STATISTICAL ISSUES IN META-ANALYSIS FOR IDENTIFYING SIGNATURE GENES IN THE INTEGRATION OF MULTIPLE GENOMIC STUDIES

by

Jia Li

BS, Dongbei University of Finance and Economics, China, 2001

MS, The University of Alabama, 2004

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jia Li

It was defended on

September 11th 2008

and approved by

Dissertation Advisor:
George C. Tseng, ScD
Assistant Professor
Department of Biostatistics
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Vanathi Gopalakrishnan, PhD
Assistant Professor
Department of Biomedical Informatics
School of Medicine
University of Pittsburgh

Committee Member:
Sati Mazumdar, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lisa Weissfeld, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

STATISTICAL ISSUES IN META-ANALYSIS FOR IDENTIFYING SIGNATURE GENES IN THE INTEGRATION OF MULTIPLE GENOMIC STUDIES

Jia Li, PhD

University of Pittsburgh, 2008

With the availability of tons of expression profiles, the need for meta-analyses to integrate different types of microarray data are obvious. For detection of differentially expressed genes, most of the current efforts are focused on comparing and evaluating gene lists obtained from each individual dataset. Several statistical meta-analysis methods, including Fisher's method and the random effects model, have been proposed but the statistcal framework is not often rigorously formulated for evaluation and comparison. In this dissertation, we attempt to formulate meta-analysis in genomic studies and develop systematic integration methods for two-class studies and multi-class studies.

First, we tackle two often-asked biological questions:"Which genes are significant in one or more data sets?" and "Which genes are significant in all data sets?". We illustrate two statistical hypothesis settings and propose an optimally weighted statistic and compare to classical Fisher's equally weighted statistic and Tippett's minimum p-value statistic. Generally there exists no uniformly most powerful test and we show that all of the three methods are admissible under simplified Gaussian assumptions. Furthermore, the optimally weighted statistic maintains advantages of the two classical methods and consistently performs well when the two methods perform poorly in respective extreme alternative hypotheses. The optimal weights provide natural categorization of the detected genes to facilitate further biological investigation. We demonstrate the comparison and advantages of optimally weighted

statistic by power analysis, simulations and two real data analyses of combining multi-tissue energy metabolism mouse data sets and prostate cancer data sets.

Second, we propose two methods for identifying biomarkers of concordant patterns across studies, when there are more than two classes in each study. So far, published meta-analysis methods for this purpose mostly consider two-class comparison. Methods for combining multi-class studies and pattern concordance are rarely explored. We first consider a natural extension of combining p-values from the traditional ANOVA model. Since p-values from ANOVA do not reflect pattern information, we propose a multi-class correlation measure (MCC) under equal-weight bivariate mixture model to specifically seek for biomarkers of concordant patterns across a pair of studies. For both approaches, we focus to identify biomarkers differentially expressed in all studies (ANOVA-maxP, min-MCC). Both ANOVA-maxP and min-MCC are evaluated by simulation studies and by applications to a multi-tissue mouse metabolism data set and a multi-platform mouse trauma data set.

Finally, we develop a "genomeMeta" R package. genomeMeta produces visualization and summarization of biomarkers identified by methods that we describe and propose in this dissertation.

This work could improve public health by providing more effective methodologies for biomarker detection in the integration of multiple genomic studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I thank my primary advisor Dr. George C. Tseng for his vision, leadership, and friendship. Without his selfless support and gentle guidance, this happy ending/begining would simply not have been possible. He introduced me to the field of microarray data analysis and inspired me to develop statistical methods to the integrative analysis of multiple microarray studies. He has elevated me from rudiments to knowledge by sharing in his wisdom.

My committee members, Drs. Sati Mazumdar, Lisa Weissfeld, and Vanathi Gopalakrishnan, were amazingly constructive in advice and accommodating in schedule. They provided insightful and critical suggestions that improved many aspects of this work.

Hearty thanks to Dr. Lan Kong, Dr. Steven Belle and Dr. Abdus Wahed, who provided financial support through my Ph.D study. The projects tremendously enriched my doctoral training. I thank them for their guidance, advice and numerous opportunities that they has provided over the last three years. I would like to express my appreciation to Dr. Zuzana Swigonova for providing one of the data sets in this dissertation.

Sincere thanks also to all the people of Bioinformatics and Statistical Learning Group lead by Dr. George Tseng. We have been working together for two years. We shared our ideas and experiences. And thanks to all the statistic staffs and friends. Their support and encouragement are indispensable.

Finally, I thank my husband, Shaochun, for supporting my decision to come to Pittsburgh and the many sacrifices he endured to accompany this decision. And I thank my parents who live thousands of miles way for their closest care and love.

1.0 INTRODUCTION

In this introductory chapter we start in section 1.1 by describing the data structure of microarray data particulary in the context of multiple microarray data sets. We further describe an example to help you understand the data structure. In section 1.2, we use this example to illustrate the idea of meta-analysis methods.

1.1 MICROARRAY DATA STRUCTURE AND AN EXAMPLE

Microarray technology allows researchers to examine the expression of thousands of genes in parallel. With microarray experiments becoming more and more mature and prevalent, the integration of experimental data sets from multiple laboratories and experimental techniques has become one of the most challenging tasks in bioinformatic and genomic research. In contrast to traditional epidemiological or psychological studies, microarrays monitor gene expression for thousands of genes simultaneously - a process that brings challenges to integrative analysis. The standard data structure for a set of microarray studies used in this dissertation is provided here.

Let y_{gsk} denote the gene expression for the gth gene in the sth sample of the kth study, $g = 1, \ldots, G, s = 1, \ldots, S_k, k = 1, \ldots, K$. Sample annotations are denoted as r_{sk}, where in any given study k, common microarray designs can include (1) $r_{sk} \in \{0, 1\}$ (two-class comparison); (2) $r_{sk} \in \{1, 2, \ldots, J\}$ (multi-class comparison); (3) $r_{sk} = (t_{sk}, c_{sk})$ (t_{sk}: survival time; c_{sk}: censoring status); (4) $r_{sk} \in R$ (time series). Given study k, suppose an appropriate test statistic T_k is chosen for the data structure $\{r_{sk}; 1 \leq s \leq S_k\}$ and the resulting p-values for each gene g (denoted as p_{gk}) can be derived from the observed expression intensities

Table 1: An example.

Tissue type (k)	Brown fat			Liver			Heart			Skeletal			total
Genotype[1] (r_{sk})	1	2	3	1	2	3	1	2	3	1	2	3	
Number of arrays	4	4	4	4	4	4	3	4	4	3	3	3	44

[a]1: Wild type; 2: VLCAD -/-; 3: LCAD -/-.

$\{y_{gsk}, 1 \leq s \leq S_k\}$.

Example: The below example of integrative analysis of microarray studies is the multi-tissue analysis on variability in gene expression profiles across four tissues (heart, skeletal muscle, liver, and brown fat) as a consequence of single enzyme deficiency in the mitochondrial-oxidation pathway. Using VLCAD-deficient (VLCAD-/-), LCAD-deficient (LCAD-/-), and wild type (VLCAD+/+, LCAD+/+) mouse models, we apply whole genome microarray methods to investigate tissue dependent responses to single defects in these mitochondrial-oxidation enzymes. Thus, in this example, each study represents a tissue ($K = 4$). Three to four individuals are sampled per genotype and tissues, accounting for 45 samples in total. One out of the 45 hybridizations (wild type heart sample) have to be excluded due to detected significant degradation of RNA (Table 1), and 14495 genes are investigated ($G = 14495$). VLCAD is involved in degradation of long- and very long-chain unsaturated substrates from C14 to C20, with optimum for palmitoyl-CoA (C16:0-CoA). It is ubiquitously expressed in human tissues with high levels in liver, heart, and skeletal muscle. To compare the variation in gene expressions between wild type and VLCAD-/-, we can set $r_{sk} \in \{1, 2\}$ and analyze a total of 29 samples. To investigate gene variation in wild type, VLCAD-/- and LCAD-/-, we can set $r_{sk} \in \{1, 2, 3\}$ analyze the entire 44 samples.

1.2 META-ANALYSIS

As the global expression analysis by microarray technology becomes more and more prevalent, systematic information integration of multiple genomic studies brings new statistical challenges. The challenge to statisticians is to develop effective methodologies for combining information from related studies. This type of analysis is often called meta-analysis.

The main aim of meta-analysis is first, to provide a more powerful statistical test of a specific study hypothesis than a test provided by each study separately, and second, to provide validation for findings from the individual study. Meta-analysis has a long history, which have been widely used in epidemiology and evidence-based medicine ever since early twenties. Meta-analysis methods can be categorized into two traditions. One is to combine statistical significance from each individual study, and the other is to combine statistical estimates particularly effect sizes from each individual study. The representative methodologies for the first tradition are Fisher's method (Fisher, 1932), Tippett's minimum p-value (Tippett,1931) and Wilkinson's maximum p-value (Wilkson, 1951). The representative methodologies for the second tradition are fixed and random effects models. Details of traditional methodologies are described in section 1 of chapter 2. There are advantages and disadvantages for both of the traditions. Briefly, methods based on combining significance are free of distributional assumptions, but do not support inferences about magnitudes and directions. On the other hand, methods based on combining estimates provide information about magnitudes and directions, but are more stringent on assumptions. And they can not be easily extended to more complicated studies such as studies with multiple classes.

With microarray experiments becoming more mature and more prevalent, the integration of experimental data sets from multiple laboratories and experimental techniques has become one of the most challenging tasks in bioinformatic and genomic research. Currently "meta-analysis" in the biological literature contains a widespread use of naive "intersection" and/or "union" operations on differentially expressed gene lists obtained from individual studies by certain criteria (e.g. False Discovery Rate ≤ 0.05) (Segal et al., 2004, Borovecki et al., 2005, Cardoso et al., 2007, Pirooznia et al., 2007, and many more). One can quickly note that intersections are often too conservative and unions are anti-conservative, especially when the number of studies increases. For instance, Table 2 lists two-sample moderated t statistics, several often used meta-analysis methods and corresponding p-values for comparison between wild type and VLCAD-/- in each tissue for several genes in the mouse energy metabolism example in section 1.1. For the purpose of illustration, we do not worry about multiple comparisons in this example. Using the intersection method, individual p-values ≤ 0.05 in all tissues, is too conservative and all of the top three genes listed in table 2 are not identified.

Table 2: Several examples for meta analysis methods.

	P-values of moderated t statistic				statistics(p-values) for meta analysis methods			
	Brown fat	Liver	Heart	Skeletal	Fisher	minP	maxP	REM
1417025_at	0.636	0.008	0.001	0.024	16.56(7E-06)	0.001(0.002)	0.636(0.181)	-0.056(0.952)
1424039_at	0.006	0.005	0.005	0.526	16.56(7E-06)	0.005(0.018)	0.526(0.092)	-1.953(0.051)
1415994_at	0.668	0.815	0.694	0.002	7.32(0.007)	0.002(0.007)	0.815(0.453)	-0.431(0.666)
1416531_at	0.001	0.008	0.008	0.005	21.69(1E-05)	0.001(0.005)	0.008(1E-20)	4.978(6E-07)

Note: P-values are obtained by permutation analysis.
Fisher: Fisher's method; minP: Tippett's minimum p-value method; maxP: Wilkinson's maximum p-value; REM: random effects model.

Several research groups have developed statistical methodologies particularly for microarray studies based on traditional meta-analysis methods. They are discussed in section 2 of chapter 2. For example, Fisher's method and the random effects model are widely used. Multiple comparisons also need to be addressed in genomic studies. The statistical framework, however, is often not rigorously formulated and a formal evaluation and comparison could not be performed. According to the individual analysis in table 2, we note that genes behave differently across tissues, because it is expected that tissue specific physiology will result in tissue dependent responses. From table 2, gene 1416531_at which has small p-values in four tissues is identified by all methods while the others genes which have small p-values in some, but not all of the tissues are identified by Fisher's method and Tippett's minimum p-value method. Thus, we can say that maxP and the random effects model tend to detect genes differentially expressed between wild type and VLCAD-/- in all tissues, while minP and Fisher's methods detect genes differentially expressed between wild type and VLCAD-/- in at least one of the tissues. A statistical framework should be rigorously defined for different methodologies. In the first part of chapter 3, two complementary hypothesis settings are outlined. And all meta-analysis methods targeted for their respective hypothesis settings are clarified. If tissue specific genes are of greater interest and we would like to know in which tissue the gene is differentially expressed, more powerful methodologies such as the optimally-weighted statistic (OW) can be developed. The details are discussed in the remaining part of chapter 3.

Returning to the example in section 1.1, if one would like to investigate the variation in gene expression among wild type, VLCAD-/- and LCAD-/-, there are three classes to

be compared in each issue. As a natural extension, an ANOVA model can be used to test the significance of variation in gene expressions across genotypes in each tissue. The corresponding p-values from F-test are then combined using different methods. In addition, genes demonstrating a consistent pattern across the four tissues might be of greater interest to investigators. So far, however, most of the methods in the literature, including those described above, focus on the two group comparison. Combining data sets with more than two groups is rarely discussed. Here, we propose a minimum multi-class correlation statistic (minMCC) to address this problem. The simulation studies and the application to the mouse energy metabolism data and mouse trauma data indicate that minMCC is a powerful test for identifying biologically relevant expression changes. The methods are described in chapter 4.

Chapter 5 describes an R package -*genomeMeta* that we developed for the visualization and implication of all possible meta-analysis methods for microarray data. We discuss how to use the package in several examples. This is followed by a conclusion and discussion in Chapter 6.

2.0 PREVIOUS META-ANALYSIS METHODS

2.1 META-ANALYSIS METHODS IN TRADITIONAL EPIDEMIOLOGIC RESEARCHES

In this section we consider K independent experiments that have been performed to detect a certain effect. Here θ_k is the parameter that characterizes the effect of study k, $k = 1, \ldots, K$. The kth experiment is designed to test the hypothesis $H_{0k} : \theta_k = 0$ against an alternative $H_{1k} : \theta_k \neq 0$ using the test statistic T_k. Significance levels and effect sizes are two common metrics utilized in meta-analysis. In the following, we introduce several methods for each metric separately.

2.1.1 Methods for combining significance

The methods of combining significance of independent tests have been discussed by a number of researchers, including Fisher (1932) and Pearson (1938). Assuming T_k follows a continuous distribution, the significance of a test is often defined as $p_k = Pr(T_k > t_k|H_{0k})$, that is, as a p-value. Note that when H_{0k} is true, p_k is uniformly distributed. Thus, combining the significance of independent tests is sometimes called omnibus or nonparametric.

2.1.1.1 Fisher's Method The best known method in this category is Fisher's (1932). Fisher's method uses the product of p-values from the test for each experiment and transforms it to chi-square scores using a $-2log$ transformation resulting in

$$V^{EW} = -\sum_{k=1}^{K} 2log(p_k) \tag{2.1}$$

For this reason it is also known as the Inverse Chi-square method. If all of the null hypotheses of the K tests are true, then $-\sum_{k=1}^{K} 2log(p_k)$ will have a χ^2 distribution with $2K$ degrees of freedom.

Fisher's method was shown in the literature to have good power under a wide range of alternative conditions and to be the most asymptotically Bahadur optimal (ABO) among several commonly used combined tests (Little and Folks, 1971, 1973). Fisher's method, however, is easily been dominated by small p-values because of its asymmetric transformation. For example, for gene 1415994_at in table 2, one of these tissues rejects the null hypothesis with $p = 0.002$, while the others do not. By Fisher's method the p-value is 0.007, and we still say the gene is differentially expressed. Another drawback of the Fisher's method can be seen from the examples, genes 1417025_at and 1424039_at clearly behave differently across tissues, but the same results are obtained by Fisher's method ($V^{EW} = 16.56, p - value = 7E - 06$). To distinguish between these two genes, an ad-hoc approach has to be added.

2.1.1.2 The Truncated Product Method Zaykin et al. (2002) of proposed a truncated version of Fisher's method, where they suggest using the product of p-values that do not exceed a value τ:

$$V^T = -\prod_{k=1}^{K} p_k^{I(p_k \leq \tau)}, \tag{2.2}$$

where $I(\cdot)$ is the indicator function. This test is less sensitive to small p-values, however, the choice of τ is arbitrary (the authors suggest a conventional value such as 0.05).

