ISSUES IN INFORMATION INTEGRATION OF OMICS DATA: MICROARRAY META-ANALYSIS FOR CANDIDATE MARKER AND MODULE DETECTION AND GENOTYPE CALLING INCORPORATING FAMILY INFORMATION

by

Lun-Ching Chang

MS, National Chung Cheng University, Taiwan, 2007

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

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Lun-Ching Chang

It was defended on

April 17th, 2014

and approved by

George C. Tseng, ScD, Associated Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Wei Chen, PhD, Assistant Professor, Department of Pediatrics, Childrens Hospital of Pittsburgh of UPMC, University of Pittsburgh

Yongseok Park, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Daniel E. Weeks, PhD, Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Etienne Sibille, PhD, Associated Professor, Department of Psychiatry, Center for Neuroscience, University of Pittsburgh

Dissertation Advisors: George C. Tseng, ScD, Associated Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh,

Wei Chen, PhD, Assistant Professor, Department of Pediatrics, Childrens Hospital of Pittsburgh of UPMC, University of Pittsburgh

ISSUES IN INFORMATION INTEGRATION OF OMICS DATA: MICROARRAY META-ANALYSIS FOR CANDIDATE MARKER AND MODULE DETECTION AND GENOTYPE CALLING INCORPORATING FAMILY INFORMATION

Lun-Ching Chang, PhD

University of Pittsburgh, 2014

Abstract:

Nowadays, more and more high-throughput genomic data sets are publicly available; therefore, performing meta-analysis to combine results from independent studies becomes an essential approach to increase the statistical power, for example, in the detection of differentially expressed genes in microarray studies. In addition to meta-analysis, researchers also incorporate pathway or clinical information from external databases to perform integrative analysis. In this thesis, I will present three projects which encompass three types of integrative analysis. First, we perform a comprehensive comparative study to evaluate 12 microarray meta-analysis methods in simulation studies and real examples by using four quantitative criteria: detection capability, biological association, stability and robustness, and we propose a practical guideline for practitioners to choose the most appropriate meta-analysis method in real applications. Second, we develop a meta-clustering method to construct co-expressed modules from 11 major depressive disorder transcriptome datasets, incorporated with GWAS and pathway information from external databases. Third, we propose a computationally feasible algorithm to call genotypes with higher accuracy by considering family information from next generation sequencing data for two purposes: (1) to propose a new genotype calling algorithm for complex families, and (2) to extend our algorithm to incorporate external reference panels to analyze family-based sequence data with a small sample

size. In conclusion, we develop several integrative methods for omics data analysis and the result improves public health significance for biomarker detection in biomedical research and provides insights to help understand the underlying disease mechanisms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 DEVELOPMENT OF OMICS TECHNOLOGY

RNA microarray is an important technology for studying gene expression, which can quantify the amount of mRNA transcripts present in a collection of cells, and has been widely used to identify differential expressed (DE) genes in biomedical research. Many microarray platforms were commonly used such as dual- label cDNA; oligonucleotide platforms; Affymetrix GeneChip or Illumina BeadChip. These high-throughput technologies provide us an opportunity to measure thousands of genes and can help identify disease patients or identify candidate genes that contribute to tumor progression, and also improve in cancer diagnostic and prognosis. However, limitations of microarray technology are the quality and amount of RNA, and can only provides gene-based information.

In the past decade, Genome-Wide Association Studies (GWAS) have generated a considerable amount of gene- and disease-related information. GWAS usually test for the association between genotype and phenotype on hundreds of thousands to millions SNPs simultaneously and provides unbiased of large scale investigation of DNA structural (SNP and other variants) changes. GWAS have been successfully identified thousands of associated SNPs for many common diseases [Hindorff et al., 2009], but heterogeneity and various sources of noise have limited the discovery of disease mechanisms. In the beginning of GWAS in the past era, many data sets have small sample sizes due to high costs, however, GWAS typically has small effect sizes, and such disease-related loci can be detected only by very large sample sizes, and most of GWAS do not replicate their finding in subsequent studies.

Next Generation Sequencing (NGS) is the most advanced technology, which can be used for systematically searching the rare variants and help people discover the disease mecha-

nisms. In 2010, the 1,000 Genome Project has generated publicly available database and provided a deep characterization of human genome sequence for people to understand the relationship between genotype and phenotype, which is the central goal in biomedical research [Abecasis et al., 2010]. Unlike GWAS (generate one genotype at each locus), NGS technology generated millions of short segments of sequence "reads" (25 - 250 base), then we need to strongly rely on comprehensive computational analysis to assemble millions of "short-reads" into full sequence. Over 95% of variants in genome regions have allele frequency of 1% or higher were accessible in NGS technology is the major advantage, however, short reads make assembly hard, and does not allow the assessment of copy number accurately. In addition, the cost of NGS data set still high and not affordable for many small labs.

1.2 INFORMATION INTEGRATION

In the past decade, more and more biological high-throughput genomic data sets were publicly available, especially for transcriptomic study of microarray experiments. Combining results from independent studies is an essential way to increase the statistical power to detect differential expressed (DE) genes because the signal from single study is weak (due to limited sample sizes or small effect sizes), especially for complex diseases such as major depressive disorder (MDD); hypertension or Diabetes. Most of published papers of integrative analysis were meta-analysis for DE gene in microarray study. In addition to meta-analysis, incorporating external database such as GWAS or pathways, and integration of clinical data were also alternative ways of integrative analysis. In this section, we will briefly introduce three ways of integrative analysis which are relevant to three topics of my dissertation: (1)genomic meta-analysis; (2)integration of external database and (3)integration of family-based data.

1.2.1 "Horizontal" and "Vertical" genomic meta-analysis

In the comprehensive literature review for microarray meta-analysis proposed by Tseng et al. [2012], the genomic information integration from transcriptomic studies can be combined "horizontally" (Figure 1(A): combines different sample cohorts for the same molecular event) or "vertically" (Figure 1(B): combines different molecular events usually in the same sample cohort, for example, transcriptome profile, genotypes, DNA copy number variation, methylation, microRNA, proteome and phenome. Such as The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/)). In addition to increase the statistical power, genomic meta-analysis can also provides robust and accurate validation across independent studies, and the result can guide future experiments. In Tseng et al. [2012], they also indicated that many meta-analysis methods have been proposed and used in published applications, but the hypothesis behind the analysis needs more attention. In chapter 2, we performed a comprehensive comparative study of microarray meta-analysis method.

1.2.2 Integration of external database

There were many external databases of integration analysis were publicly available and can be used for genomic research. Unlike microarray study, we are unable to get whole data sets or SNPs list from most of GWAS publications to replicate their results, however, there is a database called GWAS catalog collected a list of SNPs/genes reported by GWAS publications [Hindorff et al., 2009], this database contains many entries of disease- or trait-associated SNPs with p-values less than 10^{-5} from all published GWAS which has PubMed ID and SNPs/genes list were updated every six months. We can easily incorporate the information from GWAS catalog database to validate our finding in GWAS or use it as external reference to support our finding.

Pathway analysis (also known as gene set analysis) is a useful statistical tool to test DE gene sets under certain biological function from established pathway databases. A collection of annotated gene sets for pathway analysis such as molecular signatures database (MSigDB: http://www.broadinstitute.org/gsea/msigdb/index.jsp), this database contains pre-defined biological categories such as Gene Ontology (GO); the Kyoto Encyclopedia of Genes and

Genomes (KEGG); Biocarta gene sets; Reactome gene sets which can be used to perform the enrichment analysis to understand the underlying biological mechanism. The integrative analysis incorporated by external databases mentioned above (GWAS and biological pathway databases) were implemented in Chapter 3.

1.2.3 Integration of family-based data

Next generation sequencing (NGS) is the advanced technology for rare variants detection, which strongly rely on accurate genotype calls. We can incorporate the family information from family-based sequencing data because modeling inheritance of alleles can help achieve more accurate variants and reduce sequencing error in NGS platforms. Chen et al. [2013] proposed a genotype calling method by considering family structure in trios that can achieve more accurate genotype calls as compared with one without considering the family structure (reduce genotype calling error rate by 50%). In chapter 4, we proposed an algorithm to integrated family information for genotype calling method of complex families.

1.3 MICROARRAY META-ANALYSIS

In this section, I will introduce the microarray meta-analysis for detecting candidate marker and co-expression module, which motivate me to perform the comparative study of microarray meta-analysis in chapter 2 and developed a co-expression meta-clustering method chapter 3.

1.3.1 Meta-analysis for candidate marker detection

The general steps and key issues for meta-analysis are (1) collect relevant microarray studies for targeted disease hypothesis; (2) extract useful data sets (for example, raw data, p-values or effect sizes); (3) the inclusion/exclusion criteria for microarray studies; (4) use appropriate meta-analysis method to combine multiple studies and (5) analyze data sets and interpret the results. In steps (1)-(3), our first concern to collect the studies is the heterogeneity between

studies, which may be caused by different experimental settings, study design, chip platforms, or statistical method for each individual analysis. For the issue of inclusion/exclusion criteria, it can be arbitrary chosen by ad-hoc expert opinion, naïve sample size threshold or chip platforms without an objective quality control procedure. Step (4) and (5) include the selection of meta-analysis method under certain hypothesis and the result interpretation.

Many microarray meta-analysis methods have been developed and applied in the literature. First, the most common way to combine multiple studies is to combine p-values from each single study. The traditional meta-analysis method can be traced back to 1930s, Fisher [1925] and the minimum p-value method [Tippett et al., 1931], and later new methods were implemented by Stouffer [Stouffer, 1949] and maximum p-value method [Wilkinson, 1951]. It is well-known that the Fisher statistic follows the chi-square distribution, which can be dominated by single extremely significant p-value (maybe just simply due to the large sample sizes). A recent published meta-analysis called adaptive weighted (AW) Fisher's method proposed by Li and Tseng [2011] can avoid the potential bias from Fisher's method. In brief, the AW Fisher's method searches through all possible weights to find the best adaptive weight with the smallest derived p-value from multiple studies and the heterogeneity can be elucidated. Song and Tseng [2012] developed r-th ordered p-value (rOP) method to consider a robust form of maximum p-value method to identify markers differentially expressed in "majority" of studies. Second, another typical meta-analysis method is to combine effect sizes, either using fixed effect model (FEM) or random effect model (REM). The FEM assumes there is no heterogeneity between studies and the overall effect size was estimated by weighted effect size from each single study. Misuse of the FEM method to heterogeneity case will under-estimate the overall effect size (when the underlying assumption of FEM method was violated). REM method is the one takes heterogeneity between studies into consideration, which gives different effect size to each single study. The most common visualization tool for FEM and REM methods is the forest plot, which shows the mean effect size of single study and its confidence interval (usually 95%) and as well as combined effect. Third, combining statistical ranks provides another meta-analysis method. The advantage of the rank-based methods do not require normality assumption and will not be dominated by extremely small p-value because of the large sample sizes.

Many meta-analysis methods have been developed and successfully applied to multiple microarray studies, however, there is currently no clear guideline for people to choose the meta-analysis method properly. In the literature, there were two comparative studies have systematically compared multiple meta-analysis methods [Hong and Breitling, 2008; Campain and Yang, 2010], but the conclusion from these two comparative papers were suggestive with limited insights to guide practitioners. The selection of appropriate meta-analysis method depends both statistical and biological considerations, and this motivates the comparative study in chapter 2

1.3.2 Meta-analysis for detecting co-expression module

Gene co-expression study is an alternative way to look at gene changes of transcriptome studies. Genes are co-expressed if the patterns of expression are highly correlated across samples in a data set, and may reflect possible shared function by similar expression pattern between genes. It has been shown co-expressed genes may arise through multiple biological pathways including cellular co-expression and common regulatory pathways [Lee et al., 2004; Gaiteri et al., 2010]. In the literature, co-expression analysis have been used to build gene networks, and to identify communities, modules, or genes with shared functions [Dobrin et al., 2009; Elo et al., 2007]. In a gene co-expression network, nodes represent genes and nodes are connected if two corresponding genes are highly correlated (co-expressed). Zhang et al. [2005] proposed a general framework for weighted gene co-expression network analysis (WGCNA). They assigns a connection weight for each gene pair to build co-expression network instead of using a binary index of 0 (unconnected) or 1 (connected). Due to the instability of module detection in each single study, this motivates me to develop meta-clustering method to combine multiple microarray studies to construct co-expressed module in chapter 3.

1.4 OVERVIEW OF THE THESIS

1.4.1 Comprehensive study of microarray meta-analysis methods

More and more transcriptomic microarray studies have been generated and deposited in the public domain, such as ArrayExpress from EBI (http://www.ebi.ac.uk/arrayexpress/), Stanford Microarray Database (SMD, http://genome-www5.stanford.edu/), and Gene Expression Omnibus (GEO) from NCBI (http://www.ncbi.nlm.nih.gov/geo/). In genomic research, microarray meta-analysis has become popular, which is a set of statistical tools for combining results (can be p-values, effect sizes or ranks) from multiple studies target on same disease with similar hypothesis setting. In chapter 2, we proposed a comprehensive comparative analysis to evaluated 12 meta-analysis methods. First, we categorized the 12 meta-analysis methods (6 methods combine p values; 2 methods combine effect sizes and 4 methods combine rank statistics were briefly reviewed in the section 2.2.3) according to the three type of hypothesis the best tested in the simulation study (see more detail section 2.3). Second, four quantitative evaluation criteria (detection capability, biological association, stability and robustness) were used in six large-scale microarray applications (each data set contains 4 to 8 studies) to evaluate 12 meta-analysis methods were summarized in section 2.3.3. Third, we also proposed an entropy measure in section 2.3.5 to understand the data structure of p values of "homogeneity" or "heterogeneity" between studies if no priori information can be obtained. Finally in the section 2.4 we will give a guideline to help practitioners select the proper meta-analysis method in their applications.

1.4.2 Co-expression meta-clustering method and DNA variant Genome Wide Association Studies

In addition to DE gene analysis, co-expression analysis can be used to investigate the transcriptional co-regulation and gene correlations, and such results can help people build gene networks, or investigate the modules of genes set with shared biological functions by incorporating pathway database. In chapter 3, we integrated expression (mRNA level) and GWAS (SNPs in DNA level) studies of the molecular bases of major depressive disorder (MDD). Our

central hypothesis is that stable brain co-regulation modules identified by meta-analysis of multiple transcriptome studies may overlap with sets of genes and associated SNPs related to MDD. In section 3.2.2, we developed meta-clustering method of gene co-expression analysis by combining 11 transcriptome studies from postmortem brain of human subjects with major depressive disorder (MDD) and non-psychiatric controls subjects. Fifty co-expression modules were identified by clustering using penalized k-medoids [Tseng, 2007], then we performed enrichment analysis by comparing gene sets identified by GWAS (genes were identified by significant SNPs within pre-defined nucleotide distance from the coding region of each gene) for various sets of disorders. In the result and discussion sections, we also compared the meta-clustering approach by combining co-expression structures from multiple studies with the clustering result of single study and the performance was evaluated by considering various gene sets collected from GWAS. The purpose of this study is to provide insight into the biology of complex disease such as MDD. First, using robust clustering method to identify modules from large-scale disease related data sets. Second, incorporating the external resources (for example, GWAS results, pathway database, etc) to identify key network nodes (genes) from robust module may potentially target to modulate the biological function.

1.4.3 Genotype calling and haplotyping in families

Next generation sequencing (NGS) is the advanced technology which not only looks beyond the common variants (minor allele frequency $> 5\%$) detected by GWAS, but also systematically detect the rare variants, which may help us discover the underlying disease mechanisms more completely. However, next generation sequencing data strongly rely on advanced statistical and computational methods to generate accurate genotypes and haplotypes. Recent studies indicated LD-aware approach and using trio-based NGS data set can obtain more accurate genotype calls and phased haplotypes (Chen et al., 2013; Li et al., 2012). In chapter 4, we extended the current method from analyzing trios to nuclear family or family with multi-generations with affordable computational complexity. Since many sequencing projects contain limited sample sizes, we also developed a method to analyze family-based structure with small sample sizes by incorporating external references as high throughput sequencing

data sets become available in 1,000 Genomes Project [Abecasis et al., 2010]. In section 4.2.2, we focused on developing the procedure by looping over all possible parent-offspring trios to update the probability of observed genotype given the true genotype simultaneously, which is a pivotal step in the hidden Markov model (HMM). Through simulation studies (see section 4.4), we evaluate the performance by using the genotype error calling rate and phasing error (as haplotypes are provided), and we show that incorporating more offspring within family (or complex family with multiple generations) can achieve more accurate genotype calls than trios only, especially in low to modest depth in sequencing data. Specifically, our new method can help obtain more accurate genotypes by incorporating external references when analyzing sequencing data with small sample sizes.

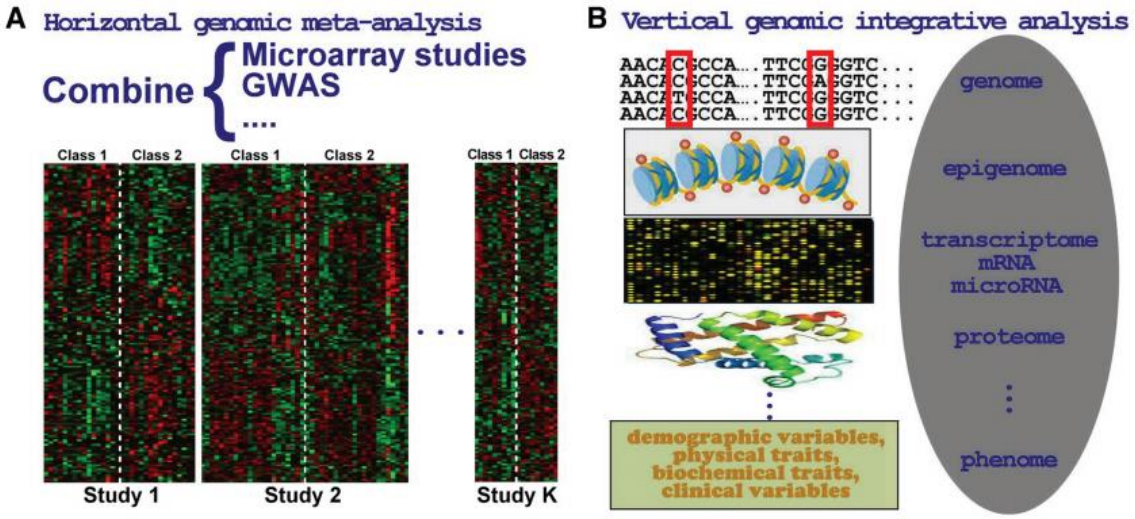

Figure 1: Types of information integration of genomic studies.

(A) Horizontal genomic meta-analysis that combines different sample cohorts for the same molecular event. **(B)** Vertical genomic integrative analysis that combines different molecular events usually in the same sample cohort [Tseng et al., 2012]. "This Figure is used with permission"

2.0 META-ANALYSIS METHODS FOR COMBINING MULTIPLE EXPRESSION PROFILES: COMPARISONS, STATISTICAL CHARACTERIZATION AND AN APPLICATION GUIDELINE

This paper has been published in BMC bioinformatics Chang et al. [2013].

2.1 INTRODUCTION

Microarray technology has been widely used to identify differential expressed (DE) genes in biomedical research in the past decade. Many transcriptomic microarray studies have been generated and made available in public domains such as the Gene Expression Omnibus (GEO) from NCBI (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress from EBI (http://www.ebi.ac.uk/arrayexpress/). From the databases, one can easily obtain multiple studies of a relevant biological or disease hypothesis. Since a single study often has small sample size and limited statistical power, combining information across multiple studies is an intuitive way to increase sensitivity. Ramasamy et al. [2008] proposed a seven-step practical guidelines for conducting microarray meta-analysis: "(i) identify suitable microarray studies; (ii) extract the data from studies; (iii) prepare the individual datasets; (iv) annotate the individual datasets; (v) resolve the many-to-many relationship between probes and genes; (vi) combine the study-specific estimates; (vii) analyze, present, and interpret results". In the first step although theoretically meta-analysis increases the statistical power to detect DE genes, the performance can deteriorate if problematic or heterogeneous studies are combined. In many applications, the data inclusion/exclusion criteria are based on ad-hoc expert opinions, a naïve sample size threshold or selection of platforms without an

objective quality control procedure. Kang et al. [2012] proposed six quantitative quality control measures (MetaQC) for decision of study inclusion. Steps (ii)-(v) are related to data preprocessing. Finally, Steps (vi) and (vii) involve the selection of the meta-analysis method and interpretation of the result and are the foci of this paper.

Many microarray meta-analysis methods have been developed and applied in the literature. According to a recent review paper by Tseng et al. [2012], popular methods mainly combine three different types of statistics: p values, effect sizes and ranks. In this chapter, we include 12 popular as well as state-of-the-art methods in the evaluation and comparison. Six methods (Fisher, Stouffer, adaptively weighted Fisher, minimum p value, maximum p value and r-th ordered p value) belonged to the p value combination category, two methods (fixed effects model and random effects model) belonged to the effect size combination category and four methods (RankProd, RankSum, product of ranks and sum of ranks) belonged to the rank combination category. Details of these methods and citations will be provided in the method section. Despite the availability of many methods, pros and cons of these methods and a comprehensive evaluation remain largely missing in the literature. To our knowledge, Hong and Breitling [2008] and Campain and Yang [2010] are the only two comparative studies that have systematically compared multiple meta-analysis methods. The number of methods compared (three and five methods, respectively) and the number of real examples examined (two and three examples respectively with each example covering 2-5 microarray studies) were, however, limited. The conclusions of the two papers were suggestive with limited insights to guide practitioners. In addition, as we will discuss in the Method section, different meta-analysis methods have different underlying hypothesis setting targets. As a result, the selection of an adequate (or optimal) meta-analysis method depends heavily on the data structure and the hypothesis setting to achieve the underlying biological goal.

In this chapter, we compare 12 popular microarray meta-analysis methods using simulation and six real applications to benchmark their performance by four statistical criteria (detection capability, biological association, stability and robustness). Using simulation, we will characterize the strength of each method under three different hypothesis settings (i.e. detect DE genes in "all studies", "majority of studies" or "one or more studies"; see Method section for more details). We will compare the similarity and grouping of the meta-analysis

methods based on their DE gene detection results (by using a similarity measure and multi-dimension scaling plot) and use an entropy measure to characterize the data structure to determine which hypothesis setting may be more adequate in a given application. Finally, we give a guideline to help practitioners select the best meta-analysis method under the choice of hypothesis setting in their applications.

2.2 METHODS

2.2.1 Real data sets

Six example data sets for microarray meta-analysis were collected for evaluation in this paper. Each example contained 4-8 microarray studies. Five of the six examples were of the commonly seen two-group comparison and the sixth example contained relapse-free survival outcome for breast cancer. We applied the MetaQC package to assess quality of the studies for meta-analysis and determined the final inclusion/exclusion criteria [Kang et al., 2012]. The principal component analysis (PCA) bi-plots and the six QC measures are summarized in Figure S1 and Table S2 and S3. Details of the data sets are available in Table S1.

2.2.2 Underlying hypothesis settings

Following the classical convention of Birnbaum [1954] and Li and Tseng [2011] (see also Tseng et al. 2012), meta-analysis methods can be classified into two complementary hypothesis settings. In the first hypothesis setting (denoted as HS_A), the goal is to detect DE genes that have non-zero effect sizes in all studies:

$$HS_A : H_0 : \bigcap_{i=1}^{K} \{\theta_k = 0\} \text{versus } H_a : \bigcap_{k=1}^{K} \{\theta_k \neq 0\} \tag{2.1}$$

where θ_k is the effect size if study k. The second hypothesis setting (denoted as HS_B), however, aims to detect a DE gene if it has non-zero effect size in "one or more" studies:

$$HS_B : H_0 : \bigcap_{i=1}^{K} \{\theta_k = 0\} \text{versus } H_a : \bigcup_{k=1}^{K} \{\theta_k \neq 0\} \tag{2.2}$$

In most applications, HS_A is more appropriate to detect conserved and consistent candidate markers across all studies. However, different degrees of heterogeneity can exist in the studies and HS_B can be useful to detect study-specific markers (e.g. studies from different tissues are combined and tissue-specific markers are expected and of interest). Since HS_A is often too conservative when many studies are combined, Song and Tseng [2012] proposed a more practical and robust hypothesis setting (namely HS_r) that targets on DE genes with non-zero effect sizes in "majority" of studies, where majority of studies is defined as, for example, more than 50% of combined studies (i.e. $r \geq 0.5 \cdot K$). The robust hypothesis setting considered was:

$$HS_r : H_0 : \bigcap_{i=1}^{K} \{\theta_k = 0\} \text{ versus } H_a : \sum_{k=1}^{K} I\{\theta_k \neq 0\} \geq r \tag{2.3}$$

A major contribution of this chapter is to characterize meta-analysis methods suitable for different hypothesis settings (HS_A, HS_B and HS_r) using simulation and real applications and to compare their performance with four benchmarks to provide a practical guideline.

2.2.3 Implementation and methods

Microarray meta-analysis implementation Assume that we have K microarray studies to combine. For study k ($1 \leq k \leq K$), denote by x_{gsk} the gene expression intensity of gene g ($1 \leq g \leq G$) in sample s ($1 \leq s \leq S_k$; S_k the number of samples in study k), and y_{sk} the disease/outcome variable of sample s. The disease/outcome variable can be binary, multi-class, continuous or censored, representing the disease state, severity or prognosis outcome (e.g. tumor versus normal or recurrence survival time). The goal of microarray meta-analysis is to combine information of K studies to detect differentially expressed (DE) genes associated with the disease/outcome variable. Such DE genes serve as candidate markers for disease classification, diagnosis or prognosis prediction and help understand the genetic mechanisms underlying a disease. In this chapter, before meta-analysis we first applied penalized t statistics to each individual study to generate p values or DE ranks for a binary outcome [Tusher et al., 2001]. In contrast to traditional t-statistics, the penalized t-statistic adds a fudge parameter s_0 to stabilize the denominator ($T = (\bar{X} - \bar{Y})/(\hat{s} + s_0)$; $\bar{X}$ and $\bar{Y}$ are means of case and control groups) and to avoid a large t-statistic due to

small estimated variance $\hat{s}$. The p values were calculated using the null distributions derived from conventional non-parametric permutation analysis by randomly permuting the case and control labels for 10,000 times [Pesarin and Salmaso, 2010]. For censored outcome variables, Cox proportion hazard models and log-rank tests were used [Cox, 1972]. Meta-analysis methods were then used to combine information across studies and generate meta-analyzed p values. To account for multiple comparisons, the Benjamini and Hochberg procedure was used to control false discovery rate (FDR) [Benjamini and Hochberg, 1995]. All methods were implemented using the "MetaDE" package in R [Wang et al., 2012a]. Data sets and all programming codes are available at http://www.biostat.pitt.edu/bioinfo/publication.htm.

Microarray meta-analysis methods According to a recent review paper [Tseng et al., 2012], microarray meta-analysis methods can be categorized into three types: combine p values, combine effect sizes and combine ranks. Below, we briefly describe 12 methods that were selected for comparison.

I. Combined p values

Fisher The Fisher's method sums up the log-transformed p values obtained from individual studies [Fisher, 1925]. The combined Fisher's statistic $\chi^2_{\text{Fisher}} = -2\sum_{i=1}^{k}\log(p_i)$ follows a χ^2 distribution with $2k$ degrees of freedom under the null hypothesis (assuming null p value are uniformly distributed). Note that we perform permutation analysis instead of such parametric evaluation for Fisher and other methods in this paper. Smaller p values contribute larger scores to the Fisher's statistic.

Stouffer Stouffer's method sums the inverse normal transformed p values. Stouffer's statistics $T_{\text{Stouffer}} = \frac{\sum_{i=1}^{k} z_i}{\sqrt{k}}$ ($z_i = \Phi^{-1}(p_i)$, where Φ is standard normal c.d.f) follows a standard normal distribution under the null hypothesis [Stouffer, 1949]. Similar to Fisher's method, smaller p values contribute more to the Stouffer's score, but in a smaller magnitude.

Adaptively weighted (AW) Fisher The AW Fisher's method assigns different weights to each individual study ($\chi^2_{\text{AW}} = -\Sigma_{k=1}^{K} w_k \cdot \log(P_k)$, $w_k = 0$ or 1) and it searches through all possible weights to find the best adaptive weight with the smallest derived p value [Li and Tseng, 2011]. One big advantage of this method is its ability to indicate which studies contribute to the evidence aggregation and elucidates heterogeneity in the meta-analysis.

For Fisher, Stouffer and minP methods targeted on HS_B, candidate markers differentially expressed in one or more studies are detected with no indication of which studies are involved in differential expression. For example, Fisher's method gives the same statistical significance for gene A with p values$= (0.1, 0.1, 0.1, 0.1)$ and gene B with p values$= (0.0001, 1, 1, 1)$; the two genes, however, have very different biological interpretations. The adaptively weighted Fisher's method (AW) was developed to improve biological interpretation and statistical power. AW considered a weighted Fisher score $U(w_1, ..., w_k) = -2\sum_k w_k \cdot \log p_k$ (where weight w_k equals 0 or 1) and the test statistic was defined as the smallest p value of all $2^k - 1$ possible weighted Fisher score (i.e. $T^{AW} = \min_{w_1,...,w_K} p(U(w1, ..., w_K)))$, where $p(U(w_1, ..., w_K))$ is the p value of $U(w_1, ..., w_k)$. The resulting best adaptive weight (i.e. $W^* = \arg\min_{w1,...,w_K} p(U(w_1, ..., w_K)))$ provides indication of which studies contribute to the statistical significance of meta-analysis. For example, $w^* = (1, 1, 1, 1)$ for gene A shows statistical significance in all four studies and $w^* = (1, 0, 0, 0)$ for gene B shows statistical significance in only the first study. AW method is admissible under classical two-sample Gaussian scenario and it generally has better statistical power than traditional Fisher and minP methods in various kinds of alternative hypothesis in HS_B. For more details, refer to Li and Tseng [2011].

Minimum p value (minP) The minP method takes the minimum p value among the K studies as the test statistic [Tippett et al., 1931]. It follows a beta distribution with degrees of freedom $\alpha = 1$ and $\beta = K$ under the null hypothesis. This method detects a DE gene whenever a small p value exists in any one of the K studies.

Maximum p value (maxP) The maxP method takes the maximum p value as the test statistic [Wilkinson, 1951]. It follows a beta distribution with degrees of freedom $\alpha = K$ and $\beta = 1$ under the null hypothesis. This method targets on DE genes that have small p values in "all" studies.

r-th ordered p value (rOP) The rOP method takes the r-th order statistic among sorted p values of K combined studies. Under the null hypothesis, the statistic follows a beta distribution with degrees of freedom $\alpha = r$ and $\beta = K - r + 1$. The minP and maxP methods are special cases of rOP. In Song and Tseng [2012], rOP is considered a robust form of maxP (where r is set as greater than $0.5 \cdot K$) to identify candidate markers differentially expressed in "majority" of studies.

II. Combined effect size

Fixed effects model (FEM) FEM combines the effect sizes across K studies by assuming a simple linear model with an underlying true effect size plus a random error in each study.

Random effects model (REM) REM extends FEM by allowing random effects for the inter-study heterogeneity in the model [Choi et al., 2003]. The meta-analysis method by combining effect sizes from several studies is a t-test based modeling approach. The effect size for a certain gene in the i^{th} study, and $i = 1, 2, ..., K$ is defined as $d_i = \frac{\bar{T}_i - \bar{C}_i}{S_i}$, where $\bar{T}_i$, $\bar{C}_i$ and S_i denote the means of treatment and control group and the estimate of the pooled standard deviation, respectively. An unbiased estimate for d_i is obtained as $d_i^{'} = d_i - 3d_i/(4(n_i - 2 - 1))$ and the estimated variance of the unbiased effect size is $\hat{\sigma}^2_{d_i} = (n_{it}^{-1} + n_{ic}^{-1}) + d_i^2((n_{it} + n_{ic}))^{-1}$, where $n_i = n_{it} + n_{ic}$ is the sample size in the i^{th} study; n_{it} and n_{ic} are the sample sizes of treatment and control group in the i^{th} study respectively. From the number of studies k, a hierarchical model is given as

$$d_i = \theta_i + \varepsilon_i, \varepsilon_i \sim N(0, s_i^2)$$
$$\theta_i = \mu + \delta_i, \delta_i \sim N(0, \tau^2)$$

where s_i^2 is the variance within certain study k; τ^2 is the variance (random effect) between studies and μ is the overall mean, which is the parameter of interest. d_i and s_i^2 given by $d_i^{'}$ and $\hat{\sigma}^2_{d_i}$ are described above. $\tau^2 = 0$ means that there is no variance between studies, hence the hierarchical model reduces to a fixed effects model (FEM), $d_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N(0, s_i^2)$. Otherwise, the hierarchical model is a random effects model (REM), $d_i = \mu + \delta_i + \varepsilon_i$, $\varepsilon_i \sim N(0, s_i^2)$ and $\delta_i \sim N(0, \tau^2)$. The $\hat{\tau}^2$ can be estimated by a method proposed by DerSimonian and Laird [1986].

