895	11.5
MD1_AMY_M	−0.1703	12.5
C_MD2_DLPFC_F	3.3032	12.6
MD3_ACC_F	0.4388	12.9
C_MD2_ACC_F	1.2998	12.4
MD3_AMY_F	−1.3581	12.9
REM	1.1959	0.2956
	p−value	0.02783

(a)   (b)

Figure 3.4: (a) the forest plot of gene SST. (b) the forest plot of gene ELP3.

DE genes. The wider spread of sex-dependent candidate markers provides an opportunity to investigate why women are more vulnerable to MDD than men [55].

3.3.5 Result of Komogorv-Smirnow test

To further evaluate the performance of individual analysis and three meta-analysis methods, Figure3.5 showed the number of MDD-related genes detected from meta-analysis and from individual study analyses under various FDR thresholds. The results showed again that MeteRG method could detect more MDD-related genes than REM method, and Fisher method detected much more MDD-related genes than REM and MetaRG because because there might be many unidentifiable confounders (surrogate variables) that Fisher could capture. The result of KS test was shown in Table 3, which showed that overall the performance
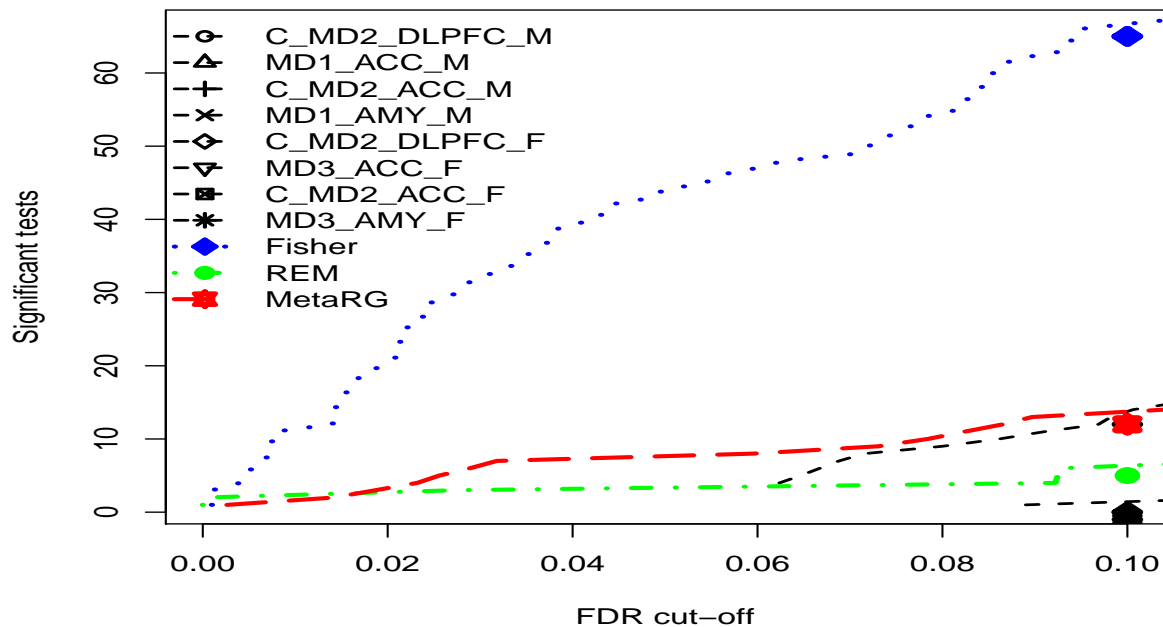
Figure 3.5: The number of genes detected from both meta-analysis and individual analyses among 156 MDD-related genes under various FDR thresholds.

of meta-analysis was better than that of individual analysis. Among three meta-analysis methods, Fisher and MetaRG had the similar performance which was better than that of REM.

3.4   DISCUSSION AND CONCLUSION

In this paper, we applied the meta-regression model on 8 MDD microarray studies, and compared it with the random effect model (REM) often used in gene expression analysis. The results showed that the meta-regression model gain some power of DE gene detection

by considering the heterogeneity potentially introduced by the three study-specific variables (i.e. sex, brain-region and array platform). More importantly, the meta-regression model has the potential ability to identify the biomarkers related to some study-specific variables, such as gene SST may be related to sex and gene ELP3 may be related to brain region.

A common situation in genomic study is that there are few studies in a meta-analysis but many possible study-level variables that might explain heterogeneity. In this situation, careful selection of appropriate covariates for inclusion into a meta-regression analysis is imperative for getting trustworthy results. In this paper, we proposed a meta-regression model with variable selection in which we allowed at most one study-level variable to be included in the meta-regression model. This provide a unified procedure to deal with the problem encountered in above situation. One of advantages of the meta-regression with variable selection is to biologically give a more appealing conclusion and interpretation, compared the univariate meta-regression, where we have a difficult to interpret the results after carrying out three mete-regressions with only one of three study-level variables. For example, based on the adaptive weights for each gene, we can figure out whether the effect sized is influenced by some specific variables.

In this paper, we compared three meta-analysis methods. Suppose that the heterogeneity can be totally explained by at least one of three study-level variables, MetaRG should yield the best result compared Fisher and REM methods. However, in this genome setting, Fisher yielded the better results in both DE gene detection and gene enrichment analysis than MetaRG and REM because because there may be many unidentifiable confounders (surrogate variables) that Fisher can capture. Therefore, a future direction is to explore methods to adjust the heterogeneity introduced by some surrogate variables, such as latent variable methods and surrogate variable analysis (SVA) proposed by Leek[38].

4.0 METADE: A R PACKAGE TO PERFORM META-ANALYSIS FOR DIFFERENTIAL EXPRESSION ANALYSIS

4.1 INTRODUCTION

Many high-throughput genomic technologies have advanced dramatically in the past decade. Microarray experiment is one example that evolved into relative maturity with generally consensus experimental protocol and data analysis strategy. Its extensive application in the biomedical field has led to an explosion of gene expression profiling studies publicly available. The noisy nature and small sample size in each dataset, however, often result in inconsistent biological conclusions [28, 92, 107]. Consequently, meta-analysis methods for combining microarray studies have been widely applied to increase statistical power and provide validated conclusions. Four major categories of statistical methods have been used to combine microarray studies in differentially expressed (DE) gene detection: combining p-values[52, 79, 80], combining effect sizes[15, 66], combining ranks [21, 48]and directly merge after normalization. In the "combining p-value" category, Fisher's method was the first analytical method applied to microarray meta-analysis. Other methods such as Stouf-fer, minimum p-value (minP), maximum p-value (maxP), rth ordered p-value (rOP), adaptively weighted Fisher (AW) and vote counting have also been widely used. In the category of 'combining effect sizes", there are two major types of statistical analysis: fixed and random effects models. For example, Choi et al (2003) combined effect sizes using weighted estimate for individual genes based on the fixed or random effects models Detailed description and comparison are given in section 1.4. In the category was initially proposed to detect differentially expressed genes for a single experiment (Breitling et al., 2004). Our package provides functions that perform most commonly used classical methods in each category as well as the proposed

65

methods from our group,such s AW method, roP methods and three meta-analysis methods with one-sided correction (minP_OC, maxP_OC, and roP_OC) which were given more detail description in subsection 4.1.1.

Despite the popularity of meta-analysis in microarray data, no comprehensive software package exists to date for easy implementation and comparison. Existing packages usually only provide limited functions to perform one or two methods; examples include GeneMeta (implements fixed and random effects models), metaMA (implements random effects model and Stouffer's method), metaArray (implements meta-analysis of probability of expression, POE), OrderedList (compares ordered gene lists), SequentialMA (determines sensitivity and decides whether more samples are needed to assure firm conclusion), RankProd (implements rank product method) and RankAggreg (implements various advanced rank aggregation methods). Methods implemented in the above packages mostly focus on binary outcomes and are not applicable to general continuous, multi-class and time-to-event outcomes. When applied to a specific microarray meta-analysis project, different algorithms generate different top ranked DE genes and q-values. Hong et al.[47] and Campain and Yang [13] are, by far, the only two comparative studies that evaluated the results of different meta-analysis algorithms. The included methods, tested examples and resulting conclusions in these two papers are, however, not yet conclusive enough to guide applications. It is very helpful that one can easily implement various methods for further comparison, assessment and selection to present results of all of the different methods. As a result, we developed the MetaDE package to provide comprehensive method selection, flexible while unified data input formats, options of different outcome types, various test statistics for DE analysis and choice of p-value calculation by fast parametric or robust permutation inferences. The goal is to provide a hands-on implementation of a given microarray meta-analysis project and easy evaluation and comparison of the results by different analytical methods. The computation is optimized by embedded C code and the open-source R environment allows extensibility for added features or methods in the future.

4.1.1 Meta-analysis methods with one-sided correction

Comparing the first two categories of meta-analysis methods in Section 1.4.1 and Section 1.4.2, combining effects sizes (e.g. random or fixed effects model) automatically identifies genes that have consistent up- or down-regulation in all studies. This may not be the case for methods combining p-values if the p-values are obtained from two-sided hypothesis testing. In this case, up- and down-regulation are treated as equally strong evidence and a gene may be detected from the meta-analysis with strong up-regulation evidence in one study but strong down-regulation evidence in another study, which leads to confusing conclusions. Theoretically, the discordance may reflect underlying biological truth due to population heterogeneity. In practice, however, such concordances are mostly results of technical artifacts such as gene annotation mistakes or cross-hybridization. A convenient solution to avoid the concordances is to generate p-values or ranks by one-sided tests. Owen [75] applied a similar Pearson one-sided test adjustment for Fisher's method. One-sided correction is helpful to guarantee identification of DE genes with concordant DE regulation directions. In this dissertation, we extended this modification to minP, maxP and roP methods which are described in following sub-sections. Under null hypothesis, the analytical cumulative distribution functions (CDF) of these three methods were derived (see Appendix D). Note that the consistent up- or down-regulation issue only exists in two-class comparison in DE gene detection and does not apply to other types of response variables (e.g. multi-class, continuous or survival).

4.1.1.1 Notations For gene g and study k, we let β_{gk} denote the effect of MDD. The Null hypothesis for β_{gk} is $H_0 : \beta_{gk} = 0$. Then, for $k = 1, 2, \cdots, K$, we can consider the hypotheses:

$$H_{0,gk} : \beta_{gk} = 0$$
$$H_{L,gk} : \beta_{gk} < 0$$
$$H_{R,gk} : \beta_{gk} > 0$$

and

$$H_{U,gk} : \beta_{gk} \neq 0,$$

based on the sign of β_{gk}. These are the null hypotheses, left- and right-sided alternatives and an undirected alternative, respectively.