2.1.1.3 Tippett's Minimum p Method This method is proposed by Tippett (1931), where

$$V^{minP} = \min_{1 \leq k \leq K} p_k. \tag{2.3}$$

Therefore, H_0 is rejected if $\min_{1 \leq k \leq K} p_k \leq 1 - (1 - \alpha)^{1/K}$, where α is the overall significance level, because $\min_{1 \leq k \leq K} p_k$ has a $Beta(1, K)$ distribution under the null hypothesis.

NOTE: This method is also known as the union-intersection method. If the rejection region for the test of H_{0k} is $\{p_k \leq \alpha\}$, then the rejection region for the union-intersection

test for $H_0 : \theta \in \bigcap_{k=1}^{K} \Theta_k$ vs. $H_1 : \theta \in \bigcup_{k=1}^{K} \Theta_k$ is $\bigcup_{k=1}^{K} \{p_k \leq \alpha\}$, which is exactly the same as $\{\min_{1 \leq k \leq K} p_k \leq \alpha\}$.

Like Fisher's test, this method is sensitive to small p values, but it is less powerful than Fisher's approach especially when all studies are significant. Again in table 2, p-values from the four tissues of gene 1415994_at are 0.668,0.815,0.694 and 0.002 respectively. By the minimum p method, the p-value of V^{minP} is 0.007. If the p-value of skeletal is changed to 0.02, the p-value of V^{minP} becomes 0.08. The conclusion is to fail to reject H_0.

2.1.1.4 Wilkinson's Method Wilkinson (1951) generalized Tippett's procedure to a more robust rth smallest p-value giving

$$V^{W} = p_{(r)}. \tag{2.4}$$

The maximum p-value is a special case and the most frequently used. It is often referred to as Wilkinson's method.

$$V^{maxP} = \max_{1 \leq k \leq K} p_k. \tag{2.5}$$

It is worth noting that the maximum P-value statistic is a special case of the intersection-union test (Berger, 1982). According to Berger's theorem for intersection-union test (IUT), the intersected region for $H_0 : \theta \in \bigcup_{k=1}^{K} \Theta_k$ vs. $H_1 : \theta \in \bigcap_{k=1}^{K} \Theta_k$ is $R_g = \bigcap_{k=1}^{K}$
$leq\alpha\} = \{\max_{1 \leq k \leq K} p_k \leq \alpha\}$.

2.1.1.5 Other Transformations As noted above, Fisher's method utilizes an inverse Chi-square transformation. Some other transformations have been considered in the literature as well. For example, the inverse Normal method proposed by Stouffer et al. (1949):

$$V^{Z} = \frac{\sum_{k=1}^{K} \Phi^{-1}(p_k)}{\sqrt{K}}. \tag{2.6}$$

This statistic is asymptotically standard normal distribution when the H_{0k} are all true. The inverse Normal transformation test takes advantage of the one-to-one mapping of the symmetric standard normal curve to the p-values. Thus it treats large and small p-values symmetrically.

Another method worth noting is the Logit transformation suggested by George (1977). It involves transforming the p-value into a logit:

$$V^L = \sum_{k=1}^{K} log\frac{p_k}{1-p_k}. \tag{2.7}$$

The distribution of V^L can be easily derived. George and Mudholkar (1977) show that it can be approximated by a t distribution with $5K + 4$ degrees of freedom.

2.1.1.6 The Weighted Method One of the limitations of the above methods is that they do not weight evidence from the studies according to their uncertainties or sample sizes. Different weights and variants to existing methods have been considered. Good (1955) investigated the distribution of the weighted version of Fisher's method

$$V^G = \prod_{k=1}^{K} p_k^{w_k}, \tag{2.8}$$

if the weights are all unequal. The probability density function of $-logV^G$ is $\sum_{k=1}^{K} \frac{W_k}{w_k} e^{-v/w_k}$ where $W_k = w_k^{K-1}/((w_k - w_1)(w_k - w_2)\ldots(w_k - w_{k+1})(w_k - w_K))$.

Zelen and Joel (1959) provide a detailed investigation of the weighted version when K is 2 and $\sum w_k = 1$, where the test is from an F test for variance ratios. They also discuss the effect of the weight factor on the type II error of the combined test. However, it is clear that the distribution is not easily calculated in general cases. In addition, in practice, it is hard to determine the weights and it is unrealistic that all weights are unequal. So the method is seldom used in practice.

Another weighted method seen in the literature is the weighted inverse Normal transformation method given by

$$V^{Zw} = \frac{\sum_{k=1}^{K} w_k \Phi^{-1}(p_k)}{\sqrt{K}}. \tag{2.9}$$

Whitlock (2005) suggests that the weights can be chosen to be the inverse of squared standard error. He further shows that the weighted method is superior to the unweighted version.

Table 3: Different indexes of effect magnitude have been used in meta-analysis

Type of response	Effect size	T_i
Continuous	Cohen's d	$\frac{\bar{y}_1-\bar{y}_2}{\sqrt{(s_1^2+s_2^2)/2}}$
	Pearson' r	$\frac{\sum_{s=1}^{S}(y_{s1}-\bar{y}_1)(y_{s2}-\bar{y}_2)}{(n-1)s_1 s_2}$
	Hedges's g	$\frac{\bar{y}_1-\bar{y}_2}{\sqrt{\frac{(n_1-1)*s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}}$
	Glass's Δ	$\frac{\bar{y}_1-\bar{y}_2}{s_1}$
Discrete	Log risk ratio	$log(\frac{P_1}{P_2})$
	Log odds ratio	$log[\frac{P_1(1-P_2)}{P_2(1-P_1)}]$

y_1 and y_2 are the means for the group 1 and 2 respectively;
s_1 and s_2 are the standard deviations of group 1 and 2 respectively;
P_1 and P_2 are the proportion of subjects in group 1 and 2 respectively.

2.1.2 Methods for combining estimates

When the studies have similar designs and measure the outcome in a similar manner, combining estimates is usually preferred to the omnibus methods in previous section as suggested by some researchers. Suppose $T_k, k = 1, \ldots, K$ are the independent estimates of effect magnitude from K studies. The corresponding population effect magnitude parameters are $\theta_k, k = 1, \ldots, K$. There are many different metrics that can be used to measure effect size. They are listed in table 3. Of these, Cohen's d is used most often and it can be transformed from Pearson's r and Hedges's g. Note that all of the measurements focus on two group comparisons, effect size for more than two groups may not be easily defined. The distribution of effect estimates for more than two groups in large sample cases also presents difficulties.

The statistical analysis procedures of combining estimates all involve large-sample theory. There are three major types of statistical analysis for combining estimates: fixed, random and mixed effects models. In the following sections, we focus on two most commonly used models: fixed and random effects models (Hedges and Vevea, 1998).

2.1.2.1 Fixed effects model Fixed effects models consider only within-study variability. The assumption is that studies use identical methods, samples, and measurements; that they should produce identical results; and that differences are only due to within-study variation. The general model is given by

$$T_k = \theta_k + \epsilon_k, \epsilon_k \sim N(0, \sigma_k^2). \tag{2.10}$$

In the fixed effects model, we assume there is a constant effect size μ for all studies, $\theta_1 = \theta_2 = \ldots = \theta_K = \mu$. Thus, $T_k \sim N(\mu, \sigma_k^2)$. The most efficient and unbiased estimator of μ is the weighted average of the estimates where the weights are determined by the inverse of their standard errors. The estimate is

$$\hat{\mu} = \frac{\sum_{k=1}^{K} W_k T_k}{\sum_{k=1}^{K} W_k}, \tag{2.11}$$

where $W_k = 1/S_k^2(T_k)$ and $S_k^2(T_k)$ is the estimated variance of T_k. The variance of $\hat{\mu}$ is then

$$Var[\hat{\mu}] = \frac{1}{\sum_{k=1}^{K} W_k}. \tag{2.12}$$

A test statistic can be constructed to test the hypothesis that $\theta_1 = \ldots = \theta_K = \mu = 0$ as follows: $Z = \frac{\hat{\mu}}{\sqrt{Var(\hat{\mu})}}$, which follows the standard normal distribution, when $\sum_{k=1}^{K} n_k \to \infty$ where $n_k, k = 1, \ldots, K$ are the sample sizes of K studies.

2.1.2.2 Random effects model An alternative approach, the random effects model, allows the study outcomes to vary accordingly to a normal distribution between studies. Many investigators consider the random effects approach to be a more natural choice than fixed effects models. The random effects model is defined as

$$\begin{aligned} T_k &= \theta_k + \epsilon_k, \epsilon_k \sim N(0, \sigma_k^2) \\ \theta_k &= \mu + \delta_k, \delta_k \sim N(0, \tau^2). \end{aligned} \tag{2.13}$$

That is, the true study effect size θ_k is no longer a constant effect, it varies across studies with mean μ and variance τ^2. Obviously, the fixed effects model is a special case of the random effects model when one assumes τ^2 to be 0.

The estimation of underlying mean μ is similar to (2.10):

$$\hat{\mu} = \frac{\sum_{k=1}^{K} W_k T_k}{\sum_{k=1}^{K} W_k}, \tag{2.14}$$

where $W_k = 1/(S_k^2(T_k) + \hat{\tau}^2)$. That is the variation comes from two sources: within study variation, $S_k^2(T_k)$, and between study variation, τ^2. There are several ways of estimating τ^2. DerSimonian & Laird (1986)'s method is among the most commonly used approach to estimate τ^2 by the method of moments. The estimate is given by

$$\hat{\tau}^2 = max[0, \frac{Q - (K-1)}{\sum_{k=1}^{K} W_k - (\sum_{k=1}^{K} W_k^2 / \sum_{k=1}^{K} W_k)}], \tag{2.15}$$

where $Q = \sum_{k=1}^{K} W_k (T_k - \hat{\mu})$.

Some more complex procedures for estimating τ^2 are given by Rubin (1980,1981), Raudenbush and Bryk (1985), and Hedges and Olkin (1985).

2.1.2.3 Choosing between fixed and random effects models As noted above, the random effects model may be more general, since both the random variation within the studies and the variation between the different studies is incorporated. However, more data are required for random effects models to achieve the same statistical power as fixed effects models. Testing how much heterogeneity there is serves as a way to determine whether fixed effects model is appropriate. Heterogeneity in meta-analysis refers to the variation in study outcomes between studies.

There are three general ways to assess heterogeneity in meta-analysis, but each has a liability for interpretation. First, one can assess the between-studies variance, τ^2, but its values depend on the particular effect size metric used, along with other factors. The second is Cochran's Q, where $Q = \sum_{k=1}^{K} W_k(T_k - \hat{\mu})$, which is the weighted sum of squared differences between individual study effects and the pooled effect across studies. Under the null hypothesis of homogeneity, Q is distributed as a chi-square statistic with $K-1$ degrees of freedom. A large Q indicates possible heterogeneity in the studies and that the fixed effects model may not be valid. However, the Q test only informs meta-analysts about the presence versus the absence of heterogeneity, but it does not report on the extent of such heterogeneity. The third approach is the I^2 index (Higgins and Thompson, 2002). It quantifies the degree of heterogeneity in a meta-analysis. $I^2 = 100\%(Q - (K-1))/Q$, if $Q > K-1$, otherwise, $I^2 = 0$. So unlike Q, it does not inherently depend upon the number of studies considered. It is interpreted as a percentage of heterogeneity. A confidence interval for I^2 is also constructed by Higgins and Thompson (2002).

2.2 META-ANALYSIS METHODS FOR MICROARRAY STUDIES

Most meta-analytic methods for microarray studies are based on extensions of the established methods used for traditional epidemiological research. For example, Rhode et al. (2002) is among the first to apply Fisher's method to microarray data. Choi et al. (2003) implement random effects models and Gentleman et al. (2008) developed a R package *GeneMeta* for the methods Choi et al. described. We demonstrate several methods in the following using

the same notation described in chapter 1; let y_{gsk} denote the gene expression for the gth gene in the sth sample of the kth study.

In addition to Fisher's method, the same group of Rhode provided three algorithms for the analysis of prostate cancer datasets (Ghosh et al., 2002). In the first algorithm, they use the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996). LASSO is a variable selection method for linear regression. It minimizes sum of squared errors subject to a constraint. In the microarray data notation, the model is

$$\sum_{k=1}^{K}\sum_{s=1}^{n_k}(Y_{gsk}-\mu_g)^2 \tag{2.16}$$

subject to the constraint that $\sum_{g=1}^{G}|\mu_g| \leq \lambda$. A gene is considered a candidate gene if the estimate of μ_g is not zero, where μ_g denotes the mean normalized difference in gene expression between the two groups (e.g. Tumor and Normal). Estimating λ is challenge in microarray analysis. The cross-validation technique is not appropriate here because of the dependence structure of genes. One could generate different gene lists for a list of different values of λ. In their second algorithm, they calculated a weighted sum of t statistics across studies for each gene, with the weights being the ratio of sample size of the individual study to the total sample size:

$$T_{g.} = \frac{1}{\sum_{k=1}^{K} n_k}\sum_{k=1}^{K} n_k T_{gk}. \tag{2.17}$$

Weighting each study according to the sample size is intuitive yet arbitrary. And last, they provide a more general model by taking into account the interaction between study and tissue types through the following model:

$$E[Y_{gsk}] = \beta_{0gk} + \beta_{1gk}X_s + \beta_{2gk}Z_k + \beta_{3gk}X_sZ_k, \tag{2.18}$$

where X_s is an indicator variable for the tissue type of sth sample, and Z_k is an indicator variable for the study. The model could be fit using ordinary least squares. All of the three methods provided by Ghosh et al. treat genes independently, so that genes are fitted using the above models independently and the false discovery rate is controlled in the end.

Choi et al. (2003) implemented the same random effects models described in section 2.1.2 for each gene. They suggested a quantile-quantile plot of the observed versus expected

Cochran's Q values over all genes to explore the homogeneity in multiple studies. If the plot approximates a straight line, especially near the center, the presumption of a Chi-square distribution is appropriate suggesting that the datasets might be sampled from a common population and that only the sampling error is present, so a simple fixed effect model is enough. If this is not the case we have to fit a random effects model. They further discussed the Bayesian interpretation for the REM model. Following the same notation from section 2.1.2, the model is written as

$$\begin{aligned} & p(\mu, \theta_1, \ldots, \theta_K, \tau^2 | T_1, \ldots, T_K) \\ & = p(\theta_1, \ldots, \theta_K | T_1, \ldots, T_K, \sigma_1^2, \ldots, \sigma_K^2) p(\mu, \tau^2 | \theta_1, \ldots, \theta_K) \\ & \propto \prod_{k=1}^{K} p(\theta_k | T_k, \sigma_k^2) p(\theta_k | \mu, \tau^2) \pi(\mu) \pi(\tau^2) \end{aligned} \tag{2.19}$$

where $\pi(\mu)$ and $\pi(\tau^2)$ are non-informative priors given as $\mu \sim N(0, 10^6)$ and $1/\tau^2 \sim gamma(0.001, 0.001)$.

Similar Bayesian hierarchical models have been suggested by Tseng et al. (2001) and Conlon and Liu (2005) for incorporating different levels of replicates information in cDNA microarray experiments. Unlike the approach of Choi et al., which estimates the gene effects first and then put the effect size into the model, Tseng et al. and Colon and Liu present Bayesian hierarchical models based on gene expression levels across studies. Let y_{gesk} denote the normalized log-expression ratios of gene g, experiment e, and slide s in study k. For each study y_{gesk} is sampled from a normal distribution of the slide effect within experiment and study. The hierarchical model is as follows:

$$\begin{aligned} y_{gesk} | \mu_{gek} & \sim N(\mu_{gek}, \tau_{gk}^2) \\ \mu_{gek} | \theta_{gk} & \sim N(\theta_{gk}, \sigma_{gk}^2) \\ \theta_{gk} | I_g = 0 & \sim N(0, \eta_{gk0}^2) \\ \theta_{gk} | I_g = 1 & \sim N(0, c_k \times \eta_{gk0}^2) \\ I_g & \sim Bernoulli(p) \\ p & \sim Uniform(0, 1). \end{aligned} \tag{2.20}$$

τ^2_{gk} explains the variation in the slide effect within study k. Further, μ_{gek} is sampled from a normal distribution with mean θ_{gk}, and σ^2_{gk} explains the variation within study k. Genes are divided into two groups, non-expressed ($I_g = 0$) and expressed ($I_g = 0$), so that $p = Pr(I_g = 1)$. Genes are determined as differentially expressed if the posterior probability $D_g = Pr(I_g = 1|data)$ is greater than a threshold.