III. Combined rank statistics

RankProd (RP) and RankSum (RS) RankProd and RankSum are based on the common biological belief that if a gene is repeatedly at the top of the lists ordered by up- or down-regulation fold change in replicate experiments, the gene is more likely a DE gene [Hong et al., 2006]. In detail, suppose there are n studies with (n_{iT}, n_{iC}) replicates, $i = 1, 2, ..., k$. Below is the algorithm of finding up-regulated differential genes from Rank Product method

proposed by Hong et al. [2006]. In the beginning, the pair-wise ratios within each study of their fold-changes were calculated (i.e., for study i, T_{ij}/C_{il}, $j = 1, 2, ..., n_{iT}$, $l = 1, 2, ..., n_{iC}$, and form $k_i = n_{iT} \times n_{iC}$ comparisons:

(1) Define the statistic of rank product $RP_g^{up} = (\Pi_i \Pi_k r_{g,i,k}^{up})^{1/k}$, where $k = k_1 + k_2 + \cdots + k_n$, and $r_{g,i,k}^{up}$ is the position of gene g in the list of genes in the i^{th} study under k^{th} comparison sorted by decreasing pair-wise ratios calculated before.

(2) Do permutations in each array independently for B times and calculate the statistics $RP_g^{up(1)}$, $RP_g^{up(2)}$, ..., $RP_g^{up(B)}$, the same in step (1).

(3) The permutation p-value and FDR assessed by permutation within each gene can be obtained by

$$p_g = (1/GB)\Sigma_b \Sigma_g I(|RP_g^{up(b)}| \leq RP_g^{up})$$
$$FDR_g = \frac{(1/B)\sum_b \sum_g I(|RP_g^{up(b)}| \leq RP_g^{up})}{\sum_g I(|RP_g^{up(b)}| \leq RP_g^{up})}$$

In rank sum (RankSum) method, the statistic $RS_g^{up} = (\sum_i \sum_k r_{g,i,k}^{up})^{1/k}$ was used to replace the statistic RP_g^{up} from the algorithm of rank product (RankProd) mentioned above. This method only considers gene ranks rather than absolute expression values, which leads to its robustness against heterogeneity across different studies.

Product of ranks (PR) and Sum of ranks (SR) These two methods apply a naïve product or sum of the DE evidence ranks across studies Dreyfuss et al. [2009]. Suppose R_{gk} represents the rank of p value of gene g among all genes in study k. The test statistics of PR and SR are calculated as $PR_g = \prod_{k=1}^{K} R_{gk}$ and $SR_g = \sum_{k=1}^{K} R_{gk}$, respectively. P values of the test statistics can be calculated analytically or obtained from a permutation analysis. Note that the ranks taken from the smallest to largest (the choice in the method) are more sensitive than ranking from largest to smallest in the PR method, while it makes no difference to SR.

2.2.4 Characterization of meta-analysis methods

MDS plots to characterize the methods The multi-dimensional scaling (MDS) plot is a useful visualization tool for exploring high-dimensional data in a low-dimensional space [Borg, 2005]. In the evaluation of 12 meta-analysis methods, we proposed a similarity measure between two ordered DE gene list (more detail in the section 2.2.6) for every pair of methods to quantify the similarity of their DE analysis results in a given example. A dissimilarity measure is then defined as one minus the adjusted DE similarity measure and the dissimilarity measure is used to generate an MDS plot of the 12 methods. In the MDS plot, methods that are clustered in a neighbourhood indicate that they produce similar DE analysis results.

Entropy measure to characterize data sets As indicated in the Section of underlying hypothesis settings, selection of the most suitable meta-analysis method(s) largely depends on their underlying hypothesis setting (HS_A, HS_B and HS_r). The selection of a hypothesis setting for a given application should be based on the experimental design, biological knowledge and the associated analytical objectives. There are, however, occasions that little prior knowledge or preference is available and an objective characterization of the data structure is desired in a given application. For this purpose, we developed a data-driven entropy measure to characterize whether a given meta-analysis data set contains more HS_A-type markers or HS_B-type markers [Martin and England, 2011]. The algorithm is described below:

1. Apply Fisher's meta-analysis method to combine p values across studies to identify the top H candidate markers. Here we use $H = 1,000$. H represents the rough number of DE genes (in our belief) that are contained in the data.
2. For each selected marker, we defined the standardized minus p value score for gene g in the k^{th} study as $l_{gk} = -\log(p_{gk}) / -\sum_{k=1}^{K} \log(p_{gk})$. Note that $0 \leq l_{gk} \leq 1$, large l_{gk} corresponds to more significant p value p_{gk}, and $\sum_{k=1}^{K} l_{gk} = 1$.
3. The entropy of gene g is defined as $e_g = -\sum_{k=1}^{K} l_{gk} \log(l_{gk})$. A box-plots of entropies of the top H genes is generated for each meta-analysis application.

Intuitively, a high entropy value indicates that the gene has small p values in all or most studies and is of HS_A or HS_r-type. Conversely, genes with small entropy have small p values in one or only few studies where HS_B-type methods are more adequate. When calculating

l_{gk} in step 2, we capped $-\log(p_{gk})$ at 10 to avoid contributions of close-to-zero p values that can generate near-infinite scores. The entropy box-plots helps determine an appropriate meta-analysis hypothesis setting if no pre-set biological objective exists.

2.2.5 Evaluation criteria

For objective quantitative evaluation, we developed the following four statistical criteria to benchmark performance of the methods.

Detection capability The first criterion considers the number of DE genes detected by each meta-analysis method under the same pre-set FDR threshold (e.g. FDR= 1%). Although detecting more DE genes does not guarantee better "statistical power", this criterion has served as a surrogate of statistical power in previous comparative studies [Wu et al., 2005]. An implicit assumption underlying this criterion is that the statistical procedure to detect DE genes in each study and the FDR control in the meta-analysis are accurate (or roughly accurate). To account for data variability in the evaluation, we bootstrapped (i.e. sampled with replacement) the samples in each study for $B = 50$ times and calculated the number of detected DE genes under the pre-set FDR threshold with 5%. Denote by r_{meb} the rank of detection power performance (the smaller the better) of method m $(1 \leq m \leq 12)$ in example e $(1 \leq e \leq 6)$ and in the b^{th} $1 \leq b \leq 50$ bootstrap simulation. The mean standardized rank (MSR) for method m and example e is calculated as $MSR_{me} = \sum_{b=1}^{B}(r_{meb}/\#\text{of methods compared})/B$ and the aggregated standardized rank (ASR) is calculated as $ASR_m = \sum_{e=1}^{6}(MSR_{me})/6$, representing the overall performance of method m across all six examples. Table S4 shows the MSR and ASR of all 12 methods and Figure 3 (in the result section) shows plot of mean number of detected DE genes with error bars of standard errors for each method ordered by ASR. We note that MSR and ASR are both standardized between 0 and 1. The standardization in MSR is necessary because in the breast cancer survival example we cannot apply FEM, REM, RankSum and RankProd as they are developed only for a two-group comparison.

Biological association The second criterion requires that a good meta-analysis method should detect a DE gene list that has better association with pre-defined "gold standard"

pathways related to the targeted disease. Such a "gold standard" pathway set should be obtained from biological knowledge for a given disease or biological mechanism under investigation. However, since it is well-known that pathway collections are always incomplete and erroneous, such prior knowledge may be missing or arguable by different experts. To facilitate this evaluation without bias, we developed a computational and data-driven approach to determine a set of surrogate disease-related pathways out of a large collection of pathways by combining pathway enrichment analysis results from each single study. Specifically, we first collected 2,287 pathways (gene sets) from MSigDB (http://www.broadinstitute.org/gsea/msigdb/): 1,454 pathways from "GO," 186 pathways from "KEGG," 217 pathways from "BIOCARTA" and 430 pathways from "REACTOME", respectively. We filtered out pathways with less than 5 genes or more than 200 genes and 2,113 pathways were left for the analysis. DE analysis was performed in each single study separately and pathway enrichment analysis was performed for all the 2,113 pathways by the Kolmogorov-Smirnov (KS) association test. Denote by p_{uk} the resulting pathway enrichment p value for pathway u ($1 \leq u \leq 2,113$) and study k ($1 \leq k \leq K$). For a given study k, enrichment ranks over pathways were calculated as $r_{uk} = \text{rank}_u(p_{uk})$. A rank-sum score for a given pathway u was then derived as $s_u = \sum_{k=1}^{K} r_{uk}$. Intuitively, pathways with small rank-sum scores indicate that they are likely associated with the disease outcome by aggregated evidence of the K individual study analyses. We chose the number of top D pathways that had the smallest rank-sum scores as the surrogate disease-related pathways and used these to proceed with the biological association evaluation of meta-analysis methods in the following.

Given the selected surrogate pathways D, the following procedure was used to evaluate performance of the 12 meta-analysis methods for a given example e ($1 \leq e \leq 6$). For each meta-analysis method m ($1 \leq m \leq M = 12$), the DE analysis result was associated with pathway u and the resulting enrichment p value by KS-test was denoted by $\widetilde{P}_{med}$ ($1 \leq d \leq D$). The rank of $\widetilde{P}_{med}$ for method m among 12 methods was denoted by $v_{med} = \text{rank}_m(\widetilde{P}_{med})$. Similar to the detection power evaluation, we calculated the mean standardized rank (MSR) for method m and example e as $MSR_{me} = \sum_{d=1}^{D}(v_{meb}/\#\text{of methods compared})/D$ and the aggregated standardized rank (ASR) as $ASR_m = \sum_{e=1}^{6}(MSR_{me})/6$, representing the overall

performance of method m. To select the parameter D for surrogate disease-related pathways, Supplement Figure S4 shows the trend of MSR_{me} (on the y-axis) versus D (on the x-axis) as D increases. The result indicated that the performance evaluation using different D only minimally impacted the conclusion when $D > 30$. We choose $D = 100$ throughout this paper.

Note that we used the KS test, instead of the popular Fisher's exact test because each single study detected variable number of DE genes under a given FDR cutoff and the Fisher's exact test is usually not powerful unless a few hundred DE genes are detected. On the other hand, the KS test does not require an arbitrary p value cutoff to determine the DE gene list for enrichment analysis.

Stability The third criterion examines whether a meta-analysis method generates stable DE analysis results. To achieve this goal, we randomly split samples into half in each study (so that cases and controls are as equally split as possible). The first half of each study was taken to perform the first meta-analysis and generate a DE analysis result. Similarly, the second half of each study was taken to perform a second meta-analysis. The generated DE analysis results from two separate meta-analyses were compared by the adjusted DE similarity measures (described in the next section). The procedure was repeated for $B = 50$ times. Denote by S_{meb} the adjusted DE similarity measure of method m of the b^{th} simulation in example e. Similar to the first two criteria, MSR and ASR are calculated based on S_{meb} to evaluate the methods.

Robustness The final criterion investigates the robustness of a meta-analysis method when an outlying study is randomly added to the meta-analysis. For each of the six real examples, we randomly picked one irrelevant study from the other five examples and added it to the meta-analysis. The adjusted DE similarity measure was calculated between the original meta-analysis and the new meta-analysis with an added outlier. A higher adjusted similarity measure shows better robustness againt inclusion of the outlying study. This procedure is repeated until all irrelevant studies are used. The MSR and ASR are then calculated based on the adjusted DE similarity measures to evaluate the methods.

2.2.6 Similarity measure between two ordered DE gene lists

To compare results of two DE detection methods (from single study analysis or meta-analysis), a commonly used method in the literature is to take the DE genes under a certain p value or FDR threshold, plot the Venn diagram and compute the ratio of overlap. This method, however, greatly depends on the selection of the FDR threshold and is unstable. Another approach is to take the generated ordered DE gene lists from two methods and compute the non-parametric Spearman rank correlation [Spearman, 1904]. This method avoids the arbitrary FDR cutoff but gives, say, the top 100 important DE genes and the bottom 100 non-DE genes equal contribution. To circumvent this pitfall, Li et al. [2011] proposed a parametric reproducibility measure for ChIP-seq data in the ENCODE project. Yang et al. [2006] introduced an OrderedList measure to quantify similarity of two ordered DE gene lists. For simplicity, we extended the OrderedList measure into a standardized similarity score for the evaluation purpose in this paper. Specifically, suppose G_1 and G_2 are two ordered DE gene lists (e.g. ordered by p values) and small ranks represent more significant DE genes. We denote by $O_n(G_1, G_2)$ the number of overlapped genes in the top n genes of G_1 and G_2. As a result, $0 \leq O_n(G_1, G_2) \leq n$ and a large $O_n(G_1, G_2)$ value indicates high similarity of the two ordered lists in the top n genes. A weighted average similarity score is calculated as $S(G_1, G_2) = \sum_{n=1}^{G} e^{-\alpha n} \cdot O_n(G_1, G_2)$, where G is the total number of matched genes and the power α controls the magnitude of weights emphasized on the top ranked genes. When α is larger top ranked genes are weighted higher in the similarity measure. The expected value (under the null hypothesis that the two gene rankings are randomly generated) and maximum value of S can be easily calculated: $E_{null}(S(G_1, G_2)) = \sum_{n=1}^{G} e^{-\alpha n} \cdot \frac{n^2}{G}$ and $\max(S(G_1, G_2)) = \sum_{n=1}^{G} e^{-\alpha n} \cdot n$ We apply an idea similar to the adjusted Rand index [Hubert and Arabie, 1985] used to measure similarity of two clustering results and define the adjusted DE similarity measure as

$$S^*(G_1, G_2) = \frac{S(G_1, G_2) - E_{null}(S(G_1, G_2))}{\max(S(G_1, G_2)) - E_{null}(S(G_1, G_2))} \tag{2.4}$$

This measure ranges between -1 to 1 and gives an expected value of 0 if two ordered gene lists are obtained by random chance. Yang et al. [2006] proposed a resampling-based ROC

method to estimate the best selection of α. Since the number of DE genes in our examples is generally high, we choose a relatively small $\alpha = 0.001$ throughout this paper. We have tested different α and found that the results were similar (Figure S7).

2.3 RESULTS

2.3.1 Simulation setting

We conducted simulation studies to evaluate and characterize the 12 meta-analysis methods for detecting biomarkers in the underlying hypothesis settings of HS_A, HS_B or HS_r. The simulation algorithm is described below:

1. We simulated 800 genes with 40 gene clusters (20 genes in each cluster) and other 1,200 genes do not belong to any cluster. The cluster indexes C_g for gene $g(1 \leq g \leq 2,000)$ were randomly sampled, such that $\sum I\{C_g = 0\} = 1,200$ and $\sum I\{C_g = c\} = 20, (1 \leq c \leq 40)$.

2. For genes in cluster $c(1 \leq c \leq 40)$ and in study $k(1 \leq k \leq 5)$, we sampled $\sum_{ck}^{'} \sim W^{-1}(\Psi, 60)$, where $\Psi = 0.5I_{20\times20} + 0.5J_{20\times20}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all elements equal 1. We then standardized $\sum_{ck}^{'}$ into $\sum_{ck}$, where the diagonal elements are all 1's.

3. 20 genes in cluster c was denoted by the index of $g_{c1,\ldots,g_{c20}}$, i.e. $C_{g_{cj}} = c$, where $1 \leq c \leq 40$ and $1 \leq j \leq 20$. We sampled gene expression levels of genes in cluster c for sample n as $(X'_{g_{c1}nk}, \ldots, X'_{g_{c20}nk})^T \sim MVN(0, \sum_{ck})$ where $1 \leq n \leq 100$ and $1 \leq k \leq 5$, and sample expression level for the gene $g \sim N(0, \sigma_k^2)$ which is not in any cluster for sample n, where $1 \leq n \leq 100$, $1 \leq k \leq 5$ and σ_k^2 was uniformly distributed from [0.8, 1.2], which indicates different variance for study k.

4. For the first 1,000 genes $(1 \leq g \leq 1,000)$, k_g (the number of studies that are differentially expressed for gene g) was generated by sampling $k_g = 1, 2, 3, 4, 5$, respectively. For the next 1,000 genes $(1,001 \leq g \leq 2,000)$, $k_g = 0$ represents non-DE genes in all five studies.

5. To simulate expression intensities for cases, we randomly sampled $\delta_{gk} \in \{0, 1\}$, such that $\sum_k \delta_{gk} = k_g$. If $\delta_{gk} = 1$, gene g in study k was a DE gene, otherwise it was a non-DE gene. When $\delta_{gk} = 1$, we sampled expression intensities μ_{gk} from a uniform distribution in the range of [0.5, 3], which means we considered the concordance effect (up-regulated) among all simulated studies. Hence, the expression for control samples are $X_{gnk} = X'_{gnk}$, and case samples are $Y_{gnk} = X'_{g(n+50)k} + \mu_{gk} \cdot \delta_{gk}$, for $1 \leq g \leq 2,000$, $1 \leq n \leq 50$ and $1 \leq k \leq 5$

In the simulation study, we had 1,000 non-DE genes in all five studies ($k_g = 0$), and 1,000 genes were differentially expressed in $1 - 5$ studies ($k_g = 1, 2, 3, 4, 5$). On average, we had roughly the same number (~ 200) of genes in each group of $k_g = 1, 2, 3, 4, 5$. See Figure S2 for the heatmap of a simulated example (red represents up-regulated genes). We applied the 12 meta-analysis methods under FDR control at 5%. With the knowledge of true k_g, we were able to derive the sensitivity and specificity for HS_A and HS_B, respectively. In HS_A, genes with $k_g = 5$ were the underlying true positives and genes with $k_g = 0-4$ were the underlying true negatives; in HS_B, genes with $k_g = 1 - 5$ were the underlying true positives and genes with $k_g = 0$ were the true negatives. By adjusting the decision cutoff, the receiver operating characteristic (ROC) curves and the resulting area under the curve (AUC) were used to evaluate the performance. We simulated 50 data sets and reported the means and standard errors of the AUC values. AUC values range between 0 and 1. AUC= 50% represents a random guess and AUC= 1 reaches the perfect prediction. The above simulation scheme only considered the concordance effect sizes (i.e. all with up-regulation when a gene is DE in a study) among five simulated studies. In many applications, some genes may have p-value statistical significance in the meta-analysis but the effect sizes are discordant (i.e. a gene is up-regulation in one study but down-regulation in another study). To investigate that effect, we performed a second simulation that consider random discordant cases. In step 5, the μ_{gk} became a mixture of two uniform distributions: $\pi_{gk} \cdot Unif[-3, -0.5] + (1 - \pi_{gk}) \cdot Unif[0.5, 3]$, where π_{gk} is the probability of gene $g(1 \leq g \leq 2,000)$ in study $k(1 \leq k \leq 5)$ to have a discordant effect size (down-regulated). We set $\pi_{gk} = 0.2$ for the discordant simulation setting.

2.3.2 Simulation results to characterize the methods

The simulation study provided the underlying truth to characterize the meta-analysis methods according to their strengths and weaknesses for detecting DE genes of different hypothesis settings. The performances of 12 methods were evaluated by receiver operating characteristic (ROC) curves, which is a visualization tool that illustrates the sensitivity and specificity trade-off, and the resulting area under the ROC curve (AUC) under two different hypothesis settings of HS_A and HS_B. Table 1 shows the detected number of DE genes under nominal FDR at 5%, the true FDR and AUC values under HS_A and HS_B for all 12 methods. The values were averaged over 50 simulations and the standard errors are shown in the parentheses.

Figure 2 shows the histograms of the true number of DE studies (i.e. k_g) among the detected DE genes under FDR= 5% for each method. It is clearly seen that minP, Fisher, AW, Stouffer and FEM detected HS_B-type DE genes and had high AUC values under HS_B criterion (0.98-0.99), compared to lower AUC values under HS_A criterion (0.79-0.9). For these methods, the true FDR for HS_A generally lost control (0.41-0.44). On the other hand, maxP, rOP and REM had high AUC under HS_A criterion (0.96-0.99) (true FDR = 0.068-0.117) compared to HS_B (0.75-0.92). maxP detected mostly HS_A-type of markers and rOP and REM detected mostly HS_r-type DE genes. PR and SR detected mostly HS_A-type DE genes but they surprisingly had very high AUC under both HS_A and HS_B criteria. The RankProd method detected DE genes between HS_r and HS_B types and had a good AUC value under HS_B. The RankSum detected HS_B-type DE genes but had poor AUC values (0.5) for both HS_A and HS_B. Table 1 includes our concluding characterization of the targeted hypothesis settings for each meta-analysis method (see also Figure S5 of the ROC curve and AUC of HS_A-type and HS_B-type in 12 meta-analysis methods). Figure S3 shows the result for the second discordant simulation setting. The numbers of studies with opposite effect size are represented by different colours in histogram plot (green: all studies with concordance effect size; blue: one study has opposite effect size with the remaining; red: two studies have opposite effect size with the remaining). In summary, almost all meta-analysis methods could not avoid inclusion of genes with opposite effect sizes. Particularly, methods utilizing p-values

from two-sided tests (e.g. Fisher, AW, minP, maxP and rOP) could not distinguish direction of effect sizes. Stouffer was the only method that accommodated the effect size direction in its z-transformation formulation but its ability to avoid DE genes with discordant effect sizes seemed still limited. Owen [2009] proposed a one-sided correction procedure for Fishers method to avoid detection of discordant effect sizes in meta-analysis. The null distribution of the new statistic, however, became difficult to derive. The approach can potentially be extended to other methods and more future research will be needed for this issue.

2.3.3 Results of the four evaluation criteria

Detection capability Figure 3 shows the number of DE genes identified by each of the 12 meta-analysis methods (FDR= 10% for MDD and breast cancer due to their weak signals and FDR= 1% for all the others). Each plot shows mean with error bars of standard error for 50 bootstrapped data sets. Table S4 shows the MSR and ASR for each method in the six examples. The methods in Figure 3 are ordered according to their ASR values. The top six methods with the strongest detection power were those that detected HS_B-type DE genes from the conclusion of Table 1: Fisher, AW, Stouffer, minP, FEM and RankSum. The order of performance of these six methods was pretty consistent across all six examples. The next four methods were rOP, RankProd, maxP and REM and they targeted on either HS_r or HS_A. PR and SR had the weakest detection capability, which was consistent with the simulation result in Table 1.

Biological association Figure 4 shows plots of mean with error bars of standard error from the pathway association p values (minus log-transformed) of the top 100 surrogate disease-related pathways for the 12 methods. Table S5 shows the corresponding MSR and ASR. We found that Stouffer, Fisher and AW had the best performance among the 12 methods. Surprisingly we found that although PR and SR had low detection capability in simulation and real data, they consistently had relatively high biological association results. This may be due to the better DE gene ordering results these two methods provide, as was also shown by the high AUC values under both hypothesis settings in the simulation.

Stability Figure 5 shows the plots of mean with error bars of standard error of stability

calculated by adjusted DE similarity measure. Table S6 contains the corresponding MSR and ASR. In summary, RankProd and RankSum methods were the most stable meta-analysis methods probably because these two nonparametric approaches take into account all possible fold change calculations between cases and controls. They do not need any distributional assumptions, which provided stability even when sample sizes were small [Breitling and Herzyk, 2005]. The maximum p value method consistently had the lowest stability in all data sets, which is somewhat expected. For a given candidate marker with a small maximum p value, the chance that at least one study has significantly inflated p values is high when sample size is reduced by half. The stability measures in the breast cancer example were generally lower than other examples. This is mainly due to the weak signals for survival outcome association, which might be improved if larger sample size is available.

Robustness Figure 6 shows the plots of mean with error bard of standard error of robustness calculated by adjusted DE similarity measure between the original meta-analysis and the new meta-analysis with an added outlier. Table S7 shows the corresponding MSR and ASR values. In general, methods suitable for HS_B (minP, AW, Fisher and Stouffer) have better robustness than methods for HS_A or HS_r (e.g. maxP and rOP). The trend is consistent in the prostate cancer, brain cancer and IPF examples but is more variable in the weak-signal MDD and breast cancer examples. RankSum was surprisingly the most sensitive method to outliers, while RankProd performs not bad.

2.3.4 Characterization of methods by MDS plots

We applied the adjusted DE similarity measure to quantify the similarity of the DE gene orders from any two meta-analysis methods. The resulting dissimilarity measure (i.e. one minus adjusted similarity measure) was used to construct the multidimensional scaling (MDS) plot, showing the similarity/dissimilarity structure between the 12 methods in a two-dimensional space. When two methods were close to each other, they generated similar DE gene ordering. The patterns of MDS plots from six examples generated quite consistent results (Figure S6). Figure 7(a) shows an aggregated MDS plot where the input dissimilarity matrix is averaged from the six examples. We clearly observed that Fisher, AW, Stouffer,

minP, PR and SR were consistently clustered together in all six individual and the aggregated MDS plot (labeled in red). This is not surprising given that these methods all sum transformed p value evidence across studies (except for minP). Two methods to combine effect sizes and two methods to combine ranks (FEM, REM, RankProd and RankSum labeled in blue) are consistently clustered together. Finally, the maxP and rOP methods seem to form a third loose cluster (labeled in green).

2.3.5 Characterization of data sets by entropy measure

From the simulation study, selection of a most suitable meta-analysis method depends on the hypothesis setting behind the methods. The choice of a hypothesis setting mostly depends on the biological purpose of the analysis; that is, whether one aims to detect candidate markers differentially expressed in "all" (HS_A), "most" (HS_r) or "one or more" (HS_B) studies. However, when no biological prior information or preference exists, the entropy measure can be objectively used to determine the choice of hypothesis setting. The analysis identifies the top 1,000 genes from Fisher's meta-analysis method and the gene-specific entropy of each gene is calculated. When the entropy is small, the p values are small in only one or very few studies. Conversely, when the entropy is large, most or all of the studies have small p values. Figure 7(b) shows the box-plots of entropy of the top 1,000 candidate genes identified by Fisher's method in the six data sets. The result shows that prostate cancer comparing primary and metastatic tumor samples had the smallest entropy values, which indicated high heterogeneity across the three studies and that HS_B should be considered in the meta-analysis. On the other hand, MDD had the highest entropy values. Although the signals of each MDD study were very weak, they were rather consistent across studies and application of HS_A or HS_r was adequate. For the other examples, we suggest using the robust HS_r unless other prior biological purpose is indicated.

2.4 CONCLUSIONS AND DISCUSSIONS

2.4.1 An application guideline for practitioners

From the simulation study, the 12 meta-analysis methods were categorized into three hypothesis settings (HS_A, HS_B and HS_r), showing their strengths for detecting different types of DE genes in the meta-analysis (Figure 2 and the second column of Table 2). For example, maxP is categorized to HS_A since it tends to detect only genes that are differentially expressed in all studies. From the results using four evaluation criteria, we summarized the rank of ASR values (i.e. the order used in Figure 3- Figure 6) and calculated the rank sum of each method in Table 2. The methods were then sorted first by the hypothesis setting categories and then by the rank sum. The clusters of methods from the MDS plot were also displayed. For methods in the HS_A category, we surprisingly see that the maxP method performed among the worst in all four evaluation criteria and should be avoided. PR was a better choice in this hypothesis setting although it provides a rather weak detection capability. For HS_B, Fisher, AW and Stouffer performed very well in general. Among these three methods, we note that AW has an additional advantage to provide an adaptive weight index that indicates the subset of studies contributing to the meta-analysis and characterizes the heterogeneity (e.g. adaptive weight (1,0,...) indicates that the marker is DE in study 1 but not in study 2, etc.). As a result, we recommend AW over Fisher and Stouffer in the HS_B category. For HS_r, the result was less conclusive. REM provided better stability and robustness but sacrificed detection capability and biological association. On the other hand, rOP obtained better detection capability and biological association but was neither stable nor robust. In general, since detection capability and biological association are of more importance in the meta-analysis and rOP has the advantage to link the choice of r in HS_r with the rOP method (e.g. when $r = 0.7 \cdot K$, we identify genes that are DE in more than 70% of studies), we recommend rOP over REM.

Below, we provide a general guideline for a practitioner when applying microarray meta-analysis. Data sets of a relevant biological or disease hypothesis are firstly identified, preprocessed and annotated according to Step (i) - (v) in Ramasamy et al. [2008]. Proper quality

assessment should be performed to exclude studies with problematic quality (e.g. with the aid of MetaQC as we did in the six examples). Based on the experimental design and biological objectives of collected data, one should determine whether the meta-analysis aims to identify biomarkers differentially expressed in all studies (HS_A), in one or more studies (HS_B) or in majority of studies (HS_r). In general, if higher heterogeneity is expected from, say, heterogeneous experimental protocol, cohort or tissues, HS_B should be considered. For example, if the combined studies come from different tissues (e.g. the first study uses peripheral blood, the second study uses muscle tissue and so on), tissue-specific markers may be expected and HS_B should be applied. On the contrary, if the collected studies are relatively homogeneous (e.g. use the same array platform or from the same lab), HS_r is generally recommended, as it provides robustness and detects consistent signals across the majority of studies. In the situation that no prior knowledge is available to choose a desired hypothesis setting or if the researcher is interested in a data-driven decision, the entropy measure in Figure 7(b) can be applied and the resulting box-plot can be compared to the six examples in this paper to guide the decision. Once the hypothesis setting is determined, the choice of a meta-analysis method can be selected from the discussion above and Table 2.

2.4.2 Conclusion

In this paper, we performed a comprehensive comparative study to evaluate 12 microarray meta-analysis methods using simulation and six real examples with four evaluation criteria. We clarified three hypothesis settings that were implicitly assumed behind the methods. The evaluaion results produced a practical guideline to inform biologists the best choice of method(s) in real applications.

With the reduced cost of high-throughput experiments, data from microarray, new sequencing techniques and mass spectrometry accumulate rapidly in the public domain. Integration of multiple data sets has become a routine approach to increase statistical power, reduce false positives and provide more robust and validated conclusions. The evaluation in this paper focuses on microarray meta-analysis but the principles and messages apply to other types of genomic meta-analysis (e.g. GWAS, methylation, miRNA and eQTL).

When next-generation sequencing technology becomes more affordable, sequencing data will become more prevalent as well and similar meta-analysis techniques will apply. For these different types of genomic meta-analysis, similar comprehensive evaluation could be performed and application guidelines should be established as well.