Using $\hat{\beta}_{gk}^{obs}$ as test statistics, we may define

$$\tilde{p}_{gk} = Pr(\hat{\beta}_{gk} \leq \hat{\beta}_{gk}^{obs} | \beta_{gk})$$

and

$$p_{gk} = Pr(|\hat{\beta}_{gk}| \geq |\hat{\beta}_{gk}^{obs}| | \beta_{gk} = 0)$$

The $p-$values for alternatives $H_{L,gk}$, $H_{R,gk}$ and $H_{U,gk}$, respectively, are $\tilde{p}_{gk}$, $1 - \tilde{p}_{gk}$ and $p_{gk} = 2\min(\tilde{p}_{gk}, 1 - \tilde{p}_{gk})$.

4.1.1.2 Pearson's method(Fisher_OC) To guarantee identification of DE genes with concordant DE direction, Owen [75] revisited Pearson's method[77] and showed that Pearson's method has proved useful in a genomic setting[105], screening for age-related genes. Let $\tilde{p}_{gk}$ denote the p-value for the test of left-sided alternative in study $k(1 \leq k \leq K)$.

$$Q_g^L \quad = -2 * \sum_{k=1}^{K} \ln \tilde{p}_{gk} \tag{4.1}$$

$$Q_g^R \quad = -2 * \sum_{k=1}^{K} \ln(1 - \tilde{p}_{gk}) \tag{4.2}$$

Then, the test statistic of Pearson is defined as

$$V_g^{Fisher\_OC} = \max\{Q_g^L, Q_g^R\} \tag{4.3}$$

Under null hypothesis, the distribution of $V_g^{Fisher\_OC}$ can not be derived analytically. A conservative p-value was suggested by Owen.

4.1.1.3 minP method with one-sided correction(min_OC) From hereafter, we will omit the subscript g. Under the null hypothesis,$\tilde{p}_k$, $1 - \tilde{p}_k$ and p_k all have the $U(0,1)$ distribution. It follows that

$$Q_L^{minP} = \min\{\tilde{p}_1, \tilde{p}_2, \cdots, \tilde{p}_K\} \tag{4.4}$$

$$Q_R^{minP} = \min\{1 - \tilde{p}_1, 1 - \tilde{p}_2, \cdots, 1 - \tilde{p}_K\} \tag{4.5}$$

and

$$Q_U^{minP} = \min\{p_1, p_2, \cdots, p_K\} \tag{4.6}$$

All have the $Beta(1, K)$ distribution under H_0.

Then, the one-sided test statistic is defined as

$$Q_C^{minP} = \min(Q_L^{minP}, Q_R^{minP}) \tag{4.7}$$

Theorem 4.1.1. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic Q_C^{minP} is given by*

$$G(z) \quad = \quad \begin{cases} 1 - (1 - 2z)^K, \textit{if } 0 \leq z \leq 0.5 \\ 1, \textit{if } 0.5 \leq z \leq 1 \end{cases} \tag{4.8}$$

4.1.1.4 maxP method with one-sided correction (maxP_OC) Similarly, Let $\tilde{p}_{gk}$ denote the p-value for the test of left-sided alternative in study $k(1 \leq k \leq K)$.

$$Q_g^L = \max_{1 \leq k \leq K} \tilde{p}_{gk} \tag{4.9}$$

$$Q_g^R = \max_{1 \leq k \leq K}(1 - \tilde{p}_{gk}) \tag{4.10}$$

Then, the test statistic of maxP_OC is defined as

$$Q_g^{maxP\_OC} = \min\{Q_g^L, Q_g^R\} \tag{4.11}$$

Under null hypothesis, the p-value associated with $Q_g^{maxP\_OC}$ can be assessed by the following Theorem, which is proved in AppendixD.

Theorem 4.1.2. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic $Q_g^{maxP\_OC}$ is given by*

$$F(z) = \begin{cases} 2z^K, \text{ if } 0 \leq z < 0.5 \\ 2z^K - (2z-1)^K, \text{ if } 0.5 \leq z \leq 1 \end{cases} \tag{4.12}$$

4.1.1.5 roP method with one-sided correction(roP_OC) Similarly, Let $\tilde{p}_{gk}$ denote the p-value for the test of left-sided alternative in study $k(1 \leq k \leq K)$.

$$Q_g^L = p_{g(r)}\{\tilde{p}_1, \tilde{p}_2, \cdots, \tilde{p}_K\} \tag{4.13}$$

$$Q_g^R = p_{g(r)}\{1 - \tilde{p}_1, 1 - \tilde{p}_2, \cdots, 1 - \tilde{p}_K\} \tag{4.14}$$

$$\tag{4.15}$$

Then, the test statistic of roP_OC is defined as

$$V_g^{roP\_OC} = \min\{Q_g^L, Q_g^R\} \tag{4.16}$$

Under null hypothesis, the p-value associated with $V_g^{roP\_OC}$ can be assessed by the following Theorem.

Theorem 4.1.3. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic $V_g^{roP\_OC}$ is given by* **Case1**: *$r \geq K - r + 1$.*

$$F(z) = \begin{cases} 2(1 - F(r-1, K, z)), \text{if } 0 \leq z < 0.5 \\ 1 - \sum_{j=K-r+1}^{r-1} \sum_{h=K-r+1}^{K-j} \frac{K!}{j!h!(K-j-h)!}(1-z)^{j+h}(2z-1)^{K-j-h}, \text{if } 0.5 \leq z \leq 1 \end{cases}$$

(4.17)

(4.18)

Case2: *$r < K - r + 1$*

$$F(z) = \begin{cases} 1 - \sum_{l=0}^{r-1} \sum_{m=0}^{r-1} \frac{K!}{l!m!(K-l-m)!} z^{m+l}(1-2z)^{K-l-m}, \text{if } 0 \leq z \leq 0.5 \\ 1, \text{if } 0.5 \leq z \leq 1 \end{cases}$$

(4.19)

4.2 IMPLEMENTATION

MetaDE package implements 12 major meta-analysis methods for differential expression analysis: Fisher, Stouffer, adaptively weighted Fisher (AW), minimum p-value (minP), maximum p-value (maxP), rth ordered p-value (rOP), vote counting, fixed effects model (FEM), random effects model (REM), rank product (rankProd), naïve rank sum/product and meta-analysis adjusted for confounding variables (MetaACV) (Table 4.1). Detailed algorithms, their restrictions and general pros and cons are discussed in the online supplement document. In addition to selecting a meta-analysis method, several other considerations are involved in the implementation. (1) Choice of test statistics: Multiple test statistics are available for each type of outcome variable. For multi-class outcomes, the minimum multi-class correlation (min-MCC) was particularly developed to capture concordant expression patterns that F-statistics can fail [50]. (2) One-sided correction: For binary outcomes, DE genes with discordant regulations (e.g. up-regulation in one study but down-regulation in another study) can often be identified if two-sided p-values are to be combined and the results are difficult to interpret. One-sided correction that was considered by Pearson (1938) is helpful to guarantee identification of DE genes with concordant DE regulation directions.

Methods		Types of outcome variable			
		binary	multi-class	continuous	survival
Step 1. choice of test statistics for individual dataset		paired t; unpaired t; moderated t	F-statistics	Pearson correlation; Spearman correlation; regression	log-rank statistics
Step 2. Choice of combining methods					
combine p-values	Fisher	√ª	√	√	√
	Stouffer	√ª	√	√	√
	AW	√ª	√	√	√
	minP	√ª	√	√	√
	maxP	√ª	√	√	√
	rOP	√ª	√	√	√
	vote counting	√ª	√	√	√
combine effect sizes	FEM*	√	×	×	×
	REM*	√	×	×	×
combine ranks	rankProd*	√	×	×	×
	naïve rank sum/prod	√ª	√	√	√
minMCC*		√ª	√		
MetaACV (can apply to all methods combining p-values)		√ª	×	√ (only applicable to regression)	√

*no need to select test statistics in these methods.
a with and without one-sided correction.

Figure 4.1: Summary of 12 microarray meta-analysis methods included.

(3) Adjustment for confounding variables: Clinical or technical variables (e.g. gender, race etc) can be important confounders that affect sen-sitivity of DE gene detection. We have adopted the Meta-Analysis Adjusted for Confounding Variables (meta-ACV) with effective gene-specific variable selection [100] and applied the approach to all p-value combination methods and all outcomes except for multi-class variables. MetaDE takes two types of uni-fied input formats: standard Ex-pressionSet objects from Bioconductor or lists of ordinary data matrixes in R. Options of gene matching across studies and gene filtering are available. Missing values are allowed if a gene is miss-ing in partial studies. Outputs of the meta-analysis results include DE gene lists with corresponding q-values and various visualization tools. A technical document, a tutorial and R help files are available online, accompanying the package.

4.2.1 Data pre-processing

Gene matching: Usually different microarray platforms use their own probe IDs. To perform metan-analysis, we need match probe IDs from different platforms to the unique official gene ID, such as ENTREZ ID or gene symbol. In this package, we focus on the gene symbol. In MetaDE package, we provide two options to match probe ID to gene symbol when multiple probes (or probe sets) matched to an identical gene symbol: one is average method in which we take the average value of expression values among multiple probe IDs to represent the corresponded gene symbol; another one is "IQR" method in which we selected the probe ID with the largest interquartile range (IQR) of expression values among all multiple probe IDs to represent the corresponded gene symbol. The procedure of gene matching can be implemented by function `Match.gene()`. The arguments of this function are

```
Match.gene(x,pool.replicate=c("average","IQR"))
```

where **x** is an eSet (Container for high-throughput assays and experimental metadata), and one column named by "GENESYMBOL" of featureData of **x** must include the gene symbols. The arguments **for pool.replicate** are then:

- "**average**": the average method mentioned as above was chosen to perform gene matching;
- "**IQR**": the "IQR" method mentioned as above was chosen to perform gene matching;

Gene filtering: If we hold an enormous number of Genes, thus raise many practical and theoretical problems in controlling the false discovery rate(FDR). Biologically, it is likely that most genes are either un-expressed or un-informative. In gene expression analysis to find DE genes, these genes contribute the false discoveries, so it is desirable to filter out these genes prior to analysis. After genes were matched across five studies, the unique gene symbols were available across all studies. Two sequential steps of gene filtering were then performed. In the first step, we filtered out genes with very low gene expression that were identified with small average expression values across majority of studies. Specifically, mean intensities of each gene across all samples in each study were calculated and the corresponding ranks were obtained. The sum of such ranks across all studies of each gene was calculated and genes with the highest $\alpha\%$ rank sum were considered un-expressed genes (i.e. small

expression intensities) and were filtered out. Similarly, in the second step, we filtered out non-informative (small variation) genes by replacing mean intensity in the first step with standard deviation. Genes with the lowest $\beta\%$ rank sum of standard deviations were filtered out. Finally, the total number of matched genes is $G \times (1-\alpha) \times (1-\beta)$, which are used for further analysis. The procedure of gene filtering can be implemented by function `Gene.filter`. The arguments of this function are

```
Gene.filter(x,DelPerc=c(alpha,beta)),
```

where **x** is the input variable,which is a list of a list datasets and a list of labels; argument **DelPerc** is a numeric vector of length 2, which specify how many percent of genes need to be filtered out during the two sequential steps of gene filtering.