Owen (2007) re-introduced Pearson's method and applies it to the AGEMAP project. In the AGEMAP project 40 mice were studied with different ages. The samples were from 16 different tissues. To investigate the association between gene expression and age, an ordinary regression can be fitted. Let β_{gk} be the age slope for gth gene in study k. The null hypothesis is $H_{0,gk} : \beta_{gk} = 0$. He define a test statistic as $V_g^{EW*} = max(V_g^{EW;L}, V_g^{EW;R})$, where $V_g^{EW;L} = -\sum_{k=1}^{K} log(\tilde{p}_{gk}), V_g^{EW;R} = -\sum_{k=1}^{K} log(1 - \tilde{p}_{gk})$ and $\tilde{p}_{gk}$ is the is the one-sided p-value $p_{gk} = Pr(|\hat{\beta}_{sk}| \leq |\hat{b}_{sk}||H_{0,gk})$ of gene g in study k. Clearly, this test statistic is a modification of Fisher's method by combining one-sided p-values. A similar modification could be done to Stouffer et al.'s method: $V_g^{Z*} = max(V_g^{Z;L}, V_g^{Z;R})$,where $V_g^{Z;L} = -\sum_{k=1}^{K} \Phi^{-1}(\tilde{p}_{gk}), V_g^{Z;R} = -\sum_{k=1}^{K} \Phi^{-1}(1 - \tilde{p}_{gk})$. He claims that V_g^{EW*} and V_g^{Z*} are sensitive to either consistently left or right sided departure. And he further investigates the relative effectiveness of $V_g*, V_g^{\cdot;R}$ and $V_g^{\cdot:L}$ in detecting alternate hypotheses that vary in a consistent direction from the null. As expected, $V_g^{\cdot;R}$ is more sensitive to positive slopes and for negative slopes, $V_g^{\cdot;L}$ is the best choice. The method, however, is still easily dominated by one or two extremely significant p-values and does not tell which studies are significant for distinguishing concordant genes and discordant genes (see Chapter 3).

Breitling et al. (2004) proposed the rank product method for detecting regulated genes in replicated microarray experiments. Hong et al. (2006) modified and extended the rank product method to accommodate the analysis of multiple microarray studies. For a study k with two groups ($r_{sk} = \{0, 1\}$), denote n_{0k} and n_{1k} as the number of replicates for group 0 and 1 respectively. They first calculate the pair-wise fold changes between two groups for each gene within the study, that is, $FC_{gkt} = (y_{gki}|r_{sk} = 1)/(y_{gkj}|r_{sk} = 0)$, $i \in (r_{sk} = 0)$ and $j \in (r_{sk} = 1)$. There are $n_{0k} \times n_{1k}$ pairs. For K studies, there are a total of $\sum_{k=1}^{K} n_{0k} \times n_{1k}$ pairs. So $t = 1, \ldots, \sum_{k=1}^{K} n_{0k} \times n_{1k}$. The rank product statistic for each gene is defined as $V^{RP} = (\prod_{t=1}^{\sum_{k=1}^{K} n_{0k} \times n_{1k}} r_{gt})^{1/K}$, where r_{gt} is the rank of FC of the gth gene under the

tth comparison. The null distribution of V^{RP} is then determined from permutation. For genes regulated in the opposite direction, they swap the the groups and perform the same algorithm once more. Since this method combines ranks instead of real gene expression, fewer assumptions are needed. They also claim that it overcomes the heterogeneity among multiple datasets.

3.0 OPTIMALLY WEIGHTED STATISTIC FOR COMBINING MULTIPLE MICROARRAY STUDIES

3.1 BACKGROUND

Integrating results from different studies is commonplace in a wide variety of biological applications. Significance levels and effect sizes are two common metrics utilized in meta analysis. The random effects model is often used to incorporate random effects caused by, for example, variations of sampling schemes in different studies. In many applications, data structures and statistical hypotheses may differ, making a direct combination of effect sizes impossible. It becomes more feasible to combine the transformed probability integrals of test statistics (usually p-values) - a procedure that depends only on the significance values of individual tests and not on the underlying data structure. The well-known method in this category was proposed by Fisher (1932). The test statistic involves the log-transformation of p-values to Chi-square scores and equally-weighted summation: $V^{EW} = -\sum_{k=1}^{K} log(p_k)$, where K studies are combined and p_k is the p-value of study k, $1 \leq k \leq K$. Assuming both independence among studies and the p-values calculated from correct null distributions in each study, $2V^{EW}$ follows a Chi-square distribution with 2K degrees of freedom. A large number of different transformations have been considered in the literature - for example, inverse normal, logit and inverse Chi-square transformation with varying degrees of freedom (Stouffer et al., 1949; Lancaster, 1961; George, 1977). While Fisher's method is not a uniformly most powerful test, it has strong support in the literature for its good power under a wide range of alternative conditions and for being the most asymptotically Bahadur optimal (ABO) among several commonly used combined tests (Little and Folks, 1971, 1973).

Different weights and variations of the test statistic have also been considered. Good (1955) suggested using unequal weights for each individual study, and Olkin and Saner (2001) proposed a trimmed version of the Fisher procedure to remove the effect of possibly aberrant extremes. Another well known method in this category is Tippetts (1931) minimum p-value statistic: $V^{minP} = min_{1 \leq k \leq K} p_k$. Wilkinson (1951) generalized Tippetts procedure to a more robust rth smallest p-value. It is worth noting that the minimum p-value (minP) and maximum p-value statistics (maxP) are special cases of the union-intersection test (UIT) (Roy, 1953) and the intersection-union test (IUT) (Berger, 1982) respectively. For a comprehensive review and comparison of different approaches, see Hedges (1992) and Cousins (2007).

Microarray technology allows researchers to examine the expression of thousands of genes in parallel. With microarray experiments becoming more mature and prevalent, the integration of homogeneous experimental data sets from multiple laboratories and experimental techniques has become imperative. In contrast to traditional epidemiological or evidence based medical studies, microarrays monitor gene expression for thousands of genes simultaneously - a process that brings challenges to integrative analysis. Currently "meta-analysis" in the biological literature consists of a widespread use of naive "intersection" and/or "union" operations on differentially expressed gene lists obtained from individual studies by certain criteria (e.g. False Discovery Rate ≤ 0.05) (Segal et al., 2004, Borovecki et al., 2005, Cardoso et al., 2007, Pirooznia et al., 2007, and many more). One can quickly note that intersections are often too conservative and unions are anti-conservative, especially as K increases. Rhode et al. (2002) was the first to apply Fisher's method to microarray data for a real sense of meta-analysis. They later introduced a weighted average of test statistics of individual tests, with the weights determined in an ad hoc manner according to sample sizes in the studies (Ghosh et al. 2003). Choi et al. (2003) pointed out that the approach in Rhode et al. "ignored the interstudy variation" and they proposed a random effects model (REM) under Gaussian assumptions and further discussed the details of a Bayesian formulation for the REM model. Similar Bayesian hierarchical models have been suggested by Tseng et al. (2001) and Conlon and Liu (2005) for incorporating different levels of replicates information in cDNA microarray experiments.

In this chapter, we propose an optimally weighted (OW) statistic for the meta-analysis

of multiple genomic studies. In contrast to the equally weighted (EW) statistic in Fishers method, the OW statistic finds the optimal weight that provides the best statistical significance. In Section 3.2, two complementary hypothesis settings (HS1 and HS2) are outlined and all meta-analysis methods targeted for their respective hypothesis settings are clarified. The remaining part of the chapter (Section 3.3-3.6) is devoted to the HS1 framework to discuss the comparison and properties of the OW statistic. In Section 3.3, we introduce the idea of the OW statistic and a detailed algorithm implementing the OW statistic and permutation test is outlined for integrating multiple genomic studies. In Section 3.4, we demonstrate the admissibility of OW, EW and minP in the traditional two sample comparison under Gaussian assumption and the power of OW is compared to EW and minP. Section 3.5 presents a simulation study and applications to a set of multi-tissue energy metabolism mouse data and a set of three prostate cancer expression profiles. Finally we summarize and discuss the advantages and limitations of the OW statistic in Section 3.6.

3.2 TWO COMPLEMENTARY HYPOTHESIS SETTINGS

In the existing meta-analysis methods applied to genomic data discussed above, no comprehensive evaluation has been performed. A major reason is the lack of rigorous formulation of hypothesis settings behind these methods. We consider meta-analysis of K gene expression profile studies: $D_1, D_2, \ldots, D_K$. In each study D_k, x_{kgs} is the gene expression intensity of gene g and sample s, where samples $s = 1, \ldots, n_k$ belong to a control group (e.g. normal patients) and $s = n_k + 1, \ldots, n_k + m_k$ belong to the target group (e.g. tumor patients). We define the first hypothesis setting as:

$$HS1: \ H_{0g} : \theta_{g1} = \theta_{g2} = \ldots = \theta_{gK} = 0 \text{ versus } H_{Ag} : \text{ at least one } \theta_{gk} \neq 0, 1 \leq k \leq K,$$

where θ_{gk} is the parameter that characterizes the effect of study k in gene g,$1 \leq g \leq G$. HS1 corresponds to the biological question Q1: "which genes are differentially expressed in one or more data sets among the K studies?". This question is often addressed when heterogenous characteristics are expected across studies (e.g., differential sample population, experimental

quality or tissues collected across studies). See the example in Section 5.2 where three tissues are examined and tissue specific physiology is expected. Another commonly asked biological question in the meta-analysis of microarray studies would be Q2: "which genes are differentially expressed in all data sets among the K studies?". The corresponding hypothesis setting is:

$$HS2: \ H_{0g}: \text{ at least one } \theta_{gk} = 0 \text{ versus } H_{Ag}: \theta_{gk} \neq 0, 1 \leq k \leq K.$$

This question is often asked when studies have homogeneous characteristics. Significance in all studies is demanded to generate highly confident biomarkers and combining multiple studies increases the statistical power.

Among the methods discussed in Section 3.1, EW, minP and the proposed OW statistic are designed to answer the HS1 problem, while REM and maxP are methods for HS2. In this chapter, we will only focus on the HS1 problem and compare the three major methods EW, minP and OW to demonstrate the properties and advantages of the proposed OW method.

3.3 OPTIMAL-WEIGHTED STATISTIC

When integrating multiple genomic studies, expression of some genes may be altered in a study specific manner (consider HS1). To uncover the pattern of altered gene expression across studies, we consider the following weighted statistic:

$$U_g(w) = -\sum_{k=1}^{K} w_k log(p_{gk}), \tag{3.1}$$

where p_{gk} is the p-value of gene g in study k, w_k is the weight assigned to the kth study and $w = (w_1, \ldots, w_k)$. Under the null hypothesis that $\theta_{gk} = 0 \forall k$ in HS1, the p-value of the observed weighted statistic, $p(u_g(w))$, can be obtained for a given gene g and weight w (see below for detailed permutation algorithm to calculate the p-value). The optimally-weighted statistic is defined as the minimal p-value among all possible weights:

$$V_g^{OW} = \min_{w \in W} p(u_g(w)), \tag{3.2}$$

where $u_g(w)$ is the observed statistic for $U_g(w)$ and W is a pre-specified search space. Our choice of search space in this chapter is $W = \{w|w_i \in \{0,1\}\}$, which results in an affordable computation of $O(2^K - 1)$ based on the norm of $K \leq 10$ in a microarray meta-analysis.

The resulting optimal weight reflects a natural biological interpretation of whether or not a study contributes to the statistical significance of a gene. We note that the OW statistic is not adequate for the traditional meta-analysis in epidemiological or evidence-based medicine research. The selection procedure in OW will introduce selection bias towards studies with concordant significant effects. The meta-analysis of genomic studies, however, is quite a different situation. The major goal is usually to screen and identify the most probable gene markers given data to facilitate future investigation. As we will show in section 3.5, the vector of optimal-weight, $w_g^* = \arg\min_{w \in W} p(u_g(w))$, actually serves as a convenient basis for gene categorization in follow-up biological interpretation and exploration.

Below we illustrate the detailed procedure for OW when applied for combining genomic studies. A permutation test is performed to assess the statistical significance and false discovery rate (FDR) is controlled at 5%. For the applications in Section 3.5, the minP and EW methods are performed using a similar permutation test.

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

1. Compute the moderated t-statistics, $\{t_{gk}$ for gene g and study k. (Efron et al., 2001; Tusher et al., 2001)
2. Permute group labels in each study for B times, and similarly calculate the permuted statistics $\{t_{gk}^{(b)}; 1 \leq g \leq G, 1 \leq k \leq K, 1 \leq b \leq B\}$.
3. Estimate p-value of t_{gk} as $p_{gk} = \sum_{b=1}^{B} \sum_{g=1}^{G} \frac{I\{t_{gk}^{(b)} \in R(t_{gk})\}}{BG}$, where $R(t_{gk})$ is the rejection region given the threshold t_{gk} and $I\{\cdot\}$ is the indicator function. Similarly given $t_{gk}^{(b)}$, compute $p_{gk}^{(b)} = \sum_{b'=1}^{B} \sum_{g'=1}^{G} \frac{I\{t_{g'k}^{(b')} \in R(t_{gk}^{(b)})\}}{BG}$

II. Calculate OW statistic:

1. Given a weight $w = (w_1, \ldots, w_k)$, calculate the weighted statistic $u_g(w) = -\sum_{k=1}^{K} w_k log(p_{gk})$ for gene g. Define $u_g^{(b)}(w) = -\sum_{k=1}^{K} w_k log(p_{gk}^{(b)})$.
2. Estimate the p-value of the observed $u_g(w)$ as $p(u_g(w)) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\{u_{g'}^{(b)}(w) \geq u_g(w)\}}{B \cdot G}$. Similarly compute $p(u_g^{(b)}(w)) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\{u_{g'}^{(b')}(w) \geq u_g^{(b)}(w)\}}{B \cdot G}$.

3. Based on II.A. (1) and II.A.(2), calculate the optimal weight as $w_g^* = \arg\min_{w \in W} p(u_g(w))$ and similarly $w_g^{(b)*} = \arg\min_{w \in W} p(u_g^{(b)}(w))$. Define the OW statistic V_g as the p-value of the optimally weighted statistic: $V_g = p(U_g(w*))$. Similarly $V_g^{(b)} = p(U_g^{(b)}(w*))$.

III. Assess p-values and q-values:

1. The p-value of V_g is calculated as $p(V_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\{V_{g'}^{(b)} \leq V_g\}}{B \cdot G}$.
2. Estimate π_0 , the proportion of null genes, as $\widehat{\pi}_0 = \frac{\sum_{g=1}^{G} I\{p(V_g) \in A\}}{G \cdot \ell(A)}$ (Storey 2002). Normally we choose $A = [0.5, 1]$ and $\ell(A) = 0.5$.
3. Estimate the FDR for each gene as $q(V_g) = \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{g'=1}^{G} I\{V_{g'}^{(b)} \leq V_g\}}{B \sum_{g'=1}^{G} I\{V_{g'} \leq V_g\}}$. The detected gene list is $G^{OW} = \{g : q(V_g) \leq 0.05\}$.

IV. Distinguish concordant and discordant genes (optional): Split the detected gene list G^{OW} into concordant and discordant gene lists. Detected genes with concordant regulation direction across significant studies are denoted as $G_{concord}^{OW} = \{g : q(V_g) \leq 0.05 \mid |\sum_{k=1}^{K} sgn(t_{gk}) \cdot w_{gk}^*| = \sum_{k=1}^{K} w_{gk}^*\}$,where $sgn(\cdot)$ is the sign function that takes the value of 1 when positive and -1 when negative. The discordant gene list is $G_{discord}^{OW} = G^{OW} \setminus G_{concord}^{OW}$.