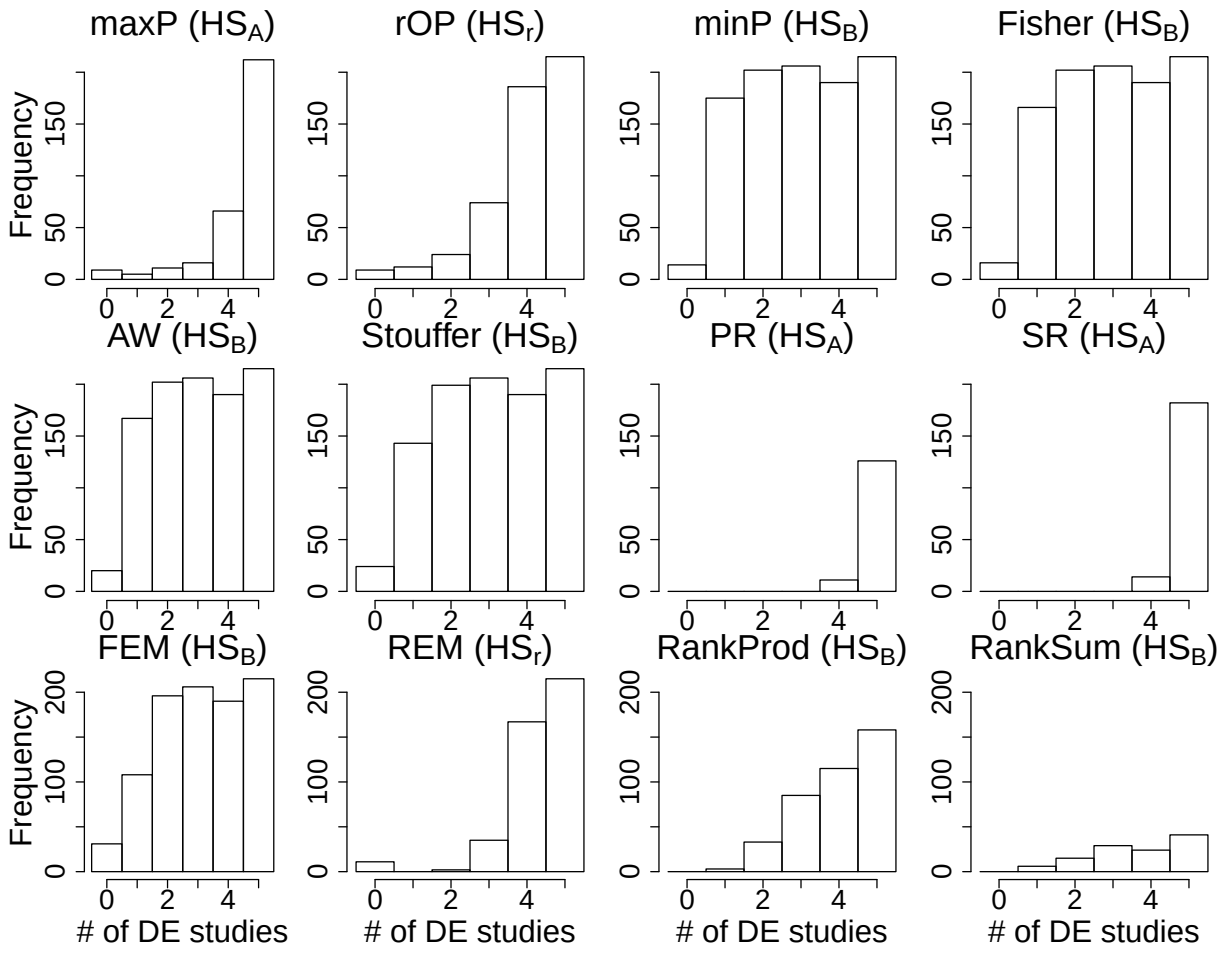

Figure 2: The histograms of the true number of DE studies were detected as DE genes under FDR $= 5\%$ in each method.

Table 1: The detected number of DE genes (at FDR= 5%), the true FDR, AUC values under HS_A and HS_B and the concluding characterization of targeted hypothesis setting of each method.

	maxP	rOP	minP	Fisher	AW	Stouffer
detected #	321	522	1005	1000	1000	974
(se)	(2.2)	(2.35)	(0.85)	(1.06)	(1.05)	(1.5)
True FDR (HS_A)	0.068	0.18	0.447	0.444	0.444	0.43
(se)	(0.008)	(0.012)	(0.0006)	(0.0007)	(0.0008)	(0.0009)
True FDR (HS_B)	0.007	0.011	0.016	0.017	0.016	0.022
(se)	(0.0005)	(0.0004)	(0.0006)	(0.0006)	(0.0007)	(0.0006)
AUC (HS_A)	0.996	0.964	0.8	0.82	0.79	0.89
(se)	(0.0003)	(0.0014)	(0.0005)	(0.0005)	(0.0005)	(0.0006)
AUC (HS_B)	0.75	0.833	0.99	0.99	0.99	0.99
(se)	(0.0013)	(0.01)	(0.0001)	(0.0001)	(0.0001)	(0.0005)
Characterization	HS_A	HS_r	HS_B	HS_B	HS_B	HS_B
	PR	SR	FEM	REM	RankProd	RankSum
detected #	136	186	948	411	391	105
(se)	(2.51)	(2.3)	(1.75)	(2.86)	(3.31)	(1.514)
True FDR (HS_A)	0.008	0.01	0.415	0.117	0.13	0.389
(se)	(0.0003)	(0.0004)	(0.0009)	(0.0015)	(0.0014)	(0.0008)
True FDR (HS_B)	0	0	0.022	0.007	0	0
(se)	(0)	(0)	(0.0007)	(0.0004)	(0)	(0)
AUC (HS_A)	0.986	0.99	0.917	0.99	0.916	0.504
(se)	(0.0003)	(0.0002)	(0.0009)	(0.0002)	(0.0011)	(0.0046)
AUC (HS_B)	0.981	0.95	0.984	0.92	0.934	0.496
(se)	(0.0004)	(0.0008)	(0.0004)	(0.0011)	(0.0012)	(0.0025)
Characterization	HS_A	HS_A	HS_B	HS_r	HS_B	HS_B

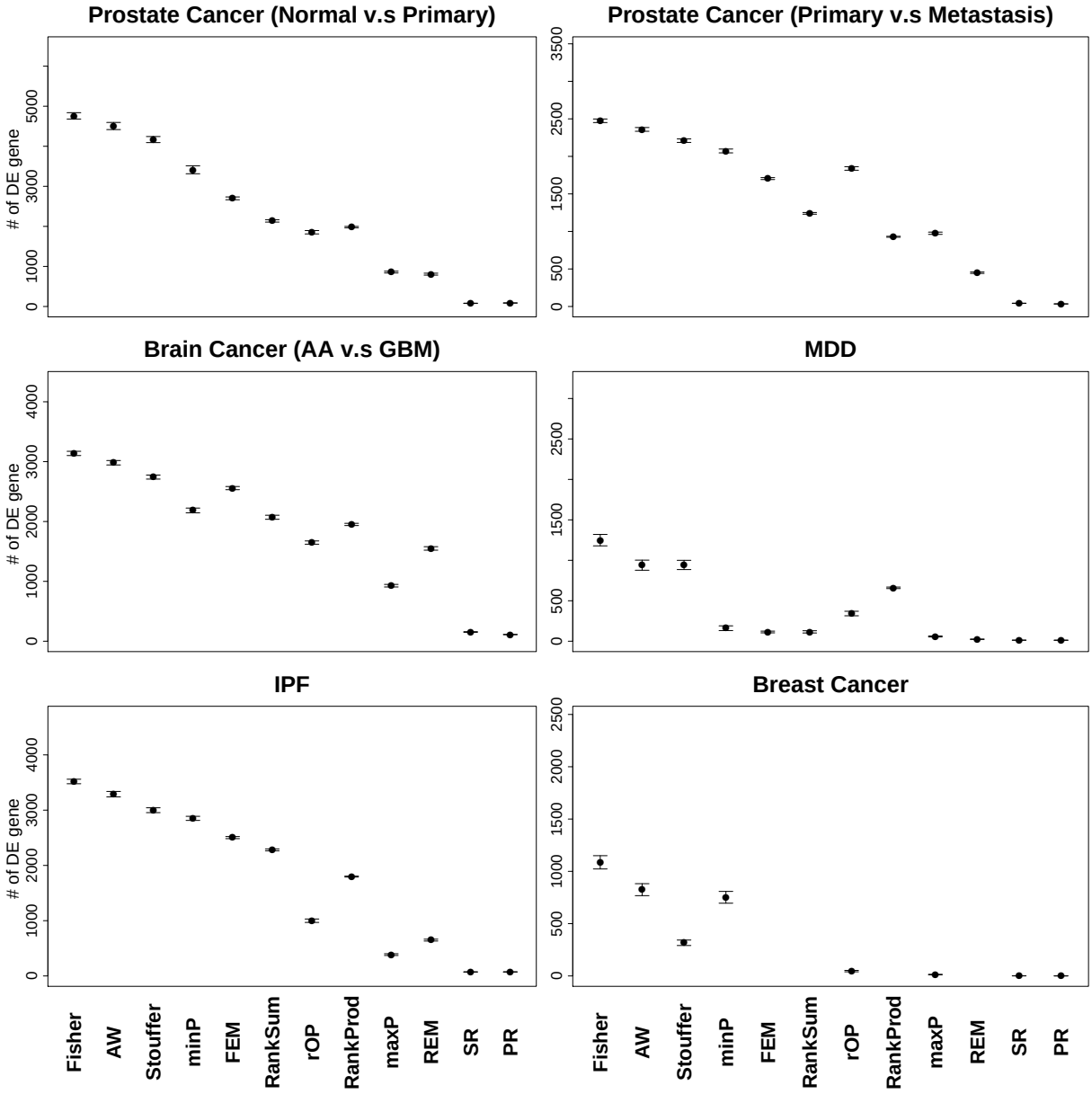

Figure 3: The plot of mean numbers of detected DE genes with error bars of standard error from 50 bootstrapped data sets for the 12 meta-analysis methods. Note that FEM, REM, RankProd and RankSum cannot be applied to survival examples.

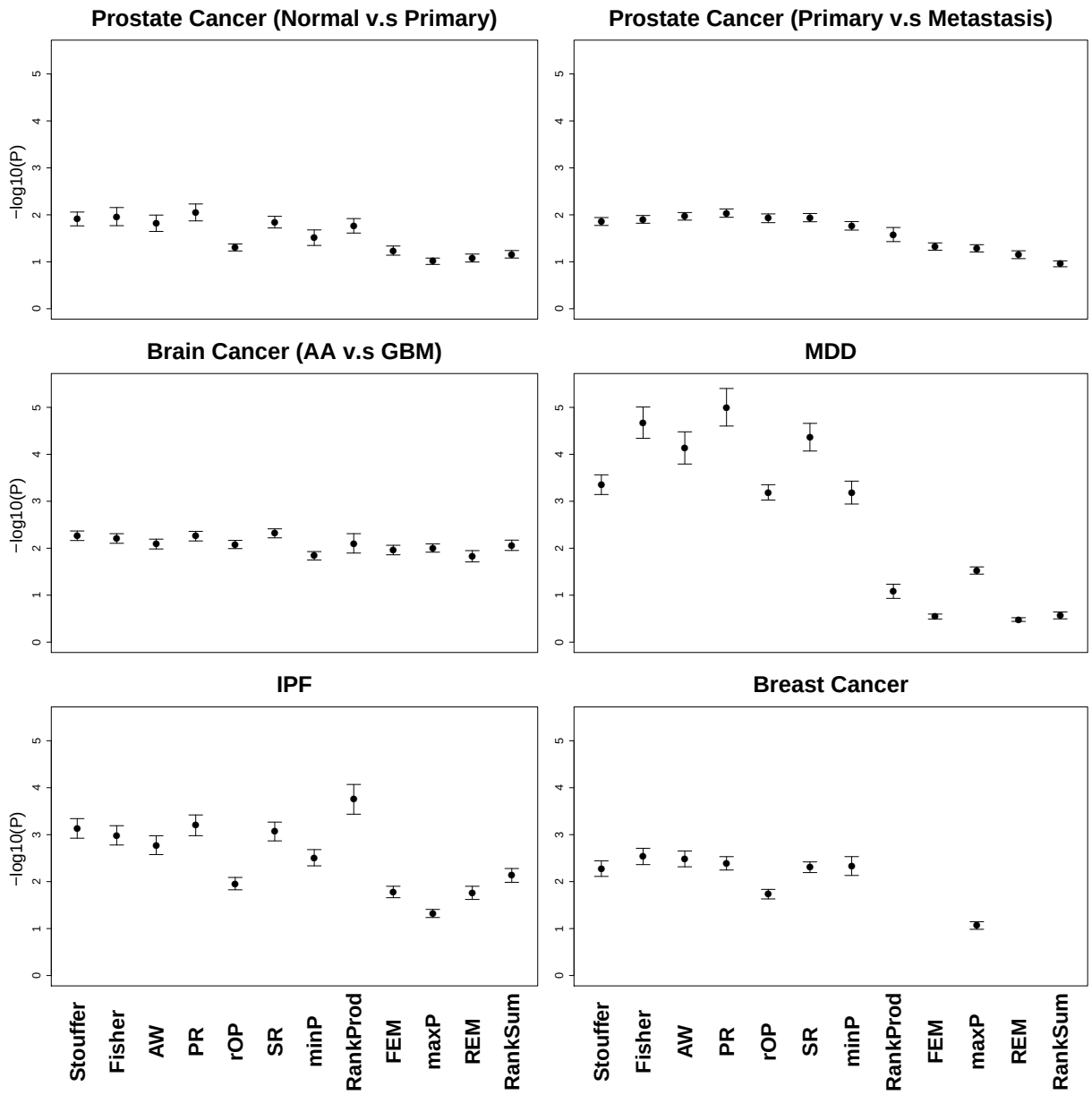

Figure 4: Plot of mean values of $-\log_{10}(p)$ with error bars of standard error from KS-test based on the top 100 surrogate pathways. Note that FEM, REM, RankProd and RankSum cannot be applied to survival examples.

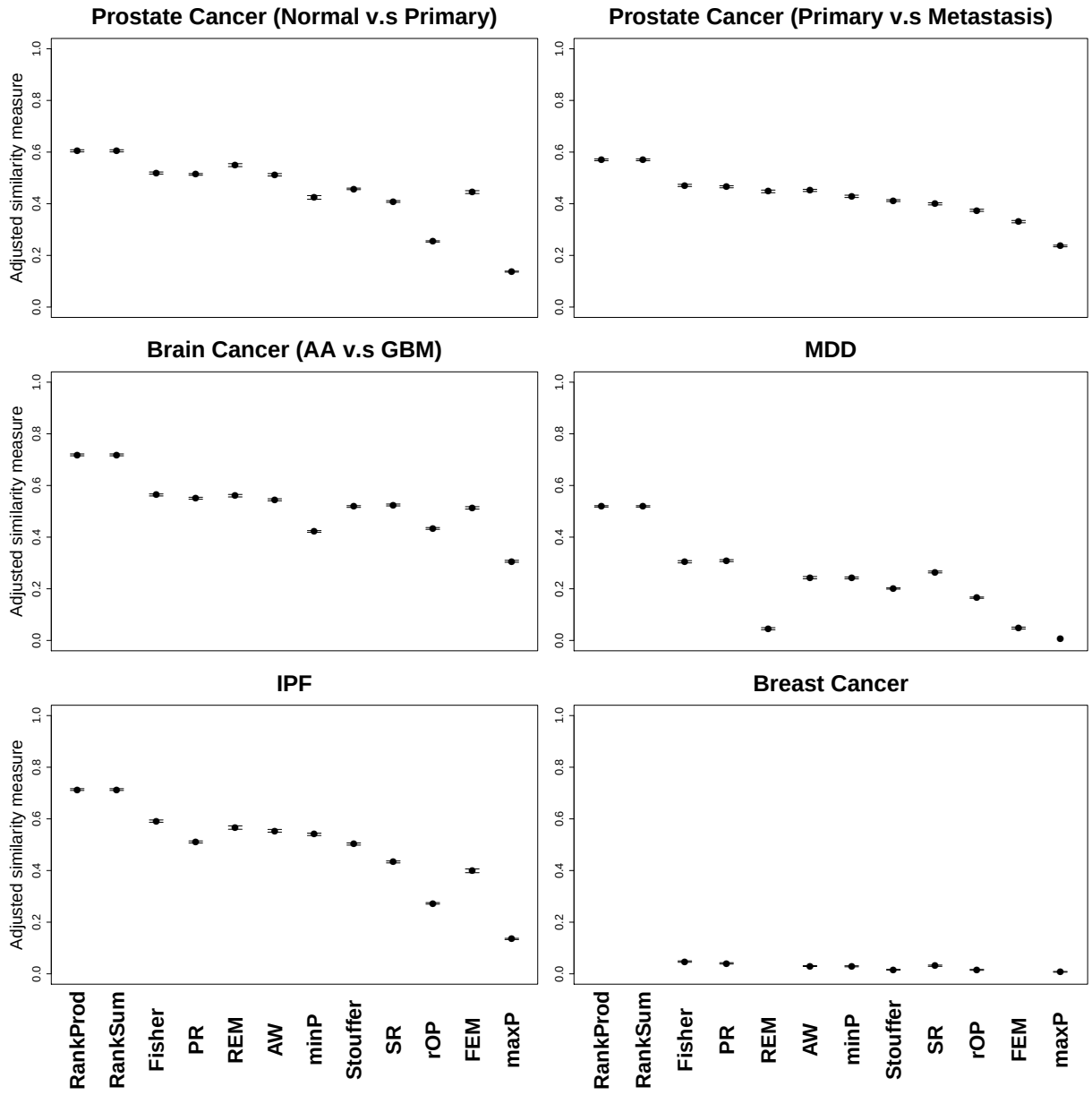

Figure 5: Plot of mean with error bars of standard error of stability in six examples based on the adjusted similarity between DE results of two randomly split data sets. Note that FEM, REM, RankProd and RankSum cannot be applied to survival examples.

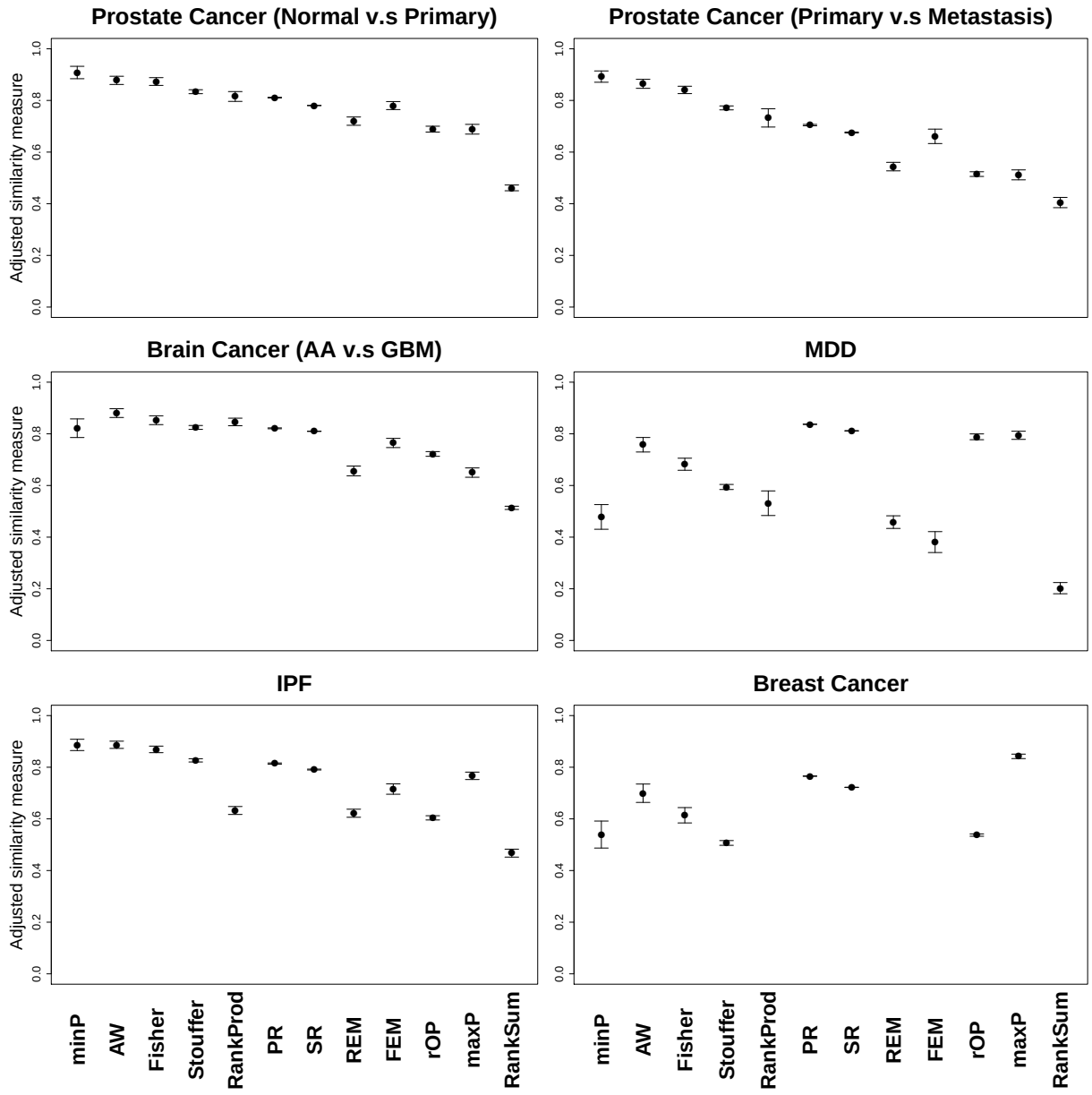

Figure 6: Plots of mean with error bars of standard error of robustness in six examples based on the adjusted similarity between DE results with/without adding one irrelevant noise study. Note that FEM, REM, RankProd and RankSum cannot be applied to survival examples.

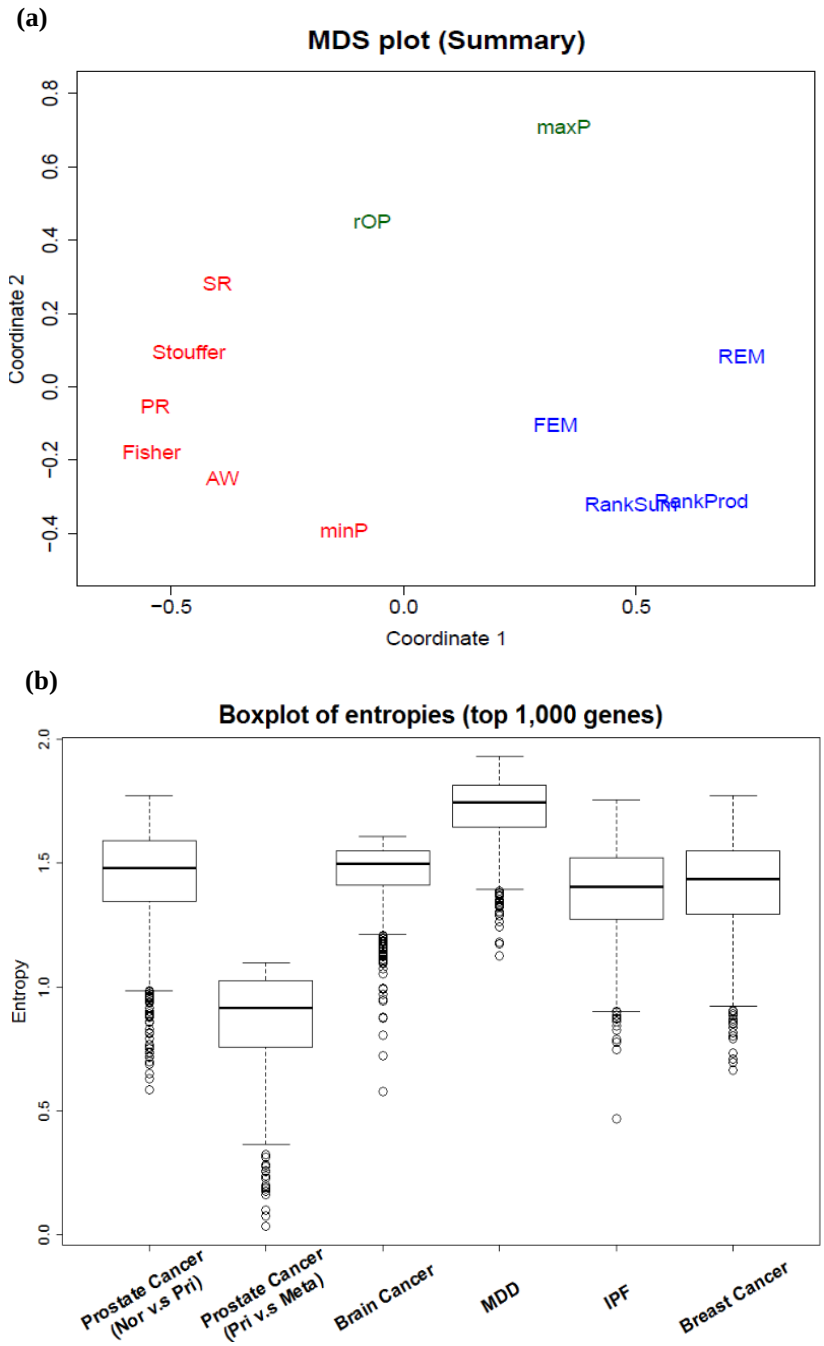

Figure 7: (a) Multi-dimensional scaling (MDS) plot of all 12 methods based on the average dissimilarity matrix of six examples. Colors (red, green and blue) indicate clusters of methods with similar DE detection ordering. (b) The boxplots of entropies in six data sets. High entropies indicate that high consistency of DE gene detection across studies (e.g. MDD). Low entropies show greater heterogeneity in DE gene detection (e.g. prostate cancer).

Table 2: Ranks of method performance in the four evaluation criteria.

	Targeted *HS*	Detection Capability	Biological Association	Stability	Robustness	Rank Sum	MDS[*1]
PR	HS_A	12	4	4	6	26	1
SR	HS_A	11	6	9	7	33	1
maxP	HS_A	9	10	12	11	42	2
rOP	HS_r	7	5	10	10	32	2
REM	HS_r	10	11	5	8	34	3
Fisher	HS_B	1	2	3	3	9	1
AW	HS_B	2	3	6	2	13	1
Stouffer	HS_B	3	1	8	4	16	1
minP	HS_B	4	7	7	1	19	1
RankProd	HS_B	8	8	1	5	22	3
RankSum	HS_B	6	12	2	12	32	3
FEM	HS_B	5	9	11	9	34	3

[*1]: same number means the methods are clustered together in MDS plot

3.0 A CONSERVED BDNF, GLUTAMATE-, GABA-ENRICHED GENE MODULE RELATED TO HUMAN DEPRESSION IDENTIFIED BY GENE COEXPRESSION META-ANALYSIS AND DNA VARIANT GENOME-WIDE ASSOCIATION STUDIES

This paper has been published in BMC bioinformatics Chang et al. [2014].

3.1 INTRODUCTION

Major depressive disorder (MDD) is a common psychiatric disease with an estimated prevalence 3% for a current episode and 5.2% for a lifetime disorder [Hasin et al., 2005],a high rate of recurrence [Mueller et al., 1999], a higher prevalence in women [Weissman et al., 1993], and a heritability of 37% (95% CI = 31% - 42%) [Sullivan et al., 2000]. Transcriptome (the set of all expressed genes in a tissue sample) and genome-wide association studies (GWAS) have separately provided clues to mechanisms of MDD, although not to the anticipated extent. Transcriptome studies mostly focus on changes in gene expression in disease states (altered expression), but also provide unique opportunities for assessing the less-investigated changes in the coordinated function of multiple genes (altered coexpression) [Gaiteri et al., 2013], 2013. GWAS seek to identify genetic markers for diseases, and have generated some findings in MDD [Rietschel et al., 2010; Shi et al., 2010; Muglia et al., 2008; Lewis et al., 2010; Shyn et al., 2009], but overall results from GWAS meta-analyses have been disappointing [Ripke et al., 2012; Hek et al., 2013], potentially due to complexity of the disease and heterogeneity of patient cohorts. GWAS and transcriptome studies are highly complementary in that they provide unbiased and large scale investigation of DNA structural (single nucleotide

polymorphisms (SNP) and other variants) and functional (RNA expression) changes across conditions, although these two approaches are only beginning to be integrated [Kupfer et al., 2011; Cristino et al., 2013].

Gene arrays allow for the unbiased quantification of expression (mRNA transcript levels) for 10,000 to 20,000 genes simultaneously. Since gene transcript levels represent the integrated output of many regulatory pathways, the study of all expressed genes provides an indirect snapshot of cellular function under diverse conditions. For instance, using postmortem brain samples, this approach has implicated dysregulated BDNF, GABA, glutamate and oligodendrocyte functions in MDD [Tripp et al., 2012; Klempan et al., 2007; Sequeira et al., 2009; Choudary et al., 2005; Guilloux et al., 2011]. However, current studies are still few, were performed in heterogeneous cohorts, and utilized early and rudimentary versions of gene arrays. Moreover, gene array studies are subject to similar limitations as early GWA studies, in that large number of genes are tested in few subjects (n=10-100). Typical analyses identify 1-10% of genes affected in the illness (differentially expressed genes), are characterized by high rates of false discovery, and may be confounded by numerous clinical (drug exposure, subtypes, duration, etc.), demographic (age, sex, race), technical parameters (RNA integrity, brain pH, postmortem interval for brain collection), or other potential co-segregating factors of unknown origin (See Kupfer et al., 2011 for discussion). Conditions of postmortem brain collection also preclude the reliable identification of acute state-dependent gene changes, but are appropriate for investigating stable long-term disease-related homeostatic adaptations.

Gene coexpression studies offer complementary perspectives on gene changes in the context of transcriptome studies. Here, two genes are defined as coexpressed in a dataset if their patterns of expression are correlated across samples. Coexpression has been shown to reflect possible shared function between these genes, and may arise through multiple biological pathways including cellular coexpression and common regulatory pathways (e.g., hormone signaling, transcription factors) [Lee et al., 2004; Gaiteri et al., 2010]. Hence, coexpression links have been used to build gene networks, and to identify communities, or modules, of genes with shared functions [Dobrin et al., 2009; Elo et al., 2007]. Notably, by incorporating multiple interactions among large number of genes, the study of gene coexpression networks

provides an approach to tackle the complexity of biological changes occurring in complex polygenic disorders [Gaiteri et al., 2010]. See Gaiteri et al., 2013 for a general review.

Concepts and methods for integrating functional (transcriptome) and structural (DNA polymorphism GWA) studies of the molecular bases of complex neuropsychiatric disorders such as MDD need to be developed to harness the potential of systematic large-scale molecular and genetic investigations of the brain. Here, our central hypothesis states that stable brain co-regulation modules identified through meta-analysis of multiple transcriptome studies may overlap with sets of genes and associated variants (SNPs) related to MDD. Based on the continuum of pathological changes between MDD and other brain disorders [Sibille and French, 2013] and co-morbidity with selected medical illnesses including cardiovascular diseases and metabolic syndrome [Pan et al., 2012; Musselman et al., 1998], we also predicted that MDD coexpression modules may be enriched in genes identified by GWAS for other psychiatric and brain disorders and potentially for medical illnesses related to depression, together identifying functionally-coherent gene sets implicated in MDD-related disease processes.

3.2 MATERIALS AND METHODS

Figure 8 illustrates the meta-clustering and validation methods of the approach. In step I, we identified 50 robust co-regulation modules in human brains by combining 11 transcriptome datasets collected from several brain regions in different cohorts of subjects with MDD and non-affected comparison subjects. Steps II and III were performed to identify MDD-related gene modules, and exclude other gene modules linked to biological functions not related to MDD. In step II, we collected different sets of genes located nearby SNPs identified by GWAS for MDD, neuropsychiatric disorders, related traits, and for systemic diseases often associated with psychiatric disorders, and perform gene set analysis to identify MDD-related gene module(s). In step III, we performed functional annotations of gene module members by using 2,334 gene sets collected from MSigDB (http://www.broadinstitute.org/gsea/msigdb/). We also organized genes identified by SNPs in published GWAS into three categories (cancer

studies, human body indices and unrelated diseases) and treated them as a non-MDD-related negative control gene sets in step IV.

3.2.1 Transcriptome data sets

Eleven MDD microarray datasets generated in our lab were used here. Cohorts and brain areas investigated are listed in Table 3 and details were provided in [Wang et al., 2012b; Sibille et al., 2004]. Among these studies, six used Affymetrix Human Genome U133 Plus 2.0 platforms (Affymetrix Inc., Santa Clara, CA), two used Affymetrix Human Genome U133A platforms, and the remaining three used Human HT-12 arrays from Illumina (Illumina Inc, San Diego, CA). For gene matching across studies, when multiple probes or probe sets match to one gene symbol, we choose the probe set with the largest variation (largest interquartile range; IQR) to represent the gene [Gentleman et al., 2005]. For preprocessing, data were log-transformed (base 2). Non-expressed (small mean intensity) and non-informative (small standard deviation) genes were filtered out. To perform such filtering for 11 studies simultaneously, we calculated the ranks of row means and row standard deviations of each gene in each single study. The ranks were summed up across 11 studies and used as criteria to filter out non-expressed and non-informative genes. Figure S8 provides a diagram and results of the transcriptome dataset preprocessing procedures.