4.2.2 Perform individual analysis

Before beginning with a meta-analysis, one must first obtain a set of p-values or effect size estimates with their corresponding sampling variances.The MeteDE package provides the `ind.analysis()` function, which can be used to perform various test statistics for DE analysis based on the type of the outcome and choice of p-value calculation by fast parametric or robust permutation inferences.For the default interface, the arguments of the function are

```
ind.analysis(x,ind.method=c("regt","modt","pairedt","pearsonr","spearmanr",
"F"), nperm,tail,...)
```

where **x** is the input variable,which is a list of a list datasets and a list of labels; argument **ind.mehtod**is a character string specifying which test statistic should be used to calculated the p-values.The options for argument **ind.method** are then:

- **regt**: The regular t-statistics.
- **modt**: The moderated-t statstics.
- **pairedt**: The paired t-statistics.
- **pearsonr**: The Pearson product correlation statistics.
- **F**: The F-statistics.
- **spearmanr**: The Spearman rank correlation statistics.

nperm is an argument to specify the choice of p-value calculation by fast parametric or robust permutation inferences. If it is `NULL`(default), the parametric method is used; If it is an integer, the permutation method is used, and the integer is the number of permutations used to infer the p-values. **tail** is a character string specifying the direction of alternative hypothesis , must be one of "low"(left-side p-value), "high"(right-sided p-value) or "abs"(two-sided p-value).

The MetaDE package also provides an function `cal.ES` to calculate various effect sizes (and the corresponding sampling variances) that are commonly used in meta-analyses.The arguments for this interface are

```
cal.ES(y,l,paired=FALSE)
```

where arguments **y**and **l** are the gene expression matrix and the vector of labels of outcome, respectively; **paired** is a logical indicating whether the experiment is paired design or not, and then the effect sizes(and corresponding sampling variances) are calculated. The output of this function is an matrix including the biased and unbiased effect size estimates (and corresponding variances) (see section 1.4.2).

4.2.3 Perform meta-analysis

The various meta-analyses can be implemented by three main functions,`MetaDE.radata()`, `MetaDE.pvalue()` and `MetaDE.ES()`, in MetaDE package. The arguments of function `MetaDE.radata()` are given by

```
MetaDE.rawdata(x, ind.method=c("modt","regt","pairedt","F","pearsonr",
"spearmanr","logrank"),meta.method=c("maxP","maxP.OC","minP","minP.OC",
"Fisher","Fisher.OC","AW","AW.OC","roP","roP.OC","vote","vote.OC",
"minMCC","rankProd","naiveranksum"),rth=NULL,nperm=NULL,ind.tail="high"
,asymptotic=FALSE,...)
```

As above,**x** is the raw data (the gene expression matrices and the labels of outcome),which is a list of a list datasets and a list of labels; the argument **ind.method** is the same as that in function `ind.analysis()`; The various meta-analysis methods described in section 1.4 that

can be specified via the **meta.method** argument are then:

- **maxP**: The maximum p-value method;
- **maxP.OC**: The maximum p-value with one-sided correction;
- **minP**: The minimum p-value method;
- **minP.OC**: The minimum p-value method with one-sided correction;
- **Fisher** The Fisher's method;
- **Fisher.OC**: The Fisher's method with one-sided correction;
- **AW**: The adaptive weight method;
- **AW.OC**: The adaptive weight method with one-sided correction;
- **roP**: The r-th ordered p-value method;
- **roP.OC**: The r-th ordered p-value method with one-sided correction;
- **vote**: The vote counting method;
- **vote.OC**: The vote counting method with one-side correction;
- **minMCC**: The the minimum multi-class correlation method [50];
- **rankProd**: The rank product method [48];
- **naiveranksum**":The naive rank summation method;

If the **meta.method** is chosen as "roP" or "roP.OC", an integer need input via argument **rth** to specify which rth ordered p-value as the statistic; If the argument **asymptotic** is TRUE, then the parametric method is used in meta-analysis to calculate the p-values; the argument **nperm** is the same as in function `ind.analysis()`.

If p-values or effect sizes (and corresponding variances) have been calculated already, for example by other methods not used in functions `ind.analysis()` or `cal.ES()` with the help of other software, then the meta-analysis can be implemented by function `MetaDE.pvalue()` or `MetaDE.ES()`. The arguments of these two functions are given by

```
MetaDE.pvalue(x,meta.method=c("maxP","minP","Fisher","Pearson","Stouffer",
"roP","AW","maxP.OC","roP.OC"),asymptotic=FALSE)
```

, where argument **x** is a list whose first object is the p-value matrix,and second object are the permutated matrices. If the second object of **x** is NULL, the parametric method is then used in meta-analysis.

```
MetaDE.ES(x,paired=FALSE,REM=TRUE,correct=TRUE)
```

, where **x**is the raw data (the gene expression matrices and the labels of outcome),which is a list of a list datasets and a list of labels; argument **paired** is a logical indicating whether the studies are paired design or not; argument **REM** is a logical specifying whether a fixed- or a random/mixed-effects model(`default`) should be fitted. Random/mixed-effects models are fitted by using "DL" method to estimate the between-study variance(see 1.4.2).

4.2.4 Draw plots

The MetaDE package provides several functions for creating plots that are frequently used in meta-analyses.For example,the `heatmap.sig.genes()` function is used to create the Heatmaps plots of the DE genes under some p-value or FDR threshhold across studies; The `forestplt()` is use to darw the Forest plots of selected genes can be generated for binary outcomes; the `draw.DEnumber()` function is used to generate the DE number plots (a plot showing the number of detected DE genes under different p-value or q-value threshold) can be shown to compare sensitivity in individual study analyses and different meta-analyses methods. To view the exact number of DE genes detected by different methods, the function `count.DEnumber()` can be used to generate the tables in which the numbers of DE genes detected by different methods under various p-value and FDR thresholds are listed. Several examples are given in next section to illustrate how such plots can be created.

4.2.5 EXAMPLE

To demonstrate the functionality of MetaDE, we performed meta-analysis to combine 4 major depressive disorder (MDD) studies. We present the results of maxP without confounder adjustment (Figure)under p-value threshold 0.001.The Figure was created with the following code.

```
>maxP.pt<-MetaDE.rawdata(x,ind.method="pairedt",meta.method="maxP",
ind.tail="abs",asymptotic=TRUE)
> heatmap.sig.genes(maxP.pt,pval.cut=0.001)
```
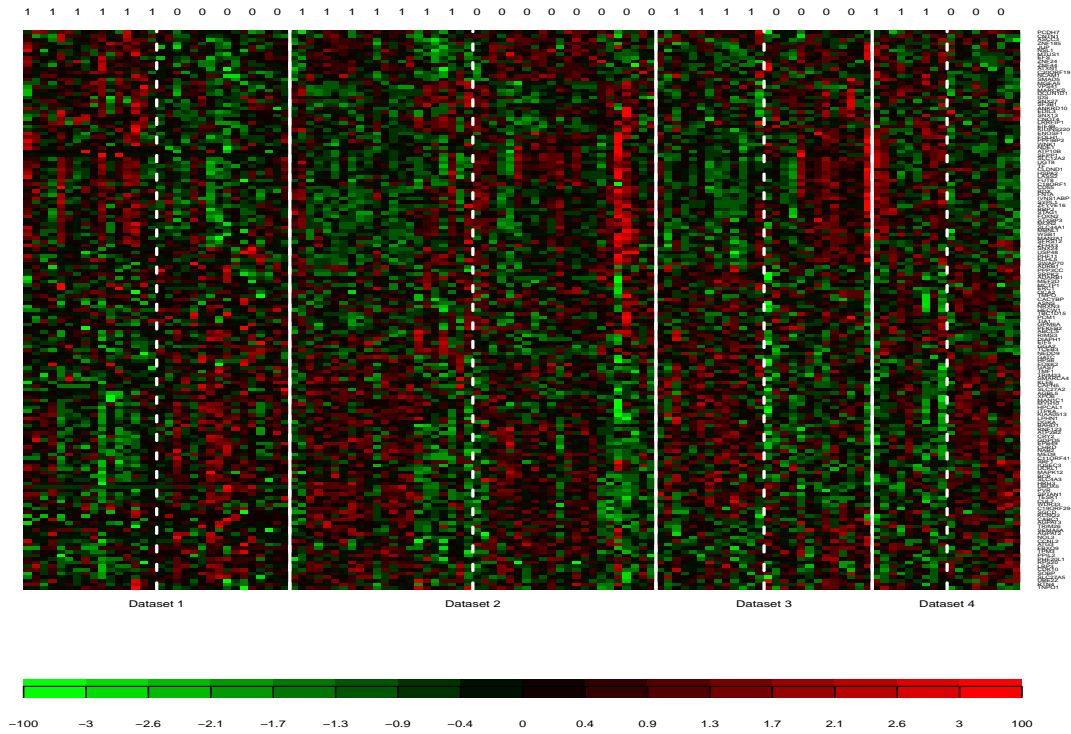
Figure 4.2: The heatmap of DE genes detected by maxP method under p-value threshold 0.001 based result of paired t-test in individual analysis.