Remarks:

1. For the application of the EW and minP method, step II.(1)-II.(3) can be skipped. Alternatively the test statistic are modified as $V_g = -\sum_{k=1}^{K} log(p_{gk})$ and $V_g^{(b)} = -\sum_{k=1}^{K} log(p_{gk}^{(b)})$ for EW or $V_g = min_{1 \leq k \leq K} p_{gk}$ and $V_g^{(b)} = min_{1 \leq k \leq K} p_{gk}^{(b)}$ for minP.
2. The sequence I., II. and III. provides an algorithm for a general framework. The statistics t_{gk} and the rejection region $R(t_{gk})$ can be replaced depending on the experimental design and hypothesis. For example, the F-statistic can be used when multiple groups of samples are available in each study.
3. When conducting two group comparisons and applying moderated t-statistics, genes detected under the general framework (sequence I., II. and III.) may contain discordant genes (i.e. a gene up-regulated in one study and down-regulated in another). The addition of step IV. provides further filtering to solve the problem. In some applications, it may be of interest to scrutinize the discordant gene list to verify whether the discordance reflects real biological discrepancy across studies (e.g. different tissues or patient populations) or is caused by artificial errors (e.g. mistakes in gene annotation). For

EW and minP, there is no direct criterion for a clear split of concordant and discordant genes. Pearson (1938) and Owen (2007) proposed a modification for EW that is sensitive to either consistent left or right sided departures: $V_g^{EW*} = max(V_g^{EW;L}, V_g^{EW;R})$, where $V_g^{EW;L} = -\sum_{k=1}^{K} log(\tilde{p}_{gk}), V_g^{EW;R} = -\sum_{k=1}^{K} log(1-\tilde{p}_{gk})$ and $\tilde{p}_{gk}$ is the is the one sided p-value of gene g in study k. The method, however, is still easily dominated by one or two extremely significant p-values and does not tell which studies are significant in distinguishing concordant versus discordant (see examples in Section 3.5.2).

4. Several forms of penalized or moderated t-statistics have been proposed and shown to outperform traditional t-statistics (Efron et al., 2001; Tusher et al., 2001; Smyth et al., 2004;). In our algorithm, we choose the penalized t-statistics used in Efron et al. (2001) and Tusher et al. (2001). The fudge parameter s0 is chosen to be the median variability estimator in the genome.

5. In III.(2), many reports have indicated the intrinsic difficulty in estimating π_0 and a poor estimator π_0 can greatly deteriorate the FDR estimation. A conservative suggestion is to set $\pi_0 = 1$ all the time and is applied in this chapter.

3.4 ADMISSIBILITY AND POWER

In this section, we drop the subscript g for genes and compare three test statistics (EW, OW and minP) for HS1 at the univariate level assuming independence between studies. To date, no best method for combining multiples studies has been identified, therefore choosing a combined statistic must reflect specific biological purposes. Birnbaum (1954) provided general conditions for evaluating combined methods, including monotonicity and admissibility. He (Birnbaum, 1955) further compared several combined test procedures by considering a one-sample test on the mean of a Gaussian distribution with known variance. Here we will use a similar two-sample test of the means of a Gaussian distribution with known variances:

$$Z_k = \frac{\overline{X}_{2k} - \overline{X}_{1k}}{\sigma_k \sqrt{1/n_{k1} + 1/n_{k2}}}.k = 1, 2, \ldots, K. \tag{3.3}$$

and the two-sided p-value $P_k = Pr(|Z| \geq |z_k||\theta = 0)$ for each study to examine the acceptance region of the different combined test procedures. We will discuss the admissibility and compare the power of the three test statistics under this simplified framework.

3.4.1 Admissibility

A test is considered admissible if it cannot be uniformly improved by any other test. It is well-known that no uniformly most powerful test exists even in the simplified scenario (3.3). Birnbaum stated a necessary and sufficient condition for a test to be admissible under this situation.

Theorem 3.4.1. *(Birnbaum 1954, 1955) Under HS1 and assumption (3.3), a necessary and sufficient condition for a combined test procedure to be admissible is that the corresponding acceptance region is convex.*

It has been noted that the acceptance regions of EW and minP are both convex and thus both methods are admissible. Figure 1 shows the acceptance regions of EW, minP and OW in the plane of a pair of Z statistic at level 0.05. When illustrating the rejection regions of several common combined tests (including EW and minP), Birnbaum showed a preference for the EW method in most cases because it appeared to be fairly sensitive in all directions. From Figure 1, we observe that OW actually shares merits from both EW and minP. It is more sensitive than the minP method when parameters from both studies depart from 0 and more sensitive than EW when only one of the parameters departs from 0. The following corollary shows that the acceptance region of OW is also convex and thus OW is also admissible.

Corollary 3.4.2. *The acceptance region of OW is convex and thus OW is admissible under HS1 and assumption (3).*

Proof: Denote by $p_k = 1 - \Phi(|z_k|)$ the two-sided p-value,where $\Phi(t) = \int_{-\infty}^{t} \phi(t)dt$, $\phi(t)$ is the density of the standard normal distribution. First we prove that $f(z_k) = -log(p_k) = -log(1 - \Phi(|z_k|))$ is convex. $f''(z) = \frac{\phi(|z|)}{[1-\Phi(|z|)]^2}\{\phi(|z|) - |z|[1 - \Phi(|z|)]\}$. It is well known that the elementary upper bound for $1 - \Phi(x)$ is $\phi(x)/x$, for $x > 0$. Thus, $f''(z) > 0$. So $f(z)$ is convex in z. Hence $f(z_1, z_2, \ldots, z_K) = -\sum_{k=1}^{n} log(p_k)$ for $\forall n \geq 1$ is convex,

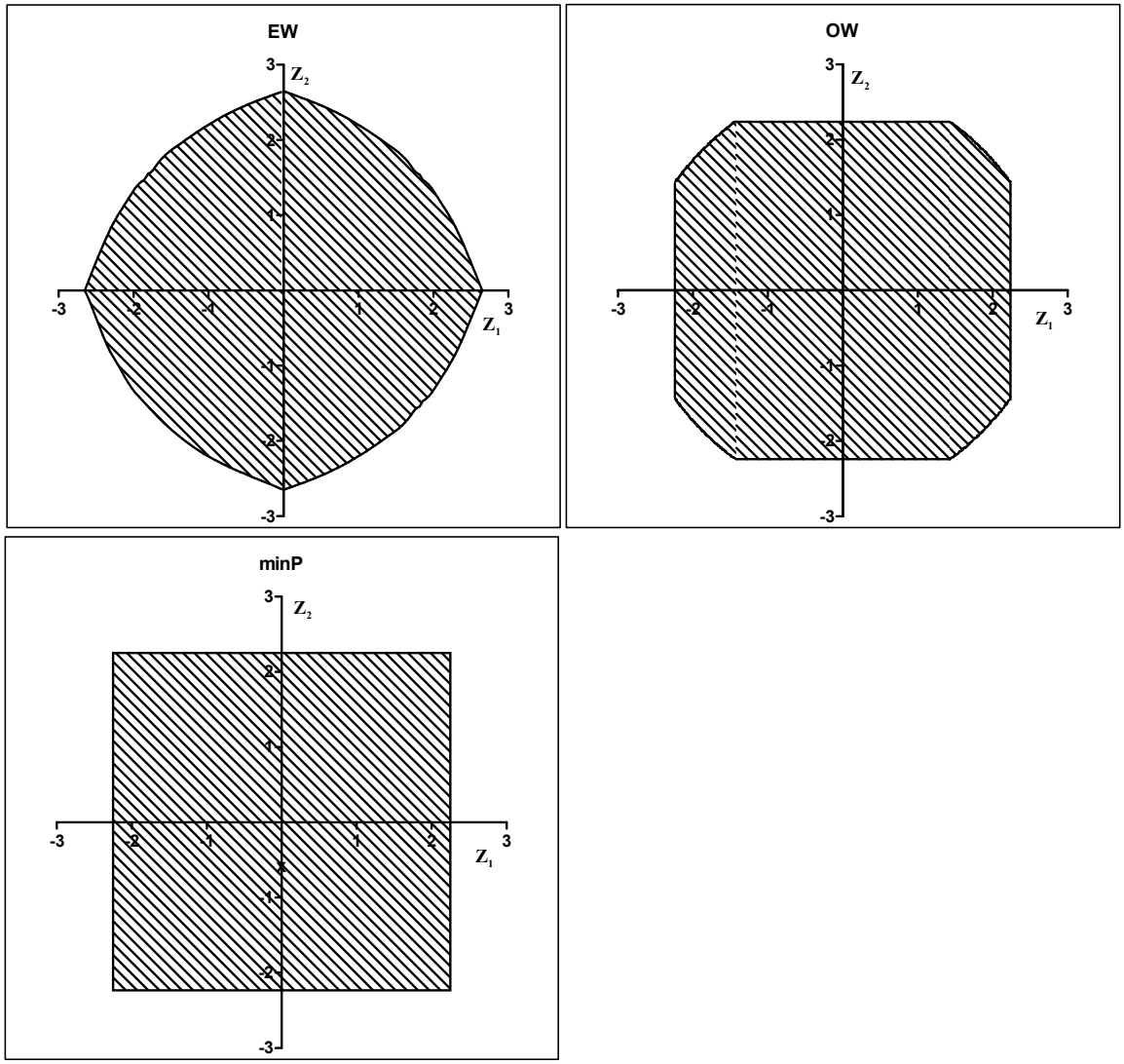

Figure 1: Acceptance regions of Equal-Weighted, Optimal-Weighted, minimum p-value and maximum p-value methods of combining p-values from two independent studies.

because the sum of convex functions is convex. For the OW statistic, the acceptance region is $\{z_1, z_2, \ldots, z_K : \min_{0\le k\le K} p(u(w)) > c\}$, where $p(u(w))$ is the right-sided p-value of $u(w)$.

$$\begin{aligned}
&\{z_1, z_2, \ldots, z_K : \min_{0\le k\le K} p(U(w)) > c\} \\
&= \bigcap_{I_k\in\{0,1\}, 1\le k\le K} \{z_1, z_2, \ldots, z_K : p(-\sum_{k=1}^{K} log[p_k^{I_k}]) > c\} \\
&= \bigcap_{I_k\in\{0,1\}, 1\le k\le K} \{z_1, z_2, \ldots, z_K : -\sum_{k=1}^{K} log[p_k^{I_k}] < \gamma_j\}, j = 1, 2, \ldots, 2^K - 1
\end{aligned} \tag{3.4}$$

Thus the acceptance region of OW is convex since the intersection of convex sets is also convex.

Note that when $K = 2$, $\{z_1, z_2 : \min_{w\in W} p(u(w)) > c\} = \{z_1, z_2 : \min(p1, p2) > c\} \bigcap \{z_1, z_2 : -log(p_1 p_2) \le \gamma\}$, the acceptance region of the OW method is an intersection of minP and EW.

3.4.2 Power comparison of EW,OW and minP under HS1

Denote by $\theta = (\theta_1, \ldots, \theta_K)$ the vector of parameters characterizing the studies and by $\beta^{OW}(\theta; \alpha)$ the power of a test controlled at level α, for the OW statistic, we have

$$\beta^{OW}(\theta; \alpha) = Pr(V^{OW} \le C_\alpha^{OW} | \theta) = 1 - \int \ldots \int_\Omega \prod_{k=1}^{K} p(P_k|\theta) dP_1 \ldots dP_K, \tag{3.5}$$

where $\Omega : \{\bigcap_{i=1}^{2^K-1} p(u(w_i)) > C_\alpha^{OW}\} = \{\bigcap_{i=1}^{2^K-1} U(w_i) < F^{-1}_{Gamma(\sum_{k=1}^{K} w_{ki}, 1)}(1 - C_\alpha^{OW})\}$ and $F^{-1}_{Gamma(\alpha,\beta)}$ is the inverse CDF of Gamma distribution with parameters α and β, $w_i = (w_{1i}, \ldots, w_{Ki})$, $w_{ki} \in \{0, 1\}, k = 1, \ldots, K$. C_α^{OW} is the solution of v to the equation $P(V^{OW} \le v | H_0) = \alpha$. If the null hypothesis of HS1 was true, it is generally accepted that the individual P_k value is a uniform random variable on [0,1]. The density of the P value under the alternative law is given by

$$p(P|\theta) = \frac{p(x|\theta)}{p(x|0)}|_{x=g(P)} \quad (0 \le P \le 1), \tag{3.6}$$

where $x = g(P)$ means the solution of $P = \int_x^1 f(x|0)dx$ (Pearson 1938).

For example, the Z test in (3.3) is used for power calculations, hence the density of P_k is

$$p(P_k|\theta_k) = \frac{1}{2}\exp\{\frac{c_k}{2}[2\Phi^{-1}(1-P_k/2)-c_k]\}+\frac{1}{2}\exp\{-\frac{c_k}{2}[2\Phi^{-1}(1-P_k/2)+c_k]\}, k = 1,\ldots,K, \tag{3.7}$$

where $c_k = \frac{\theta_k}{\sigma_k\sqrt{1/n_{k1}+1/n_{k2}}}$. We consider $K = 2$ and 8 , $n_{k1} = n_{k2} = 5, \sigma_k = 1$, so that the effect size is represented by θ_k. Power is evaluated with varying effect sizes.

Figure 2 illustrates the power curves under different alternative hypotheses when combining eight studies. As expected, when only one of the eight studies are significant, minP is much more powerful than EW (Figure 2.(a)) and when all of the eight studies are significant, EW is much more powerful than minP (Figure 2.(d)). On the other hand, OW performs stably near the best in the two extreme situations and is the top performer when two or three out of eight studies are significant (Figure 2.(b) and (c)).

3.5 APPLICATIONS

3.5.1 Simulation study

We conducted a simulation study for combining four datasets to assess the performance of our proposed optimally-weighted statistic (OW), Fisher's equally-weighted statistic (EW) and Tippett's minimum p-value method (minP); and naive union (union of DE gene lists from each individual study controlled at FDR=5%) was also included for comparison. For each dataset, we simulated five normal samples from a standard normal distribution and five case samples from a $N(\theta, 1)$. A total of $g1$ genes (category I) were differentially expressed across all four data sets. In contrast, $g2 = 400 - g1$ genes were differentially expressed only in the fourth data set (category II), and 1600 genes were considered null. The FDR was controlled at 5% for each method, and each simulation was repeated 1000 times.

Summaries of the resulting FDR and average number of genes identified in each category under three different scenarios are presented as follows: in Table 4, 0 category I and 400

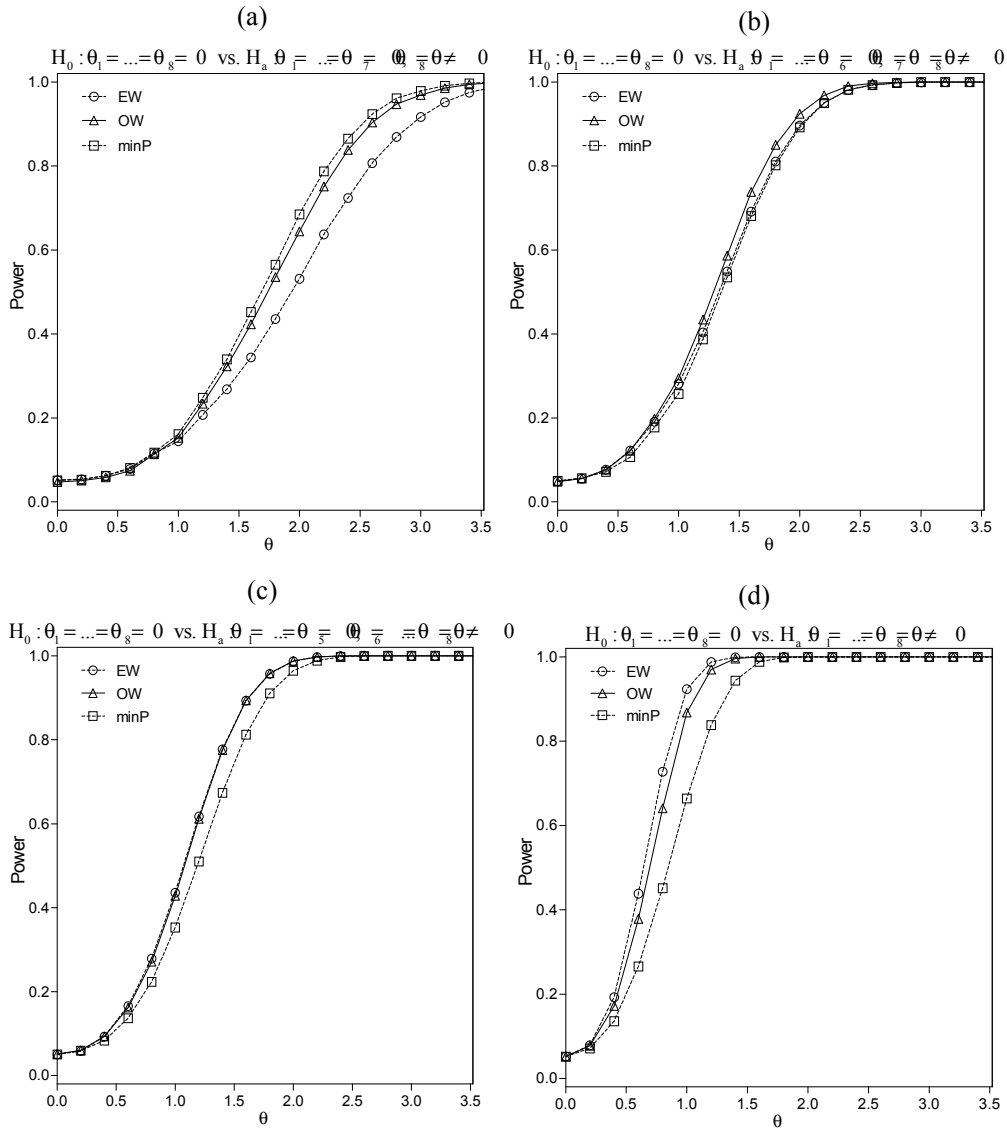

Figure 2: Power analysis of EW, OW, minP under HS1. We compare power curves of the three methods combining eight studies under different alternative hypotheses: (a) One of the eight hypotheses is false; (b) Two of the eight hypotheses are false; (c) Three of the eight hypotheses are false; (d) All of the eight hypotheses are false.