3.2.2 Meta-clustering of transcriptomic data to construct co-expression gene modules

The 11 transcriptome studies were combined to construct co-expression gene modules using a meta-clustering technique described below. Denote by X_{gsk} the gene expression intensity of gene g, sample s and study k, and $X_{gk} = (X_{g1k}, \ldots, X_{gSk})$ the vector of gene expression intensities of gene g and study k. Define the dissimilarity measure between gene i and gene j for a given study k as $d_{i,j}^{(k)} = 1 - |\text{cor}(X_{ik}, X_{jk})|$, where $\text{cor}(X_{ik}, X_{jk})$ is the Pearson correlation of the two gene vectors. To combine the dissimilarity information of the $K = 11$ studies, we took mean of meta-dissimilarity measure between gene i and gene j as $d(g_i, g_j) = \text{Mean}(d_{ij}^{(1)}, d_{ij}^{(2)}, \ldots, d_{ij}^{(K)})$. Given the meta-dissimilarity measure, the Penalized K-medoids

clustering algorithm was then applied to construct co-expression gene modules [Tseng, 2007]. The target function to be minimized by Penalized K-medoids is shown below:

$$L(C) = \sum_{i=1}^{G} \sum_{g_i \in C_h} d(g_i, \overline{g}_h) + \lambda \cdot |S| \tag{3.1}$$

where the clustering result $C = (C_1, \ldots, C_H, S)$ contains H non-overlapping gene clusters (i.e. H gene modules $C_1, \ldots, C_H$) and a set of scattered genes S that cannot be clustered into any of the tight gene modules, $\overline{g}_h$ denotes the medoid gene of cluster h such that its average dissimilarity to all other genes in the cluster is minimal, $|S|$ is the size of the scattered gene set S and λ is a tuning parameter controlling tightness of detected gene modules and the number of scattered genes discarded to S. The first term of the target function $L(C)$ calculates the total sum of within-cluster dispersion and is essentially the K-medoids algorithm (an extended form of K-means using arbitrary non-Euclidean dissimilarity measure). The second penalty term allows scattered genes not to be clustered into any gene module. For example, if the distances of a gene gi to all cluster medoids are greater than λ, minimizing $L(C)$ will assign the gene into the scattered gene set S, instead of into any gene cluster. Intuitively, smaller λ generates tighter clusters and allow more genes into scattered gene set S. The rationale for the choice of this approach was based on finding in the literature, where comparative studies show that many genes are not tightly co-expressed with any gene clusters and methods that allow scattered gene assignment generates tighter gene modules that are biologically more informative [Thalamuthu et al., 2006].

3.2.3 Parameter selection and evaluation of meta-clustering

We tested different parameter settings of $H = 50$ or 100 modules, and λ such that $\beta = 0\%, 25\%$ or 50% of genes are left to scattered gene set S. In all performance of the $2 \times 3 = 6$ combinations for the meta-clustering method, a biological validation was performed using biological pathway information. We searched ten keywords ("GABA", "Insulin", "Diabetes", "Immune", "Thyroid", "Estrogen", "Depression", "Alzheimer", "Parkinson" and "Huntington") in MSigDB and finally obtained 98 MDD-related pathways. In each clustering result, Fishers exact test was applied to each module to correlate with each of the 98 MDD-related

pathways and eight GWAS gene lists and the p-values were generated. Wilcoxon signed rank test was used to compare any pair of clustering results (from different parameter setting) so that the best parameter setting could be determined.

3.2.4 Evaluation of robustness and stability of meta-clustering method

To evaluate the robustness of the meta-clustering results, we used the Adjusted Rand Index (ARI) as a measurement of consistency between two clustering results [Hubert and Arabie, 1985]. We randomly selected a subset of studies from 11 MDD studies and calculated the ARI to assess the similarity of the obtained modules compared to those obtained using the 11 MDD studies. The procedure was repeated 100 times and the averaged ARI was calculated. For the stability of meta-clustering method, the mean and standard deviation of ARIs were obtained by bootstrapping method [Efron, 1979], where the 11 MDD studies were bootstrapped 100 times.

3.2.5 Genome-wide association studies (GWAS)-related gene categories

Eight neuropsychiatry-related candidate gene lists and three gene lists from presumably unrelated disorders or traits were identified from relevant GWAS. Individual genes were identified by the presence of GWAS significant SNPs within a given nucleotide distance from the coding region of that gene.

I. The first gene list was obtained from a published GWAS for neuroticism [van den Oord et al., 2008]. Neuroticism is a personality trait that reflects a tendency toward negative mood states, and that is linked to several internalizing psychiatric conditions. That GWAS involved 1,227 healthy individuals with self-report of no diagnosis of or treatment for schizophrenia, schizoaffective disorder or bipolar disorder and personality measures of neuroticism. In van den Oord et al. [2008],Genotyped data were generated from Affymetrix GeneChip Human Mapping 500K using BRLMM algorithm. 449 SNPs were selected by p value less than 0.001, and 155 genes were identified to have contained one or more selected SNPs in the 10 kilobases (kb) up- and down-stream extension of the coding regions.

II. The second gene list was obtained from the MDD 2000+ project that included a meta-analysis of MDD studies with 2,431 MDD cases and 3,673 controls [Wray et al., 2010]. Similarly, 532 SNPs with p value less than 0.001 were mapped to gene coding regions (including 10kb upstream and downstream regions) and 159 genes were identified.

III. The third gene list was obtained from a mega-analysis of GWAS for MDD [Ripke et al., 2012]. The associated 202 SNPs' p values were less than 10^{-5} and 52 genes were identified using the University of California Santa Cruz Human Genome Browser, hg18 assembly (UCSC hg18) with build 36.3. Gene symbols from the build version 36.3 in the National Center for Biotechnology Information (NCBI) database were used.

IV. The fourth candidate gene list was obtained from a mega-GWAS of bipolar disease which contained 7,481 patients and 9,250 controls [Sklar et al., 2011]. 6,887 SNPs were identified when p value less than 0.001. By mapping the SNPs to gene coding region using SNPnexus software (http://snp-nexus.org/), 602 genes were obtained.

V. For the fifth to eighth gene lists, we interrogated the Catalog of Published Genome-Wide Association Studies [Hindorff et al., 2009] (http://www.genome.gov/gwastudies/). The database (as of 01/31/13; time of the latest data analysis update) contained 10,183 entries of disease- or trait-associated SNPs with p values smaller than 10^{-5} in 1,491 GWAS studies. We manually regrouped the disorders and traits into 4 categories: (1) all MDD-related studies, (2) all neuropsychiatric disorder studies, (3) all neurological disorder and brain phenotypes studies, (4) all medical illnesses sharing increased risk with MDD. Note that list #3 was included in list #2 and list #2 was included in list #1. Lists #4 is independent and non-overlapping with others. The associated four gene lists were then compiled, and genes were uniquely included when the mapped SNP was within the gene region including a 100 kb upstream and downstream.

VI. As negative controls, we identified in the catalog of published GWAS three gene sets presumably not related to psychiatric diseases: (a) 65 publications (270 genes) of cancer GWAS studies; (b) 42 publications (459 genes) of human body indices GWAS studies (HBI: genetic phenotypes for human, for example: height, weight, eye color, etc.); and (c) 33 publications (187 genes) of GWAS studies for common disease traits not related to brain function or major mental illnesses.

3.2.6 Meta-analysis to aggregate evidence of association of each module with the GWAS gene lists

We performed Fisher's exact test to examine the significance of the association of genes within each co-expresion module with individual GWAS-derived gene lists, using the 10,000 genes evaluated in transcriptome meta-analysis (Figure S8) as background. To assess statistical significance of association of each identified module from meta-clustering method, we applied the Stouffer's method to combine the p values obtained from Fisher's exact test of the association between gene modules and eight GWAS gene sets. The Stouffer's statistics $T_{\textbf{Stouffer}} = \frac{\Sigma_{i=1}^{k} \phi^{-1}(P_i)}{\sqrt{k}}$ where ϕ the cumulative distribution function of a standard normal distribution [Stouffer, 1949]. The p values were assessed for each of the 50 modules by conventional permutation analysis (B=500).

3.2.7 Pathway analysis and enrichment analysis of GWAS gene lists

For biological association, 2,334 annotated pathways (gene sets) were obtained from MSigDB (www.broadinstitute.org/gsea/msigdb/), which consists of 880 canonical pathways (217 Biocarta gene sets, 180 KEGG gene sets, 430 Reactome gene sets and 53 other gene sets) and 1,454 pathways from Gene Ontology (GO). For each of the gene module, gene set (pathway) analysis was performed for the 2,334 pathways and 11 GWAS gene lists (including 3 negative controls). Fisher's exact test was performed to assess the biological association between gene modules and given gene sets. To account for multiple comparisons, Benjamini and Hochberg procedure was used to control the false discovery rate (FDR).

3.3 RESULTS

3.3.1 Data preprocessing and parameter determination

16,443 genes were retained after gene matching across the 11 studies. Cohorts 10 and 11 were from older platforms with fewer probesets representing only 12,703 genes (Figure S8).

In order to minimize the loss of information from gene matching, we allowed 20% missing values during matching, i.e., we kept genes with at least 9 existing measurements out of 11 studies. 13,500 genes were retained after filtering out lower sum rankings of median row means, and 10,000 genes after further filtering out lower sum rankings of median row standard deviations. We then tested different parameter settings for the number of modules ($H = 50$ or 100), and genes (tuned the λ values for controlling tightness of detected gene modules and the number of scattered genes set) for β =0%, 25% or 50% of genes left out of the gene set S. In all tests of the Penalized K-medoids meta-clustering method ($2 \times 3 = 6$ combinations), we performed a validation by biological pathway information content. For all clustering results, Fisher's exact test was applied to each module to correlate with each of the 98 MDD pathways and eight GWAS gene lists described in the methods, and p values were generated. The Wilcoxon signed rank test was used to compare any pair of clustering results (from different parameter settings) so that the best parameter setting could be determined. The result shows that there was no significant difference (by Wilcoxon signed rank test) between $H = 50$ and $H = 100$ cluster except $\beta = 0\%$ (i.e., keep all genes), and the minimum p value of gene set analysis in $H = 50$ was always lower than that in $H = 100$ in $\beta = 25\%$ and $\beta = 50\%$. It is reasonable to set the noise level in clustering method because noise will increase if we combined more studies. We chose $H = 50$ because the mean of the $-\log 10(p)$ in 50 modules (3.2793) was higher than 100 modules (3.0224) in $\beta = 25\%$, and the mean of the $-\log 10(p)$ in 50 modules (3.1896) was higher than 100 modules (3.0588) in $\beta = 50\%$. 50 modules also provide adequate number and sizes of gene modules for the purpose of further analyses. Given $H = 50$, we compared the performance with different choices of β. $\beta = 25\%$ performed better than $\beta = 0\%$ (p= 0.0004 using Wilcoxon signed rank test), and there was no significant difference between $\beta = 25\%$ and $\beta = 50\%$ (p= 0.0856). Finally, we selected $H = 50$ and tuning parameter λ such that $\beta = 25\%$ genes are left to scattered gene set S throughout this paper.

3.3.2 Construction of 50 meta-modules from 11 MDD studies

Using the parameters determined above, we performed a meta-analysis of module gene membership to identify the top 50 conserved meta-modules across 11 MDD transcriptome studies. A total of 10,000 genes were clustered using the Penalized K-Medoid method. 7,797 genes were clustered into $K = 50$ modules and 2,203 genes ($\beta =\sim 25\%$) were determined as scattered genes with no conserved expression pattern. We performed subsampling and bootstrap methods to assess the stability of the resulting clusters. Subsets ($n = 8$, 9 or 10) of the 11 studies were randomly selected and the meta-clustering procedure was similarly applied. The resulting meta-modules were compared with the meta-modules obtained using the 11 MDD studies using adjusted Rand index (ARI= 0.47, 0.52 and 0.63 for $n = 8$, 9, 10). We also generated bootstrapped samples in each study and repeated the meta-clustering procedures. Comparison of meta-modules generated from bootstrapped samples with original samples generated an average ARI= 0.45 (standard deviation 0.025) in 100 repeated bootstrapping simulations. In the literature, an ARI of $\sim$0.5 is interpreted as reproducible clustering result [Thalamuthu et al., 2006], hence demonstrating good stability under data perturbation (subsampling and bootstrapping) for the 50 meta-modules obtained by combining 11 studies.

3.3.3 Association of meta-modules with eleven GWAS-determined gene lists

We examined association of the 50 meta-modules with the eight GWAS gene lists using Fisher's exact test. The results are shown in Supplementary Table S8. Module #35 is found to have significant associations (p<0.05) with the six psychiatric disorder related GWAS gene sets (p= 0.03 for the neuroticism GWAS gene set; p= 0.03 for MDD 2000+ project; p=0.0001 for Mega-GWAS MDD; p=0.03 for Mega-GWAS of bipolar disorder; p=0.008 for the catalog of GWAS studies of neuropsychiatric disorder; p=0.03 for the catalog of GWAS studies of neurological disorders and brain phenotypes) and two studies with borderline p values (p=0.05 for the catalog of MDD-related GWAS studies; p=0.05 for the catalog of GWAS studies of Medical illnesses sharing clinical risk with MDD). We combined the p values of the eight psychiatric disorder related GWAS gene sets by Stouffer meta-analysis method. The p value of module #35 is 4×10^{-5} after the permutation test. In contrast, there

was no association with cancer (p=1.00), human body indices (p=0.18) and other control diseases (p=0.46) GWAS gene sets. Figure 9(a) shows the heatmaps of log-transformed p values from pathway analysis for the 50 modules obtained from MDD cases and controls combined analysis and 50 modules obtained from controls only analysis. It shows that module #35 (highlighted in green) from the combined cases and controls analysis is enriched in genes contained in six MDD-related GWAS gene sets, but not enriched in the three negative control GWAS gene sets. None of the other 49 modules showed such consistent pattern.

3.3.4 Pathway analysis of meta-module #35

Many GWAS-hit genes (overlapping genes between 88 genes in module #35 and 8 GWAS lists) were related to synaptic function, signal transduction, and neuronal development and morphogenesis (Table 4). Of specific interest, and consistent with current hypotheses for the molecular pathology of MDD, was the inclusion of brain-derived neurotrophic factor (*BDNF*) and other factors implicated in development and maintenance of cell circuits (Ephrin receptors *EPHA3* and *EPHA5*; Netrin G1 (*NTNG1*); *SLITRK3* and *SLITRK5*), of GABA-related genes (*GABBR2*, *GABRA4* and *CALB1*), glutamate receptors (*GRM1* and *GRM7*) and other signaling neuropeptides previously implicated in mechanisms of psychiatric disorders [reelin (*RELN*) and gastrin-releasing peptide (*GRP*)]. Together, these results suggest that module #35 may include multiple components of functionally-relevant local cell circuits (Table 5)

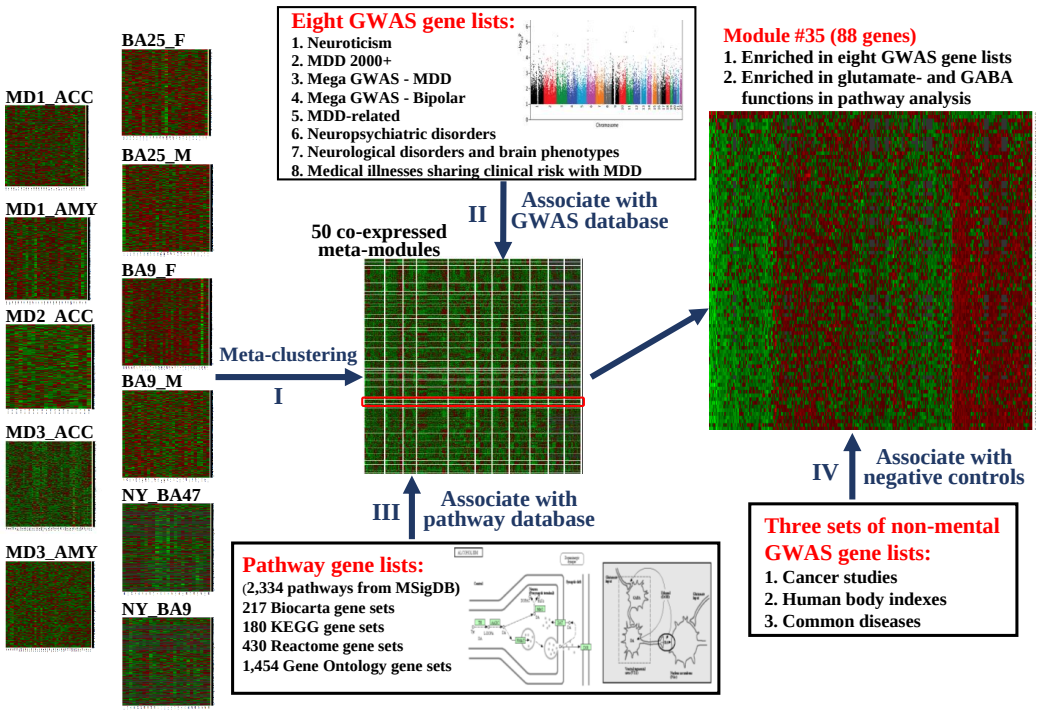

Figure 8: Overall analytical strategy.

In step I, 50 co-regulation modules were generated using meta-clustering of gene clusters identified by the penalized K-medoids method across 11 transcriptome MDD and matched controls studies. In step II modules enriched from most of selected GWAS studies related to MDD, neuropsychiatric disorder and traits, including systemic disease linked to psychiatric disorders were identified. In step III, the biological functions represented by genes included in each module were defined by pathway analysis from 2,334 gene sets of MSigDB (www.broadinstitute.org/gsea/msigdb). In step IV, SNPs from the Catalog of GWAS were organized into three categories: cancer GWAS, human body indices GWAS and GWAS for common diseases and medial illnesses unrelated to MDD or other brain function. Three additional categories were defined as non-MDD-related negative control gene sets. (Note: In order to have better performance of heatmap in module #35, we first performed the hierarchical clustering with complete agglomeration method to aggregated samples with similar expression among all 88 genes, and the genes were sorted by the correlation from high to low of selected gene in the top)

Table 3: Description of cohorts in 11 MDD microarray platforms

Cohort	Region	Code	Platform	# of probes	# of genes	# of subjects
1	ACC	MD1_ACC	Affymetrix Human Genome U133 Plus 2.0	40,610	19,466	32
2	AMY	MD1_AMY	Affymetrix Human Genome U133 Plus 2.0	40,610	19,621	28
3	ACC	MD2_ACC	Illumina HumanHT 12 (v3)	48,803	25,159	20
4	ACC	MD3_ACC	Illumina HumanHT 12 (v3)	48,803	25,159	50
5	AMY	MD3_AMY	Affymetrix HumanHT 12 (v3)	48,803	25,159	42
6	ACC	BA25_F	Affymetrix Human Genome U133 Plus 2.0	53,596	19,572	26
7	ACC	BA25_M	Affymetrix Human Genome U133 Plus 2.0	53,596	19,572	26
8	DLPFC	BA9_F	Affymetrix Human Genome U133 Plus 2.0	53,596	19,572	32
9	DLPFC	BA9_M	Affymetrix Human Genome U133 Plus 2.0	53,596	19,572	28
10	OFC	NY_BA47	Affymetrix Human Genome U133A	20,338	12,703	24
11	DLPFC	NY_BA9	Affymetrix Human Genome U133A	20,338	12,703	26

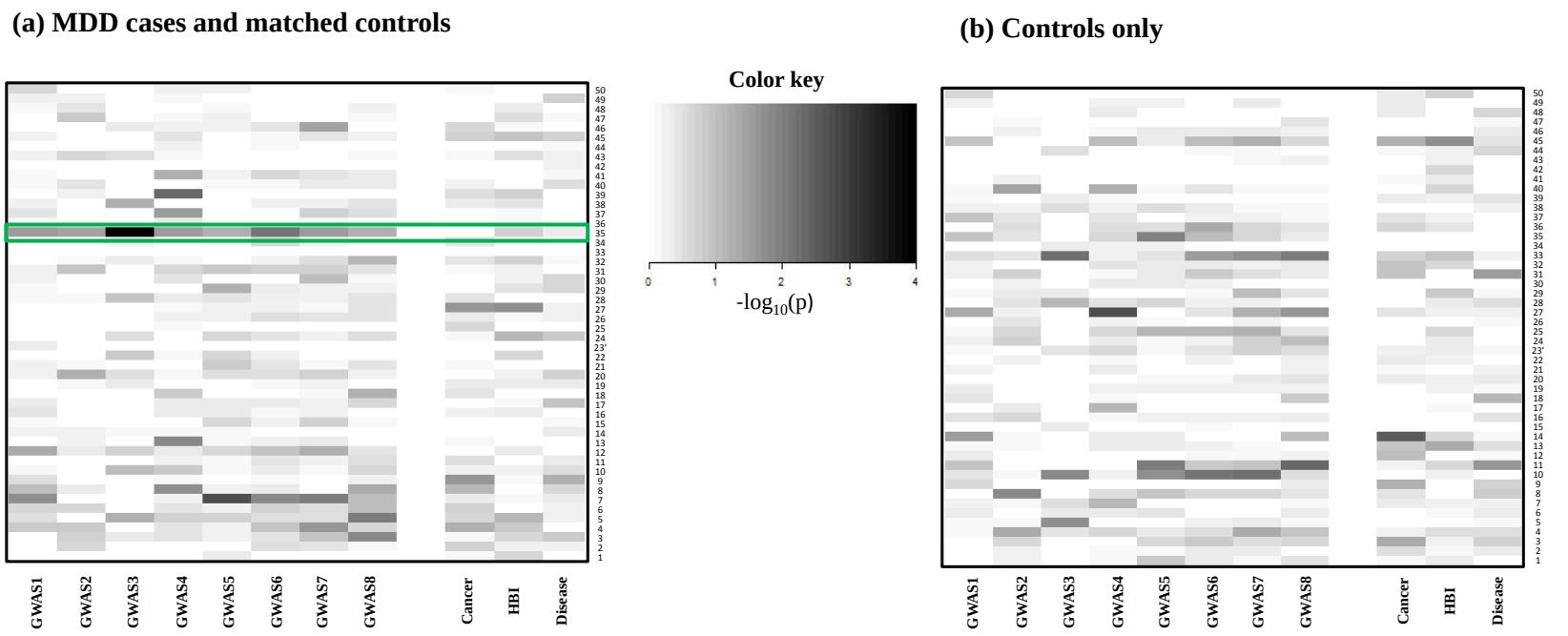

Figure 9: Consistent association of genes in module #35 with MDD-related gene categories.

(a) Heatmap of $\log_{10}$-transformed p values from Fisher's exact test for 50 modules obtained from MDD cases and matched controls and 8 MDD related GWAS and 3 negative controls. (b) Heatmap of $\log_{10}$-transformed p values from Fisher's exact test for 50 modules obtained from controls and 8 MDD related GWAS and 3 negative controls. The green rectangle identifies module #35.

Table 4: Functional groups of 88 genes in module #35

Functional groups	Gene Symbols
Transmembrane cellular localization	*CLSTN2, SYT4, LRRC8B, GPR6, TMEM158 ST8SIA3, GABBR2, NRN1, ST6GALNAC5 GLT8D2, MPPE1, GNPTAB, PVRL3, SLC35B4 SLC35F3, KCNG3, SLC30A9, PTGER4, CYP46A1 GABRA4, UST, LOC646627, NTNG1, TMEM200A TMEM70, RFTN1, GRM1, TMEM132D, KCNV1 EPHA3, CDH12, EPHA5, BEAN, SLITRK3 FREM3, GRM7, CD82, SLITRK5, VLDLR*
Neuronal development and morphogenesis	*BDNF, SLITRK3, RPGRIP1L, MAEL, NTNG1, RELN, LAMB1, SLITRK5, MYCBP2d*
GABA and glutamate	*GRM1, GRM7, GABBR2, GABRA4*
Cell adhesion	*PPFIA2, CDH12, FREM3, CLSTN2, PVRL3, RELN, LAMB1*
Transcription regulation	*EGR3, DACH1, HDAC9, ATOH7, SLC30A9 ATF7IP2, ZNF436, MYCBP2*

Table 5: Top 15 enriched pathways in module #35

Pathways	P value
METABOTROPIC_GLUTAMATE_GABA_B_LIKE_RECEPTOR_ACTIVITY	0.0003
REACTOME_CLASS_C3_METABOTROPIC_GLUTAMATE_PHEROMONE_RECEPTORS	0.0005
G_PROTEIN_SIGNALING_COUPLED_TO_CAMP_NUCLEOTIDE_SECOND_MESSENGER	0.002
CAMP_MEDIATED_SIGNALING	0.002
GLUTAMATE_RECEPTOR_ACTIVITY	0.003
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.003
G_PROTEIN_SIGNALING_COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	0.008
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	0.01
NEUROPEPTIDE_HORMONE_ACTIVITY	0.015
REACTOME_GPCR_LIGAND_BINDING	0.02
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.03
G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	0.03
SECOND_MESSENGER_MEDIATED_SIGNALING	0.04
HORMONE_ACTIVITY	0.04
REACTOME_EICOSANOID_LIGAND_BINDING_RECEPTORS	0.04

3.3.5 Control studies

To demonstrate the improvement of meta-clustering versus single study clustering, we compared the histograms of p values obtained under those different conditions. In Figure 10 , the histogram of the minus log-transformed p values of the Stouffer statistic was first plotted for the 50 meta-modules obtained from the case and control combined analysis. Module #35 with 88 genes is shown to have an aggregated minus log-transformed p value at 4.4 (i.e. p= 4×10^{-5}). We then applied the penalized K-medoid method with the same parameter setting ($K = 50$ clusters and 25% of scattered genes) for each single study. The 11 single study histograms of Stouffer p values showed overall much weaker statistical significance than for module #35. Particularly, none of the 550 modules from 11 single study cluster analysis was enriched (p value threshold 0.05) in more than three GWAS results (Figure 10). Only four out of the 550 modules had more than 14 genes (15% of the 88 genes; indicated by blue arrows in Figure 10) that overlapped with module #35. Hence, the meta-clustering approach efficiently combined weak signals in single studies to identify a stable and biologically more meaningful gene module. In other words, module #35 would not have been discovered without combining 11 studies.

We next tested the meta-clustering approach using transcriptomic data from control subjects only (i.e., removing all MDD subjects) from the same 11 studies. Out of the 50 modules generated, no module was enriched in more than two GWAS studies (p value threshold 0.05) among the eight GWAS results (see heatmap in Figure 9(b)), indicating that the inclusion of the MDD cases was necessary for the detection of significant module/GWAS overlap (i.e., module #35). We also tested the meta-clustering approach using transcriptomic data from MDD subjects only (i.e., removing all control subjects) from the same 11 studies. Among the 50 modules generated, one module (module #15 with 169 genes) was enriched in six out of the 8 GWAS categories ($p<0.05$) but notably not in the gene set corresponding to the Mega GWAS MDD ($p=0.29$) and to MDD-related studies ($p=0.43$) in the catalog of GWAS (data not shown here). This module only has 3 genes overlapped with the 88 genes (*ST8SIA3*, *GRM7* and *MYCBP2*) of module #35 extracted from the case and control combined analysis. Pathway analysis of this module indicated an over-representation of

signal transduction pathways. Overall, the statistical significance of results using MDD data only was lower and potentially inconclusive (i.e., at background noise level)

Together these results indicate that combining MDD and control subjects in meta-clustering approaches increased the significance and robustness of the results, as demonstrated by the identification of the tight module of 88 genes with high relevance to current biological knowledge about MDD.

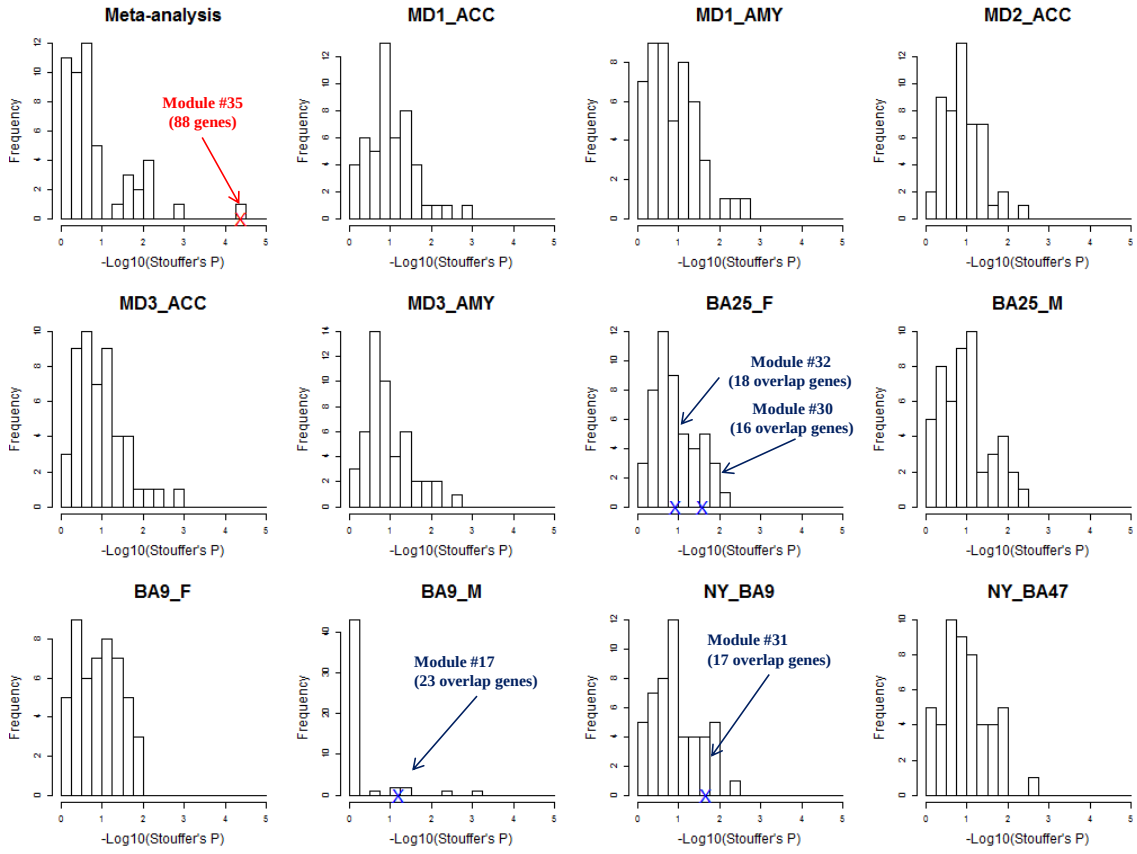

Figure 10: Histograms of the $-\log_{10}(\mathbf{p})$ of the Stouffer statistic from 50 modules of meta-analysis of 11 MDD studies and each single study.

Module #35 with 88 genes (red arrow and double-cross) have largest $-\log_{10}$ transformed p value of Stouffer's statistic 4.4. The other four blue arrows and double crosses indicated that these four modules in all single studies have more than 14 (15% of the 88 genes in module #35) overlapped with module #35. See detailed description in text.

3.4 DISCUSSION

Using methods we developed to identify conserved co-expression modules across transcriptome datasets, we report the identification of a module consisting of 88 genes that is significantly enriched in genetic variants located nearby genes otherwise associated with major depression and related phenotypes. The finding of a significant intersection of two unbiased large-scale approaches (transcriptome and GWAS) provide robust evidence for the putative recruitment and contribution to molecular and cellular mechanisms of MDD of a biological module that is formed by the identified gene set. This module includes numerous genes encoding proteins implicated in neuronal signaling and structure, including glutamate metabotropic receptors (*GRM1*, *GRM7*), GABA-related proteins (*GABRA2*, *GABRA4*, *CALB1*), and neurotrophic and development-related molecules [e.g., *BDNF*, reelin (*RELN*), Ephrin receptors (*EPHA3*, *EPHA5*)]. These findings are consistent with current hypotheses of molecular mechanisms of MDD, notably with the GABA, glutamate and neurotrophic hypotheses of depression [Sibille and French, 2013; Luscher et al., 2010; Belmaker and Agam, 2008; Nestler et al., 2002; Duman and Monteggia, 2006]. This biological "internal validation", combined with control studies showing that these results could not be achieved using single studies (due to weak signal) demonstrates that integrating transcriptome data, gene co-expression modules and GWAS results can provide a novel and powerful framework to improve understanding of MDD and other complex neuropsychiatric disorders. This approach also provided here a set of putative interacting molecular partners, potentially reflecting a core biological module that is recruited and implicated in biological mechanisms of MDD.