Figure 4.3(A) and (B) shows the DE number plots of both meta-analysis results and compared with individual study analyses. The result shows that individual study results had very weak signal. Meta-analysis improved the statistical power and provided validated conclusions. The Figures were created with the following code.

```
#Figue (A)
minP.pt<-MetaDE.rawdata(x,ind.method="pairedt",meta.method="minP",
ind.tail="abs",
asymptotic=TRUE)
fisher.pt<-MetaDE.rawdata(x,ind.method="pairedt",meta.method="Fisher",
ind.tail="abs",
```

```
asymptotic=TRUE)
stouf.pt<-MetaDE.rawdata(x,ind.method="pairedt",meta.method="Stouffer",
ind.tail="abs",
asymptotic=TRUE)
REM.pt<-MetaDE.ES(x,paired=TRUE,REM=TRUE,correct=TRUE)
pm.pt<-cbind(maxP.pt$ind.p,maxP.pt$meta.analysis$pval,
fisher.pt$meta.analysis$pval,REM.pt$pval)
method<-c(c("MD1_ACC_M","MD3_ACC_F","C_MD2_ACC_F","C_MD2_ACC_M",
"maxP","Fisher","REM"))
mlwd<-rep(c(2,4),c(4,4))
#Figue (B)
x<-list()
x$p<-p.m[,c("MD1_ACC_M","MD3_ACC_F","C_MD2_ACC_F","C_MD2_ACC_M")]
x$bp<-NULL
maxP.acv<-MetaDE.pvalue(x,meta.method="maxP",asymptotic=T)
fisher.acv<-MetaDE.pvalue(x,meta.method="Fisher",asymptotic=T)
ES.acv<-ES.m[,c("MD1_ACC_M","MD3_ACC_F","C_MD2_ACC_F","C_MD2_ACC_M")]
Var.acv<-Var.m[,c("MD1_ACC_M","MD3_ACC_F","C_MD2_ACC_F","C_MD2_ACC_M")]
REM.acv<-get.REM(ES.acv,Var.acv)
pm.acv<-cbind(x$p,maxP.acv$pval,fisher.acv$pval,REM.acv$pval)
method<-c(c("MD1_ACC_M","MD3_ACC_F","C_MD2_ACC_F","C_MD2_ACC_M","maxP",
"Fisher","REM"))
mlwd<-rep(c(2,4),c(4,4))
count.DEnumber(pm.acv,c(0.001,0.005),c(0.01,0.05,0.1),method)
#draw Figure
par(mfrow=c(1,2))
draw.DEnumber(pm.pt,0.005,0.004,method,mlwd)
title("(A) Paired t-test")
draw.DEnumber(pm.acv,0.005,0.004,method,mlwd)
title("(B) MetaACV")
```

We also can use `count.DEnumber()` function to list the exact number of DE genes under various p-value or FDR thresholds. For example, the tables can be created by the following code:

```
> #paired t-test
> count.DEnumber(pm.pt,c(0.001,0.005),c(0.01,0.05,0.1),method)
$pval.table
        MD1_ACC_M MD3_ACC_F C_MD2_ACC_F C_MD2_ACC_M maxP Fisher REM
p=0.001         2        14           8          23  155     73  13
p=0.005        22        85          52         154  425    291  59
$FDR.table
         MD1_ACC_M MD3_ACC_F C_MD2_ACC_F C_MD2_ACC_M maxP Fisher REM
FDR=0.01         0         0           0           0    0      0   0
FDR=0.05         0         0           0           1  169      5   0
FDR=0.1          0         0           0           3  572     96   0


> #MetaACV
> count.DEnumber(pm.acv,c(0.001,0.005),c(0.01,0.05,0.1),method)
$pval.table
        MD1_ACC_M MD3_ACC_F C_MD2_ACC_F C_MD2_ACC_M maxP Fisher REM
p=0.001        12        26          22          28  122     90  71
p=0.005        50       131         107         144  323    301 169
$FDR.table
         MD1_ACC_M MD3_ACC_F C_MD2_ACC_F C_MD2_ACC_M maxP Fisher REM
FDR=0.01         0         0           0           0    3      0   7
FDR=0.05         0         0           0           0   72      8  36
FDR=0.1          0         1           0           0  264    171  65
```

In the on-line technical document, we presented two other examples: prostate cancer studies (multi-class outcome) and breast cancer (survival outcome).
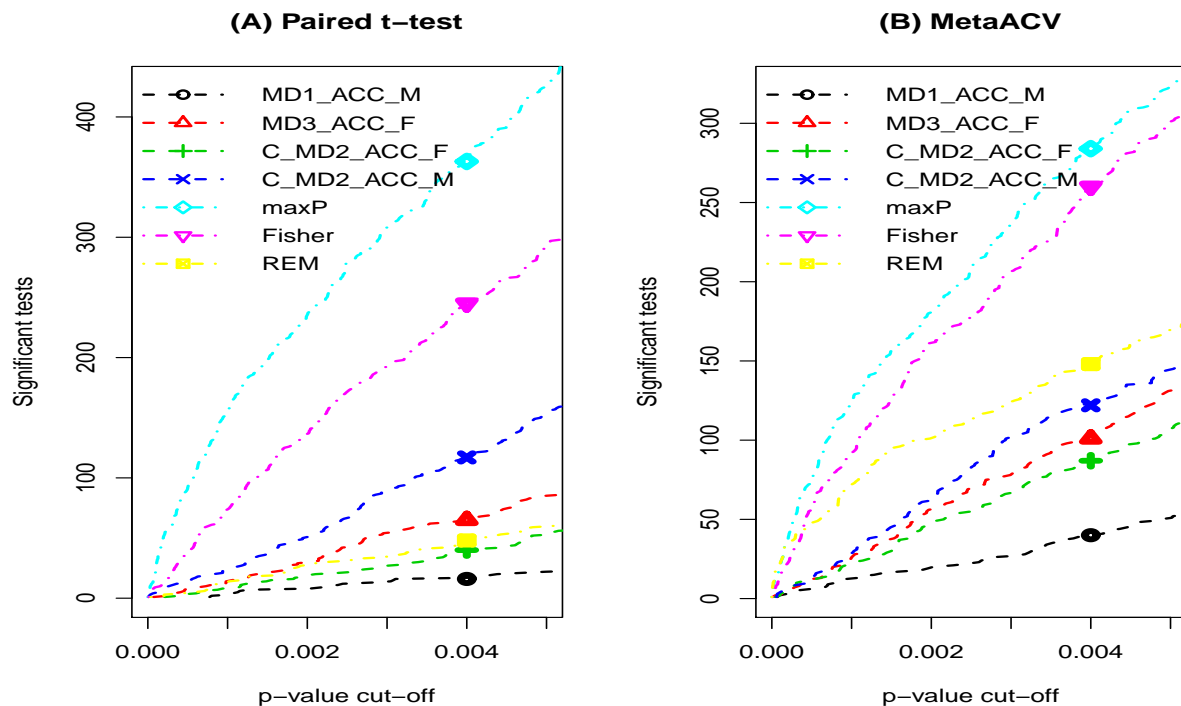
Figure 4.3: (A)The DE number plot of paired t-test.(B) The DE number plot of MetaACV.

4.3 DISCUSSION AND CONCLUSION

The MetaDE package provides a wide collection of classical and emerging meta-analysis methods for identifying DE genes. Comparing to other packages, MetaDE offers much wider options of analysis methods for both individual dataset analysis and meta-analysis. It is suitable to researchers who want to easily obtain an analysis and tailor their choices to the biological questions of interest. For example, if one is interested in finding genes that are differentially expressed between cases and controls in all datasets. One could select "moderated t-test" from the individual analysis and select "maxP" from the meta-analysis to combine the p-values for moderated t-test. This would form the "modt+maxP" method for the whole process. One also could select "REM" as the meta-analytic method for this

purpose. In this case, the effect sizes are generated during the stage of individual analysis and they are combined through the REM as described in previous section. This setting is highly suitable for those who want more flexibility. Moreover, a detailed online tutorial will guide the user to make a choice on the methods that are suitable for their research questions.

Users should be aware of the limitations of the methods implemented in the MetaDE package. First, the Bayesian approaches have not been implemented. Second, we assumed that all studies contain identical matched gene list with no missing values. In real practice, separate studies to be combined usually come from different microarray platforms. Requiring an identical matched gene list and no missing values will exclude many important genes that appear in certain studies but not in others, thus requiring an extension that allows for missing values.

While we focused on combining multiple microarray studies in this paper, the package can also be used to identify differentially expressed biomarkers from similar data types, for example, multiple genomic, epigenomic and/or proteomic datasets.

The MetaDE package provides R functions to perform meta-analysis for differential expression analysis.

5.0   CONCLUSIONS AND FUTURE WORKS

5.0.1   CONCLUSIONS

Meta-analysis or information integration of multiple genomic studies helps to increase sta-
tistical power of biomarker detection. However, the results of meta-analysis are easily biased
due to failure to incorporate important covariates at either the study or person level. For
example, in MDD studies, many clinical variables (sample-level or study-level), such as sex,
age, alcohol, antidepressant drug, or death by suicide, have been shown to be potential factors
characterizing subtypes of MDD. If the statistical models employed to identify differentially
expressed genes fail to incorporate these sources of heterogeneity, not only can this reduce the
statistical power, but also it will introduce sources of spurious signals to the gene detection.
In this dissertation, firstly, we proposed a statistical approach for meta-analysis to tackle
weak signal expression profiles that have small sample size, case-control paired design and
confounding covariates in each study. The results showed increased statistical power from
confounding variable adjustment, paired design modelling and meta-analysis in this genomic
setting and more profound biological findings have been discovered in MDD neurobiology.
Secondly, to adjust the effect of study-level variables, we extended the idea of random effects
method and gene-specific variable selection to meta-regression (MetaRG) approach, which
has an appealing property to identify confounded study-level covariates and obtain a more
accurate disease effect size after adjustment. To our knowledge, this is the first systematic
investigation in this area, which systematically considers the critical elements in the data
structure in order to obtain more accurate DE gene and pathway detection. The framework
is general and can be applied to microarray meta-analysis of other complex diseases with
similar data structure. Finally, we developed the MetaDE package to provide comprehen-

sive method selection, flexible while unified data input formats, options of different outcome types, various test statistics for DE analysis and choice of p-value calculation by fast parametric or robust permutation inferences. This provided a hands-on implementation of a given microarray meta-analysis project and easy evaluation and comparison of the results by different analytical methods.

5.0.2 FUTURE WORKS

hierarchical meta-analysis model: MDD, Bipolar and Schizophrenia are three kinks of highly correlated, but different, brain diseases. To study the linkage among these three brain disorders,a biology question of interest is whether some genes are related to all of these three diseases, or partially related to them. To addresse this question, we proposed a hierarchical meta-analysis model,which combines two complementary meta-analysis methods, rOp and AW, to detect the biomarkers partially associated with MDD, Bipolar and Schizophrenia. The hierarchical meta-analysis is given in Figure 5.1. Specifically,in Figure 5.2, we illustrate how two meta-analysis designs might be combined for an integrated hierarchical meta-analysis to detect biomarkers related to brain disorder subtypes. In the first layer of the hierarchical design, we combine studies with the same brain disease (MDD, Bipolar or schizophrenia) and control samples using rOp or maxP method to get the consistent biomarkers associated with each disease. In the second layer, AW method is used to combine the p-values obtained from the first layer to identify biomarkers associated or partially associated with these three brain disorders based on the adaptive weights. However, I did not get time to carry out this investigation during during my Ph.D study. I will continue to finish this project.