Table 4: Average number of genes detected in each category (I. 0 common DE genes; II. 400 4th-dataset-specific DE genes; Null. 1600 random noise genes), and average FDR, according to different θ.

Methods	$\theta = 2.0$				$\theta = 2.5$				$\theta = 3.0$			
	I	II	Null	FDR	I	II	Null	FDR	I	II	Null	FDR
OW	0.0	7.5	0.5	7.0%[a]	0.0	35.5	1.6	4.7%[a]	0.0	142.4	7.6	5.0%[a]
EW	0.0	3.4	0.3	7.1%[a]	0.0	13.2	0.7	5.4%[a]	0.0	65.8	3.7	5.4%[a]
minP	0.0	9.3	0.6	6.5%[a]	0.0	66.6	3.8	5.8%[a]	0.0	173.5	9.1	5.5%[a]
Union	0.0	117.2	6.2	4.8%[a]	0.0	265.0	13.2	4.6%[a]	0.0	352.2	17.7	4.7%[a]

[a]estimated FDR (mean of Null/(I+II+Null)) for HS1.

Table 5: Average number of genes detected in each category (I. 200 common DE genes; II. 200 4th-dataset-specific DE genes; Null. 1600 random noise genes), and average FDR, according to different θ.

Methods	$\theta = 1.5$				$\theta = 2.0$				$\theta = 2.5$			
	I	II	Null	FDR	I	II	Null	FDR	I	II	Null	FDR
OW	109.7	10.7	6.2	4.8%[a]	189.2	37.4	11.2	4.8%[a]	199.7	72.2	13.2	4.6%[a]
EW	154.0	11.6	8.6	4.9%[a]	197.4	28.2	11.5	4.8%[a]	200.0	49.6	12.5	4.7%[a]
minP	3.6	0.9	0.4	6.1%[a]	76.1	22.6	4.8	4.5%[a]	168.0	72.5	11.6	4.5%[a]
Union	8.9	4.7	1.1	7.3%[a]	91.9	58.0	9.6	6.3%[a]	185.8	132.9	24.1	7.0%[a]

[a]estimated FDR (mean of Null/(I+II+Null)) for HS1.

Table 6: Average number of genes detected in each category (I. 400 common DE genes; II. 0 4th-dataset-specific DE genes; Null. 1600 random noise genes), and average FDR, according to different θ.

Methods	$\theta = 1.5$				$\theta = 2.0$				$\theta = 2.5$			
	I	II	Null	FDR	I	II	Null	FDR	I	II	Null	FDR
OW	268.5	0.0	13.9	4.9%[a]	386.6	0.0	19.4	4.8%[a]	399.6	0.0	18.8	4.5%[a]
EW	339.6	0.0	17.9	5.0%[a]	397.3	0.0	20.6	4.9%[a]	400.0	0.0	19.8	4.7%[a]
minP	22.1	0.0	1.3	4.8%[a]	230.6	0.0	11.4	4.6%[a]	359.9	0.0	16.7	4.4%[a]
Union	36.9	0.0	2.6	6.1%[a]	298.9	0.0	24.1	7.4%[a]	394.8	0.0	51.9	11.6%[a]

[a]estimated FDR (mean of Null/(I+II+Null)) for HS1.

category II genes; Table 5, 200 category I and 200 category II genes; Table 6, 400 category I and 0 category II genes. The result is consistent with the power calculation in Section 4.2. In Table 4, minP is much more powerful than EW. When $\theta = 3$, minP detects an average of 173.5 genes and EW detects only 65.8 genes. OW, on the other hand, finds 142.4 genes, which is quite close to minP. Similarly in Table 6, EW (397.3 genes detected when $\theta = 2$) is more powerful than minP (230.6 detected genes) and OW (386.6 detected genes) is close to EW in performance. Overall OW performed stably well in the two extreme situations. We note that the naive union method generally loses control of FDR as expected.

3.5.2 Energy metabolism in mouse model

The deficiencies of very long-chain acyl-coenzyme A dehydrogenase (VLCAD) are associated with an energy metabolism disorder in children. Two genotypes of mouse model (wild type (VLCAD +/+) and VLCAD-deficient (VLCAD -/-)) are studied with four mice in each genotype group. For each of the 8 mice, three types of tissues (brown fat, liver and heart) were applied separately in microarray experiments to study the expression changes across genotypes. In this study, one of the hypotheses is that tissue-specific physiology triggers tissue-dependent responses, with precise pools of differentially expressed genes being confounded to the tissue in question. This hypothesis is aimed at understanding signature genes that are significant for tissue subsets, an analysis that corresponds to HS1.

In Figure 3, OW, EW, Pearsons method (described in Remark 3 of Section 3.3) and minP are implemented. The modified algorithm of OW for filtering out discordant genes (step IV in Section 3.3) is also implemented. It discards all genes that contain discordance among studies contributing to the optimal weight. We note that the modified algorithm is not applicable to EW, Pearson's method and minP since these methods do not provide indication of which studies to consider for concordance/discordance evaluation. Overall the general OW detects 203 genes in Figure 3(a) and among them, 28 genes have conflict up or down regulations (i.e. Figure 3(b) detects 175 genes). The optimal weights provide natural grouping of identified genes. For example, 55 genes with (1,1,1) optimal weight are differentially expressed in all three tissues in Figure 3(b) and 27 genes with (0,1,1) weight are statistically significant only

in liver and heart tissues, but not in brown fat. Results of the EW method are presented in Figure 3(c). It detects more genes than OW (329 genes). The identified gene list is, however, difficult to interpret and investigate even after gene reordering by hierarchical clustering. We note that minP seems to be much less powerful in this application. We note that EW, minP and OW are based on summarization of p-values across studies. The methods alone do not distinguish discordant genes that have differential expression across studies. The method presented by Owen (2007), previously proposed by Pearson (1934), was developed to modify EW for this purpose. Table 7 shows five example genes. The first two genes (1419182_at and 1423407_a_at) contain clear discordant regulation across studies. All of the four methods except for minP identify them as differentially expressed genes. Pearson's method, designed for detecting only concordant genes, failed to achieve the goal to exclude such discordant genes. In the OW method, the optimal weight enables a post-hoc approach (Step IV in Section 3.3) to filter out these discordant genes. For EW, minP and Pearson's method, such a post-hoc procedure is not feasible without indication of which studies are differentially expressed. For example, in the last two examples of Table 7 (1449015_at and 1416415_a_at), the OW method with concordance filtering will still call it a concordant DE gene although the regulation of the non-significant study (brown fat) contradicts the other two significant studies.

3.5.3 The Prostate cancer studies

We applied the algorithm to three publicly available prostate cancer gene expression datasets (Dhanasekaran et al., 2001; Luo et al., 2001; Welsh et al., 2001) independently collected in separate medical centers: two representing cDNA technology and one oligo-based technology. Probes of the three datasets were matched by UniGene ID. Gene expression comparisons were made between clinically localized prostate cancer and benign prostate tissue. The results presented in Figure 4 reflect the methods similar characteristics as seen from the mouse data. With an exception, minP does not perform as poorly as in Section 3.5.2 and a larger number of genes are detected. The OW method shows much clearer patterns than the other methods. Of the 722 genes in Figure 4(a), 618 genes show consistent regulation across studies

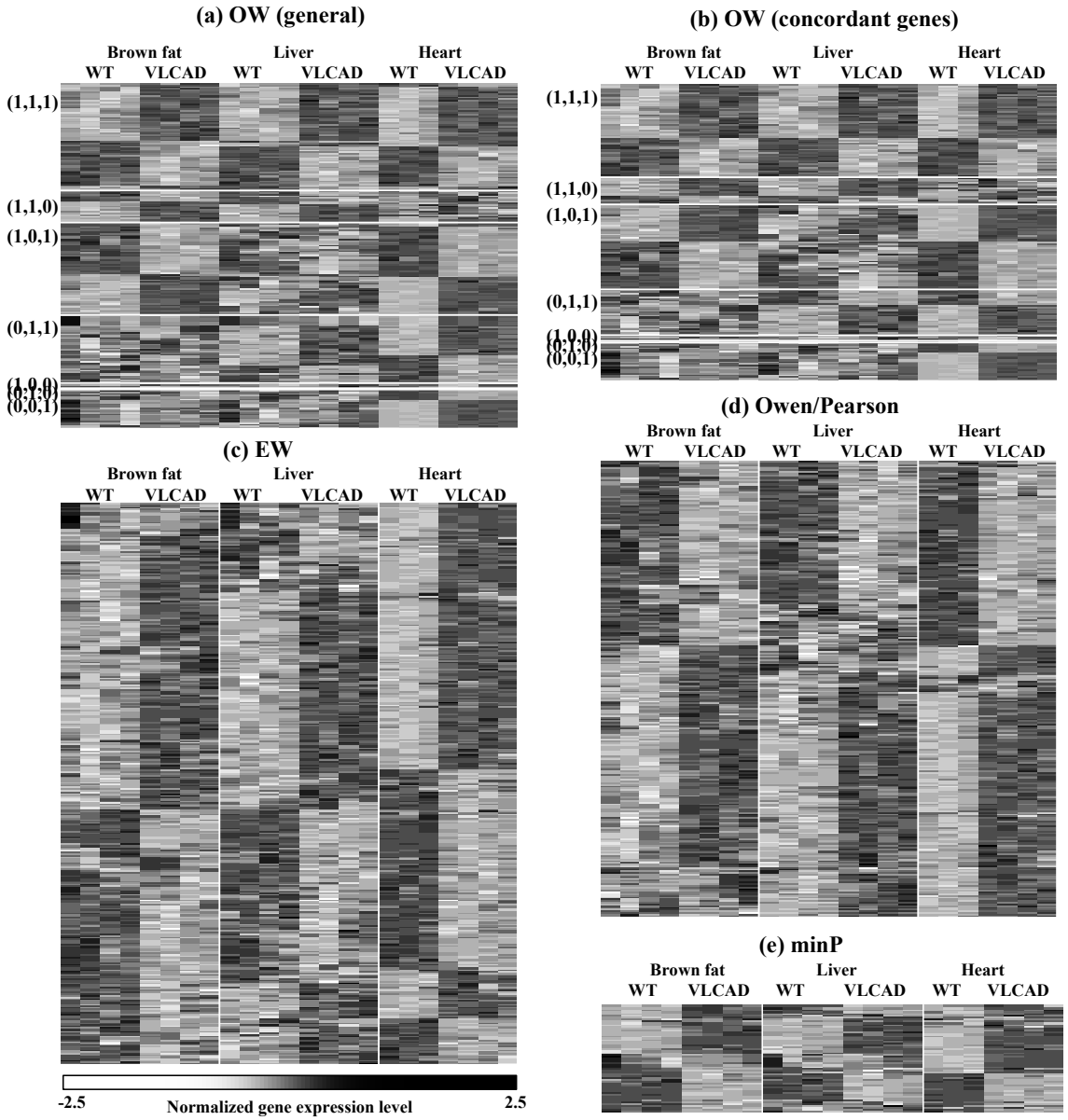

Figure 3: Heatmap of gene expressions for differentially expressed genes identified according to different methods in analyzing mouse data. Each row represents a single gene, each column represents a sample. The data were normalized within each tissue. In OW, genes are ordered by hierarchical clustering in each category formed by gene specific best weight. Genes identified from the other methods are ordered by hierarchical clustering.

Table 7: Five example genes from the mouse energy metabolism data

Gene	moderated-t statistic (p-value)			p-value(q-value) for meta analysis methods			
	Brown fat	Liver	Heart	OW	EW	minP	Pearson
1419182_at	-3.3(0.0004)	1.6(0.0190)	-1.9(0.0109)	0.0005	0.00001	0.0011	0.0004
$w*$	1	1	1	(0.030)	(0.001)	(0.088)	(0.012)
1423407_a_at	2.2(0.0027)	1.7(0.0121)	-3.7(0.0014)	0.0004	0.0001	0.0041	0.0018
$w*$	1	1	1	(0.031)	(0.001)	(0.112)	(0.034)
1418429_at	3.6(0.0003)	1.1(0.0672)	-3.2(0.0020)	0.0001	0.0001	0.0008	0.0010
$w*$	1	0	1	(0.028)	(0.001)	(0.088)	(0.023)
1449015_at	0.4(0.4558)	-3.3(0.0009)	-1.8(0.0110)	0.0008	0.0007	0.0017	0.0005
$w*$	0	1	1	(0.020)	(0.016)	(0.056)	(0.012)
1416415_a_at	-0.80(0.1496)	2.24(0.0026)	2.58(0.0023)	0.0012	0.0003	0.007	0.0005
$w*$	0	1	1	(0.035)	(0.011)	(0.089)	(0.012)

and are shown in Figure 4(b). About 14% discordant genes are observed. Causes of such discordant genes may include mistaken gene annotation in old array platforms (Dai et al., 2005), differential probe efficiencies, heterogeneous sample populations across studies and non-specific cross hybridizations. The findings suggest that results obtained from individual microarray studies require careful interpretation, and that synthesized analyses provide more powerful tests and appropriate validation.