The meta-clustering approach in this paper has the following novelty and advantages. (1) *Meta-analysis*: Our result indicated that a meta-analysis of gene clustering to combine multiple transcriptome studies can identify more accurate and robust gene modules, since the same clustering method applied to single studies did not lead to the identification of any significant and/or neuropsychiatry-related module. (2) *Cluster analysis allowing "scattered genes"*: Gene co-expression modules were identified by penalized K-medoid. This clustering technique searches for tight gene modules and allows some genes to be scattered. This means that they are not included in the final set of modules/clusters, unlike other traditional clus-

tering methods, such as hierarchical clustering, K-means or self-organizing maps that force all genes into clusters. In genomic applications, it was shown that allowing scattered genes can improve clustering performance with better biological knowledge discovery [Thalamuthu et al., 2006]. (3) *Integration and validation with external databases*: Integration with rich GWAS and pathway knowledge databases for biological and disease interpretation identified a robust module with 88 genes that is consistent with current knowledge about depression, hence providing some level of "internal control" for the methods. (4) *Case and control combined co-expression analysis*: We showed that the combination of case and control co-expression analysis was necessary to reveal the co-expression perturbation originating from the disease. This is an important observation as co-expression studies rely on subtle differences in expression patterns compared to differential expression between two groups. Hence disease-related co-expression modules could have been predicted to be unique to the disease groups and "diluted" when combined with control data. However, we show that the opposite is true, resulting in increased power in the combined dataset. For technical validation, we have performed the following: First, we fine-tuned the parameters to be used in the final meta-clustering analysis (i.e., number of modules, percentage of allowed scattered genes in penalized K-medoid method) and tested those parameters in three studies using "surrogate" information, i.e., gene families and biological pathways broadly associated with psychiatric disorders (See Methods section). Second, subsampling and bootstrap simulation were applied to investigate the stability of the identified gene modules. Third, three non-psychiatric related GWAS gene sets (cancer, human body indexes and disease traits unrelated to mental functions) served as negative controls.

Co-expression links between genes are inferred from microarray expression studies but do not refer to any specific mechanism underlying these correlations. In fact, any mechanism that synchronously regulates transcription of multiple genes may potentially generate co-expression relationships, including biophysical sources (e.g., transcription factors, spatial configuration of chromosomes, mRNA degradation, miRNA or other upstream regulation, histone acetylation and methylation patterns), technical effects (e.g., batch processing, RNA quality), cell biological sources (e.g., cellular admixture of the sampled tissue, brain region), and importantly synchronized activities across cells under homeostatic equilibria correspond-

ing to "control" states, trait conditions, or chronic disease states for instance. Here, results in module #35 identify a set of genes whose products are distributed across cell types, cellular compartments and biological processes (Tables 4-5) that together contribute to various and potentially complementary biological processes, and whose collective function may be related to pathological processes implicated in depression.

The biological content of the identified gene module is notable in that it brings together multiple genes that have been otherwise associated with depression and other neuropsychiatric disorders through multiple studies both in humans and animal models, in addition to the genetic links (i.e., GWAS) that were used here to identify them. Such commonly associated genes include those coding for BDNF, and GABA- and glutamate receptors, for instance [Tripp et al., 2012; Klempan et al., 2007; Sequeira et al., 2009; Choudary et al., 2005; Guilloux et al., 2011]. Prior findings often refer to differential expression, e.g. reduced BDNF [Guilloux et al., 2011], or reduction in calbindin (*CALB1*) positive GABA neurons [Rajkowska et al., 2006]. Here, reports of conserved co-regulated patterns between these genes suggests that changes in the fine-tuning and synchronization of the function of these gene products across cells and pathways may contribute to pathophysiological mechanisms related to brain dysfunction in MDD. The fact that these results implicate genes that are likely to be expressed across cell types or to regulate ensembles of cells (i.e. neurotrophic and neuro-maintenance factors) is consistent with mechanisms expected for polygenic com- plex disorders. Moreover, the identification of module #35 through overlap with GWAS findings for traits (i.e., neuroticism) and other neuropsychiatric disorders (Figure 9) also suggests that those genes may participate in basic cellular functions that are implicated in a continuum of biological states (i.e., from normal to disease brain functions), consistent with a dimensional understanding of biological mechanisms of brain disorders. The fact that borderline significance in gene overlap was also observed for categories of disorders sharing clinical risk with MDD (i.e., cardiovascular diseases, inflammation and metabolic syndrome) suggest that the same gene sets may also contribute to dysfunctions in peripheral organs through pleiotropic functions of common genes, hence providing putative biological links for the clinical and symptom co-morbidity. Follow-up studies of co-expression patterns obtained in datasets across these disorders may be necessary to further investigate these interesting hints.

So while these studies provide insight into the biology of complex disorders, one may reasonably ask how they may contribute to the generation of novel hypotheses and predictions. Two directions are worth mentioning. First, for the purpose of therapeutic development and target identification, the application of graph theory and other network analysis may help identify critical genes within the identified module or upstream factors, as potential mediators of the function of this module in disease state. Preliminary analyses of the network properties of module #35 did not provide clear insight into hub genes or other parameters of interest (data not shown); however these studies may be confounded by circular analyses within the same datasets. Thus, testing these hypotheses in other large-scale disease related datasets are needed to, firstly, refine gene membership into the identified module, in view of the reasonable and significant conservation of module structure across datasets, although not to absolute levels; and, secondly, to identify key network nodes with conserved cross-studies functions, as potential targets to modulate the functional outcome of the identified gene module. Finally, an additional and important outcome of these studies is that they provide a focused set of genes, which can be used for follow-up genetic association studies, hence potentially mitigating the problem of reduced statistical power of large scale genome-wide studies.

There are several limitations to this study. First, there is a bias when selecting gene sets from the catalog of published GWAS results since the targeted markers (SNPs) are updated every six months, and many more SNPs were reported in the past five years when GWAS have achieved greater sample size (including studies with more than 10,000 participants) and detection of markers with very small effect size. However, large sample sizes will also introduce a bias towards false positive markers. A related limitation is that the choice of markers (or gene) was based on fixed and arbitrary thresholds (i.e., p value and genomic distance). Moreover, we used only a small fraction of the datasets and pre-defined pathways related to psychiatric disease to decide on the number of clusters and sets of scattered genes during the method development phase, so the result of the clustering approaches may still show some instability and may vary based on different numbers of clusters and applied thresholds. Indeed, although we performed extensive validation analyses to select the

clustering parameters and increase stability of modules, the 88 genes in module #35 will inevitably vary slightly under additional data perturbation (e.g., when adding additional MDD or related studies). An additional limitation is that generating gene co-expression modules using cluster analyses is known to be sensitive to small data perturbation. To mitigate these effects, we combined multiple studies and concentrated on tight modules by leaving out scattered genes. While this approach increased the power of the meta-clustering method, it also meant combining datasets from different brain regions, hence potentially diluting the effects of local co-regulation patterns that may be important for disease mechanisms. So these results should be considered proof-of-concept, rather than experimentally and biologically optimized. Finally, it is important to note that changes in gene co-expression are difficult to confirm by independent measures. Indeed co-expression links rely on large sample size and we previously showed that the sample-to-sample variability in array-based measures of expression is typically lower than the variability obtained using alternate measures such as quantitative PCR [Gaiteri et al., 2010], so the ultimate test of the added value of these meta-co-expression studies will need to come from additional independent studies. Nonetheless, this study allowed the identification of a focused set of genes for use in future genetic association studies, and together demonstrates the importance of integrating transcriptome data, gene co-expression modules and GWAS results, paving the way for novel and complementary approaches to investigate the molecular pathology of MDD and other complex brain disorders.

4.0 THE ANALYSIS OF FAMILY-BASED SEQUENCE DATA

4.1 INTRODUCTION

Next generation sequencing is an advanced technology which can identify common variants (minor allele frequency $> 5\%$) as well as those done by typical GWAS, and systematically search for rare variants. Recently, the 1,000 Genome Project has provided characterization of human genome sequence variation for us to understand the relationship between genotype and phenotype [Abecasis et al., 2010]. Family-based sequencing studies have unique advantages and strengths in controlling population stratification, studying parent-of-origin effects, identifying rare causal variants and detecting *de novo* mutations [Ott et al., 2011; Laird and Lange, 2006; Ng et al., 2010a; Ng et al., 2009; Ng et al., 2010b; Zhu et al., 2010]. Sequencing has also been proven successful in studying Mendelian disorders in families [Ng et al., 2009; Roach et al., 2010; Boileau et al., 2012]. Numerous family-based sequencing projects (often in the design of large number of trios/nuclear families or a mixture of unrelated individuals and small families) have been carried out or launched to study complex diseases [Sanders et al., 2012; Neale et al., 2012; ORoak et al., 2012; Boomsma et al., 2013; Pilia et al., 2006]. Many ongoing sequencing projects include nuclear families (two parents with one or more offspring) or multi-generational families. Chen et al. [2013] proposed a method of genotype calling by considering family structure in trios that can achieve more accurate genotype calls in great amounts as compared with the one without considering the family structure (reduces genotype calling error rate by 50%). However, to our knowledge, there is no existing method that jointly models family constraints and LD patterns in complex pedigree (nuclear and extended families). Existing approaches include methods that focus on single sites or methods that split pedigrees into trios or treat all sequenced samples as unrelated

individuals. Among the few existing methods for genotype calling of family-based sequence data, most methods consider family constraints at each marker [Li et al., 2012; Peng et al., 2013].

In this chapter, we described a novel method and developed a software which called "FamLDCaller" (Genotype calling method incorporated by LD and family structure. We used "FLDC" for abbreviation) for genotype calling and phasing in nuclear and extend families. Firstly, we extend the current method from analyzing trios to nuclear family or family with multi-generations in a computationally efficient manner. Here we focus on developing the procedure by looping over all possible parent-offspring trios to update the probability of observed genotype given the true genotype simultaneously, which is a pivotal step in the hidden Markov model (HMM). Through two simulated studies, which are with/wothout alignment and experimental errors. We evaluate the performance by using the genotype error calling rate and phasing error (as haplotypes are provided), and we show that incorporating more offspring within family (or complex family with multiple generations) can have more accurate genotype calls than trios only, especially in low to modest depth in sequencing data. Secondly, we extend the method to analyze a small number of samples using the external reference panels. This is motivated by many pilot projects, which often include a limited number of samples (e.g. one or two trio) and LD information is not available in the study population. External reference panels (e.g. the 1,000 Genome Project) will be useful in this scenario to facilitate genotype calling and phasing if the LD pattern in the study population is well captured. Through both simulated and real studies, we show that our methods outperform the existing methods that do not use LD information or ignore the complex family constraints.

4.2 METHODS

4.2.1 Describing chromosomes as imperfect mosaics

Li and Stephens [2003] indicated that the haplotypes of each individual can be described as imperfect mosaics of other haplotypes in the sample of using hidden Markov model (HMM);

in specific, unrelated samples for same ethnicity are always sharing short stretch of the chromosomes, so each samples is a "mosaic" of haplotypes. And this approach has been successfully applied to genotype imputation and haplotype reconstruction [Li et al., 2010; Marchini et al., 2007; Scheet and Stephens, 2006]. This approach has also been used in genotype calling for sequence data. In this section, we briefly review the HMM method for model unrelated samples for the sequence data. More details can be found in our early paper [Chen et al., 2013]. First, allele from individual haplotype was sampled from reference panels consistent with observed data at each position. Second, using HMM method to update the haplotype for each individual, and the pair of haplotypes can be described as an imperfect mosaic of other reference panels.

Suppose all markers are bi-allelic, the first step is to calculate $P(R_i|G_i)$, the likelihood of observed read R_i given an underlying true genotype G_i at position i for all candidate variant sites. $P(R_i|G_i)$ can be defined by the following formula by assuming independent errors:

$$\begin{cases} P(R_i &= \mathbf{B}, \mathbf{E}|G_i = \{\text{A,A}\}) \\ &= \prod_j (1-e_j)^{I(b_j=1)}(\frac{1}{3}e_j)^{I(b_j\neq 1)} & \text{for homozygous genotype A/A} \\ P(R_i &= \mathbf{B}, \mathbf{E}|G_i = \{\text{A,B}\}) \\ &= \prod_j \{\frac{1}{2}(1-e_j)^{I(b_j=1)}(\frac{1}{3}e_j)^{I(b_j\neq 1)} \\ &+\frac{1}{2}(1-e_j)^{I(b_j=2)}(\frac{1}{3}e_j)^{I(b_j\neq 1)}\} & \text{for heterozygous genotype A/B.} \end{cases}$$

where $\mathbf{B}$ and $\mathbf{E}$ represent the vectors of base calls and corresponding error probabilities for position i and allele j ($j = 1$: first allele and $j = 2$: second allele) in each subject (b_j and e_j are corresponding elements of $\mathbf{B}$ and $\mathbf{E}$). $b_j = 1$ means the j-th allele in observed reads is identical with the j-th allele from underlying true genotype and I is an indicator function.

We then define the probability the probability of an underlying true genotype G_i given the mosaic state S_i, $P(G_i|S_i)$. The function $T(S_i)$ was defined as the number of different

alleles for genotype G_i. So $P(G_i|S_i)$ was defined by:

$$\begin{cases} (1-\varepsilon_i)^2 & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } T(S_i)=T(G_i) \\ \varepsilon_i(1-\varepsilon_i) & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } |T(S_i)-T(G_i)|=1 \\ \varepsilon_i^2 & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } |T(S_i)-T(G_i)|=2 \\ (1-\varepsilon_i)^2+\varepsilon_i^2 & T(S_i)=1 \text{ and } T(S_i)=T(G_i) \\ 2\varepsilon_i(1-\varepsilon_i) & T(S_i)=1 \text{ and } T(S_i)\neq T(G_i) \end{cases}$$

where ε_i is the cumulative effects of mutation and gene conversion (we called it mosaic error rate here) at marker i. Then we can calculated the emission probability of $P(R_i|S_i)$ as:

$$P(R_i|S_i) = \Sigma_{G_i} P(R_i|G_i) \times P(G_i|S_i) \tag{4.1}$$

Finally the transition probability $P(S_{i+1}|S_i)$ in the HMM was defined by:
$P(S_{i+1}=(w,v)|S_i=(x,y))$

$$= \begin{cases} \frac{\theta_i^2}{H^2} & w\neq x \text{ and } y\neq v \\ \frac{(1-\theta_i)\theta_i}{N}+\frac{\theta_i^2}{H^2} & w\neq x \text{ and } y=v \text{ or } w=x \text{ and } y\neq v \\ (1-\theta_i)^2+\frac{2(1-\theta_i)\theta_i}{H}+\frac{\theta_i^2}{H^2} & w=x \text{ and } y=v \end{cases}$$

where θ_i is the mosaic transition rate from position $i-1$ to position i, and H is the number of haplotypes in the reference panel. Our goal is to calculate $P(G_i|\mathbf{R})$, the probability of a genotype at position i conditional on all sequence reads:

$$P(G_i|\mathbf{R}) = \Sigma_{S_i} P(G_i|S_i) \times P(S_i|\mathbf{R}) \tag{4.2}$$

by looping all possible state S_i. Baum's forward-backward algorithm was used to calculate $P(S_i|\mathbf{R})$ and $P(G_i|\mathbf{R})$ [Rabiner, 1989].

4.2.2 Procedure for modeling nuclear family

Chen et al. [2013] proposed a strategy for parent-offspring trios with computationally efficient modeling of LD and the constraint due to Mendelian inheritance. We extended their proposed algorithm from analyzing trios to nuclear families by looping over all possible trios within each family. Consistent with their paper, we denote R_{fk}, R_{mk} and R_{ck} as the read data from k-th possible trio within a nuclear family, G_{fk}, G_{mk} and G_{ck} as the underlying true genotype for the father, mother and child, and the genotype likelihood was denoted by $P(R_{fk}|G_{fk})$, $P(R_{mk}|G_{mk})$ and $P(R_{ck}|G_{ck})$. The procedure for each iteration was described below:

I. At position i, we randomly select a child in family and corresponding parents, denoted by $\bar{R}_{i1} = (R_{f(i)1}, R_{m(i)1}), R_{c(i)1})$.

II. First update parental haplotypes by sampling a mosaic state $S_{f(i)1}$ for father, then emission probability can be written as $P(\bar{R}_{i1}|S_{f(i)1}) = \sum_g P(\bar{R}_{i1}|G_{f(i)1} = g) \times P(G_{f(i)1} = g|S_{f(i)1})$, and $P(\bar{R}_{i1}|G_{f(i)1} = g) = \frac{P(\bar{R}_{i1}, G_{f(i)1} = g)}{P(G_{f(i)1} = g)} = \sum_{g_m} P(R_{f(i)1}|G_{f(i)1} = g) \times P(R_{m(i)1}|G_{m(i)1} = g_m) \times P(R_{c(i)1}|G_{c(i)1} = \text{transmit}(g_f, g_m))$, where $\text{transmit}(g_f, g_m)$ returns the genotype for child conditional on ordered parental genotypes G_f and G_m.

III. Updates maternal haplotypes at position i conditional on the sampled genotype for the first parent. $P(\bar{R}_{i1}|S_{i1}, G_{f(i)1} = g_f) = \sum_g P(R_{f(i)1}|G_{f(i)1} = g_f) \times P(R_{m(i)1}|G_{m(i)1} = g_m) \times P(R_{c(i)1}|G_{c(i)1} = \text{transmit}(g_f, g_m))$.

IV. Randomly select second child ($R_{c(i)2}$) and corresponding parents updated from previous trio loop, and repeat step I - step III until all children ($R_{c(i)k}, k = 1, 2, \ldots, n_l$ where n_l is number of children in family l) are used in each family.

V. Update next family and repeat step I - step IV until all families are used.

Each round of updates generates a new ordered haplotypes for each family (can be unrelated individual, parent-offspring trio, nuclear family or family with multiple generations), the consensus haplotype was generated by assigning the most frequently sampled allele at each position. Figure 11 illustrates the example of updating haplotypes for each iteration in a nuclear family with three offspring. For each iteration, we randomly selected one offspring to form a trio ("random" here means we try to avoid keep using first offspring to update parents' haplotype in each iteration), and update the haplotypes of parents and offspring

(step II. and III. of the procedure). We than randomly selected second offspring to form a trio with parents' haplotypes updated from previous step, repeated step II. and III. until all possible trios were looped in each family. This method can also be applied to multi-generational family in a similar manner by looping through all offspring in a random order in each iteration.

4.2.3 Use of phased reference panels

Public reference panels (e.g. 1KG Project and HapMap Project) can provide extra LD information for genotype calling and have been successful in facilitate imputation. For genotyping sequence data, most existing software do not use the information from reference panel. Our method and implementation can incorporate phased reference panels efficiently into our genotype calling procedure. It has two advantages: 1) we will be able to call a small number of sequenced families/individuals using LD information from a similar population with phased haplotypes available; 2) the computation will be efficient because we don't have to call all individuals but only sequenced individuals. This approach is particularly useful for sequencing studies with a small sample sizes.

4.2.4 Simulated data

In the first simulation scheme, we considered 80 nuclear families, and each family has two founders and four offspring. To be realistic, we generated 12 regions with 1 Mb length of haplotypes, and each region contains 10,000 haplotypes generated from a coalescent model to mimicking the LD pattern, population demographic history and local recombination rates of European ancestry samples [Schaffner et al., 2005]. We randomly sampled haplotypes for founders in each family and simulate the Mendelian transmission for the haplotypes of offspring. The short read were simulated by assuming depth at each site follows a Poisson distribution and defined per-based sequencing error rate. Each sample was sequenced at depth 2x, 6x and 10x by assuming per base error (denoted as "BE" in all Tables and Figures throughout this chapter) rate of 0.01 (Phread scaled base quality of Q20). In order to compare with the "TrioCaller" software proposed by Chen et al. [2013], we considered the

following procedure when calling genotypes in each nuclear family: we selected first child to form into a trio and treated other three children as unrelated subjects (result from "Trio-Caller"), and then included the second child into consideration at a time until all children were used.

In the second simulation scheme, we further considered sequencing and alignment errors using the 1000 Genome Project (1000GP) data. We simulated founders entire genomes by randomly selecting a pair of haplotypes from the 1000GP CEU population (March 2012 Phase 1 release). For non-founders, we simulated cross-overs in the parental haplotypes based on the genetic map in the HapMap data, and then generate offspring genotypes by randomly selecting one haplotype from each parent. We then simulated paired-end 100bp reads according to Poisson distribution on the genome, with a mean insertion size of 400bp and a standard deviation of 50bp, and a sequencing error rate of 0.01 per base. We used BWA to align simulated reads to the reference of hg19 and carried out standard procedure for variant calling using Genome Analysis Toolkit (GATK) [McKenna et al., 2010] including indel-realignment and base quality realignment. The list of known indels from 1000GP was provided to GATK for re-alignment prior to variant calling in different depths 5x, 10x, 20x and 30x with 3,005,070 sites on chromosome 1. There are five families, and each family has 14 members (see pedigree in Figure S9). We considered the simulation settings similar to our first simulation scheme: we selected nuclear family (parents and three offspring) in each big family, then we selected first child to form into a trio and treated another two children as unrelated subjects, and included the second child into consideration at a time until all children were used. In addition, we also selected complex family with three generations from each big family.

Next, we investigate if the reference panels can help increase genotyping accuracy. We designed a simulation study by considering 2, 3 and 4 parent-offspring trios with depth 2x, 6x and 10x with per-base error rate of 0.01(Q20). For reference panels, we considered 10, 20, 40 and 60 founders from 1000 genome project.

4.3 EVALUATION CRITERIA

First, we evaluated the performance of genotype calls using *genotype mismatch rate* between genotypes estimated by our proposed algorithm and surrogated gold standard genotypes from simulated data, especially in heterozygous sites, which is more sensitive case in genotype accuracy. Second, we calculated the number of *mismatched alleles* between estimated haplotypes by our proposed algorithm and haplotypes from simulated data to evaluate the haplotyping accuracy. Third, we also evaluated the Mendelian error by calculating the *number of mismatching alleles* between each offspring and corresponding parents.

4.4 SIMULATION RESULTS

4.4.1 Overall performance of genotype accuracy

We evaluated the performance of our proposed algorithm for genotype calling method in simulation studies and real data analysis. We have two goals: (1) extended the existing method for analyzing trio-based data sets to handle complex family with multiple offspring and/or generations; (2) proposed a function to analyze a small number of family-based samples incorporating the external reference panels, such as subjects from 1,000 Genome Project. For goal one, We first evaluated the genotype accuracy when adding more offspring in each family. Figure 12 shows the mean of the genotype mismatch rate of heterozygous calls and SNP with minor allele frequency (MAF) $< 5\%$ summarized from twelve simulated haplotypes. It shows the clear pattern that adding more offspring per family can reduce the genotype mismatch rate (see also Table 6), especially in low depth (2x). The genotype mismatch rate of heterozygous calls can reduced from 4.5% to 4.38% to 4.18% to 3.94% when one, two, three and all four offspring were considered, respectively. Sequencing depth also contributed to genotype accuracy: as 80 trios and 240 unrelated samples were sequenced, the genotype mismatch rates of heterozygous calls reduced from 4.48% to 0.875% to 0.257% as depth increase from 2x to 6x to 10x. The advantage of our proposed method makes clear that adding more

offspring can achieve more accurate genotype calls, especially at low sequencing depths. The second simulation scheme considered alignment and experimental errors also target on the first purpose. Table 7 shows the genotype mismatch rate of heterozygous calls. In general, GATK has high genotype mismatch rate, especially with depth 5x (16.4%) and 10x (2.7%) and our proposed method greatly outperform the results from GATK. When all three offspring were all considered in our algorithm, the genotyping errors of heterozygous SNPs can reduced from 16.4% and 2.7% to 0.9% and 0.4% at 5x and 10x coverage, respectively. Genotype mismatch rate will keep decreasing when adding more offspring when using our proposed method, especially at low depth 5x. The genotype discordance error rate can be reduced from 0.92% to 0.84% to 0.77% by considering one, two and three offspring in each family at 5x coverage. Furthermore, Our proposed method can handle the complex family structure (genotype mismatch rate are 0.86%, 0.37%, 0.24% and 0.25% at depths 5x, 10x, 20x and 30x, respectively)

4.4.2 Performance of haplotyping

Haplotype reconstruction plays an important role for follow-up analysis such as genotype imputation; and studying the population history. The phasing error rates were calculated by the mean number of mismatched alleles between reconstructed haplotypes by using our proposed algorithm and haplotypes from simulated data (we assumed the simulated haplotypes was underlying truth). The first simulation results from twelve simulated haplotypes were summarized in Figure 13 and Table 8. In summary, at low depth 2x, adding more offspring in each family can keep reducing the genotype mismatch rate. For instance, the phasing error rate can reduce 25% when all four offspring were taken into consideration compared with trio-based (only considers one offspring). Similar to genotype accuracy, sequencing depth contributed to phasing error rate, but our proposed algorithm still showing its advantage to lower the phasing error when adding more offspring.

4.4.3 Performance on Mendelian errors

Since our proposed method considered the family constraint (we considered trio at a time within whole family structure), we can also lower the Mendelian errors. We calculated the Mendelian errors by calculating the total number of Mendelian inconsistent genotypes divided by the total number of offspring in simulated data set. In our first simulation study without considering alignment and experimental errors, the mean number of Mendelian errors of our proposed method when considering all four offspring compared with trio-based method in simulated data can be dropped from 13.86 to 9.04, 3.74 to 2.37 and 1.42 to 0.63 at 2x, 6x and 10x coverage, respectively (see Figure 14 and Table 9). The second simulation result was summarized in Table 10 which showed the mean number of Mendelian errors for each offspring with considering alignment and experimental errors. As compared with the results from GATK, our proposed method can reduce the mean number of Mendelian error from 28212.6 and 6475.1 to 718.3 and 227.73 SNP at 5x and 10x coverage, respectively. In addition, when adding more offspring into consideration, our algorithm can achieve lower Mendelian errors, especially with low depth 5x: the mean number of Mendelian errors can be reduced from 1118.1 to 962.2 to 718.3 when considering one, two and three offspring in each family, respectively.

4.4.4 Performance of incorporating reference panels

Next, we proceed to evaluate the genotype mismatch rates, phasing errors and Mendelian errors by incorporating external references when sequencing data with small sample sizes for our second purpose (see the simulation results summarized in Figure 15 to Figure 17 and Table 11 to Table 13). In summary, for limited number of sequencing data, our proposed algorithm by incorporating external references can also provide the accurate genotypes, and reduce the phasing errors and Mendelian errors. For example, the genotype mismatch rates dropped from 7% to 4% to 2.8% to 2.4%; the phasing error rates dropped from 0.2% to 0.12% to 0.08% to 0.07% and the mean number of Mendelian errors dropped from 5.92 to 3.21 to 1.92 to 1.46 when incorporating 10, 20, 40 and 60 founders from 1,000 Genome Project for 2 sequenced trios at 2x coverage. Sequencing depth is also a key factor for genotype

accuracy, and we found that increasing the number of external references (founders) can be a good way to compensate the depth. For example, the genotype mismatch rate at coverage 2x incorporated by 60 founders is 2.4%, which is similar to the genotype mismatch rate at coverage 6x incorporated by 10 founders (∼2%); the genotype mismatch rate at coverage 6x incorporated by 60 founders is 0.55%, which is similar to the genotype mismatch rate at coverage 10x incorporated by 10 founders (0.056%).

4.5 PERFORMANCE ON REAL DATA

We applied our methods to an ongoing sequencing project, which has a total of 2,499 sample and includes 623 families with an average depth 10X (unpublished data). We focused on chromosome 20 and calculated the mismatch rate between the called genotypes from our method and the available genotypes from DNA microarray chips. Then, we compared the mismatch rate using our methods with that using other existing method BEAGLE [Browning and Browning, 2009]. In summary, our method outperforms the results from BEAGLE. For all SNPs, the genotype mismatch rate of BEAGLE and our method are 1.012×10^{-3} and 7.75×10^{-4}; For heterozygous SNPs, the genotype mismatch rate of BEAGLE and our method are 1.863×10^{-3} and 1.539×10^{-3}. We will continue our investigation when more sequence data are available. Specifically, we also investigated few small regions on chromosome 20 using different states (100, 200 and 400), as a result, the genotype mismatch rate for heterozygous calls can reduced from 1.53×10^{-3} to 1.22×10^{-3} to 1.04×10^{-3} when 100, 200 and 400 states were used.

We also applied our method to the 1,000 Genomes Project on deep sequenced trios for our second purpose to incorporate external panels when analyzing family-based sequencing data with small sample size. There are two trios with one trio from CEU and the other from YRI. These two trios have been genotyped on OMNI chip. For CEU trio, the genotype mismatch rate are 1.483×10^{-3} and 1.776×10^{-3} for all and heterozygous SNPs, respectively; for YRI trio, the genotype mismatch rate are 1.886×10^{-3} and 2.345×10^{-3} for all and heterozygous SNPs, respectively.

4.6 IMPLEMENTATION AND SOFTWARE AVAILABILITY

We have implemented our methods efficiently in a C++ program FamLDCaller, which is available from http://genome.sph.umich.edu/wiki/FamLDCaller.

4.7 DISCUSSION

In this chapter, we proposed a computationally feasible algorithm to call genotypes more accurately by considering multiple offspring in family-based next generation sequencing data set. In the simulation studies, we showed our proposed algorithm can obtain more accurate genotype calls, lower phasing errors and Mendelian errors compared with the result from trios-based method. In each iteration of MCMC step, we updated each parent multiple times by incorporating multiple offspring within each family. Our proposed method outperforms the results from Genome Analysis Toolkit (GATK) proposed by McKenna et al. [2010], which is the most popular tools for site discovery and generate genotype likelihoods for each sample, but they did not consider the family structure, which plays an important role for reducing genotyping error and Mendelian errors, especially for the data with low depth. In addition, our proposed algorithm provides a function to incorporate the external panels from 1,000 genome project when analyzing small number of family-based NGS data set (e.x. $2 \sim 4$ trios). With the high cost of NGS data set and such price still not affordable in many labs, the only way is to sacrifice the sample sizes or depths with limited budget. Our proposed method in the simulation study showed that we can achieve satisfying result by incorporating more founders if possible and the performance was as well as the same data set with high depth.

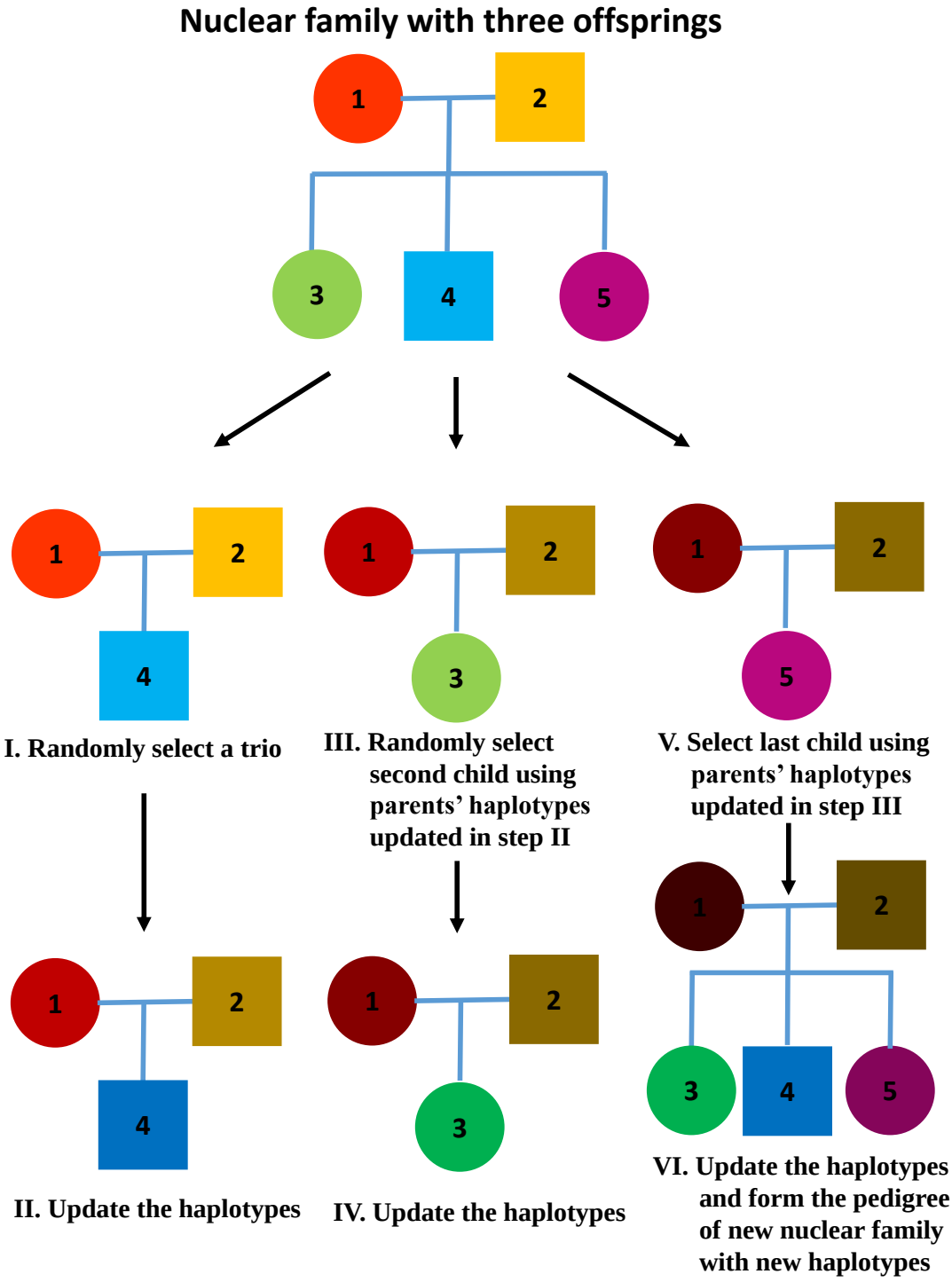

Figure 11: Example of updating haplotypes for each iteration in one nuclear family with three offspring.