Evaluation of meta-analysis methods: Among the many microarray meta-analysis methods used in the literature, most methods have their pros and cons depending on the data structure and biological goal. However,so far, there is no any rigorous criteria and method to evaluate the performance of each method. Therefore, it will be very useful to develop some methods to evaluate the performance of the meta-analysis methods.
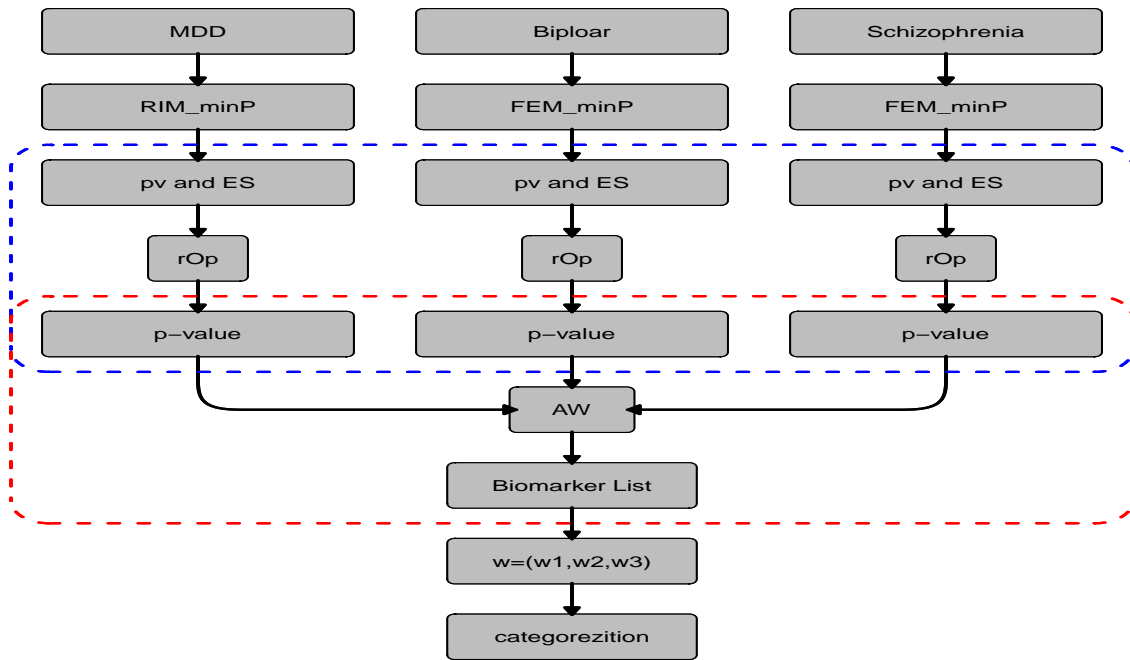
Figure 5.1: The flow chart of a hierarchical meta-analysis.

Improvement of the MetaDE Package: The MetaDE package is developing. We need to improve its functionality, for example, to include more methods in this package.
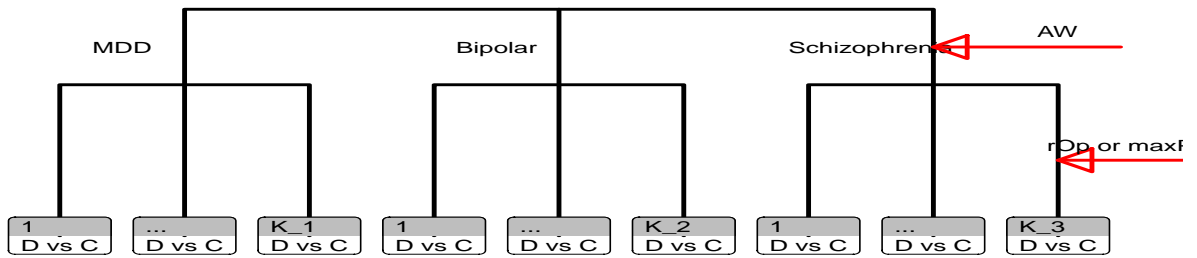
Figure 5.2: The diagram of hierarchical meta-analysis.

APPENDIX A

ALGORITHM OF PERMUTATION ANALYSIS

The procedure of permutation with minP variable selection method::

- Step1: For a give gene g, fit all possible RIM (FEM) models that include at most r (0,1,or r) clinical variables;

- Step2: Select the model RIM_minP(FEM_minP) with minimum p-value associated with LRT for testing $H_0 : \beta_{g0} = 0$. Denote the resulting minimum p-values as $p_g^{(o)}$;

- Step3: Permute the labels of disease and control within each pair (or among all samples) B times. For the bth permutation, repeat step1-2 to get minimum p-value, $p_g^{(b)}$ ($1 \leq b \leq B, 1 \leq g \leq G$);

- Step4: The corrected p-value for gene g is calculated by
  $p_g = \frac{\sum_{b=1}^{B} I(p_g^{(b)} \leq p_g^{(o)})}{B}$ $1 \leq b \leq B, 1 \leq g \leq G)$,
  where $I(.)$ is an indicator function, which takes values ones when the statement is true and zero otherwise.

Remark: Similarly, the above procedures can be used to correct the p-values associated with the RIM_BIC (FEM_BIC) models. Note that BIC is used to choose the RIM_BIC (FEM_BIC) models in step 2-3.

APPENDIX B

ALGORITHM OF CONCORDANCE TEST

Procedure to test consistency of covariate effects in detected DE genes. We denote by e_{glk} the effect of covariate $X_l(1 \leq l \leq L)$ on the gene expression of gene g in the kth study if the covariate is selected by RIM_minP model selection. When covariate X_l is note selected by RIM_minP, e_{glk} is not defined. Define

$$C_{gl}(i,j) = \begin{cases} 1, if\ e_{gli}.e_{glj} > 0 \\ -1, if\ e_{gli}.e_{glj} < 0 \\ 0, if\ e_{gli}\ or\ e_{glj}\ is\ not\ defined \end{cases} \tag{B.1}$$

$C_{gl}(i,j)$ takes value 1 if covariate $X_l(1 \leq l \leq L)$ appears in RIM_minP modes in study i and j, and both effects have the same direction (both positive or both negative). In this situation, the effect $X_l(1 \leq l \leq L)$ in study i and j are consistent. On the contrary, $C_{gl}(i,j)$ takes value -1 when covariate $X_l(1 \leq l \leq L)$ appears in the RIM_minP models in study i and j, and have discordant effect sizes. When the covariate $X_l(1 \leq l \leq L)$ does not appear in the RIM_minP model of either study i or j, $C_{gl}(i,j)$ takes value 0. To test whether the covariates are selected by common covariates across studies more frequently than random, we calculate the total number of times a covariate is selected by RIM_minP among all pairs of studies for a given gene set G' and denote as test statistics T_1 below

$$T_1(G') = \sum_{g \in G'} \sum_{l=1}^{L} \sum_{1 \leq i < < K} |C_{gl}(i,j)| \tag{B.2}$$

To test the concordance of covariate effects across studies, we count only the concordant cases in T_1:

$$T_1(G') = \sum_{g \in G'} \sum_{l=1}^{L} \sum_{1 \leq i < \leq K} I(C_{gl}(i,j) = 1) \tag{B.3}$$

The concordance rate is then defined as $R(G') = T_2(G')/T_1(G')$. To test whether $T_1(G')$ and $R(G'$ is larger than obtained by random with statistically significance. Permutation analysis below is performed:

Algorithm :Test the concordance::

- Step 1: Given a table of gene set G',we calculate the observed statistics of $T_1(G')$ and $R(G')$ and denote them as $T_1^{(o)}$ and $R^{(o)}$, respectively;
- Step 2: Randomly permute the observed (0,1,-1) values across clinical variables for a given gene and a give study for B times. For each permuted data, calculate the $T_1(G')$ and $R(G')$ similarly to obtained $T_1^{(b)}$ and $R^{(b)}$.
- Step 3: Calculate the p-values associated with $T_1(G')$ and $R(G'$ as $p_{T_1} = \frac{\sum_{b=1}^{B} I(T_1^{(b)} \geq T_1^{(o)})}{B}$ $p_R = \frac{\sum_{b=1}^{B} I(R^{(b)} \geq R^{(o)})}{B}$.

APPENDIX C

ALGORITHM OF META-REGRESSION ANALYSIS

Algorithm :to assess the p-value associated with MetaRG_BIC::

- Step 1: Fit the RIM_BIC models in each individual analysis, and estimate the observed effect sizes Y_{gk} and within-study variance σ_{gk}^2.
- Step 2: Fit all the MetaRG models that include at most one (0 or 1) study-specific variable,and then calculate the adaptive weight w^*, and then calculate the p-value : $P[Z(w^*)]$, denoted as $p^{(o)}$
- Step 3: permute the MDD and control labels within each pair in each study. For b^{th} permutation, repeat step 1 and step 2. Denote the p-value in b^{th} permutation as $p_g^{(b)}$
- Step 4: The resulting unbiased p-value for gene g is calculated as
$p_g = \frac{\sum_{g'=1}^{G} \sum_{b=1}^{B} I(p_g^{(b)} \leq p_g^{(o)})}{B}$, where $I(.)$ is an indicator function,

  which takes value one when the statement is true and zero otherwise.

APPENDIX D

THE PROOF OF ONE-SIDED CORRECTION METHODS

Theorem   D.0.1. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic Q_C^{minP} is given by*

$$
G(z) \quad = \quad \begin{cases} 1 - (1-2z)^K, \textit{if } 0 \leq z \leq 0.5 \\[2mm] 1, \textit{if } 0.5 \leq z \leq 1 \end{cases} \tag{D.1}
$$

Proof. Let $\tilde{p}_{[r]}$ denote the rth order statistics of $\tilde{p}_1, \cdots, \tilde{p}_K$. Then, we have

$$
S = Q_L^{minP} = \tilde{p}_{[1]}
$$

and

$$
T = Q_R^{minP} = 1 - \tilde{p}_{[K]}
$$

Therefore, we have

$$
X = \tilde{p}_{[1]} = S
$$

and

$$
Y = \tilde{p}_{[K]} = 1 - T
$$

By theorem **??**, we have the joint pdf of X and Y:

$$g(x, y) = K(K-1)(y-x)^{K-2}$$

if $0 < x \leq y < 1$, and zero otherwise. The Jacobian is J=-1, so the joint pdf of S and T is given by

$$f(s, t) = K(K-1)(1-s-t)^{k-2} \tag{D.2}$$

If $1-s-t > 0$, $0 < s < 1$ and $0 < t < 1$, and zero otherwise.