3.6 DISCUSSION

In this chapter, we propose an optimally weighted statistic for combining multiple studies and applied it for combining microarray studies. The evaluation of meta-analysis methods heavily depends on the biological question being investigated and the corresponding statistical hypothesis settings. We formulated two hypothesis settings, HS1 and HS2, to identify differentially expressed genes considered significant in either partial data sets or in all data sets. The classical EW, minP and proposed OW methods were targeted on HS1 and are discussed in this chapter. We showed that OW, as well as EW and minP, were all admissible under a simplified scenario considered. In the power analysis, EW was more powerful when all data sets were significant and minP was more powerful when only one or few data sets were significant. The OW method, which can be viewed as a compromise between EW and minP, performed close to the better performing method in either extreme alternative

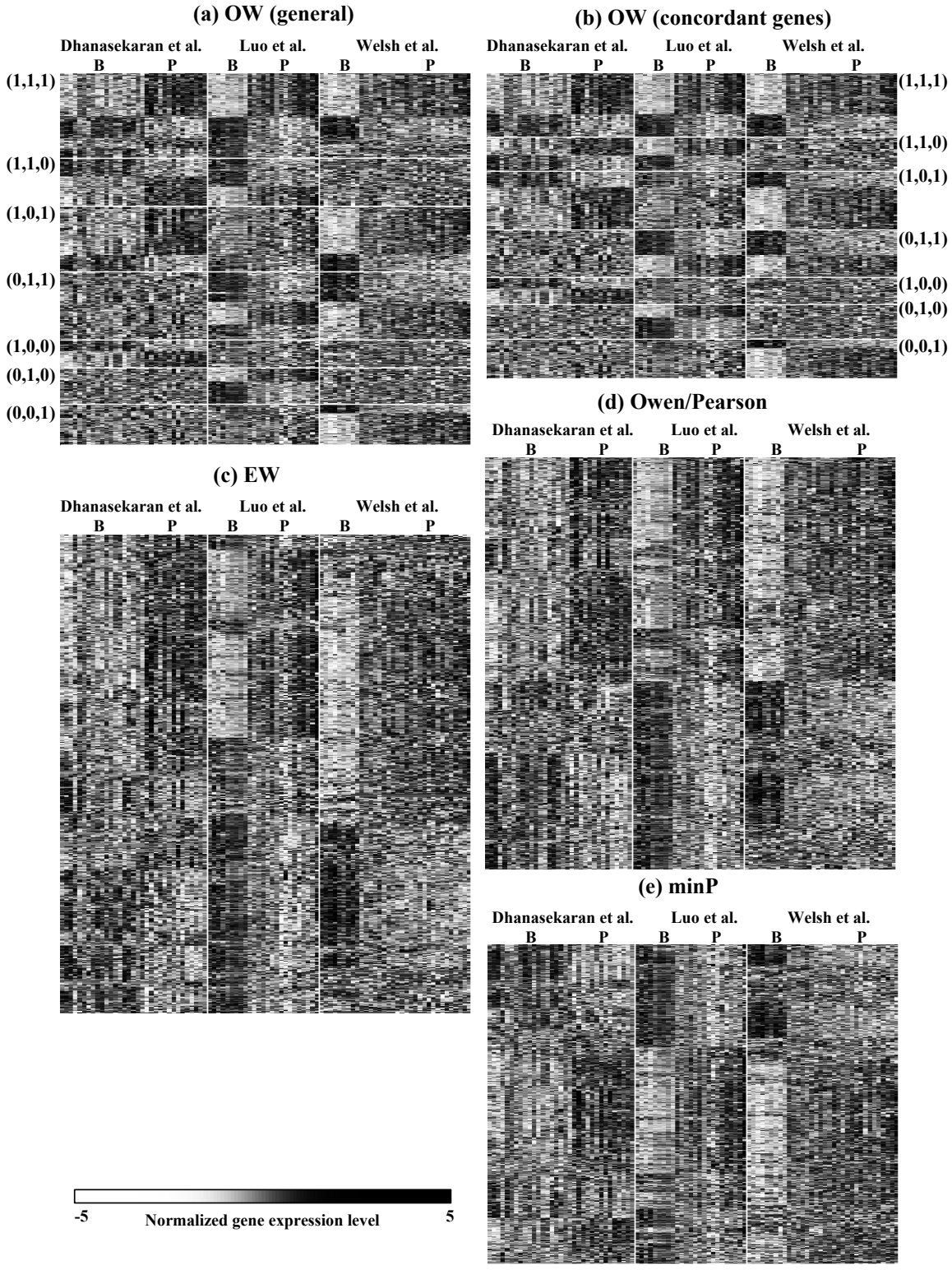

Figure 4: Heatmap of gene expressions for differentially expressed genes identified according to different methods in analyzing prostate cancer data. Each row represents a single gene, each column represents a prostate sample. The data were normalized within each study. In OW, the genes are ordered by hierarchical clustering in each category formed by gene specific best weight. Genes identified from the other methods are ordered by hierarchical clustering.

hypothesis settings (Figure 2). Simulation results also confirmed this robust property of OW (Table 4-6). In the applications, OW expressed an additional advantage of categorizing differentially expressed genes by their optimal weights, thus providing a practical basis for further biological exploration. The modified algorithm in Section 3.3 procedure II.B. avoided detection of genes with discordant regulation direction and was appealing for the specific biological purpose.

There are a few limitations and possible future extensions for this work. Firstly we assumed all studies contain identical matched gene list and no missing values. In practice, studies to be combined usually come from different microarray platforms. Requiring an identical matched gene list and no missing values will exclude many important genes appearing in one study but not in another. An extension allowing missing values will be necessary. Secondly we focused on the two-group comparison in this chapter and provided a modification for detecting only genes with concordant expression changes. For comparison of more than two groups, the F-statistic and its variations can be applied and the resulting p-values can be combined similarly. Small p-values across studies, however, do not guarantee concordant expression patterns. We are currently developing a pair-wise correlation approach to replace p-values for this problem.

While this chapter only considered combining multiple microarray studies, the methods can easily be extended for combination of multiple genomic, epigenomic and/or proteomic studies for instance, data sets from SNP arrays, genome arrays, methylation arrays, proteomic experiments and ChIP-on-chip experiments.

4.0 MINIMUM MULTI-CLASS CORRELATION STATISTIC WHEN COMBINING MULTIPLE MULTI-CLASS MICROARRAY STUDIES

4.1 BACKGROUND

Microarray technology provides an opportunity for global monitoring of gene expression activities. As the technology matures and becomes prevalent in biomedical research, many data sets have been accumulated in the public internet domain; for example, the NCBI Gene Expression Omnibus (Edgar et al., 2002), the EBI ArrayExpress (Parkinson et al., 2005) and the Stanford Microarray Database (Sherlock et al., 2001). The development of effective information integration of multiple microarray studies has gained increasing attention. Among the various types of microarray statistical analysis, detection of differentially expressed (DE) genes is one of the most important goals. Samples under two different conditions (e.g. normal versus diseased patients) are examined. Many statistical methods have been proposed for detecting biomarkers differentially expressed across the two classes (Breitling et al., 2005; Efron et al., 2001; Newton et al., 2004; Tusher et al., 2001). When multiple microarray studies are available, meta-analysis is expected to increase statistical power for DE gene detection. Rhode et al. (2002) was among the first to apply the traditional Fisher's method (Fisher, 1948) for combining multiple microarray studies. Many other approaches have been proposed later, including a lasso-based method (Ghosh et al., 2003), random effects models (Choi et al., 2003; Stevens and Doerge, 2005), Bayesian methods (Tseng et al., 2001; Jung et al., 2006; Conlon et al., 2007), rank-based approaches (Breitling et al., 2005; Hong et al., 2006) and others. Recently Li and Tseng (2008) elucidated two statistical hypothesis settings behind two separate biological goals in combining multiple array studies: (1) HS1: detecting biomarkers statistically significant in one or more studies, and (2) HS2: detecting

highly confident biomarkers differentially expressed in all studies. An optimally-weighted statistic was modified from the Fisher's score and was proposed for the former hypothesis setting. The optimal weights provided natural categorization of the detected biomarkers for further biological investigation. So far, most of the methods in the literature, including those described above, focus on the two-class "disease-versus-normal" setting. Combining data sets with more than two classes is rarely discussed.

In general, two metrics are commonly combined in the meta-analysis. The first is to combine the effect sizes of each study to generate a conclusion of overall effect size and its confidence interval, which is commonly seen in the research of evidence-based medicine. Random effects model, for example, is a popular method in this category. The second metric combines significance levels or their transformation scores. The famous Fisher's method belongs to the category that sums up the log-transformed p-values. Many other statistics including trimmed version of Fisher's method (Olkin and Saner, 2001), minimum p-value (Tippett, 1931) and Wilkinsons rth smallest p-value (1951) have also been considered. It is worth noting that the former metric is only valid in the two-class comparison. When more than two classes are considered, no single effect size can be computed and, instead, the patterns (between-group fluctuation of the group means) become the concern. The latter category is, however, extensible to data sets with more than two classes. The F-statistic (or equivalently ANOVA model) and its variants can be applied and p-values can be assessed and combined across studies.

In this chapter, we explore the method of ANOVA-maxP, which detects biomarkers with a significant pattern (large between-group variation versus small within-group variations) in all studies. We also note that small p-values (equivalently large F-statistics) in all studies do not guarantee a consistent pattern across studies. Figure 5 shows two example genes detected by ANOVA-maxP where the patterns across studies are not consistent.

To overcome this issue, we develop a pairwise multi-class correlation (MCC) measure. The correlation measure is defined through an equal-weight bivariate mixture model from the multi-class observations. A min-MCC algorithm is extended for combining multiple studies and the method guarantees the detection of only concordant pattern biomarkers. The performance of ANOVA-maxP and min-MCC is assessed through simulation and applications

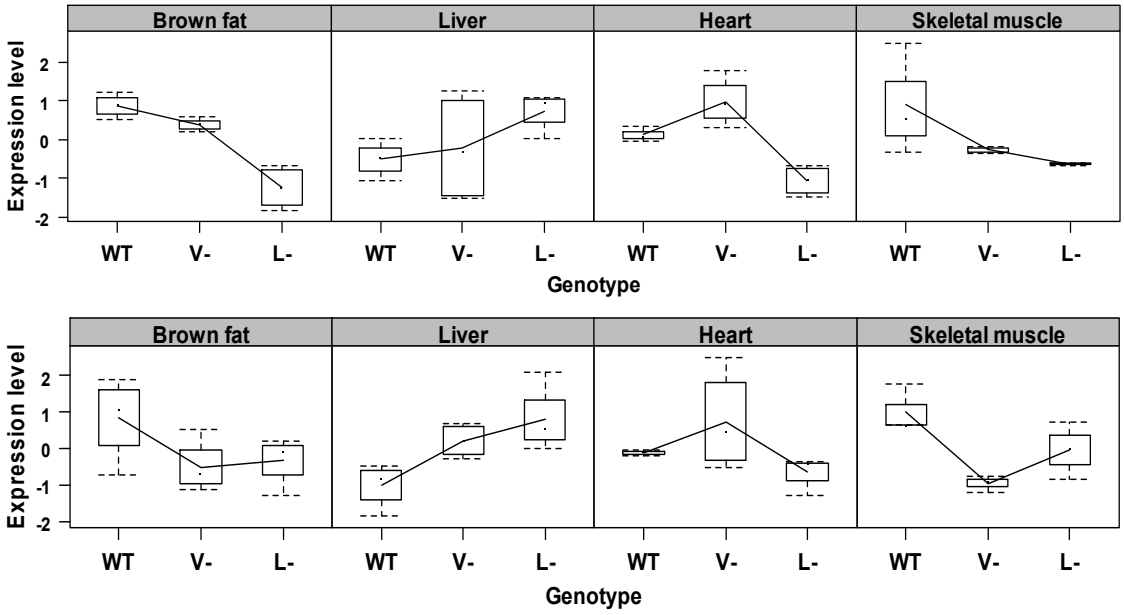

Figure 5: Two genes from the mouse metabolism data and detected by ANOVA-maxP. The patterns across tissues are lowly or negatively correlated. Box-plots of each genotype in each tissue are plotted and the mean expression levels are connected. Upper: expression pattern of Acsl5 (involved in fatty acid degradation pathway). Lower: expression pattern of gene Scd1 (involved in fatty acid synthesis pathway).

Table 8: Sample information of the mouse metabolism data set.

Tissue type	Brown fat			Liver			Heart			Skeletal			total
Genotype[a]	1	2	3	1	2	3	1	2	3	1	2	3	
Number of arrays	4	4	4	4	4	4	3	4	4	3	3	3	44

[a] 1: Wild type; 2: VLCAD -/-; 3: LCAD -/-.

to a multi-platform mouse trauma data set and a multi-tissue mouse energy metabolism data set. The result shows that ANOVA-maxP detects both concordant and discordant genes and min-MCC detects only concordant genes. The two methods are complementary and serve different biological purposes.

4.2 METHODS

4.2.1 Data description and notation

Two real data sets are used to evaluate proposed methods. The first data set involves samples from three genotype mice: wild-type (WT), LCAD knock-out (LCAD -/-) and VLCAD knock-out (VLCAD -/-). Deficiency of very long chain acyl-CoA dehydrogenase (VLCAD) is known to be related to a common energy metabolism disorder in children. On the other hand, LCAD (long-chain acyl-CoA dehydrogenase) deficient mice are known to have impaired fatty acid oxidation and develop a disease similar to other disorders of mitochondrial fatty acid oxidation. For each of the 12 mice (four mice in each genotype), four types of tissues (brown fat, skeletal, liver and heart) were applied to microarray experiment separately to study the expression changes across genotypes. For duplicate spots, the mean of the duplicates was used. Data from the four tissues were combined and log2 transformed. Genes with little information content (average log2-scaled means ≤ 7 or average log2-scaled standard deviations ≤ 0.4) are filtered out. A total of 4,288 genes are left for meta-analysis. Among the 48 arrays performed, four arrays were identified with a quality defect and were deleted from further analysis. The detailed sample information is described in Table 8.

Table 9: Sample information of the mouse trauma data set.

Array platform	Codelink					Affymetrix					total
Experimental conditions[a]	I	II	III	IV	V	I	II	III	IV	V	
Number of arrays	4	3	4	4	4	4	3	4	4	4	38

[a]I:no manipulation; II: 1.5h HS; III: 1.5h HS+BF+1h R; IV: 1.5h HS+BF+4.5h R; V: 1.5h HS+BF+6h R.

The second data set applied is about mouse trauma experiments. Victims of trauma-hemorrhagic shock (T-HS) (for example those due to car accident etc) often die due to severe, complex and uncontrollable physiological disturbances that occur in many organs, especially the liver. The progress of T-HS and resuscitation (R) is examined by well controlled murine systems to identify gene expression profiles that are characteristic of this stress. Specifically five groups of mice experiments were performed: (I) non-manipulated mice to serve as the negative control group; (II) 1.5h of Hemorrhagic Shock without resuscitation (1.5hHS) served as the positive control group; (III) 1.5h of hemorrhagic shock + bone fracture, followed by one hour of fluid resuscitation (1.5hHS+BF+1hR); (IV) Similar to group III except for 4.5h of fluid resuscitation (1.5hHS+BF+4.5hR); (V) Similar to group III except for 6h of fluid resuscitation (1.5hHS+BF+6hR). Four mice are in each group with the liver samples applied to microarray experiments (a total of 20 mice). The array experiments are done twice by both Codelink and Affymetrix platforms. One array of group II in Codelink and one array of group II in Affymetrix was of problematic quality and was removed from further analysis. Table 9 describes the experimental details of the multi-platform data. After some standard preprocessing procedures, 19,132 genes from Affymetrix platform and 26,063 genes from Codelink platform were matched by GeneCruiser, resulting in 6,338 common genes for the meta-analysis in this chapter.

The first data set contains a multi-tissue design. The application of HS1 helps to detect tissue-specific biomarkers and HS2 identifies consistent tissue-invariant biomarkers. Both hypothesis settings are of biological interest. On the other hand, the second data set contains a multi-platform design. HS2 is of interest to generate highly confident biomarkers confirmed by both platforms while HS1 becomes of less biological interest. The detected platform-

specific biomarkers from HS1 may help to identify technical issues across the two platforms.

In general, we consider S studies to be combined ($S = 4$ in mouse metabolism data and $S = 2$ in mouse trauma data). Among each study, K groups of samples are measured with nsk replicates for study s and group k. Denote by x_{sgki} the expression intensity of gene $g(1 \leq g \leq G)$, study $s(1 \leq s \leq S)$, group $k(1 \leq k \leq K)$ and replicated sample $i(1 \leq i \leq n_{sk})$. In this chapter, we particularly consider the situation when $K > 2$.

4.2.2 ANOVA-maxP for multiple studies

ANOVA-maxP is a natural extension of the traditional p-value based meta-analysis method. The method is to take, for each gene, the maximum p- value observed over the S studies as the test statistic. As a result, a biomarker is conservatively detected only if the p-values for all studies are small. In the multi-class data structure considered in this chapter, the ANOVA model is first used to test the significance of variation in gene expressions across phenotype groups in each study. The corresponding p-values from the F-test are then combined by taking the maximum (Box 1).

Box 1. Procedures for ANOVA-maxP

1. Compute the F-statistic, F_{gs}, for gene g in the sth study.
2. Permute group labels in each study for B times, and similarly calculated the permuted statistics, $F_{gs}^{(b)}$, where $1 \leq g \leq G$, $1 \leq s \leq S$ and $1 \leq b \leq B$.
3. Estimate the p-value of F_{gs} as $p_{gs} = \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(F_{g's}^{(b)} \geq F_{gs})}{(B \cdot G)}$, where $I(\cdot)$ is the indicator function that takes value one when the statement is true and zero otherwise. Similarly given $F_{gs}^{(b)}$, compute $p_{gs}^{b} = \dfrac{\sum_{b'=1}^{B}\sum_{g'=1}^{G} I(F_{g's}^{(b')} \geq F_{gs}^{(b)}}{(B \cdot G)}$.
4. The maximum p-value statistic is defined as $V_g = max_{1 \leq s \leq S} p_{gs}$. Similarly define $V_g^{(b)} = max_{1 \leq s \leq S} p_{gs}^{(b)}$.
5. Estimate the p-value of V_g by $p(V_g) = \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(V_{g'}^{B} \leq V_g)}{(B \cdot G)}$.
6. Estimate the q-value for each gene as $q(V_g) = \hat{\pi}_0 \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(V_{g'}^{B} \leq V_g)}{(B \sum_{g'=1}^{G} I(V_{g'} \leq V_g))}$.