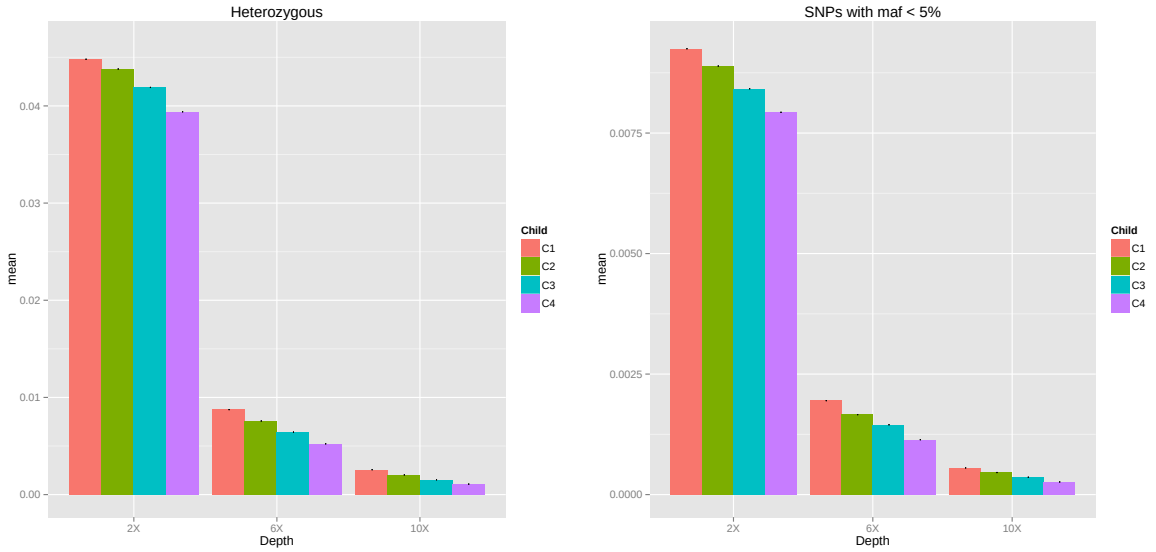

Figure 12: Genotype mismatch rate of heterozygous calls and SNPs with maf < 5% (Simulation I). C1: trios; C2: nuclear families of two offspring; C3: nuclear families of three offspring and C4: nuclear families of four offspring.

Table 6: Genotype mismatch rate of heterozygous calls and SNPs with maf < 5% (Simulation I)

		2x	6x	10x
Heterozygous calls	80 trios, 240 unrelated	0.0448	0.00875	0.00257
	80 nuclear families (twooffspring) , 160 unrelated	0.0438	0.00758	0.00203
	80 nuclear families (three offspring), 80 unrelated	0.0419	0.00643	0.00153
	80 nuclear families (four offspring)	0.0394	0.00523	0.00108
SNPs with maf < 5%	80 trios, 240 unrelated	0.00925	0.00195	0.000554
	80 nuclear families (twooffspring) , 160 unrelated	0.00889	0.00166	0.000458
	80 nuclear families (three offspring), 80 unrelated	0.00842	0.00145	0.00036
	80 nuclear families (four offspring)	0.00793	0.00114	0.000263

Table 7: Genotype discordance rate of heterozygous calls (Simulation II)

Depth	5	5	10	10	20	20	30	30
Method	GATK	FLDC	GATK	FLDC	GATK	FLDC	GATK	FLDC
F3	0.164	0.0092	0.0277	0.0042	0.00454	0.0026	0.00313	0.00256
F4	0.164	0.0084	0.0277	0.0037	0.00454	0.0025	0.00313	0.00255
F5	0.164	0.0077	0.0277	0.0032	0.00454	0.0024	0.00313	0.00253
F6	0.1638	0.0086	0.0276	0.0037	0.00453	0.0024	0.00312	0.00250

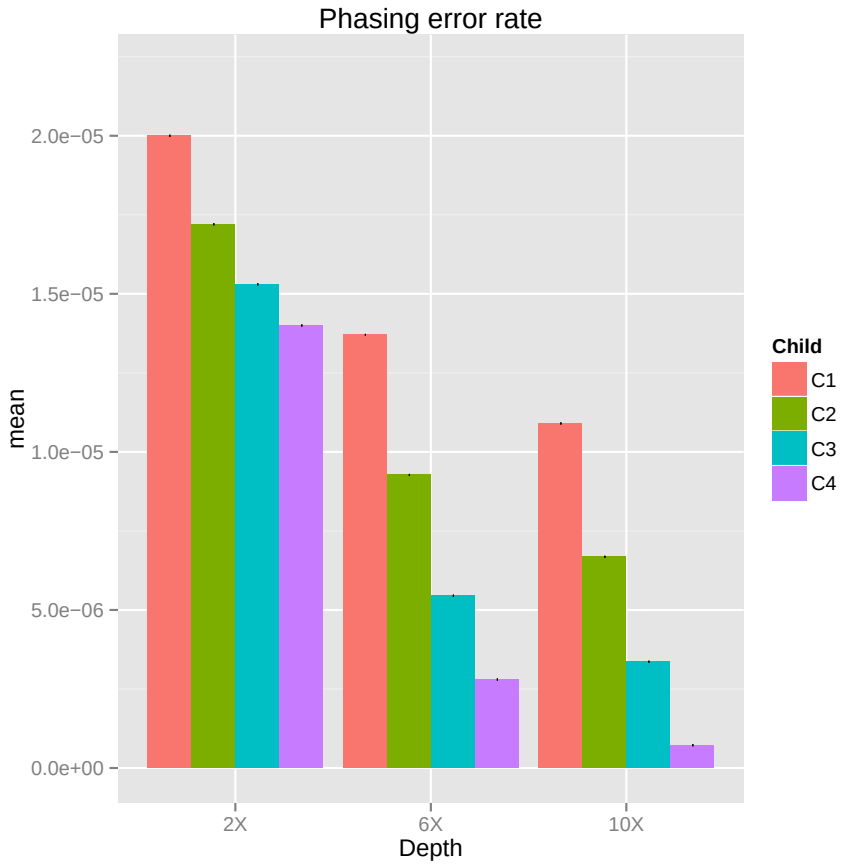

Figure 13: Phasing error rate (Simulation I). C1: trios; C2: nuclear families of two offspring; C3: nuclear families of three offspring and C4: nuclear families of four offspring.

Table 8: Phasing error rate (Simulation I)

		2x	6x	10x
BE = 20	80 trios, 240 unrelated	2e-05	1.37e-05	1.09e-05
	80 nuclear families (twooffspring) , 160 unrelated	1.72e-05	9.27e-06	6.68e-06
	80 nuclear families (three offspring), 80 unrelated	1.53e-05	5.45e-06	3.36e-06
	80 nuclear families (four offspring)	1.4e-05	2.8e-06	7.22e-07

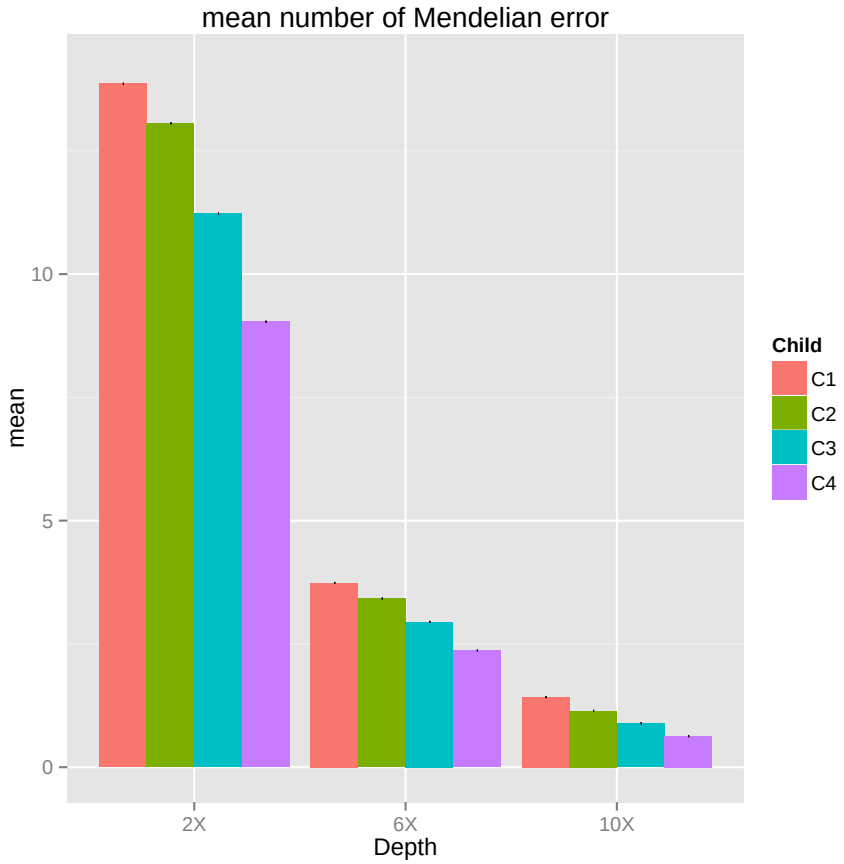

Figure 14: Mendelian error (Simulation I). C1: trios; C2: nuclear families of two offspring; C3: nuclear families of three offspring and C4: nuclear families of four offspring.

Table 9: Mendelian error (Simulation I)

		2x	6x	10x
BE = 20	80 trios, 240 unrelated	13.86	3.737	1.422
	80 nuclear families (twooffspring) , 160 unrelated	13.06	3.422	1.142
	80 nuclear families (three offspring), 80 unrelated	11.23	2.946	0.8898
	80 nuclear families (four offspring)	9.035	2.366	0.6297

Table 10: Mendelian error (Simulation II)

Depth	5	5	10	10	20	20	30	30
Method	GATK	FLDC	GATK	FLDC	GATK	FLDC	GATK	FLDC
F3	28212.6	1118.1	6475.1	483.27	927.267	182.2	628.2	163.5
F4	28212.6	962.2	6475.1	350.13	927.267	134.6	628.2	126.3
F5	28212.6	718.3	6475.1	227.73	927.267	93.87	628.2	86.1
F6	33427	350.6	7420.6	99.3	1161.2	46	804.5	39.1

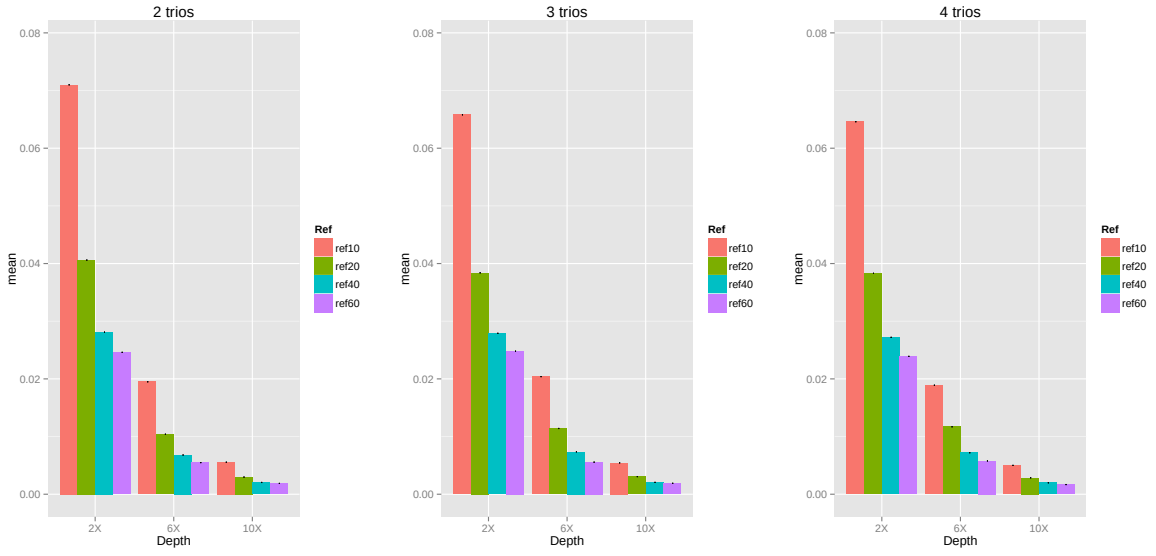

Figure 15: Genotype discordance rate of heterozygous calls (Simulation III). ref10: 10 founders; ref20: 20 founders; ref40: 40 founders and ref60: 60 founders.

Table 11: Genotype discordance rate of heterozygous calls (Simulation III)

	reference (# of founders)	10	20	40	60
BE = 20 2 trios	2x	0.071	0.0406	0.0281	0.0246
	6x	0.0195	0.0104	0.0068	0.00549
	10x	0.00555	0.00297	0.00204	0.00187
BE = 20 3 trios	2x	0.0658	0.0384	0.0279	0.0248
	6x	0.0204	0.0114	0.00736	0.00556
	10x	0.00544	0.00306	0.00205	0.00189
BE = 20 4 trios	2x	0.0646	0.0383	0.0272	0.0239
	6x	0.0189	0.0117	0.0072	0.00578
	10x	0.005	0.00285	0.00199	0.00168

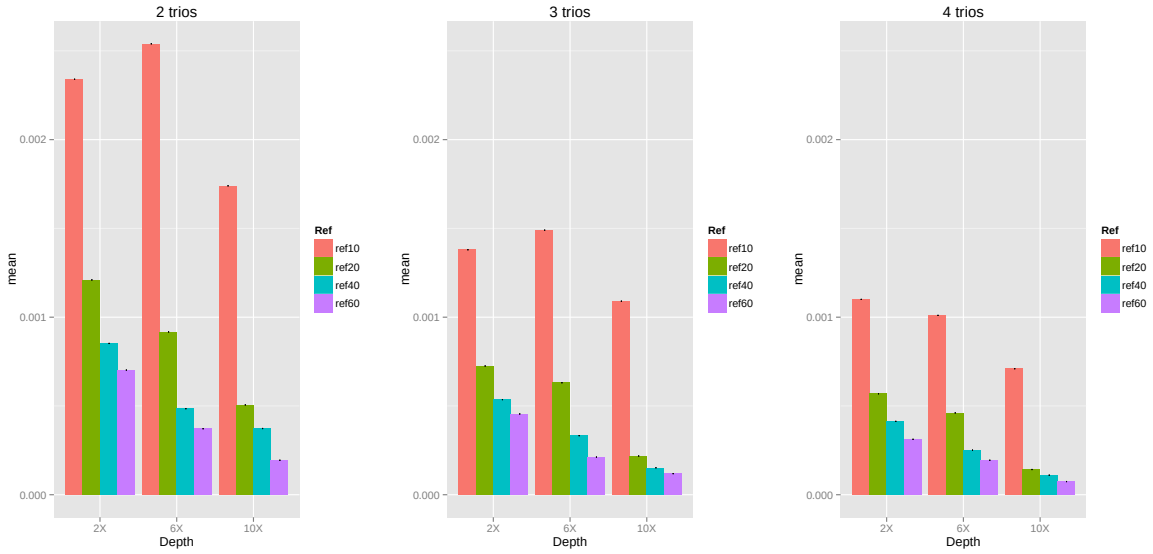

Figure 16: Phasing error rate (Simulation III). ref10: 10 founders; ref20: 20 founders; ref40: 40 founders and ref60: 60 founders.

Table 12: Phasing error rate (Simulation III)

	reference (# of founders)	10	20	40	60
BE = 20 2 trios	2x	0.00234	0.00121	0.000852	0.000702
	6x	0.00254	0.000917	0.000485	0.000372
	10x	0.00174	0.000506	0.000373	0.000194
BE = 20 3 trios	2x	0.00138	0.000725	0.000536	0.000455
	6x	0.00149	0.000631	0.000333	0.000212
	10x	0.00109	0.000218	0.000152	0.000119
BE = 20 4 trios	2x	0.0011	0.000569	0.000413	0.000312
	6x	0.00101	0.000461	0.000251	0.000194
	10x	0.00071	0.000142	0.00011	7.43e-05

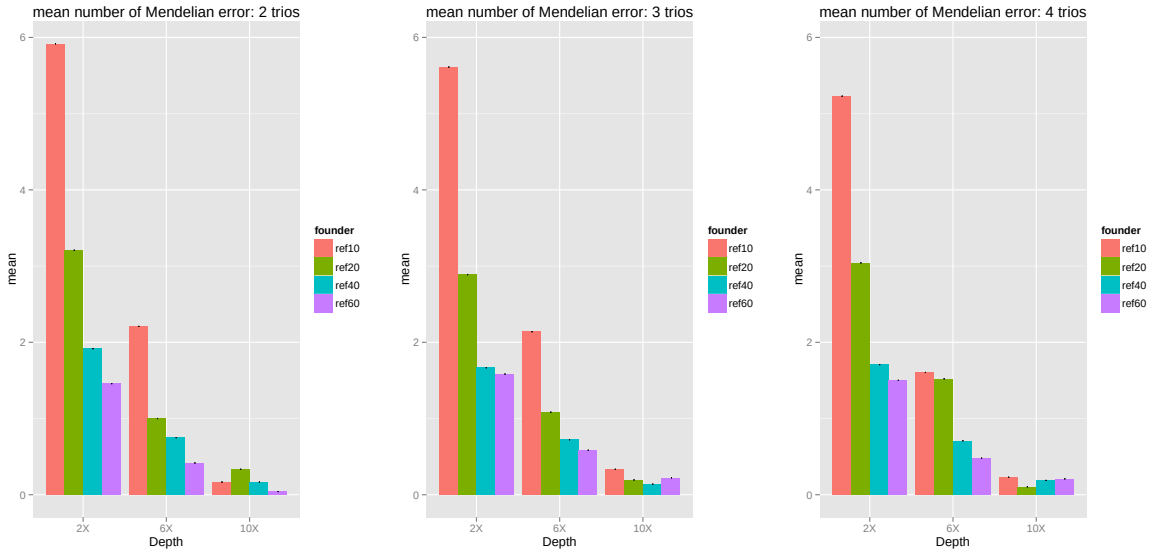

Figure 17: Mendelian error (Simulation III). ref10: 10 founders; ref20: 20 founders; ref40: 40 founders and ref60: 60 founders.

Table 13: Mendelian error (Simulation III)

	reference (# of founders)	10	20	40	60
BE = 20 2 trios	2x	5.917	3.208	1.917	1.458
	6x	2.208	1	0.75	0.4167
	10x	0.1667	0.3333	0.1667	0.04167
BE = 20 3 trios	2x	5.611	2.889	1.667	1.583
	6x	2.139	1.083	0.7222	0.5833
	10x	0.3333	0.1944	0.1389	0.2222
BE = 20 4 trios	2x	5.229	3.042	1.708	1.5
	6x	1.604	1.521	0.7083	0.4792
	10x	0.2292	0.1042	0.1875	0.2083

5.0 CONCLUSIONS AND FUTURE DIRECTIONS

5.1 SUMMARY OF CONTRIBUTIONS

In chapter 2, we provided a practical guideline for users to choose the most appropriate meta-analysis method when combining microarray data sets. For instance, if we are looking at the study-specific markers (e.g. microarray data sets with different tissues target on same disease trait, which means we expected "heterogeniety" between studies), the meta-analysis methods target on HS_B (DE genes has non-zero effect size in "one or more" studies) can identify tissue-specific DE genes. Among the meta-analysis methods target on HS_B we compared in this project, we will suggest to use adaptive weighted (AW) Fisher's method because this method provides an additional information of adaptive weight index (0: non-significant or 1: significant) and its performance is comparable to Fisher's and Stouffer's methods (see Table 2). When there is no prior information can be obtained, one can also uses our proposed entropy measure to understand the data structure (see Figure 7 (b)). In addition, HS_A or HS_r-typed meta-analysis methods are more appropriate to detected conserved and consistent DE genes across all studies.

In chapter 3, we developed a meta-clustering method to identified modules consistently co-expressed in all 11 transcriptomic MDD studies. Around 7,500 genes (we filtered out 25% scattered genes) were clustered into 50 co-expressed modules, and integrated with external databases, such as pathway database (MSigDB) and catalog of GWAS database (see overall analytical strategy in Figure 8). One robust module with 88 genes was significantly enriched in eight lists of genetic markers located nearby genes associated with major depression and related phenotype such as neuropsychiatric disorders; brain or neurological functions; disease sharing clinical risk with MDD (Diabetes, Hypertension, etc) (more detail was described in

section 3.2.5 and Figure 9). In pathway analysis, these 88 genes was enriched in GABA or Gultamate-related pathways, which may share the similar biological function of brain (more detail was described in sections of 3.2.7, 3.3.4 and Table 5). We also showed that our meta-analyzed co-expression modules can achieve more accurate and robust gene modules (see section 3.3.5 and Figure 10).

In chapter 4, we proposed a computationally efficient algorithm and developed a software "FamLDCaller" to call genotypes of next generation sequencing (NGS) data sets incorporated by family structures of nuclear family (multiple offspring) or complex family (more than two generations) (Objective 1). We showed that we can achieve more accurate genotype calls and reduced the Mendelian and phasing errors by adding more offspring in each family from the results of simulation studies and real data analysis, especially with low coverage data sets. In addition, Our proposed software "FamLDCaller" includes a function to incorporate samples from reference panels such as 1,000 Genome Project to call genotypes of family-structured NGS data sets with small sample sizes (Objective 2). We concluded more accurate genotypes can be achieved when incorporating more references.

In conclusion, the thesis first performed a comprehensive comparative study of twelve micoarray meta-analysis methods, which can be categorized according to three types of hypothesis settings they best tested (see simulation result in section 2.3.2) and we provided an application guideline for practitioners based on our proposed four quantitative evaluation criteria applied in six real examples (see discussion in section 2.4.1). Second, the thesis then presented a meta-clustering method to combine 11 MDD microarray studies to construct conserved co-expressed modules incorporated by GWAS result and pathway databases. Third, we developed a software "FamLDCaller" (http://genome.sph.umich.edu/wiki/FamLDCaller) which can be analyzed (1) in NGS data set with family structures of nuclear or complex families; (2) in NGS data set with small sample sizes incorporated by publicly available reference panels from 1,000 Genome Project. Taken together, this thesis provides several advantages of integrative analysis of omics data: (1) good summary of meta-analysis methods of microarray studies; (2) the needs of meta-clustering method to generate robust co-expressed modules and further integrate with GWAS and pathway databases; (3) Genotype callings method integrating family structure.

5.2 FUTURE DIRECTIONS

Below I will briefly discuss possible future direction from this thesis.

5.2.1 Consistency of differential expression (DE) changes in Adaptive weighted Fisher

In chapter 2, only Stouffer's method (belongs to HS_B) and REM (belongs to HS_r) methods can avoid detecting markers with discordance effect sizes when combining multiple microarray studies from the simulation study target on the case discordance effect sizes (see simulation result in section 2.3.2 and Figure S3). However, Stouffer's method became unstable as number of studies increased and is too sensitive to extremely small p-values in few studies; REM method is not robust enough to detect all the concordance markers. As a result, most developed meta-analysis methods were not designed to handle the case of discordance effect sizes, which is a common issue in real applications. In order to detect concordant DE genes, we can simply perform a post hoc filtering procedure by removing any detected biomarkers with discordant directions based on significant studies defined by adaptive weight Fisher's method (only check studies with "1" from the adaptive weighted vector w^*, see how w^* was generated from AW Fisher's method in section 2.2.3). Alternatively, we may modify the hypothesis setting and define a new concordant-based AW statistic. We expect the later approach will perform better and will be our future direction.

5.2.2 MetaClustering-clusters

In chapter 3, we applied the meta-clustering method "PPAM" by taking the mean of dissimilarity measure matrix (dissimilarity measure between gene i and gene j for a given single study k was defined by $d_{i,j}^{(k)} = 1 - |\text{cor}(X_{ik}, X_{jk})|$, where $\text{cor}(X_{ik}, X_{jk})$ is the Pearson correlation of the two gene vectors.) from K ($K = 11$) MDD transcriptome studies (see Figure 18 (A): MetaClustering-Distances). Here we aim at detecting conserved co-expressed pattern in most or all studies. There is an alternative way to perform meta-clustering method at clusters level in co-expression analysis. We can apply clustering method to each single

study separately and combine the clustering results in the following step (see Figure 18 (B): MetaClustering-Clusters). Here we will combine the clustering results constructed from each studies, hence my proposed future direction will aim at detecting co-expressed gene modules in "majority" of studies.

5.2.3 Allowing for non-autosomal genotype calling and short indels

Trio-based genotype calling method proposed by Chen et al. [2013], our developed method "FamLDCaller" can not only handles nuclear or complex family structures, but also and allows people to use external database (e.x. 1,000 Genome Project) as reference panels to obtain more accurate genotype calls when analyzing family-structured NGS data sets with small sample sizes. In the future, our method can be modified to handle X and Y chromosomes. For males, the model can be reformulated to handle a haploid case where each hidden state is one haplotype rather than a pair. In the same principle, for females, we can calculate $P(R_i|S_i)$, where $S_i = (S_{i,f}, S_{i,m})$, $S_{i,f}$ is a pair of reference haplotype and $S_{i,m}$ is a reference haplotype. The transmission and emission probabilities need to be modified accordingly. In addition, although our methods focus on SNPs and haplotypes, they can be modified to accommodate short insertions or deletions by reconstructing $P(G_i|S_i)$ and $P(R_i|G_i)$, to include a modified error model and the information about read depth.

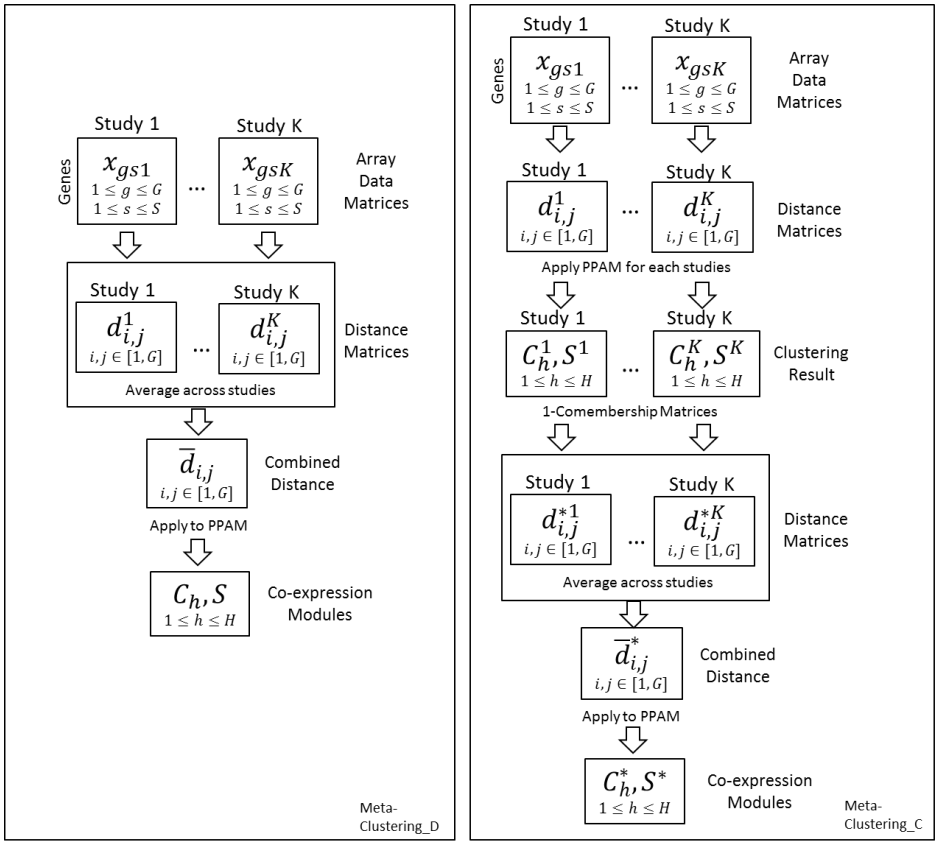

Figure 18: General workflow of meta-clustering methods to combine co-expressed genes in different approaches. A. Meta-clustering Distance; B. Meta-clustering Clusters. "This Figure is used with permission by Rui Chen's in his Doctoral Thesis proposal proposed in 2014"

APPENDIX: SUPPLEMENTARY FIGURES AND TABLES

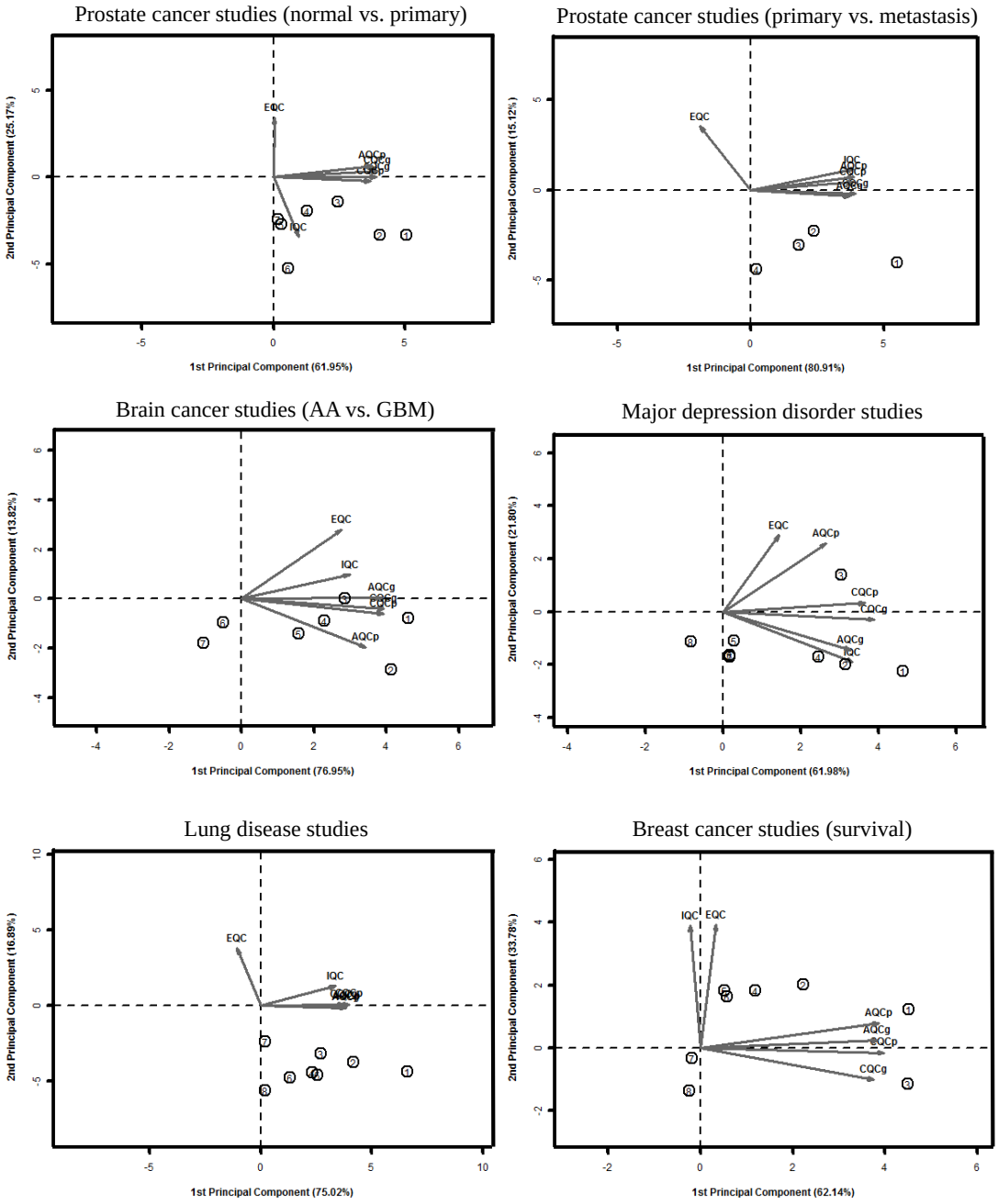

Figure S1: Meta QC.