Thus, we have

$$G(z) = Pr(Q_C^{minP} \leq z)$$
$$= Pr(\min(S, T) \leq z)$$
$$= 1 - Pr(\min(S, T) > z)$$
$$= 1 - Pr(S > z, T > z)$$

if $0 < z < 0.5$

$$Pr(S > z, T > z) = \int_z^{1-z} \int_z^{1-s} K(K-1)(1-s-t)^{K-2} dt ds$$
$$= (1-2z)^K$$

if $0.5 \leq z \leq 1$

$$Pr(S > z, T > z) = 0$$

Then, we have

$$G(z) = \begin{cases} 1 - (1-2z)^K, \textit{if } 0 \leq z < 0.5 \\ 1, \textit{if } 0.5 \leq z \leq 1 \end{cases}$$

$\square$

Theorem   D.0.2. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic $Q_g^{maxP\_OC}$ is given by*

$$F(z) = \begin{cases} 2z^K, \text{ if } 0 \le z < 0.5 \\ 2z^K - (2z-1)^K, \text{ if } 0.5 \le z \le 1 \end{cases} \tag{D.3}$$

Proof. Let $\tilde{p}_{[r]}$ denote the rth order statistics of $\tilde{p}_1, \cdots, \tilde{p}_K$. Then, we have

$$S = Q_L^{maxP} = \tilde{p}_{[K]}$$

and

$$T = Q_R^{maxP} = 1 - \tilde{p}_{[1]}$$

Therefore, we have

$$X = \tilde{p}_{[1]} = 1 - T$$

and

$$Y = \tilde{p}_{[K]} = S$$

By theorem **??**, we have the joint pdf of X and Y:

$$g(x,y) = K(K-1)(y-x)^{K-2}$$

if $0 < x \le y < 1$, and zero otherwise. The Jacobian is J=-1, so the joint pdf of S and T is given by

$$f(s,t) = K(K-1)(s+t-1)^{k-2} \tag{D.4}$$

If $s + t - 1 > 0$, $0 < s < 1$ and $0 < t < 1$, and zero otherwise.

93

Thus, we have

$$G(z) = Pr(Q_C^{maxP} \le z)$$

$$= Pr(\min(S, T) \le z)$$

$$= 1 - Pr(\min(S, T) > z)$$

$$= 1 - Pr(S > z, T > z)$$

if $0 < z < 0.5$

$$Pr(S > z, T > z) = \int_z^{1-z} \int_{1-s}^1 K(K-1)(s+t-1)^{K-2} \mathrm{d}t \mathrm{d}s + \int_{1-z}^1 \int_z^1 K(K-1)(s+t-1)^{K-2} \mathrm{d}t \mathrm{d}s$$

$$= 1 - 2z^K$$

if $0.5 \le z \le 1$

$$Pr(S > z, T > z) = \int_z^1 \int_z^1 K(K-1)(s+t-1)^{K-2} \mathrm{d}t \mathrm{d}s$$

$$= 1 - 2z^K + (2z-1)^K$$

Then, we have

$$G(z) = \begin{cases} 2z^K, \textit{if } 0 \le z < 0.5 \\ 2z^K - (2z-1)^K, \textit{if } 0.5 \le z \le 1 \end{cases}$$

$\square$

Theorem   D.0.3. *If $\tilde{p}_1, \cdots, \tilde{p}_K$ independently and identically follow $U(0,1)$ distribution, then the cumulative distribution function (CDF) of statistic $V_g^{roP\_OC}$ is given by* ***Case1****:* $r \geq K - r + 1$.

$$
F(z) = \begin{cases} 2(1 - F(r-1, K, z)), \, if \, 0 \leq z < 0.5 \\ 1 - \sum_{j=K-r+1}^{r-1} \sum_{h=K-r+1}^{K-j} \frac{K!}{j!h!(K-j-h)!}(1-z)^{j+h}(2z-1)^{K-j-h}, \, if \, 0.5 \leq z \leq 1 \end{cases}
$$

(D.5)

(D.6)

Case2*:* $r < K - r + 1$

$$
F(z) = \begin{cases} 1 - \sum_{l=0}^{r-1} \sum_{m=0}^{r-1} \frac{K!}{l!m!(K-l-m)!}z^{m+l}(1-2z)^{K-l-m}, \, if \, 0 \leq z \leq 0.5 \\ 1, \, if \, 0.5 \leq z \leq 1 \end{cases}
$$

(D.7)

Proof. Let $X_{[r]} = X_{[r]}\{X_1, \cdots, X_K\}$ denote the rth order statistic of K random variables $X_1, \cdots, X_K$, Under the null hypothesis, $\tilde{p}_k$, $1 - \tilde{p}_k$ and p_k all have the $U(0,1)$ distribution. It follows that

$$Q_L^{rop} = X_{[r]}\{\tilde{p}_1, \tilde{p}_2, \cdots, \tilde{p}_K\} \tag{D.8}$$

$$Q_R^{rop} = X_{[r]}\{1 - \tilde{p}_1, 1 - \tilde{p}_2, \cdots, 1 - \tilde{p}_K\} \tag{D.9}$$

and

$$Q_U^{rop} = X_{[r]}\{p_1, p_2, \cdots, p_K\} \tag{D.10}$$

All have the $Beta(r, K - r + 1)$ distribution under H_0.

Then, the one-sided test statistic is defined as

$$Q_C^{rop} = \min(Q_L^{rop}, Q_R^{rop}) \tag{D.11}$$

Let $\tilde{p}_{[r]}$ denote the rth order statistics of $\tilde{p}_1, \cdots, \tilde{p}_K$. Then, we have

$$S = Q_L^{rop} = \tilde{p}_{[r]}$$

and

$$T = Q_R^{rop} = 1 - \tilde{p}_{[K-r+1]}$$

Therefore, we have

$$X = \tilde{p}_{[K-r+1]} = 1 - T$$

and

$$Y = \tilde{p}_{[r]} = S$$

Case1: $r > K - r + 1$. Then, by theorem, we have the joint pdf of X and Y:

$$g(x, y) = \frac{x^{K-r}(y - x)^{2r-K-2}(1 - y)^{K-r}}{B(r, K - r + 1)B(K - r + 1, 2r - K - 1)}$$

if $0 < x < y < 1$, and zero otherwise, and $B(., .)$ is the Beta function. The Jacobian is J=-1, so the joint pdf of S and T is given by

$$f(s, t) = \frac{(1 - t)^{K-r}(s + t - 1)^{2r-K-2}(1 - s)^{K-r}}{B(r, K - r + 1)B(K - r + 1, 2r - K - 1)} \tag{D.12}$$

If $s + t - 1 > 0$, $0 < s < 1$ and $0 < t < 1$, and zero otherwise.

Thus, we have

$$G(z) = Pr(Q_C^r op \le z)$$
$$= Pr(\min(S, T) \le z)$$
$$= 1 - Pr(\min(S, T) > z)$$
$$= 1 - Pr(S > z, T > z)$$

96

if $0 < z < 0.5$

$$
\begin{aligned}
Pr(S > z, T > z) &= \int_z^{1-z} \int_{1-s}^1 \frac{(1-t)^{K-r}(s+t-1)^{2r-K-2}(1-s)^{K-r}}{B(r, K-r+1)B(K-r+1, 2r-K-1)} \mathrm{d}t\mathrm{d}s \\
&\quad + \int_{1-z}^1 \int_z^1 \frac{(1-t)^{K-r}(s+t-1)^{2r-K-2}(1-s)^{K-r}}{B(r, K-r+1)B(K-r+1, 2r-K-1)} \mathrm{d}t\mathrm{d}s \\
&= \frac{\int_z^{1-z}(1-s)^{K-r}s^{r-1}\mathrm{d}s}{B(r, K-r+1)} + (F(r-1, K, z) - F(K-r, K, z)) \\
&= (F(r-1, K, z) - F(r-1, K, 1-z)) + (F(r-1, K, z) - F(K-r, K, z)) \\
&= 2F(r-1, K, z) - F(r-1, K, 1-z) - F(K-r, K, z) \\
&= 2F(r-1, K, z) - 1
\end{aligned}
$$

,whre $F(x, K, p) = Pr(X \le x)$ is the CDF of random variable X from a binomial distribution with the sample size K and the "probability of success",p

if $0.5 \le z \le 1$

$$
\begin{aligned}
Pr(S > z, T > z) &= \int_z^1 \int_z^1 \frac{(1-t)^{K-r}(s+t-1)^{2r-K-2}(1-s)^{K-r}}{B(r, K-r+1)B(K-r+1, 2r-K-1)} \mathrm{d}t\mathrm{d}s \\
&= \frac{1}{B(r, K-r+1)} \int_z^1 (1-s)^{K-r}s^{r-1}I_{\frac{1-z}{s}}(K-r+1, 2r-k-2)\mathrm{d}s \\
&= \frac{1}{B(r, K-r+1)} \sum_{j=K-r+1}^{r-1} \frac{(r-1)!}{j!(r-1-j)!}(1-z)^j \int_z^1 (1-s)^{K-r}(s-1+z)^{r-1-j}\mathrm{d}s \\
&= \frac{1}{B(r, K-r+1)} \sum_{j=K-r+1}^{r-1} \frac{(r-1)!}{j!(r-1-j)!}(1-z)^j z^{K-j} \int_0^{\frac{1-z}{z}} v^{K-r}(1-v)^{r-1-j}\mathrm{d}v \\
&= \sum_{j=K-r+1}^{r-1} \sum_{h=K-r+1}^{K-j} \frac{K!}{j!h!(K-j-h)!}(1-z)^{j+h}(2z-1)^{K-j-h}
\end{aligned}
$$

Then, we have

$$
G(z) = \begin{cases} 2(1 - F(r-1, K, z)), \textit{if } 0 \le z < 0.5 \\ 1 - \sum_{j=K-r+1}^{r-1} \sum_{h=K-r+1}^{K-j} \frac{K!}{j!h!(K-j-h)!}(1-z)^{j+h}(2z-1)^{K-j-h}, \textit{if } 0.5 \le z \le 1 \end{cases}
$$

Case2: $r < K - r + 1$, similarly, we have

$$g(x, y) = \frac{x^{r-1}(x-y)^{K-2r}(1-x)^{r-1}}{B(r, K-r+1)B(K-2r+1, r)}$$

and

$$g(s, t) = \frac{s^{r-1}(1-s-t)^{K-2r}(t)^{r-1}}{B(r, K-r+1)B(K-2r+1, r)}$$

if $1 - t - s > 0$, $0 < t < 1$ and $0 < s < 1$, and zero otherwise.