4.2.3 Multi-class correlation (MCC) for a pair of studies

Below we describe our proposed pairwise multi-class correlation measure (MCC), given a gene, in two studies. For simplicity, we drop the subscript of gene g and studies s. Consider $x_{kj}(1 \leq k \leq K, 1 \leq j \leq n_k)$ to represent the expression intensity of class k, sample j for the first study and $x_{kj}(1 \leq k \leq K, 1 \leq j \leq m_k)$ for the second study. A naive measure to quantify the correlation of the expression patterns across two studies may be the direct sample correlation of $(x_{11}, \ldots, x_{1n_1}, \ldots, x_{K1}, \ldots, x_{Kn_k})$ and $(y_{11}, \ldots, y_{1m_1}, \ldots, y_{K1}, \ldots, y_{Km_k})$ if $n_k = m_k, \forall k$. However, since $n_k \neq m_k$ in general and this naive definition ignores the exchangeability within $(x_{11}, \ldots, x_{1n_k})$ and $(y_{11}, \ldots, y_{1m_k})$ for a given $1 \leq k \leq K$, we need to develop a better-defined correlation measure.

Assume we know the underlying distribution X_k and Y_k, where x_{kj} are i.i.d. from X_k, y_{kj} are i.i.d. from Y_k, $E(X_k) = \mu_{X_k}$, $E(Y_k) = \mu_{Y_k}$, $Var(X_k) = \sigma^2_{X_k}$ and $Var(Y_k) = \sigma^2_{Y_k}$. Also assume X_k's and Y_k's are independent. Define a bivariate distribution (X, Y) to be the equal mixture of bivariate distributions (X_k, Y_k) such that

$$F_{(X,Y)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X_k,Y_k)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X_k)}(s)F_{(Y_k)}(t), \tag{4.1}$$

where $F_X(\cdot)$ represent the cumulative distribution function of X. We define the multi-class correlation measure of $(X_1, \ldots, X_K)$ and $(Y_1, \ldots, Y_K)$ to be $cor(X, Y)$. It can be easily shown that

$$\begin{aligned}
&\rho((X_1, \ldots, X_K), (Y_1, \ldots, Y_K)) \\
&= \frac{E(XY) - EX \cdot EY}{\sqrt{Var(X) \cdot Var(Y)}} \\
&= \frac{\frac{1}{K}\sum_{k=1}^{K} \mu_{X_k}\mu_{Y_k} - (\frac{1}{K}\sum_{k=1}^{K} \mu_{X_k})(\frac{1}{K}\sum_{k=1}^{K} \mu_{Y_k})}{\sqrt{[\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{X_k} + \frac{1}{K}\sum_{k=1}^{K} (\mu_{X_k} - \bar{\mu}_X)^2][\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{Y_k} + \frac{1}{K}\sum_{k=1}^{K} (\mu_{Y_k} - \bar{\mu}_Y)^2]}}
\end{aligned} \tag{4.2}$$

where $\bar{\mu}_X = \frac{1}{K}\sum_{k=1}^{K} \mu_{X_k}$ and $\bar{\mu}_Y = \frac{1}{K}\sum_{k=1}^{K} \mu_{Y_k}$. This correlation measure takes values between -1 and 1. A large positive correlation indicates a similar pattern between two studies for a given gene.

In practice, distributions of (X_k, Y_k) are unknown. Instead we are given a set of observations $(\tilde{x}, \tilde{y})$, where $\tilde{x} = \{x_{kj}, 1 \le k \le K, 1 \le j \le n_k\}$,$\tilde{y} = \{y_{kj}, 1 \le k \le K, 1 \le j \le m_k\}$. Denote by X'_k the empirical distribution of $\{x_{kj}, 1 \le j \le n_k\}$ such that $F_{X'_k} = \sum_{j=1}^{n_k} I(x_{kj} \le t)$ and similarly $F_{Y'_k} = \sum_{j=1}^{n_k} I(y_{kj} \le t)$. Define (X', Y') to be an equal mixture of bivariate distribution (X'_k, Y'_k) such that

$$F_{(X',Y')}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X'_k,Y'_k)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X'_k)}(s)F_{(Y'_k)}(t). \tag{4.3}$$

The multi-class correlation (MCC) based on the observed $(\tilde{x}, \tilde{y})$ becomes

$$\begin{aligned}
&\rho(\tilde{x}, \tilde{y}) \\
&= \frac{E(X'Y') - EX' \cdot EY'}{\sqrt{Var(X') \cdot Var(Y')}} \\
&= \frac{\frac{1}{K}\sum_{k=1}^{K} \mu_{X'_k}\mu_{Y'_k} - (\frac{1}{K}\sum_{k=1}^{K} \mu_{X'_k})(\frac{1}{K}\sum_{k=1}^{K} \mu_{Y'_k})}{\sqrt{[\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{X'_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{X'_k} - \bar{\mu}'_X)^2][\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{Y'_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{Y'_k} - \bar{\mu}'_Y)^2]}},
\end{aligned} \tag{4.4}$$

where $\mu_{X'_k} = \sum_{j=1}^{n_k} x_{kj}/n_k$, $\mu_{Y'_k} = \sum_{j=1}^{n_k} y_{kj}/m_k$, $\sigma^2_{X'_k} = \sum_{j=1}^{n_k}(x_{kj} - \mu_{X'_k})^2/n_k$ and $\sigma^2_{Y'_k} = \sum_{j=1}^{n_k}(y_{kj} - \mu_{Y'_k})^2/m_k$.

Detailed procedures for the application of MCC for finding biomarkers with consistent patterns in two studies are listed in Box 2.

Box 2. Procedures of MCC for combining two studies

1. Compute MCC statistic, MCC_g for given gene g.
2. Permute group labels in each study for B times, and similarly calculated the permuted statistics, $MCC_g^{(b)}$, where $1 \le g \le G$, $1 \le s \le S$ and $1 \le b \le B$.
3. Estimate the p-value of MCC_g as
   $p(MCC_g) = \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(MCC_{g'}^{(b)} \ge MCC_g)}{(B \cdot G)}$.
4. Estimate the q-value for each gene of MCC_g as
   $q(MCC_g) = \hat{\pi}_0 \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(MCC_{g'}^{(b)} \le MCC_g)}{(B\sum_{g'=1}^{G} I(MCC_{g'} \le MCC_g))}$.

4.2.4 Min-MCC for multiple studies

The MCC measure described above measures correlation between two given studies. It can be extended for identifying genes with a consistent pattern across more than two studies. The minimum MCC is defined as $min - MCC_g = min_{1\leq u\leq v\leq S}(MCC_{g(u)(v)})$, where $MCC_{g(u)(v)}$ is the MCC measure for gene g and between study u and study v. With slight modification of the algorithm in Box 2, the algorithm for min-MCC is provided in Box 3.

Box 3. Procedures of min-MCC for combining multiple studies

1. Compute the MCC statistic, $MCC_{g(u)(v)}$ for given gene g and for a pair of studies u and v.
2. Permute group labels in each study for B times, and similarly calculated the permuted statistics, $MCC^{(b)}_{g(u)(v)}$ for study u and v, where $1 \leq g \leq G$, $1 \leq s \leq S$ and $1 \leq b \leq B$.
3. Calculate $min - MCC_g = min_{1\leq u\leq v\leq S}(MCC_{g(u)(v)})$ and $min - MCC^{(b)}_g = min_{1\leq u\leq v\leq S}(MCC^{(b)}_{g(u)(v)})$.
4. Estimate the p-value of $min - MCC_g$ as
$$p(min - MCC_g) = \frac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(min - MCC^{(b)}_{g'} \geq min - MCC_g)}{(B \cdot G)}.$$
5. Estimate the q-value for each gene of $min - MCC_g$ as
$$q(min - MCC_g) = \hat{\pi}_0 \frac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(min - MCC^{(b)}_{g'} \leq min - MCC_g)}{(B\sum_{g'=1}^{G} I(min - MCC_{g'} \leq min - MCC_g))}.$$

4.3 RESULTS

4.3.1 Simulation studies

We conducted two simulation scenarios for combining two and three genomic studies to assess the performance of our proposed min-MCC method and ANOVA-maxP, where the MCC method for combining two studies is a special case of min-MCC. Denote by x_{sgki} the expression intensity of study $s(1 \leq s \leq S)$, gene $g(1 \leq g \leq G)$, sample class $k(1 \leq k \leq K)$ and replicated sample $i(1 \leq i \leq n_{sk})$. In the first simulation scenario, we simulated two studies ($S = 2$) and the performance of MCC and ANOVA-maxP was compared. In the

Table 10: Settings of the first simulation scenario.

		Scenario 1	
Effect size	N	Study 1	Study2
		$(n_{11}, n_{12}, n_{13}) = (10, 5, 8)$ $(\mu_{11}, \mu_{12}, \mu_{13}), \sigma_1$	$(n_{21}, n_{22}, n_{23}) = (5, 8, 10)$ $(\mu_{21}, \mu_{22}, \mu_{23}), \sigma_2$
0.5	I	(1,3,5),3.5	(2,4,6),3.1
	II	(1,3,5),3.5	(6,4,2),3.1
	NULL	(0,0,0),3.5	(0,0,0),3.1
0.6	I	(1,3,5),2.9	(2,4,6),2.6
	II	(1,3,5),2.9	(6,4,2),2.6
	Null	(0,0,0)2.9	(0,0,0),2.6
0.7	I	(1,3,5),2.5	(2,4,6),2.2
	II	(1,3,5),2.5	(6,4,2),2.2
	Null	(0,0,0),2.5	(0,0,0),2.2
0.8	I	(1,3,5),2.2	(2,4,6),1.9
	II	(1,3,5),2.2	(6,4,2),1.9
	Null	(0,0,0),2.2	(0,0,0),1.9

I: genes with concordant patterns across studies. II: genes with discordant patterns across studies. Null: null genes.

second scenario, three studies were simulated ($S = 3$) and we compared ANOVA-maxP with min-MCC. Each study had three subclasses ($K = 3$). The numbers of replicates, n_{sk} ($1 \leq s \leq S$ and $1 \leq k \leq K$), were different among each subclass of each study. A total of G=2000 genes in each study were simulated. Among these 2000 genes, 300 genes displayed concordant patterns across all studies (category I) and 100 genes were discordant(category II). The expression intensities were simulated from $x_{sgki} \sim N(\mu_{sk}, \sigma_s^2)$. For genes with concordant pattern, mean vectors $\mu_s = (\mu_{s1}, \ldots, \mu_{sk})$ across studies had pair-wise correlation one. For discordant genes, the pair-wise correlation of mean vectors was low or negative. For the rest of the 1600 null genes, $(\mu_{s1}, \ldots, \mu_{sk}) = (0, \ldots, 0)$. The effect size was defined as the ratio of the standard deviation of the mean vectors to the within group standard deviation, σ_s. We chose effect sizes to be 0.5, 0.6, 0.7 and 0.8. The false discovery rate (FDR) was controlled at 0.05 for each method, and each simulation was repeated 200 times. The details of the simulation settings are described in Tables 10 and 11.

The average number of genes identified in each category under the two different scenarios are presented in Tables 12 and 13 respectively. As expected, the ANOVA-maxP detects both category I and category II genes because the p-values in ANOVA do not reflect the pattern information for each individual study. On the other hand, min-MCC method detects almost

Table 11: Settings of the second simulation scenario.

Effect size	N	Scenario 2		
		Study 1	Study 2	Study 3
		$(n_{11}, n_{12}, n_{13}) = (10, 5, 8)$ $(\mu_{11}, \mu_{12}, \mu_{13}), \sigma_1$	$(n_{21}, n_{22}, n_{23}) = (5, 8, 10)$ $(\mu_{21}, \mu_{22}, \mu_{23}), \sigma_2$	$(n_{31}, n_{32}, n_{33}) = (8, 10, 5)$ $(\mu_{31}, \mu_{32}, \mu_{33}), \sigma_3$
0.5	I	(1,3,5),3.5	(2,4,6),3.1	(1,4,7),4.4
	II	(1,3,5,3.5	(6,4,2),3.1	(1,7,4),5.3
	III	(0,0,0),3.5	(0,0,0),3.1	(0,0,0),4.4
0.6	I	(1,3,5),2.9	(2,4,6),2.6	(1,4,7),3.7
	II	(1,3,5),2.9	(6,4,2),2.6	(1,7,4),4.4
	Null	(0,0,0),2.9	(0,0,0),2.6	(0,0,0),3.7
0.7	I	(1,3,5),2.5	(2,4,6),2.2	(1,4,7),3.2
	II	(1,3,5),2.5	(6,4,2),2.2	(1,7,4),3.8
	Null	(0,0,0),2.5	(0,0,0),2.2	(0,0,0),3.2
0.8	I	(1,3,5),2.2	(2,4,6),1.9	(1,4,7),2.8
	II	(1,3,5),2.2	(6,4,2),1.9	(1,7,4),3.3
	Null	(0,0,0),2.2	(0,0,0),1.9	(0,0,0),2.8

I: genes with concordant patterns across studies. II: genes with discordant patterns across studies. Null: null genes.

only concordant genes (category I) and is much more powerful than the ANOVA-maxP if the goal is to identify only concordant genes. For instance, when the effect size equals 0.6 and three studies are combined, min-MCC detects an average of 265.2 (out of 300) genes of true concordant pattern genes, while ANOVA-maxP only identifies 206.1 together with 67.7 discordant genes. If the biological goal is to identify only genes with concordant patterns, min-MCC provides a better performance (false discovery rate should be calculated as FDR1=(II+null)/(I+II+null) in this case). On the other hand, if genes with discordant genes are also of biological interests, the ANOVA-maxP should be the method of choice (false discovery rate calculated as $FDR2 = (null)/(I + II + null)$). We note that the simulation results are all slightly conservative (true $FDR \cong 0.04$ when nominal FDR is controlled at 0.05). This is because the parameter π_0, the proportion of null genes, is set to equal 1 throughout all of the methods (Boxes 1-3). The estimation of π_0 has been a difficult issue and it has been shown that improper estimation can lead to poor inference. To always set $\pi_0 = 1$ can be conservative (i.e. few biomarkers are claimed as it should be). But in the situation that most genes in the genome are non-differentially expressed, which is true in many applications, the effect will be minimal.

Table 12: Evaluation of ANOVA-maxP and MCC methods by simulation in the first scenario.

Effect size	Methods	I	II	Null	$FDR1 = \frac{II+Null}{I+II+Null}$	$FDR2 = \frac{Null}{I+II+Null}$
0.5	ANOVA-maxP	41.9	13.6	2.4	0.28	0.04 [a]
	min-MCC	113.1	0.0	4.7	0.04 [b]	0.04
0.6	ANOVA-maxP	113.5	37.4	6.5	0.28	0.04 [a]
	min-MCC	205.6	0.0	8.1	0.04 [b]	0.04
0.7	ANOVA-maxP	174.2	57.4	9.8	0.28	0.04 [a]
	min-MCC	260.5	0.1	10.5	0.04 [b]	0.04
0.8	ANOVA-maxP	228.8	76.1	13.2	0.28	0.04 [a]
	min-MCC	286.7	0.1	11.4	0.04 [b]	0.04

I: genes with concordant patterns across studies. II: genes with discordant patterns across studies. Null: null genes.

[a]ANOVA-maxP detects both concordant and discordant genes. FDR2 is a better measure for false discoveries.

[b]MCC detects only concordant genes. FDR1 is a better measure for false discoveries.

Table 13: Evaluation of ANOVA-maxP and min-MCC methods by simulation in the second scenario.