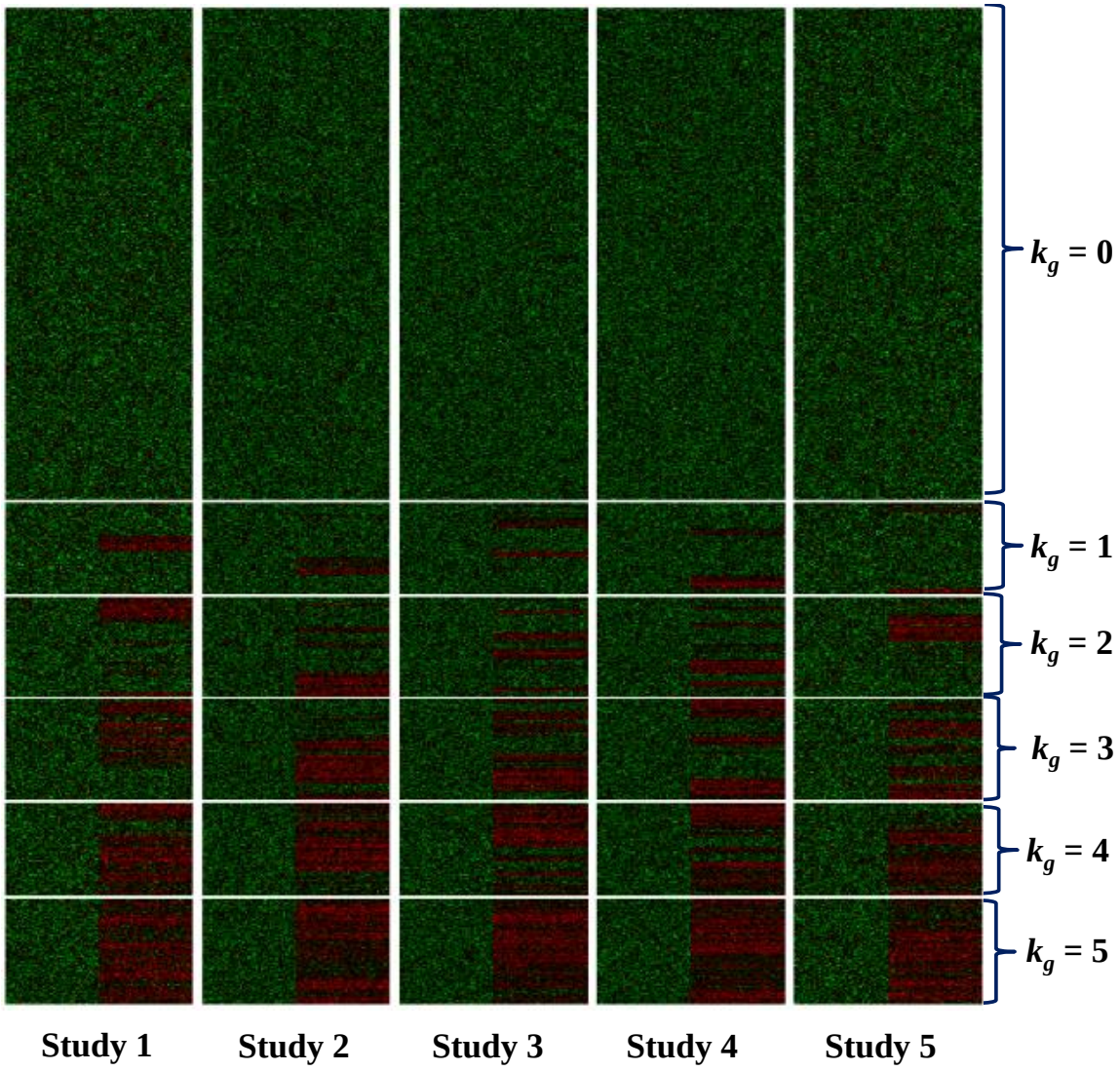

Figure S2: Heatmap of simulated example (red color represents up-regulated genes).

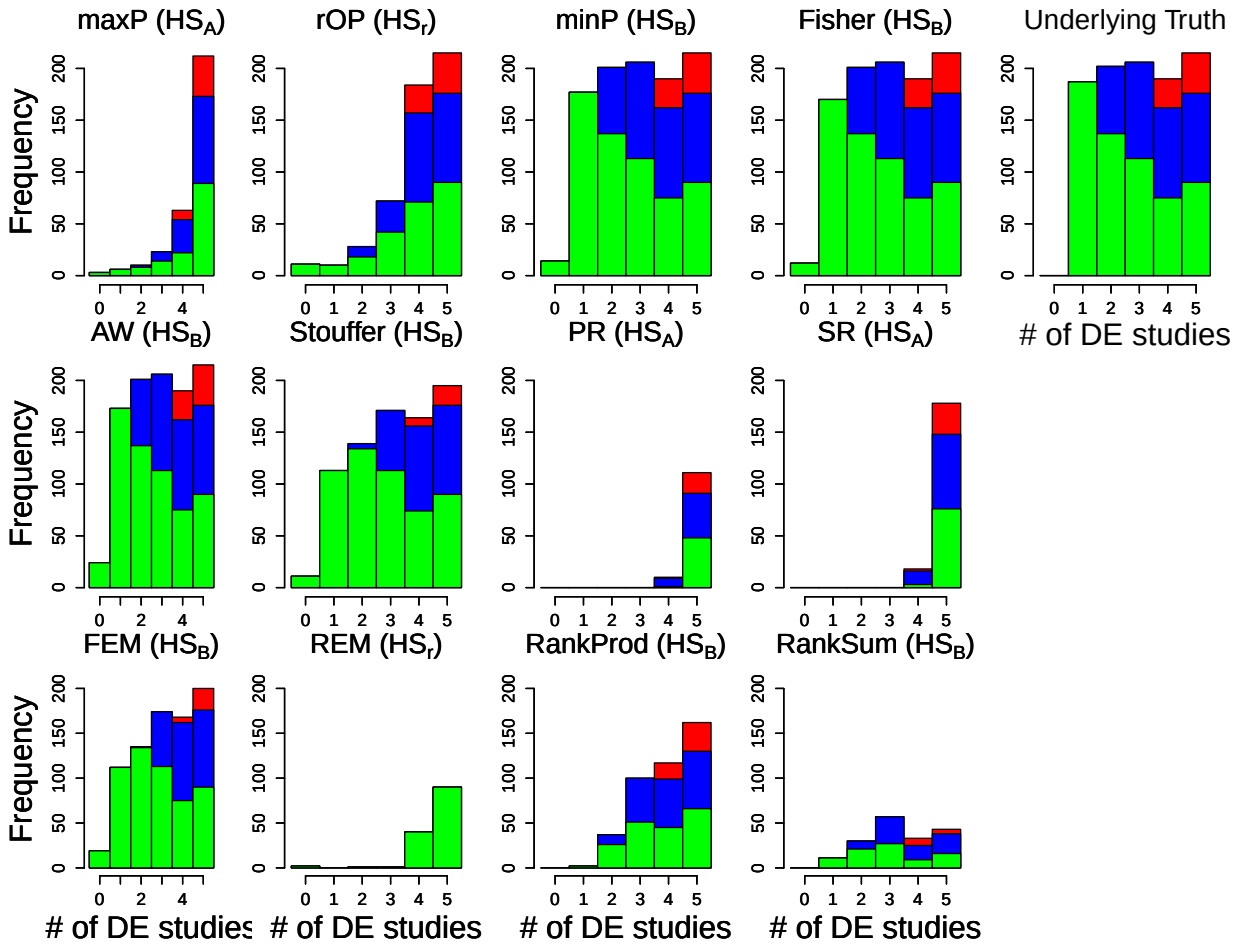

Figure S3: The histograms of the true number of DE studies among detected DE genes under FDR=5% in each method for discordance case (green color represents all concordance effect sizes; blue color represents one study has opposite effect size and red color represents two studies have opposite effect size).

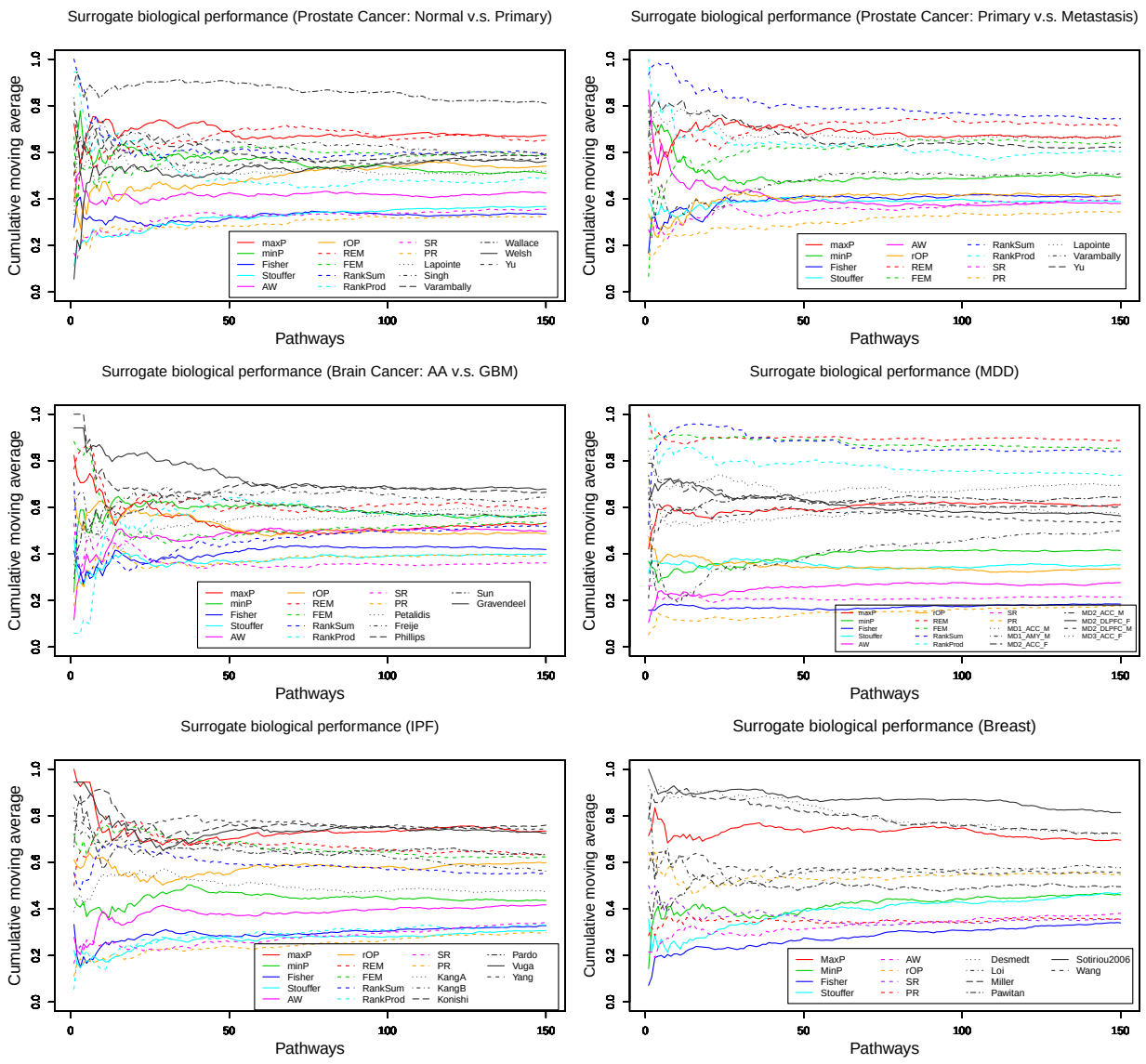

Figure S4: Cumulative moving average to determine $D = 100$.

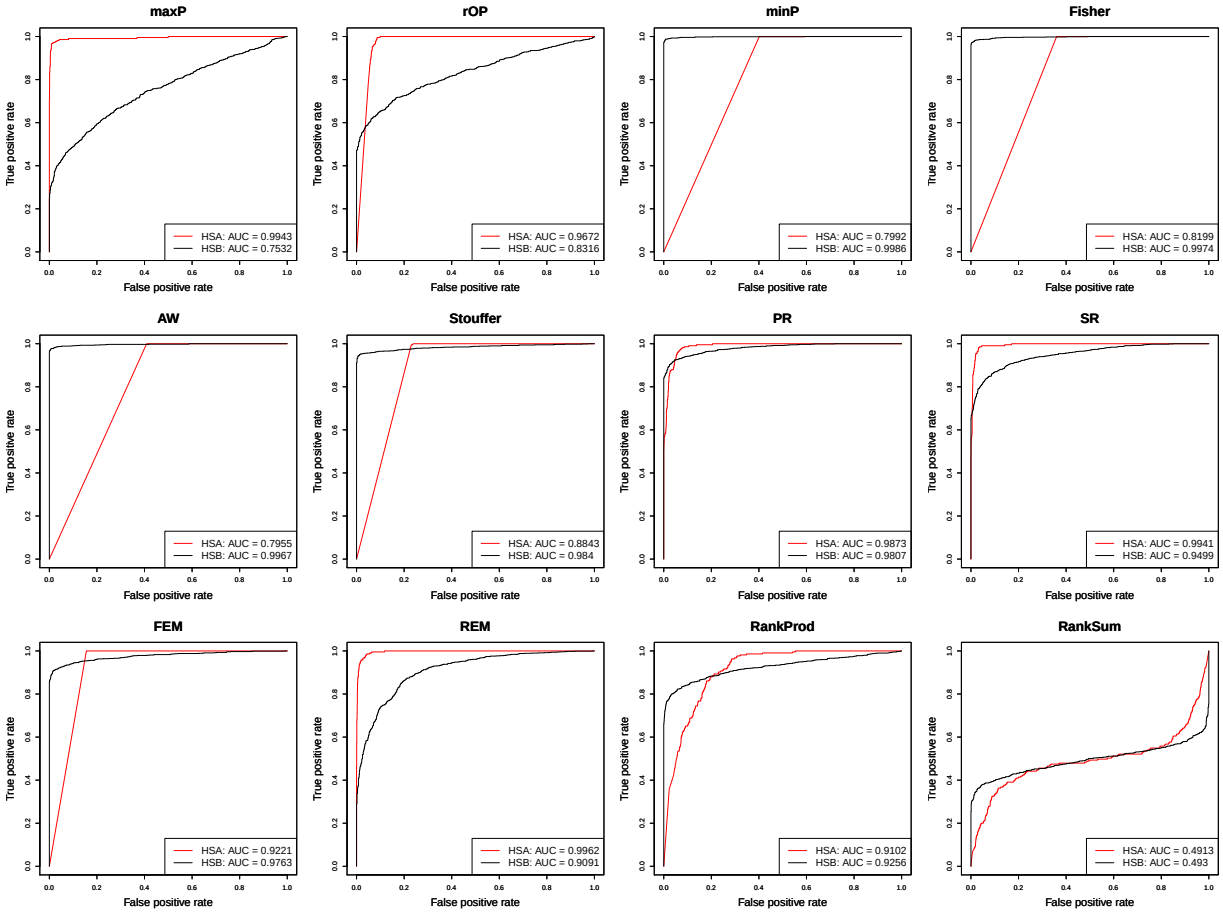

Figure S5: The ROC curves and AUC for the hypothesis settings of HS_A-type and (red line) HS_B-type (black line) in each meta-analysis method.

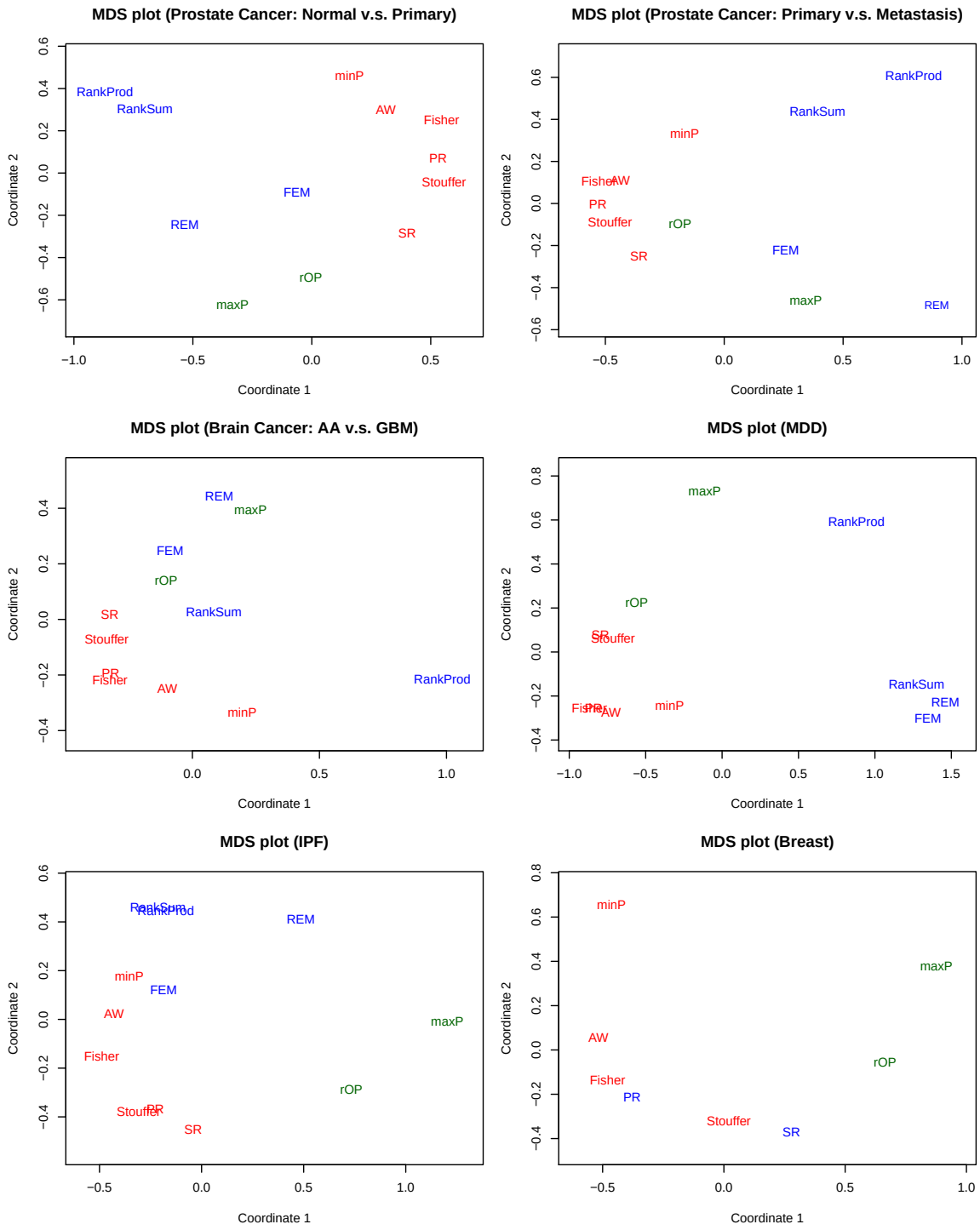

Figure S6: Multidimensional scaling (MDS) plots of individual data sets.

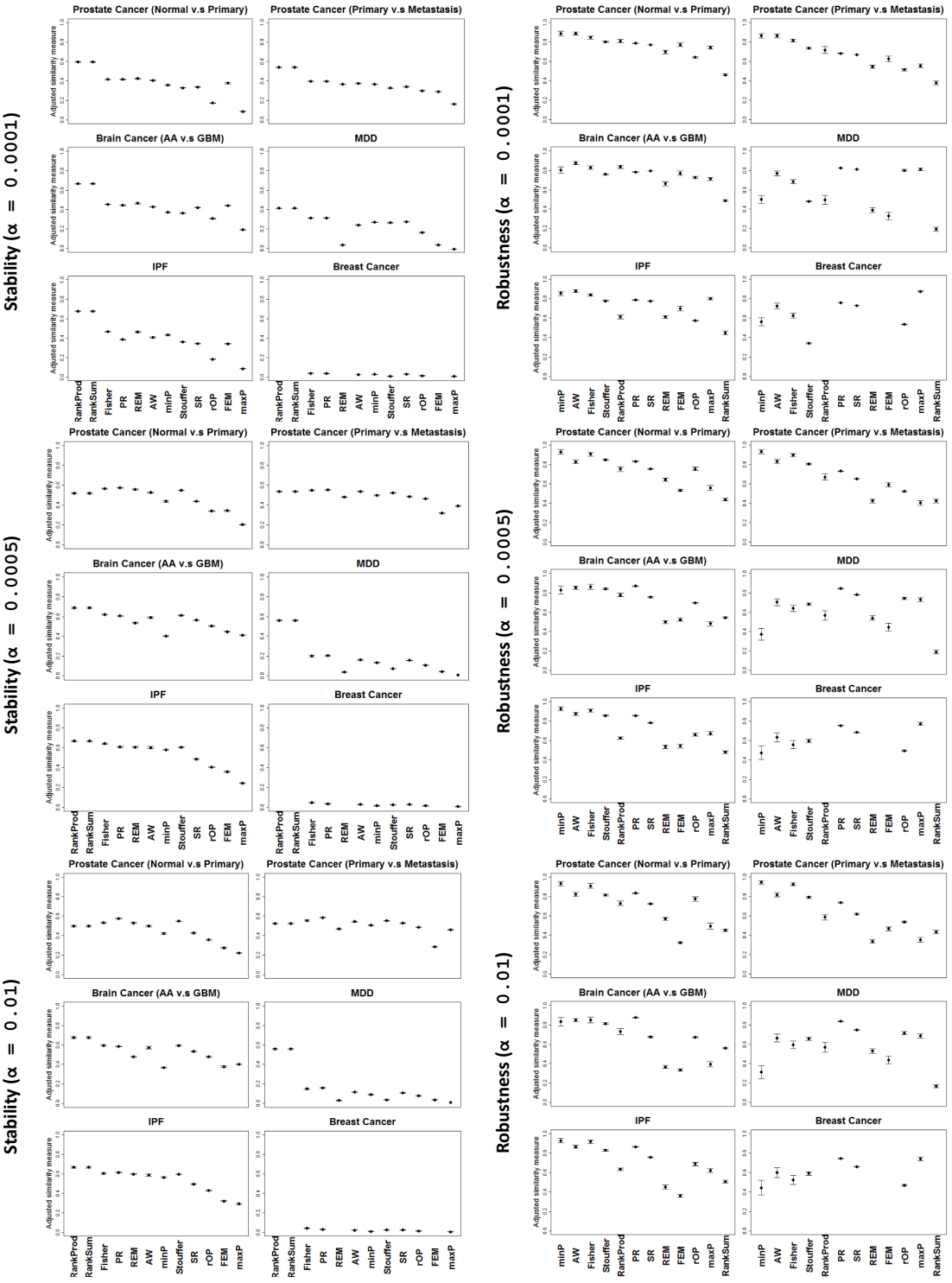

Figure S7: Stability and Robustness plot for $\alpha = 0.0001$, 0.005 and 0.01.

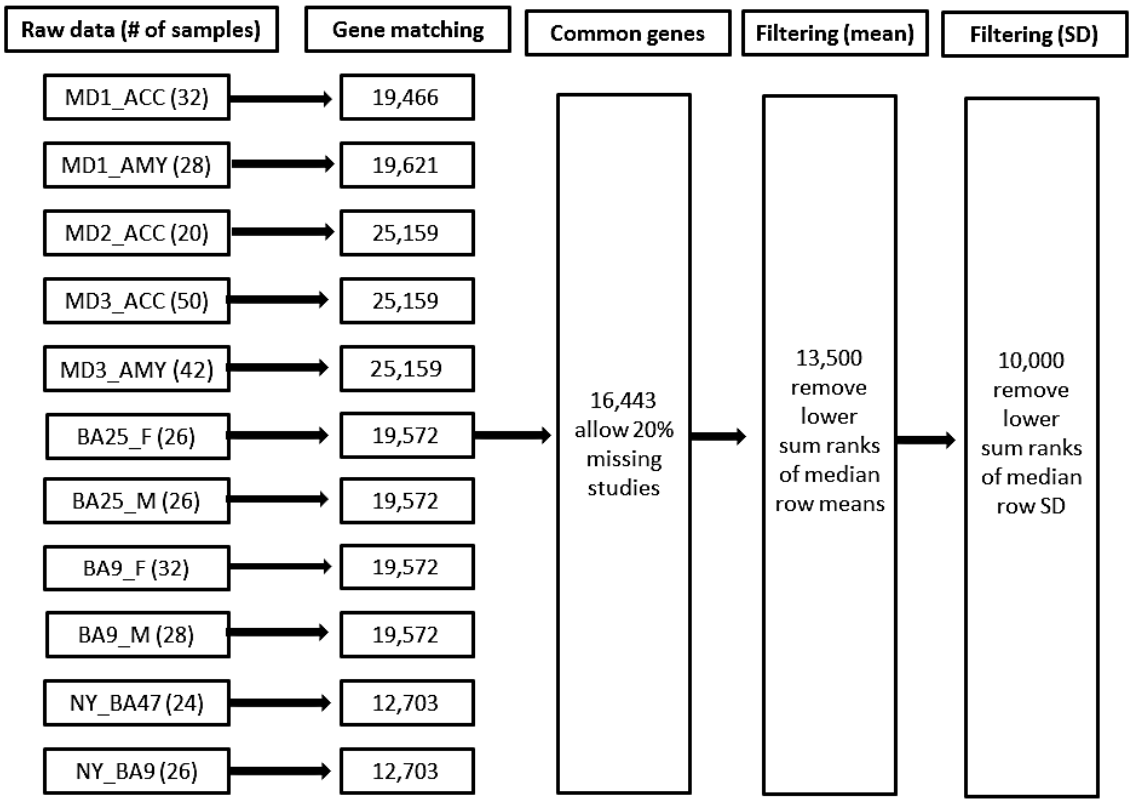

Figure S8: Diagram of pre-processing procedure of 11 MDD transcriptome data sets.

Number of samples and number of matched genes in each single (MDD) study. In matching step, we allowed 20% missing studies, then 16,443 genes were identically matched among 11 studies. 13,500 genes were kept by filtering out lower sum ranks of median row means; 10,000 genes were kept by filtering out lower sum ranks of median row standard deviations.

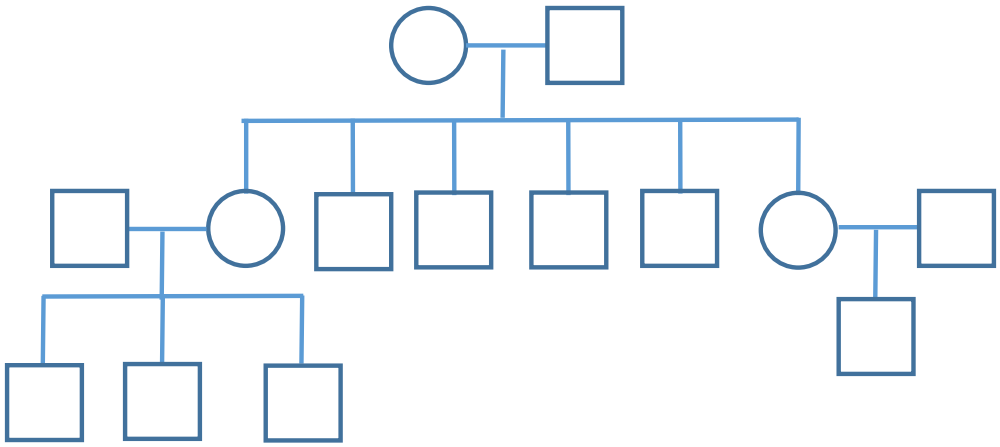

Figure S9: Pedigree of complex family simulated from 1,000 genome project.