Then, if $0 \leq z < 0.5$, we have

$$Pr(S > z, T > z) = \int_z^{1-z} \int_z^{1-s} \frac{s^{r-1}(1-s-t)^{K-2r}t^{r-1}}{B(r, K-r+1)B(K-2r+1, r)} dt ds$$

$$= \sum_{j=K-2r+1}^{K-r} \sum_{h=j+1}^{r+j} \frac{K!}{(K-r-j)!h!(r+j-h)!} z^{K-h}(1-2z)^h$$

if $0.5 \leq z < 1$, we have $Pr(S > z, T > z) = 0$

$$G(z) = \begin{cases} 1 - \sum_{j=K-2r+1}^{K-r} \sum_{h=j+1}^{r+j} \frac{K!}{(K-r-j)!h!(r+j-h)!} z^{K-h}(1-2z)^h, & \text{if } 0 \leq z \leq 0.5 \\ 1, & \text{if } 0.5 \leq z \leq 1 \end{cases}$$

$$= \begin{cases} 1 - \sum_{l=0}^{r-1} \sum_{m=0}^{r-1} \frac{K!}{l!m!(K-l-m)!} z^{m+l}(1-2z)^{K-l-m}, & \text{if } 0 \leq z \leq 0.5 \\ 1, & \text{if } 0.5 \leq z \leq 1 \end{cases}$$

□

APPENDIX E

TABLE OF SIMULATIONS

(X: disease state; Y: gene expression; Z: clinical variables)

Scenarios	Assumptions	Simulation scheme
Scenario I Z Y X	$Z_1, \ldots, Z_{10} \sim BIN(1,0.5)$ Y is linked to X , Z_1 and Z_2 by equation: $\quad Y = \beta_0 + \beta_1 * X + \beta_2 * Z_1 + \beta_3 * Z_2 +$ ε, Where $\varepsilon \sim N(0, \sigma^2)$	Step 1: Simulate $Z_1, \ldots, Z_{10}$ i.i.d. from $BIN(1,0.5)$ Step 2: Simulate microarray data • 100 DE genes: Cases: (the first 25 samples, X=1) $Y = \beta_0 + \beta_1 + \beta_2 * Z_1 + \beta_3 * Z_2 + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ Controls: (the last 25 samples, X=0) $Y = \beta_0 + \beta_2 * Z_1 + \beta_3 * Z_2 + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ • 900 Non-DE genes $Y = N(\beta_0, \sigma^2)$ for all 50 samples. $\beta_0 = 0, \beta_1 = 0.8, \beta_2 = 0.6,$ $\beta_3 = -0.3, \sigma = 1.$
Scenario II Z Y X	$Z_1, \ldots, Z_{10} \sim BIN(1,0.5)$ Y is linked to X alone by equation: $Y = \beta_0 + \beta_1 * X + \varepsilon$, Where $\varepsilon \sim N(0, \sigma^2)$	Step 1: Simulate $Z_1, \ldots, Z_{10}$ i.i.d. from $BIN(1,0.5)$ Step 2: Simulate microarray data • 100 DE genes: Cases: (the first 25 samples, X=1) $Y = \beta_0 + \beta_1 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ Controls: (the last 25 samples, X=0) $Y = \beta_0 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ • 900 Non-DE genes $Y = N(\beta_0, \sigma^2)$ for all 50 samples. $\beta_0 = 0, \beta_1 = 1.5, \sigma = 1.$
Scenario III Z Y X	Z_1 and X are linked to each other by equation: $\quad logit(\text{Pr}(Z_1 = 1)) = \gamma_0 + \gamma_1 * X$ $Z_2, \ldots, Z_{10} \sim BIN(1,0.5)$ Y is linked to X and Z by equation: $\quad Y = \beta_0 + \beta_1 * X + \varepsilon$, Where $\varepsilon \sim N(0, \sigma^2)$	Step 1: Simulate $Z_2, \ldots, Z_{10}$ i.i.d. from $BIN(1,0.5)$ Step 2: Simulate Z_1 by $\quad Z_1\|_X \sim BIN(1, \dfrac{1}{1 + e^{-\gamma_0 - \gamma_1 \cdot x}})$ Step 3: Simulate microarray data • 100 DE genes: Cases: (the first 25 samples, X=1) $Y = \beta_0 + \beta_1 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ Controls: (the last 25 samples, X=0) $Y = \beta_0 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ • 900 Non-DE genes $Y = N(\beta_0, \sigma^2)$ for all 50 samples. $\gamma_0 = 0, \gamma_1 = 2, \beta_0 = 0, \beta_1 = 1,$ $\sigma = 1.$

Figure E1: Simulation scheme of three correlation structures in Scenario I, II and III. (X: disease state; Y: gene expression; Z: clinical variables)

APPENDIX F

TEN MDD RELATED GENES

	Age					Alcohol					Antidep					pH					PMI					Suicide					Co-appearance T_1	Concordance T_2	Ratio R $=T_2/T_1$
	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY	MD1_ACC	MD2_ACC	MD3_ACC	MD1_AMY	MD3_AMY			
VGF	0	0	-1	0	-1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	7	7	
SST	0	0	-1	0	-1	1	1	0	1	0	0	-1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	7	7	
CNP	0	1	0	1	0	-1	0	0	1	0	-1	0	1	0	-1	0	-1	0	0	0	0	0	0	0	-1	0	0	1	0	0	5	2	
NPY	0	-1	-1	0	-1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	-1	1	1	1	10	7	
TAC1	0	-1	-1	0	-1	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	5	5	
MBP	0	0	1	0	1	0	-1	0	0	0	-1	0	0	1	0	0	0	1	1	0	0	-1	0	0	0	-1	0	0	0	-1	4	3	
MOBP	0	0	0	1	1	0	0	0	0	0	-1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	-1	0	1	0	-1	7	3	
RGS4	-1	0	0	0	-1	0	1	-1	1	0	1	0	-1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	6	3	
HTR2A	0	-1	0	1	0	1	1	0	0	-1	1	0	-1	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	10	3	
HTR2C	-1	0	-1	0	-1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	5	5	
total																															T_1 =66	T_2 =45	R= T_2/ T_1 =0.68
p-value																															0.17		0.02

Figure F1: The direction of covariates effect in RIM_minP models for 10 MDD related genes from literature.

BIBLIOGRAPHY

[1] Altar, C.A. et al. Electroconvulsive seizures regulate gene expression of distinct neurotrophic signaling pathways. *J Neurosci* 24, 2667-77 (2004).

[2] Assaf P.Oron, Zhen, Jiang & Rober Gentleman. gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, 24(22) 2586-2591, 2008.

[3] Aston C, Jiang L, Sokolov BP. Transcriptional profiling reveals evidence for signaling and oligodendroglial abnormalities in the temporal cortex from patients with major depressive disorder. *Mol Psychiatry*, 10(3):309-322, 2005.

[4] Ayuso-Gutiḷerrezd JL (2005) Depressive subtypes and efficacy of antidepressive pharmacotherapy. *World J Biol Psychiatry.* 6 Suppl 2: 31-37.

[5] Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI(2009): Understanding heterogeneity in meta-analysis: the role of meta-regression. *Int J Clin Pract* , 63(10):1426-1434.

[6] Bauer M, Whybrow PC, Angst J, Versiani M, Muler H-J, of Societies of Biological Psychiatry Task Force on Treatment Guidelines for Unipolar Depressive Disorders WF. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 2: Maintenance treatment of major depressive disorder and treatment of chronic depressive disorders and subthreshold depressions. *World J Biol Psychiatry* 3: 69-86,2002.

[7] Belmaker RH, Agam G. Major depressive disorder. *N Engl J Med* , 358(1):55-68,2008.

[8] Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289-300,1995.

[9] Becker, B. J. Synthesizing standardized meanchange measures. *British Journal of Mathematical and Statistical Psychology*, 41, 257-278,1988.

[10] Berkey CS, Hoaglin DC, Mosteller F, Colditz GA: A random-effects regression model for meta-analysis. *Stat Med* , 14(4):395-411, 1995.

[11] Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA(1998) Meta-analysis of multiple outcomes by regression with random effects. *Stat Med*, 17(22):2537-2550.

[12] Gabriel F. Berriz, Oliver D. King, Barbara Bryant, Chris Sander and Frederick P. Roth (2003) Characterizing gene sets with FuncAssociate, *Bioinformatics*, 19, 2502-2504.

[13] Campain, A. and Yang, Y.H. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, 11, 408.

[14] Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A, Poulton R. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* 301: 386-389, 2003.

[15] Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84-90,2003.

[16] Choudary, P.V. et al. Altered cortical glutamatergic and GABAergic signal transmission with glial involvement in depression. *Proc Natl Acad Sci U S A* 102, 15653-8 2005. Cipriani2005

[17] Cipriani A, Brambilla P, Furukawa T, Geddes J, Gregis M, Hotopf M, Malvini L, Barbui C (2005) Fluoxetine versus other types of pharmacotherapy for depression. *Cochrane Database Syst Rev*: CD004185.

[18] Cochran W.G. The combination of Estimates from Different Expreriments. *Biometerics*,10(1): 101-129, 1954.

[19] Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, Mosteller F(1994) Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA*, 271(9):698-702.

[20] Cohen, J. Statistical Power Analysis for Behavioural Sciences, 2nd edition. Erlbarum, hillsdale, Nj.

[21] DeConde, R.P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B. and Etzioni, R. Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol*, **5**, Article15,2006.

[22] DerSimonian R. and Laird N.M. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177-188, 1986.

[23] Demidenko,Eugene. Mixed model: Theory and Application. *WILEY-INTERSCIENCE*,2004.

[24] Kam D. Dahlquist, Nathan Salomonis, Karen Vranizan, Steven C. Lawlor and Bruce R. Conklin(2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, *Nat Genet*, 31, 19-20.

[25] Sorin Draghici,a Purvesh Khatri, Pratik Bhavsar, Abhik Shah, Stephen A. Krawetz, and Michael A. Tainsky(2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate, *Nucl. Acids Res.*, 31, 3775-3781.

[26] Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods* 1996, 1(2):170-177, 1996.

[27] Efron B., Tibshirani, R., Storey J. D., and Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96, 1151-1160, 2001.

[28] Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178, 2005.

[29] Evans, S.J. et al. Dysregulation of the fibroblast growth factor system in major depression. *Proc Natl Acad Sci U S A* 101, 15506-11, 2004.

[30] Fava M, Kendler KS Major depressive disorder. *Neuron* 28: 335-341,2000.

[31] Fisher R. Statistic Methods for Research works. 4th. (Edinburgh, Oliver and Boyd), 1932.

[32] Fisher R. Combining independent tests of significance. *American Statistician*, 2(5):30 1948.

[33] Fleiss JL(1993): The statistical basis of meta-analysis. *Stat Methods Med Res*, 2(2):121-145.

[34] Glorioso, C., Oh, S., Douillard, G.G. and Sibille, E. Brain molecular aging, promotion of neurological disease and modulation by Sirtuin5 longevity gene polymorphism. *Neurobiol Dis*, 41, 279-290.

[35] Fleiss,J.L. Measures of effect size for categorical data. In the handbook of Research Synthesis. Sage, New York, NY. 245-260.

[36] Gaynes BN, Rush AJ, Trivedi MH, Wisniewski SR, Spencer D, Fava M. The STAR*D study: treating depression in the real world. *Cleve Clin J Med* 75: 57-66,2008.

[37] Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006, *Nucl. Acids Res.*, 34, D322-326.

[38] Greenland S: Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev* , 9:1-30, 1987.

[39] Glorioso, C. and Sibille, E. Between destiny and disease:Genetics and molecular pathways of human central nervous system aging. *Prog Neurobiol*, 93, 165-181.

[40] Colditz GA, Burdick E and Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. *Am J Epidemiol* 142(4):371-382, 1995.

[41] Hamidi, m., drevets, w.c. and price, j.l. glial reduction in amygdala in major depressive disorder is due to oligodendrocytes, *biol psychiatry*, 55, 563-569, 2004.

[42] Hardy RJ, Thompson SG: Detecting and describing heterogeneity in meta-analysis. *Stat Med* , 17(8):841-856, 1998.

[43] Hedges,L.V. Distribution theory for glassa̐́rs estimator of effect size and related estimators. *J. Educ. Stat.*, 6, 107ÍC128, 1981.

[44] Hedges L, Olkin I. Statistcal Methods for meta-analysis. London: Academeic Press, 1985.

[45] Higgins JP, Thompson SG(2004) Controlling the risk of spurious findings from meta-regression. *Stat Med*,23(11):1663-1682.

[46] Holden C. Mental health. Global survey examines impact of depression. *Science* 288: 39-40, 2000.

[47] Hong, F. and Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**, 374-382.

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

[49] Huizenga HM, Visser I, Dolan CV(2011): Testing overall and moderator effects in random effects meta-regression. *Br J Math Stat Psychol*, 64(Pt 1):1-19.

[50] Lu, S., Li, J., Song, C., Shen, K. and Tseng, G.C. Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26, 333-340, 2010.

[51] Iwamoto, K., Kakiuchi, C., Bundo, M., Ikeda, K. and Kato, T. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol Psychiatry* 9, 406-16 (2004).

[52] Li J and Tseng,G.C. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Annals of Applied Statistics*. accepted.

[53] Kang HJ, Adams DH, Simen A, Simen BB, Rajkowska G, Stockmeier CA. Overholser JC, Meltzer HY, Jurjus GJ, Konick LC et al. Gene expression profiling in postmortem prefrontal cortex of major depressive disorder. *J Neurosci*, 27(48):13329-13340,2007.

[54] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucl. Acids Res.*, 28, 27-30.

[55] Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS, Replication NCS The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA 289*: 3095-3105, 2003.

[56] Kendler KS, Hettema JM, Butera F, Gardner CO, Prescott CA. Life event dimensions of loss, humiliation, entrapment, and danger in the prediction of onsets of major depression and generalized anxiety. *Arch Gen Psychiatry.* 60: 789-796,2003.

[57] Kendler KS, Liu X-Q, Gardner CO, McCullough ME, Larson D, Prescott CA. Dimensions of religiosity and their relationship to lifetime psychiatric and substance use disorders. *Am J Psychiatry.*160: 496-503,2003.

[58] Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 5: e45,2008.

[59] Klempan, T.A., Ernst, C., Deleva, V., Labonte, B. & Turecki, G. Characterization of QKI gene expression, genetics, and epigenetics in suicide victims with major depressive disorder. *Biol Psychiatry* 66, 824-31 2009.

[60] Knapp G, Hartung J: Improved tests for a random effects meta-regression with a single covariate. *Stat Med*, 22(17):2693-2710,2003.

[61] Kui Shen and George C Tseng. (2010) Meta-analysis for pathway enrichment analysis when combining multiple microarray studies. *Bioinformatics.* 26:1316-1323.

[62] Law MR, Wald NJ, Thompson SG(1994) By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ*, 308(6925):367-372.

[63] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM(1999) Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 282(11):1061-1066.

[64] Miklos GL, Maleszka R. Microarray reality checks in the context of a complex disease. *Nat Biotechnol*,**22(5)**:615-621,2004.

[65] Depression *National Institute of Mental Health (NIMH)*,2008,

[66] Marot, G., Foulley, J.L., Mayer, C.D. and Jaffrezic, F. Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics*, **25**, 2692-2699, 2009.

[67] Mistry, M. and Pavlidis, P. A cross-laboratory comparison of expression profiling data from normal human postmortem brain. *Neuroscience*, **167**, 384-395,2010.

[68] Morris, S. B.. Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, 53, 17ÍC29,2000.

[69] Mosteller F, and Bush RR . Selected quantitative techniques. *In: Lindzey G, ed. Handbook of Social Psychology*, Cambridge, Mass: Addison-Wesley, p 289-334,1954.

[70] Mueller TI, Leon AC, Keller MB, Solomon DA, Endicott J, Coryell W, Warshaw M, Maser JD Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *Am J Psychiatry* 156: 1000-1006,1999.

[71] Naudet F, Maria AS, Falissard B: Antidepressant response in major depressive disorder(2011) a meta-regression comparison of randomized controlled trials and observational studies. *PLoS One*, 6(6):e20811

[72] Nestler EJ, Barrot M, DiLeone RJ, Eisch AJ, Gold SJ, Monteggia LM. Neurobiology of depression. *Neuron*, 34(1):13-25, 2002.

[73] Michael a. Newton, Fernando a. Quintana, Johan a. Denboon, Srikumar Sengupta and Paul Ahlquist. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, *Ann. Appl. Stat.*, 1, 85-106.

[74] Oquendo MA, Barrera A, Ellis SP, Li S, Burke AK, Grunebaum M, Endicott J, Mann JJ (2004) Instability of symptoms in recurrent major depression: a prospective study. *Am J Psychiatry* 161: 255-261.

[75] Art B. Owen(2009) KARL PEARSON'S META-ANALYSIS REVISITED, *The Annals of Statistics*,37(6B): 3867-3892, 2009.

[76] Park, T., Yi, S.G., Shin, Y.K. and Lee, S. Combining multiple microarrays in the presence of controlling variables. *Bioinformatics*, **22**, 1682-1689, 2006.

[77] Pearson, K. On a new method of determining ařgoodness of fit.aś *Biometrika* 26:425ÍC442, 1934.

[78] Ramasamy, A., Mondry, A., Holmes, C.C. and Altman, D.G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*, 5, e184,2008.

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

[80] Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101, 9309-9314,2004.

[81] Rosenthal, R. Parametric measures of effect size. In The handbook of Research synthesis. *Sage, New York,NY.* 231-244.

[82] Sequeira A, Gwadry FG, Ffrench-Mullen JM, Canetti L, Gingras Y, Casero RA, Jr., Rouleau G, Benkelfat C, Turecki G. Implication of SSAT by gene expression and genetic variation in suicide and major depression. *Arch Gen Psychiatry*,63(1):35-48,2006.

[83] Sequeira, A. et al. Patterns of gene expression in the limbic system of suicides with and without major depression. *Mol Psychiatry* 12, 640-55 2007.

[84] Schwarz, G.E. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464, 1978.

[85] Sequeira, A. et al. Global brain gene expression analysis links glutamatergic and GABAergic alterations to suicide and major depression. *PLoS One* 4, e6585 2009.

[86] Shen, K. and Tseng, G.C. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26, 1316-1323,2010.

[87] Sibille E. Wang Y, Joeyen-Waldorf J, Gaiteri C, Surget A, Oh S, Belzung C, Tseng GC. Lewis DA: A molecular signature of depression in the amygdala. *Am J Psychiatry.* , **166**(9).15444-13549, 2005.

[88] Sibille E, Arango V, Galfalvy HC, Pavlidis P, Erraji-Benchekroun L, Ellis SP, John Mann J. Gene expression profiling of depression and suicide in human prefrontal cortex. *Neuropsychopharmacol*, 29(2):351-361,2004.

[89] Smith TC, Spiegelhalter DJ, Thomas A(1995): Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med*, 14(24):2685-2699.

[90] Stouffer, S., Suchman,E., DeVinnery,L., Star,S.,, and Wiliams,J.. The American Soldier,volumn I: Adjustment during Army Life. *Princeton University Press*, 1949.

[91] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102, 15545-15550.

[92] Tan, P.K., Downey, T.J., Spitznagel, E.L., Jr., Xu, P., Fu, D., Dimitrov, D.S., Lempicki, R.A., Raaka, B.M. and Cam, M.C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, **31**, 5676-5684, 2003.

[93] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*; 309: 1351Í C5, 1994.

[94] Thompson SG, Sharp SJ: Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* , 18(20):2693-2708, 1999.

[95] Thompson SG, Higgins JP: How should meta-regression analyses be undertaken and interpreted? *Stat Med*, 21(11):1559-1573, 2002.

[96] Lu Tian, Steven A. Greenber, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane and Peter J. Park(2005) Discovering statistically significant pathways in expression profiling studies, *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13544-13549.

[97] L.H.C. Tippett. The Methods in Statistics. *Williams and Norgate, Ltd.*, **1** edition, 1931.

[98] Tochigi, M. et al. Gene expression profiling of major depression and suicide in the prefrontal cortex of postmortem brains. *Neurosci Res* 60, 184-91 (2008).

[99] Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98: 5116-5121, 2001.

[100] Xingbin Wang, Yan Lin, Chi Song, Etienne Sibille and George C Tseng (2011). A statistical framework to integrate weak-signal microarray studies adjusted for confounding variables with application to major depressive disorder. *bioinformatics*. submitted.

[101] Whitlock, M.C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*. 18(5):1368-1373, 2005.

[102] Wilcoxon,Frank Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80-83, 1945.

[103] Wilkinson B. A statistical consideration in psychological research. *Psychol Bull*, 48(3):156-158,1951.

[104] Zheng Jiang & Robert Gentleman. Extensions to gene set enrichment. *bioinformatics*, 23(3): 306-313,2007.

[105] Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., Lakatta, E. G., Boheler, K. R., Xu, X., Mattson, M. P., Falco, G., Ko, M. S. H., Schlessinger, D., Firman, J., Kummerfeld, S. K., Wood, W. H., III, Zonderman, A. B., Kim, S. K. and Becker, K. G. . AGEMAP: A gene expression database for aging in mice. *PLOS Genetics*. 3:2326ÍC2337,2007.

[106] Barry R Zeeberg,Weimin Feng,Geoffrey Wang, May D Wang, Anthony T Fojo, Margot Sunshine, Sudarshan Narasimhan, David W Kane, William C Reinhold, Samir Lababidi,Kimberly J Bussey, Joseph Riss, J Carl Barrett, and John N Weinstein (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biol*, 4, R28.

[107] Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C. and Guo, Z. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, **25**, 1662-1668, 2009.

[108] Zhong, S., Li, C. and Wong, W.H. (2003) ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis, *Nucl. Acids Res.*, 31, 3483-3486.