Table S1: Detailed data sets description

	Author	Year	Platform	Sample Size (Case/Controls)	Source
Prostate Cancer Studies (Normal v.s Primary)	Welsh	2001	HG-U95A	34(25/9)	public.gnf.org/cancer/
	Singh	2002	HG-U95Av2	102(52/50)	www.broad.mit.edu
	Lapointe	2004	cDNA	103(62/41)	GSE3933
	Yu	2004	HG-U95Av2	83(65/18)	GSE6919
	Varambally	2005	HG-U133 Plus 2	13(7/6)	GSE3325
	Wallace	2008	HG-U133A2	89(69/20)	GSE6956
	Nanni	2006	HG-U133A	30(23/7)	GSE3868
Prostate Cancer Studies (Primary v.s Metastasis	Lapointe	2004	cDNA	71(62/9)	GSE3933
	Varambally	2005	HG-U133 Plus 2	13(7/6)	GSE3325
	Yu	2004	HG-U95Av2	90(65/25)	GSE6919
	Tomlins	2006	cDNA	49(30/19)	GSE6099
Brain Cancer Studies	Freije	2004	HG-U133A,B	85(59/26)	GSE4412
	Phillips	2006	HG-U133A,B	100(76/24)	GSE4271
	Sun	2006	HG-U133 Plus 2	100(81/19)	GSE4290
	Petalidis	2008	HG-U133A	58(39/19)	GSE1993
	Gravendeel	2009	HG-U133 Plus 2	175(159/16)	GSE16011
	Paugh	2010	HG-U133 Plus 2	42(33/9)	GSE19578
	Yamanaka	2006	Agilent	29(22/7)	GSE4381
MDD Studies	MD1_AMY	2009	HG-U133 Plus 2	28(14/14)	Dr. Sibille
	MD1_ACC	2009	HG-U133 Plus 2	32(16/16)	Dr. Sibille
	MD3_ACC	2009	HumanHT-12	44(22/22)	Dr. Sibille
	MD2_ACC_M	2010	HG-U133 Plus 2	18(9/9)	Dr. Sibille
	MD2_ACC_F	2010	HG-U133 Plus 2	26(13/13)	Dr. Sibille
	MD2_DLPFC_M	2010	HG-U133 Plus 2	28(14/14)	Dr. Sibille
	MD2_DLPFC_F	2010	HG-U133 Plus 2	32(16/16)	Dr. Sibille
	MD3_AMY	2009	HumanHT-12	42(21/21)	Dr. Sibille
Lung Disease Studies (IPF)	Pardo	2005	Codelink	24(13/11)	GSE2052
	Yang	2007	Agilent 43K	29(20/9)	GSE5774
	Vuga	2009	Codelink	7(4/3)	GSE10921
	Konishi	2009	Agilent 4x44K	38(23/15)	GSE10667
	KangA	2011	Agilent 4x44K	63(52/11)	Dr. Kaminski
	KangB	2011	Agilent 8x60K	96(75/21)	Dr. Kaminski
	Larsson	2008	HG-U133 Plus 2	12(6/6)	GSE11196
	Emblom	2010	cDNA	58(38/20)	GSE17978
Breast Cancer Studies	Loi	2007	HG-U133A	125	GSE6532
	Miller	2005	HG-U133A,B	236	GSE3494
	Pawitan	2005	HG-U133A,B	159	GSE1456
	Sotiriou2006	2006	HG-U133A	187	GSE2990
	Desmedt	2007	HG-U133A	198	GSE7390
	Wang	2005	HG-U133A	286	GSE2034
	Sotiriou2003	2003	cDNA	110	
	vantVeer	2002	cDNA	97	

Table S2: MetaQC results

Data set	Study	IQC	EQC	CQC_g	CQC_p	AQC_g	AQC_p	Rank
Prostate Cancer Studies (Normal v.s Primary)	1. Welsh	4.38	0.53*	54.63	64.08	18.9	39.09	2.25
	2. Yu	6.64	0.9*	46.91	55.48	14.84	26.2	2.33
	3. Lapointe	2.1*	1.33*	27	53.98	6.28	18.29	3.17
	4. Singh	1.14*	0.95*	14.67	19.21	3.85	18.34	4.17
	5. Varambally	4.38	1.06*	8.7	3.29	2.55	2.41	4.92
	6. Wallace	7.86	0.27*	0*	27.05	0*	3.69	5.33
	7. Nanni	0.75*	0.7*	0.88*	4.2	0.63*	11.45	5.83
Prostate Cancer Studies (Primary v.s Metastasis)	1. Varambally	6.4	0.27*	16.88	23.86	5.5	11.66	1.5
	2. Yu	4.74	0.94*	6.77	13.73	1.43*	6.4	2
	3. Lapointe	3.3	0.8*	2.91	4.08	2.95	5.95	2.67
	4. Tomlins	1.3*	0.51*	0.21*	0.21*	0.1*	0.41*	3.83
Brain Cancer Studies	1. Sun	4.96	2.64	151.63	128.5	61.12	48.82	1.5
	2. Petalidis	4.24	1.17*	148.97	122.39	56.74	75.83	2.83
	3. Freije	5.27	2.52	89.34	68.09	43.31	20.49	3
	4. Phillips	4.81	1.73*	84.93	56	37.22	25.31	3.83
	5. Gravendeel	6.27	1.13*	38.53	48.98	11.9	35.74	4.17
	6. Paugh	1.51*	1.26*	1.62*	0.17*	1.7*	1.77*	6
	7. Yamanaka	0.1*	0.56*	0.92*	0.94*	1.85*	0.31*	6.67
MDD Studies	1. MD2_ACC_F	8.48	1.08*	34.48	54.49	11.9	10.6	1.83
	2. MD2_DLPFC_F	7.87	1.13*	34.58	32.29	6.33	6.91	2.67
	3. MD2_DLPFC_M	2.55	2.08*	24.36	46.97	3.54	20.54	3
	4. MD1_ACC_M	5.03	0.45*	23.25	50.38	4.29	10.74	3.67
	5. MD3_ACC_F	0.74*	1.05*	9.33	9.31	4.8	4.62	5.5
	6. MD2_ACC_M	2.99	1.04*	7.41	9.4	3.36	0.96*	5.83
	7. MD1_AMY_M	1.97*	0.11*	5.47	23.76	1.93*	7.83	6.17
	8. MD3_AMY_F	1.56*	0.96*	0.96*	0.15*	0.38*	2.31	7.33
Lung Disease Studies (IPF)	1. KangA	6.64	0.34*	140.41	85.47	39.01	40.71	2.17
	2. KangB	5.46	0.64*	94.08	45.06	27.4	22.56	2.33
	3. Konishi	6.76	0.77*	17.99	31.45	5.99	21.42	3
	4. Yang	4.07	0.44*	26.61	23.7	9.57	18.41	4.17
	5. Pardo	4.44	0.35*	15.6	29.98	14.56	17.09	4.5
	6. Vuga	2.28	0.39*	1.41*	17.32	1.02*	14.5	6
	7. Larsson	1.85*	1.32*	0.54*	4.83	0.12*	1.26*	6.33
	8. Emblom	0.03*	0.19*	1.68*	0.07*	0.68*	0.56*	7.5
Breast Cancer Studies	1. Pawitan	3.63	4	29.79	116.82	21.99	83.85	2.25
	2. Loi	6.64	4	13.9	66.34	7.32	62.05	2.58
	3. Sotiriou2006	1.3*	0.38*	49.91	134.6	14.3	72.17	3.33
	4. Miller	6.14	4	7.64	47.17	4.24	30.76	3.58
	5. Desmedt	6.26	4	5.09	14.94	3.24	17.21	4.42
	6. Wang	6.03	3.52	0.75*	25.72	2.48	26.01	5.17
	7. Sotiriou2003	2.15*	1.37*	0.28*	4.62	0.07*	2.58	6.83
	8. vantVeer	0.03*	0.26*	0.14*	1.83*	0.15*	0.8*	7.83

Table S3: Data sets and number of matched genes

Disease	# of studies	# of studies passed MetaQC	Comparison	# of matched genes
Prostate cancer	7	6	Binary (normal vs. primary)	6,940
Prostate cancer	4	3	Binary (primary vs. metastastasis)	4,260
Brain cancer	7	5	Binary (AA vs. GBM) (AA vs. GBM)	6,019
Major Deoressive Disorder (MDD)	8	6	Binary (Normal vs. MDD)	6,000
Idiopathic Pulmonary Fibrosis (IPF)	8	6	Binary (Normal vs. IPF)	5,481
Breast cancer	8	6	Survival time (Relapse free survival)	10,688

Based on the QC, the study "Nanni" was removed from 7 prostate cancer studies comparing normal and primary cancer patients; the study "Tomlins" was removed from 4 prostate cancer studies com-paring primary cancer patients and metastasis cancer patients. In the 7 brain cancer studies, the "Paugh" and "Yamanaka" studies were removed. In the case of major depression disorder (MDD) studies, we removed the study "MD3 AMY F". Studies "Larsson" and "Em-blom" were removed from 8 lung disease studies. In breast cancer survival data sets, two cDNA data sets "Sotiriou2003" and "vantVeer" were removed.

Table S4: Mean standardized rank (MSR) and aggregated standardized rank (ASR) for detection capability

	Fisher	AW	Stouffer	minP	FEM	RankSsum	rOP	RankProd	maxP	REM	SR	PR
Prostate cancer (normal v.s. primary)	0.08	0.17	0.25	0.34	0.42	0.51	0.63	0.61	0.78	0.80	0.98	0.93
Prostate cancer (primary v.s. metastasis)	0.08	0.17	0.25	0.33	0.55	0.45	0.51	0.71	0.70	0.83	0.92	1.00
Brain cancer (AA v.s. GBM)	0.08	0.17	0.27	0.40	0.40	0.54	0.57	0.65	0.75	0.75	0.92	1.00
MDD	0.17	0.22	0.26	0.51	0.55	0.48	0.49	0.52	0.72	0.83	0.85	0.90
IPF	0.08	0.17	0.27	0.32	0.42	0.49	0.67	0.60	0.81	0.76	0.97	0.95
Breast Cancer	0.13	0.33	0.41	0.38	NA	NA	0.63	NA	0.75	NA	0.99	0.89
Aggregated standardized ranks	0.11	0.20	0.29	0.38	0.47	0.49	0.58	0.62	0.75	0.79	0.94	0.95

Table S5: Mean standardized rank (MSR) and aggregated standardized rank (ASR) for biological association

	Stouffer	Fisher	AW	PR	rOP	SR	minP	RankProd	FEM	maxP	REM	RankSum
Prostate cancer (normal v.s. primary)	0.42	0.38	0.38	0.41	0.44	0.53	0.49	0.58	0.63	0.69	0.75	0.80
Prostate cancer (primary v.s. metastasis)	0.36	0.38	0.39	0.42	0.36	0.52	0.49	0.55	0.68	0.72	0.78	0.86
Brain cancer (AA v.s. GBM)	0.44	0.45	0.40	0.54	0.42	0.64	0.47	0.50	0.56	0.65	0.72	0.73
MDD	0.27	0.29	0.40	0.28	0.42	0.28	0.56	0.68	0.80	0.84	0.81	0.88
IPF	0.35	0.41	0.36	0.40	0.45	0.48	0.52	0.63	0.66	0.72	0.76	0.78
Breast Cancer	0.45	0.45	0.48	0.42	0.52	0.56	0.71	NA	NA	0.92	NA	NA
Aggregated standardized ranks	0.38	0.39	0.40	0.41	0.43	0.50	0.54	0.59	0.67	0.75	0.76	0.81

Table S6: Mean standardized rank (MSR) and aggregated standardized rank (ASR) for stability

	RankProd	RankSum	Fisher	PR	REM	AW	minP	Stouffer	SR	rOP	FEM	maxP
Prostate cancer (normal v.s. primary)	0.08	0.17	0.40	0.38	0.27	0.46	0.70	0.72	0.71	0.92	0.70	1.00
Prostate cancer (primary v.s. metastasis)	0.08	0.17	0.38	0.43	0.35	0.39	0.58	0.66	0.73	0.84	0.90	1.00
Brain cancer (AA v.s. GBM)	0.08	0.17	0.37	0.40	0.30	0.50	0.74	0.68	0.77	0.67	0.83	1.00
MDD	0.04	0.04	0.23	0.19	0.82	0.45	0.45	0.58	0.36	0.67	0.78	0.90
IPF	0.08	0.17	0.38	0.44	0.32	0.52	0.49	0.61	0.76	0.92	0.82	1.00
Breast Cancer	NA	NA	0.18	0.32	NA	0.50	0.51	0.59	0.70	0.80	NA	0.89
Aggregated standardized ranks	0.07	0.14	0.32	0.36	0.41	0.47	0.58	0.64	0.67	0.80	0.81	0.97

Table S7: Mean standardized rank (MSR) and aggregated standardized rank (ASR) for robustness

	minP	AW	Fisher	Stouffer	RankProd	PR	SR	REM	FEM	rOP	maxP	RankSum
Prostate cancer (normal v.s. primary)	0.20	0.27	0.23	0.34	0.56	0.52	0.54	0.59	0.65	0.80	0.79	1.00
Prostate cancer (primary v.s. metastasis)	0.17	0.21	0.24	0.33	0.47	0.53	0.65	0.65	0.70	0.81	0.79	0.94
Brain cancer (AA v.s. GBM)	0.24	0.29	0.31	0.40	0.52	0.54	0.59	0.63	0.57	0.76	0.71	0.94
MDD	0.55	0.51	0.56	0.59	0.33	0.47	0.37	0.55	0.52	0.49	0.58	0.97
IPF	0.21	0.31	0.29	0.35	0.47	0.49	0.62	0.51	0.74	0.76	0.73	1.00
Breast Cancer	0.57	0.60	0.62	0.51	NA	0.49	0.62	NA	NA	0.40	0.70	NA
Aggregated standardized ranks	0.32	0.37	0.37	0.42	0.47	0.51	0.57	0.59	0.64	0.67	0.72	0.97

Table S8: Meta modules and GWAS gene lists(cases and controls)

Cluster	# of genes	Neuroticism	MDD2000+	Mega MDD	Mega bipolar	MDD	Neuropsychiatric Disorder	Neurological disorders and brain phenotypes	Medical illnesses sharing clinical risk with MDD
1	94	1.0000	1.0000	1.0000	0.9055	0.4510	0.8924	0.8395	0.8220
2	114	1.0000	0.2349	1.0000	0.9529	1.0000	0.2977	0.3431	0.7862
3	187	1.0000	0.1919	0.4421	0.3674	0.5863	0.3231	0.1025	0.0138
4	165	0.1287	0.1483	1.0000	0.3726	0.5012	0.1193	0.0206	0.2418
5	132	0.2638	1.0000	0.0616	0.1788	0.1648	0.2689	0.2783	0.0085
6	118	0.2244	0.2469	1.0000	0.3611	0.5722	0.1871	0.2630	0.0997
7	203	0.0183	1.0000	1.0000	0.6054	0.0016	0.0147	0.0086	0.0925
8	233	0.0991	0.4377	1.0000	0.0160	0.5230	0.4444	0.8360	0.0448
9	130	0.2581	1.0000	1.0000	0.7913	1.0000	0.7638	0.9700	0.5708
10	161	0.7094	1.0000	0.0870	0.1290	0.7387	0.4564	0.6628	0.2194
11	215	1.0000	1.0000	1.0000	0.5324	0.6805	0.3474	0.5406	0.3031
12	251	0.0406	0.4572	0.1799	0.4642	0.2199	0.1237	0.0533	0.3904
13	94	1.0000	0.5357	1.0000	0.0152	0.7858	0.5150	0.4018	0.8220

Table S8 (Continued)

14	369	0.5340	0.5419	0.6882	0.7633	0.7186	0.8008	0.8453	0.8263
15	146	0.6735	1.0000	1.0000	0.9446	0.2104	0.5232	0.1982	0.6833
16	154	0.3256	1.0000	1.0000	0.4530	0.4555	0.7367	0.6037	0.9340
17	85	0.4775	1.0000	1.0000	0.4685	0.4008	0.4368	0.6153	0.2051
18	32	1.0000	1.0000	1.0000	0.1441	1.0000	1.0000	0.7005	0.0600
19	204	1.0000	0.6810	0.4713	0.8537	0.8469	0.7052	0.5611	0.9968
20	109	0.5656	0.0577	0.2871	0.6651	0.2581	0.2634	0.1874	0.5807
21	263	0.5969	0.7250	1.0000	0.9644	0.1331	0.3495	0.6833	0.3448
22	53	1.0000	1.0000	0.1512	0.6498	0.2110	0.6272	1.0000	0.9213
23	189	0.4209	1.0000	1.0000	0.9863	0.8148	0.9823	0.9754	0.9793
24	97	1.0000	1.0000	0.2598	0.9837	0.2066	0.3399	0.5916	0.2994
25	68	1.0000	1.0000	1.0000	0.7781	1.0000	1.0000	0.9671	0.8335
26	293	0.8965	1.0000	1.0000	0.5548	0.5155	0.2725	0.3434	0.3939
27	126	1.0000	1.0000	1.0000	0.7707	0.6082	0.5633	0.7316	0.3741
28	181	0.7512	0.6583	0.1061	0.4739	0.3356	0.5812	0.5972	0.3378
29	131	0.6334	1.0000	1.0000	0.9744	0.0607	0.4185	0.5165	0.5784
30	231	0.5258	1.0000	1.0000	0.3574	0.9781	0.9838	0.0930	0.6435
31	224	0.5092	0.1067	1.0000	0.2154	0.1542	0.1782	0.1621	0.3536
32	178	1.0000	0.6557	0.4260	0.7507	0.9468	0.5633	0.2519	0.0726
33	117	1.0000	1.0000	1.0000	0.8679	0.8535	0.8488	0.9786	0.9964
34	236	0.8380	1.0000	0.5223	0.8624	0.7401	0.2262	0.5955	0.8679

Table S8 (Continued)

35	88	0.0287	0.0339	0.0001	0.0284	0.0537	0.0078	0.0293	0.0524
36	149	0.6810	1.0000	1.0000	0.4190	0.9138	0.8481	0.6814	0.9228
37	156	0.3312	1.0000	1.0000	0.0277	0.9233	0.8721	0.1851	0.3050
38	122	0.6070	1.0000	0.0536	0.9645	0.8651	0.5343	0.5629	0.3433
39	91	1.0000	0.5241	1.0000	0.0041	1.0000	0.9753	0.9930	0.9874
40	142	0.6633	0.3189	1.0000	0.9373	0.6731	0.6692	0.4926	0.4969
41	192	0.7716	1.0000	1.0000	0.0542	0.6043	0.2308	0.3575	0.4075
42	113	1.0000	1.0000	1.0000	0.9512	1.0000	0.9900	0.9922	0.9714
43	114	0.5820	0.2349	0.2981	0.6992	1.0000	0.9904	0.9343	0.7862
44	186	1.0000	1.0000	1.0000	0.5050	0.8077	0.7529	0.9428	0.9934
45	112	0.5755	1.0000	1.0000	0.3167	0.8409	0.6540	0.3230	0.6051
46	198	1.0000	1.0000	0.4611	0.5769	0.6253	0.3853	0.0378	0.9066
47	167	1.0000	0.1521	1.0000	0.6939	0.5093	0.7979	0.8841	0.6691
48	148	0.6786	0.3368	1.0000	0.9863	0.6951	0.8444	0.9883	0.5413
49	119	0.5978	0.6220	1.0000	0.7307	1.0000	0.9922	0.9990	0.9776
50	117	0.2216	1.0000	1.0000	0.5338	0.5676	0.8488	0.8777	0.9964

BIBLIOGRAPHY

GR Abecasis, David Altshuler, A Auton, LD Brooks, RM Durbin, Richard A Gibbs, Matt E Hurles, Gil A McVean, DR Bentley, A Chakravarti, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

RH Belmaker and Galila Agam. Major depressive disorder. *New England Journal of Medicine*, 358(1):55–68, 2008.

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

Allan Birnbaum. Combining independent tests of significance*. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

Catherine Boileau, Dong-Chuan Guo, Nadine Hanna, Ellen S Regalado, Delphine Detaint, Limin Gong, Mathilde Varret, Siddharth K Prakash, Alexander H Li, Hyacintha d'Indy, et al. Tgfb2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of marfan syndrome. *Nature genetics*, 44(8):916–921, 2012.

Dorret I Boomsma, Cisca Wijmenga, Eline P Slagboom, Morris A Swertz, Lennart C Karssen, Abdel Abdellaoui, Kai Ye, Victor Guryev, Martijn Vermaat, Freerk van Dijk, et al. The genome of the netherlands: design, and project goals. *European Journal of Human Genetics*, 2013.

Ingwer Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

Rainer Breitling and Pawel Herzyk. Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *Journal of bioinformatics and computational biology*, 3(05):1171–1189, 2005.

Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

Anna Campain and Yee H Yang. Comparison study of microarray meta-analysis methods. *BMC bioinformatics*, 11(1):408, 2010.

Lun-Ching Chang, Hui-Min Lin, Etienne Sibille, and George C Tseng. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC bioinformatics*, 14(1):368, 2013.

Lun-Ching Chang, Stephane Jamain, Chien-Wei Lin, Dan Rujescu, George C Tseng, and Etienne Sibille. A conserved bdnf, glutamate-and gaba-enriched gene module related to human depression identified by coexpression meta-analysis and dna variant genome-wide association studies. *PloS one*, 9(3):e90980, 2014.

Wei Chen, Bingshan Li, Zhen Zeng, Serena Sanna, Carlo Sidore, Fabio Busonero, Hyun Min Kang, Yun Li, and Gonçalo R Abecasis. Genotype calling and haplotyping in parent-offspring trios. *Genome research*, 23(1):142–151, 2013.

Jung Kyoon Choi, Ungsik Yu, Sangsoo Kim, and Ook Joon Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90, 2003.

PV Choudary, M Molnar, SJ Evans, H Tomita, JZ Li, MP Vawter, RM Myers, WE Bunney, H Akil, SJ Watson, et al. Altered cortical glutamatergic and gabaergic signal transmission with glial involvement in depression. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15653–15658, 2005.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

AS Cristino, SM Williams, Z Hawi, JY An, MA Bellgrove, CE Schwartz, L da F Costa, and C Claudianos. Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. *Molecular psychiatry*, 2013.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Radu Dobrin, Jun Zhu, Cliona Molony, Carmen Argman, Mark L Parrish, Sonia Carlson, Mark F Allan, Daniel Pomp, Eric E Schadt, et al. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*, 10(5):R55, 2009.

Jonathan M Dreyfuss, Mark D Johnson, and Peter J Park. Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. *Molecular cancer*, 8(1):71, 2009.

Ronald S Duman and Lisa M Monteggia. A neurotrophic model for stress-related mood disorders. *Biological psychiatry*, 59(12):1116–1127, 2006.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Laura L Elo, Henna Järvenpää, Matej Orešič, Riitta Lahesmaa, and Tero Aittokallio. Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process. *Bioinformatics*, 23(16):2096–2103, 2007.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing, 1925.

Chris Gaiteri, Jean-Philippe Guilloux, David A Lewis, and Etienne Sibille. Altered gene synchrony suggests a combined hormone-mediated dysregulated state in major depression. *PloS one*, 5(4):e9970, 2010.

Chris Gaiteri, Ying Ding, George C. Tseng, and Etienne Sibille. Beyond modules & hubs: Investigating pathogenic molecular mechanisms of brain disorders through gene coexpression networks. Submitted, 2013.

Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. *Bioinformatics and computational biology solutions using R and Bioconductor*, volume 746718470. Springer New York, 2005.

Jean-Philippe Guilloux, Gaelle Douillard-Guilloux, Rama Kota, Xingbin Wang, AM Gardier, Keri Martinowich, George C Tseng, David A Lewis, and Etienne Sibille. Molecular evidence for bdnf-and gaba-related dysfunctions in the amygdala of female subjects with major depression. *Molecular psychiatry*, 17(11):1130–1142, 2011.

Deborah S Hasin, Renee D Goodwin, Frederick S Stinson, and Bridget F Grant. Epidemiology of major depressive disorder: results from the national epidemiologic survey on alcoholism and related conditions. *Archives of general psychiatry*, 62(10):1097, 2005.

Karin Hek, Ayse Demirkan, Jari Lahti, Antonio Terracciano, Alexander Teumer, Marilyn C Cornelis, Najaf Amin, Erin Bakshis, Jens Baumert, Jingzhong Ding, et al. A genome-wide association study of depressive symptoms. *Biological psychiatry*, 2013.

Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, 2008.

Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, 2006.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985.

Dongwan D Kang, Etienne Sibille, Naftali Kaminski, and George C Tseng. Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic acids research*, 40(2):e15–e15, 2012.

Timothy A Klempan, Adolfo Sequeira, L Canetti, Aleksandra Lalovic, C Ernst, et al. Altered expression of genes involved in atp biosynthesis and gabaergic neurotransmission in the ventral prefrontal cortex of suicides with and without major depression. *Molecular psychiatry*, 14(2):175–189, 2007.

David J Kupfer, Jules Angst, Michael Berk, Faith Dickerson, Sophia Frangou, Ellen Frank, Benjamin I Goldstein, Allison Harvey, Fouzia Laghrissi-Thode, Marion Leboyer, et al. Advances in bipolar disorder: selected sessions from the 2011 international conference on bipolar disorder. *Annals of the New York Academy of Sciences*, 1242(1):1–25, 2011.

Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385–394, 2006.

Homin K Lee, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14(6):1085–1094, 2004.

Cathryn M Lewis, Mandy Y Ng, Amy W Butler, Sarah Cohen-Woods, Rudolf Uher, Katrina Pirlo, Michael E Weale, Alexandra Schosser, Ursula M Paredes, Margarita Rivera, et al. Genome-wide association study of major recurrent depression in the uk population. *American Journal of Psychiatry*, 167(8):949–957, 2010.

Bingshan Li, Wei Chen, Xiaowei Zhan, Fabio Busonero, Serena Sanna, Carlo Sidore, Francesco Cucca, Hyun M Kang, and Gonçalo R Abecasis. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS genetics*, 8(10): e1002944, 2012.

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

Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

Bernhard Luscher, Qiuying Shen, and Nadia Sahir. The gabaergic deficit hypothesis of major depressive disorder. *Molecular psychiatry*, 16(4):383–406, 2010.

Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

Nathaniel FG Martin and James W England. *Mathematical theory of entropy*, volume 12. Cambridge University Press, 2011.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Timothy I Mueller, Andrew C Leon, Martin B Keller, David A Solomon, Jean Endicott, William Coryell, Meredith Warshaw, and Jack D Maser. Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *American Journal of Psychiatry*, 156(7):1000–1006, 1999.

P Muglia, F Tozzi, NW Galwey, C Francks, R Upmanyu, XQ Kong, A Antoniades, E Domenici, J Perry, Stéphane Rothen, et al. Genome-wide association study of recurrent major depressive disorder in two european case–control cohorts. *Molecular psychiatry*, 15 (6):589–601, 2008.

Dominique L Musselman, Dwight L Evans, and Charles B Nemeroff. The relationship of depression to cardiovascular disease: epidemiology, biology, and treatment. *Archives of general psychiatry*, 55(7):580, 1998.

Benjamin M Neale, Yan Kou, Li Liu, Avi Maayan, Kaitlin E Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, 2012.

Eric J Nestler, Michel Barrot, Ralph J DiLeone, Amelia J Eisch, Stephen J Gold, and Lisa M Monteggia. Neurobiology of depression. *Neuron*, 34(1):13–25, 2002.

Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1): 30–35, 2009.

Sarah B Ng, Abigail W Bigham, Kati J Buckingham, Mark C Hannibal, Margaret J McMillin, Heidi I Gildersleeve, Anita E Beck, Holly K Tabor, Gregory M Cooper, Heather C Mefford, et al. Exome sequencing identifies mll2 mutations as a cause of kabuki syndrome. *Nature genetics*, 42(9):790–793, 2010a.

Sarah B Ng, Deborah A Nickerson, Michael J Bamshad, and Jay Shendure. Massively parallel sequencing and rare disease. *Human molecular genetics*, 19(R2):R119–R124, 2010b.

Brian J ORoak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D Smith, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397): 246–250, 2012.

Jurg Ott, Yoichiro Kamatani, and Mark Lathrop. Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12(7):465–474, 2011.

Art B Owen. Karl pearson's meta-analysis revisited. *The annals of statistics*, pages 3867–3892, 2009.

An Pan, NaNa Keum, Olivia I Okereke, Qi Sun, Mika Kivimaki, Richard R Rubin, and Frank B Hu. Bidirectional association between depression and metabolic syndrome a systematic review and meta-analysis of epidemiological studies. *Diabetes Care*, 35(5): 1171–1180, 2012.

Gang Peng, Yu Fan, Timothy B Palculict, Peidong Shen, E Cristy Ruteshouser, Aung-Kyaw Chi, Ronald W Davis, Vicki Huff, Curt Scharfe, and Wenyi Wang. Rare variant detection using family-based sequencing analysis. *Proceedings of the National Academy of Sciences*, 110(10):3985–3990, 2013.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. Wiley. com, 2010.

Giuseppe Pilia, Wei-Min Chen, Angelo Scuteri, Marco Orrú, Giuseppe Albai, Mariano Dei, Sandra Lai, Gianluca Usala, Monica Lai, Paola Loi, et al. Heritability of cardiovascular and personality traits in 6,148 sardinians. *PLoS genetics*, 2(8):e132, 2006.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Grazyna Rajkowska, Gillian O'Dwyer, Zsofia Teleki, Craig A Stockmeier, and Jose Javier Miguel-Hidalgo. Gabaergic neurons immunoreactive for calcium binding proteins are reduced in the prefrontal cortex in major depression. *Neuropsychopharmacology*, 32(2): 471–482, 2006.

Adaikalavan Ramasamy, Adrian Mondry, Chris C Holmes, and Douglas G Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, 5(9):e184, 2008.

Marcella Rietschel, Manuel Mattheisen, Josef Frank, Jens Treutlein, Franziska Degenhardt, René Breuer, Michael Steffens, Daniela Mier, Christine Esslinger, Henrik Walter, et al. Genome-wide association-, replication-, and neuroimaging study implicates¡ i¿ homer1¡/i¿ in the etiology of major depression. *Biological psychiatry*, 68(6):578–585, 2010.

Stephan Ripke, Naomi R Wray, Cathryn M Lewis, Steven P Hamilton, Myrna M Weissman, Gerome Breen, Enda M Byrne, Douglas HR Blackwood, Dorret I Boomsma, Sven Cichon,

et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4):497–511, 2012.

Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328 (5978):636–639, 2010.

Stephan J Sanders, Michael T Murtha, Abha R Gupta, John D Murdoch, Melanie J Raubeson, A Jeremy Willsey, A Gulhan Ercan-Sencicek, Nicholas M DiLullo, Neelroop N Parikshak, Jason L Stein, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012.

Stephen F Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J Daly, and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583, 2005.

Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

Adolfo Sequeira, Firoza Mamdani, Carl Ernst, Marquis P Vawter, William E Bunney, Veronique Lebel, Sonia Rehal, Tim Klempan, Alain Gratton, Chawki Benkelfat, et al. Global brain gene expression analysis links glutamatergic and gabaergic alterations to suicide and major depression. *PLoS One*, 4(8):e6585, 2009.

J Shi, JB Potash, JA Knowles, MM Weissman, W Coryell, WA Scheftner, WB Lawson, JR DePaulo, PV Gejman, AR Sanders, et al. Genome-wide association study of recurrent early-onset major depressive disorder. *Molecular psychiatry*, 16(2):193–201, 2010.

Si I Shyn, J Shi, JB Kraft, JB Potash, JA Knowles, MM Weissman, HA Garriock, JS Yokoyama, PJ McGrath, EJ Peters, et al. Novel loci for major depression identified by genome-wide association study of sequenced treatment alternatives to relieve depression and meta-analysis of three studies. *Molecular psychiatry*, 16(2):202–215, 2009.

Etienne Sibille and Beverly French. Biological substrates underpinning diagnosis of major depression. *The international journal of neuropsychopharmacology/official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)*, pages 1–17, 2013.

Etienne Sibille, Victoria Arango, Hanga C Galfalvy, Paul Pavlidis, Loubna Erraji-Benchekroun, Steve P Ellis, and J John Mann. Gene expression profiling of depression and suicide in human prefrontal cortex. *Neuropsychopharmacology*, 29(2):351–361, 2004.

Pamela Sklar, Stephan Ripke, Laura J Scott, Ole A Andreassen, Sven Cichon, Nick Craddock, Howard J Edenberg, John I Nurnberger, Marcella Rietschel, Douglas Blackwood, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nature genetics*, 43(10):977, 2011.

Chi Song, George C Tseng, et al. Hypothesis setting and order statistic for robust genomic meta-analysis. *The Annals of Applied Statistics*, 8(2):777–800, 2014.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

Samuel A Stouffer. A study of attitudes. *Scientific American*, 180(5):11, 1949.

Patrick F Sullivan, Michael C Neale, and Kenneth S Kendler. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, 157(10): 1552–1562, 2000.

Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

Leonard Henry Caleb Tippett et al. The methods of statistics. an introduction mainly for workers in the biological sciences. *The Methods of Statistics. An Introduction mainly for Workers in the Biological Sciences.*, 1931.

Adam Tripp, Hyunjung Oh, Jean-Philippe Guilloux, Keri Martinowich, David A Lewis, and Etienne Sibille. Brain-derived neurotrophic factor signaling and subgenual anterior cingulate cortex dysfunction in major depressive disorder. *The American journal of psychiatry*, 169(11):1194, 2012.

George C Tseng. Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, 23(17):2247–2255, 2007.

George C Tseng, Debashis Ghosh, and Eleanor Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9): 3785–3799, 2012.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

Edwin JCG van den Oord, Po-Hsiu Kuo, Annette M Hartmann, B Todd Webb, Hans-Jurgen Moller, John M Hettema, Ina Giegling, József Bukszár, and Dan Rujescu. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Archives of general psychiatry*, 65(9):1062, 2008.

Xingbin Wang, Dongwan D Kang, Kui Shen, Chi Song, Shuya Lu, Lun-Ching Chang, Serena G Liao, Zhiguang Huo, Shaowu Tang, Ying Ding, et al. An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, 28(19):2534–2536, 2012a.

Xingbin Wang, Yan Lin, Chi Song, Etienne Sibille, and George C Tseng. Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder. *BMC bioinformatics*, 13(1):52, 2012b.

Myrna M Weissman, Roger Bland, Peter R Joyce, Stephen Newman, J Elisabeth Wells, and Hans-Ulrich Wittchen. Sex differences in rates of depression: cross-national perspectives. *Journal of affective disorders*, 29(2):77–84, 1993.

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

NR Wray, ML Pergadia, DHR Blackwood, BWJH Penninx, SD Gordon, DR Nyholt, S Ripke, DJ Macintyre, KA McGhee, AW Maclean, et al. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular psychiatry*, 17(1):36–48, 2010.

Wei Wu, Nilesh Dave, George Tseng, Thomas Richards, Eric Xing, and Naftali Kaminski. Comparison of normalization methods for codelink bioarray data. *BMC bioinformatics*, 6 (1):309, 2005.

Xinan Yang, Stefan Bentink, Stefanie Scheid, and Rainer Spang. Similarities of ordered gene lists. *Journal of Bioinformatics and Computational Biology*, 4(03):693–708, 2006.

Bin Zhang, Steve Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

Xiaofeng Zhu, Tao Feng, Yali Li, Qing Lu, and Robert C Elston. Detecting rare variants for complex traits using family and unrelated data. *Genetic epidemiology*, 34(2):171–187, 2010.