Effect size	Methods	I	II	Null	$FDR1 = \frac{II+Null}{I+II+Null}$	$FDR2 = \frac{Null}{I+II+Null}$
0.5	ANOVA-maxP	126.6	42.0	7.2	0.28	0.04 [a]
	min-MCC	210.2	0.2	9.2	0.04 [b]	0.04
0.6	ANOVA-maxP	206.1	67.7	11.6	0.28	0.04 [a]
	min-MCC	265.2	0.05	11.5	0.04 [b]	0.04
0.7	ANOVA-maxP	258.5	85.9	15.2	0.28	0.04 [a]
	min-MCC	289.2	0.02	12.	0.04 [b]	0.04
0.8	ANOVA-maxP	286.0	95.0	15.9	0.28	0.04 [a]
	min-MCC	297.6	0.0	12.8	0.04 [b]	0.04

I: genes with concordant patterns across studies. II: genes with discordant patterns across studies. Null: null genes.

[a]ANOVA-maxP detects both concordant and discordant genes. FDR2 is a better measure for false discoveries.

[b]min-MCC detects only concordant genes. FDR1 is a better measure for false discoveries.

4.3.2 Application to mouse metabolism data

We applied both the ANOVA-maxP and the min-MCC methods to the mouse metabolism data. The first biological goal was to identify biomarkers that have a consistent pattern across all four tissues (i.e. reliable tissue-invariant biomarkers). To achieve this goal, the min-MCC was better. A total of 394 genes were identified and are displayed in Figure 6A. It is clearly seen that these genes have a clear concordant pattern across all four tissues. A simple cluster analysis can further group them into six major patterns for further biological investigations. For the second biological goal, we were also interested in genes with clear but discordant patterns across tissues. These tissue-dependent biomarkers reflected tissue-specific biological changes under VLCAD and LCAD mutations. The ANOVA-maxP identified 676 genes and the heatmap of detected biomarkers is shown in Figure 6B. Figure 6C shows a histogram of min-MCC (minimum of pair-wise multi-class correlation measure) of the 676 detected biomarkers. Also 431 of the 676 genes (63.76%) had negative min-MCC (i.e. with discordant pattern in at least a pair of tissues). Figure 6D shows a heatmap of these 431 discordant genes and they are potential targets to identify tissue-specific regulators in the mutations. For instance, Figure 5 shows the box-plots and patterns of two genes, Acsl5 and Scd1, among the 722 genes. Acsl5 is known to be involved in fatty acid degradation pathway. When VLCAD or LCAD is knocked out, this gene is down-regulated in brown fat tissue, suggesting the loss of degradation and metabolism activity of fatty acid. On the other hand, Scd1 is known to help synthesize fatty acid. Its up-regulation in the liver tissue when VLCAD or LCAD is deleted seems to suggest that the defect of metabolism in these mice has increased the activity of fatty acid synthesis in liver.

4.3.3 Application to mouse trauma data

We similarly applied the ANOVA-maxP and the min-MCC to the mouse trauma data. We note that the two studies to be combined were from two commercial platforms, Affymetrix and Codelink. Ideally both array platforms measure identical samples and there should not exist any discordant pattern biomarker. Combining the two data sets should increase statistical power and detect more concordant pattern genes. Indeed, by controlling the FDR

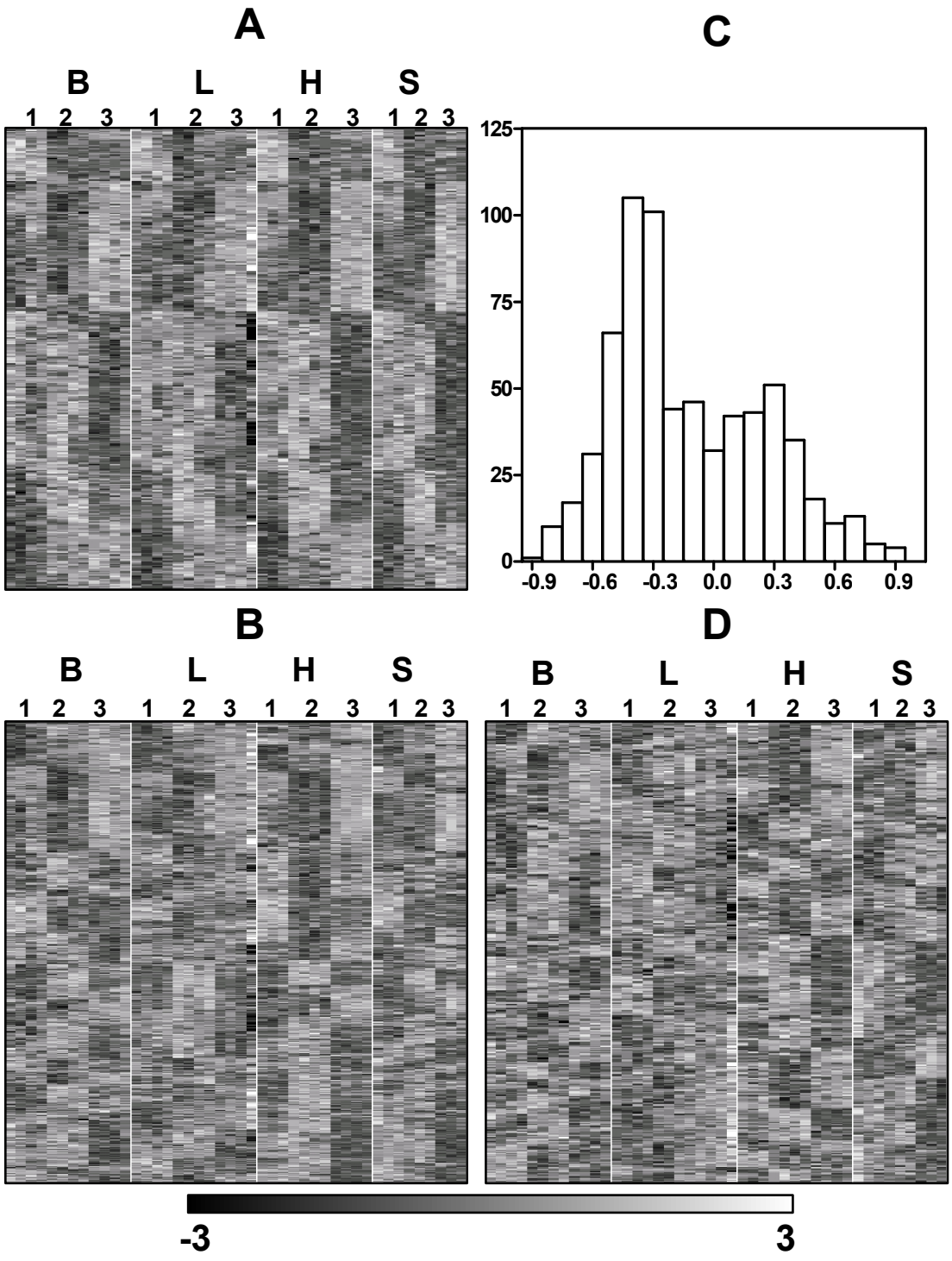

Figure 6: Application to mouse metabolism data. A: heatmap of 394 min-MCC genes, B: heatmap of 676 ANOVA-maxP genes, C: histogram of min-MCC of the 676 ANOVA-maxP genes, D: heatmap of 431 discordant ANOVA-maxP genes. (B: Brown fat; L: Liver; H: Heart; S: Skeletal muscle. 1: Wild type; 2: VLCAD -/-; 3: LCAD -/-.).

at 0.05, 4,004 genes were identified using the min-MCC (figure 7A). These highly-reliable biomarkers confirmed by both platforms were used for further cluster analysis and pathway analysis to understand the biological changes under different levels of severity of trauma (manuscript in preparation). On the other hand, the ANOVA-maxP detected 3,587 genes (heatmap shown in Figure 7B) and 208(5.8%) genes showed discordant patterns of negative MCCs across the two platforms (Figure 7C). The higher proportion of genes with concordant patterns confirmed that the two array platforms are highly reproducible. The 208 discordant genes, however, need further investigation (heatmap in Figure 7D). The discordances are possibly due to mistaken gene annotation, differential hybridization efficiencies caused by different probe selection criteria or non-specific cross-hybridization.

4.4 DISCUSSION

Systematic information integration helps to increase the statistical power of biomarker detection. The evaluation of performance and choice of method depend on the ultimate biological goal. Many meta-analysis methods have been proposed for combining two-class genomic studies. In this chapter, we consider the issue of combining multi-class microarray studies. Two methods are proposed and evaluated. The statistic ANOVA-maxP takes the maximum of the ANOVA p-values in all studies. The method detects genes with clear class-wise patterns (i.e. large between-class variation and small within-class variation). It is easily seen that p-values of the ANOVA method cannot provide pattern information and genes detected may have discordant patterns across studies. To identify concordant pattern biomarkers across studies, we develop a novel multi-class correlation measure (MCC) via a bivariate equal-weight mixture distribution from the observations of a pair of studies. The method is extended to taking the minimum of the MCC of all pairwise studies (min-MCC) when more than two studies are considered. We showed that the ANOVA-maxP and min-MCC are complementary methods for combining multiple multi-class microarray studies depending on the ultimate biological purposes. When study variations are expected and both study-invariant (concordant pattern) and study-specific (discordant pattern) genes are of biological inter-

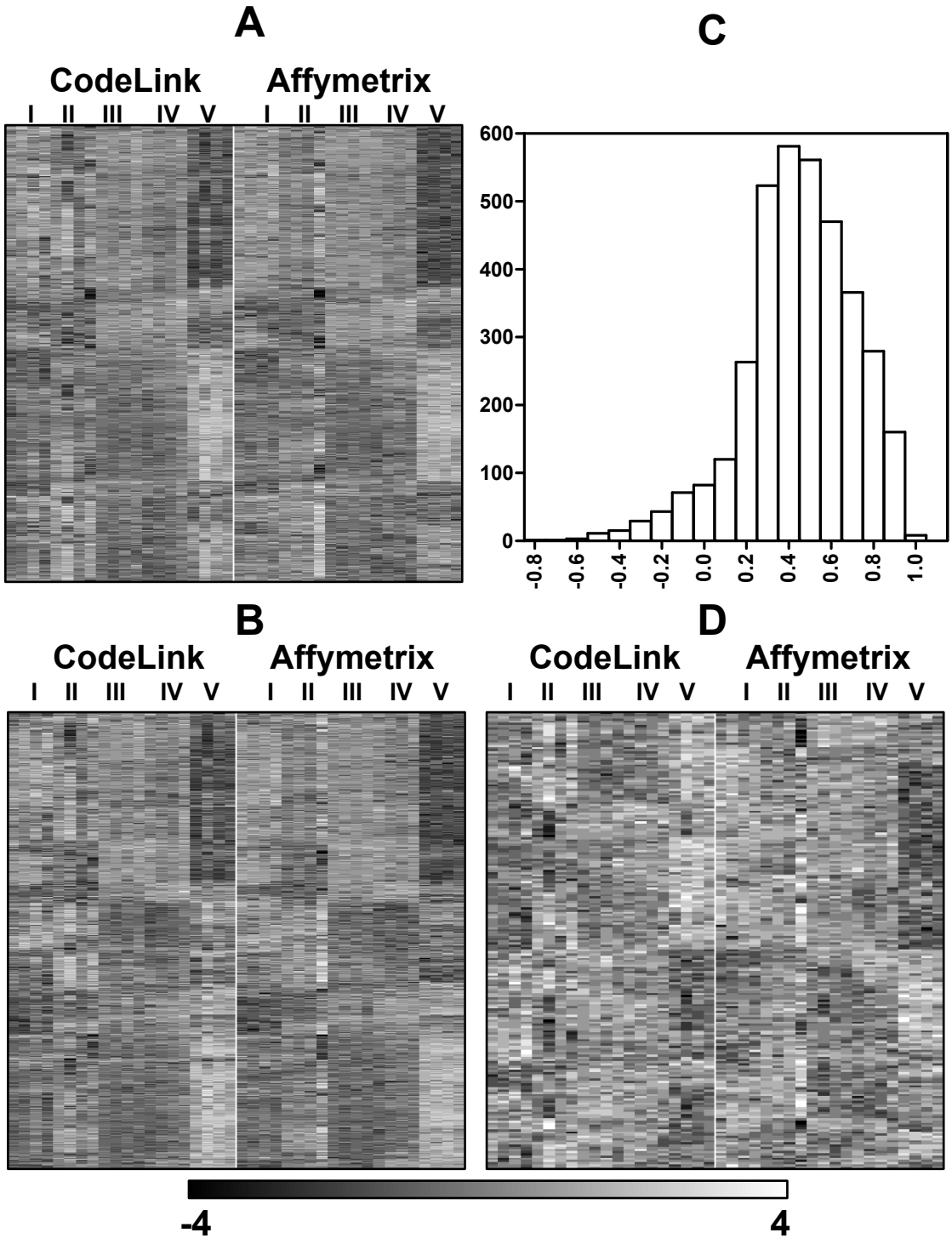

Figure 7: Application to mouse trauma data. A: heatmap of 4,004 min-MCC genes, B: heatmap of 3,587 ANOVA-maxP genes, C: histogram of min-MCC of the 3,587 ANOVA-maxP genes, D: heatmap of 208 discordant ANOVA-maxP genes.

ests, the ANOVA-maxP performs well. When detecting highly reliable biomarkers (i.e. only study-invariant or concordant pattern genes) is the goal, the min-MCC is a better choice and provides better statistical power.

There are a few possible extensions and future directions. The two proposed methods utilize the most conservative and extreme statistic (maximum of ANOVA p-values and minimum of pairwise MCC). This may be too stringent and sensitive to outliers especially when K is large. A quick modification may be to use the rth ranked statistic instead.

Currently we consider all studies to have identical K classes. Both ANOVA-maxP and min-MCC can be extended for studies containing mismatched and different number of classes. For min-MCC, the pairwise MCC can be defined using only the overlapping classes across a pair of studies.

5.0 GENOMEMETA PACKAGE

5.1 INTRODUCTION

With microarray technology becoming more prevalent in recent years, it is now common for several laboratories to employ the same microarray technology to identify differentially expressed genes that are related to the same phenomenon in the same species. Summarizing information from multiple studies is becoming increasingly popular in this field. Meta-analysis is an important component and often uses statistical methods to combine results from a series of related studies. Increasing statistical power in detecting differentially expressed genes and confidence in inferences from individual studies are two major goals of meta-analysis. The need of meta-analysis tools for combining microarray studies is obvious.

We develop a R package, *genomeMeta*, partly done in C which contains classes and methods useful for analyzing and visualizing multiple microarray data sets. In *genomeMeta*, we implement many of the tools described in previous chapters including the two statistics we propose in chapter 3 and 4. Inputs are a list of microarray datasets with class information for each dataset. *genomeMeta* produces meta-analysis results according to the different methods and visualization of genes identified by the different methods.

This chapter provides an overview of how *genomeMeta* organizes data and describes the methods utilized. And finally an example shows how the package is used.

5.2 DESCRIPTION

5.2.1 Data structure

We assume that as a first step all studies have been appropriately pre-processed and genes in each study have been matched and merged by a common key (e.g. Unigene ID). We require that the data object contains two lists. The first is a list of datasets from multiple studies. Each dataset is a matrix with rows representing genes or variables, and columns representing arrays or samples. The second is also a list which contains labels for each dataset. Each label is an object of class "factor".

5.2.2 genomeMeta methods

In *genomeMeta*, we implement many of the tools described in Chapters 2, 3 and 4. We provide methods for the combination of datasets based on both two sample comparisons and multiple-class comparisons. The available methods are shown in Figure 8. If each study contains two classes, the traditional two-sample t statistic with unequal variance or the moderated t statistic described in chapter 3 can be performed. The p-values are then computed by permutation. For combining p-values, four methods are available to choose from minP, maxP, Fisher and OW. As discussed and compared in previous chapters, maxP aims to identify genes differentially expressed in all studies. The others methods aim to identify genes differentially expressed in at least one of the studies. OW further divides genes into several categories showing those in which study the gene is significant. minP is not recommended for its poor power and unstable performance in general as shown in chapter 3. If a study has more than two classes, the F statistic with either equal or unequal variances in analysis of variance is performed. Similarly, the p-values of the observed F statistics are calculated from permutation. The p-values are then combined in the same manner.

Finally, genes with consistent patterns in all studies might be of more interest to researchers. We provide a powerful method min-MCC which is derived based on an equal mixture of bivariate normal distributions.

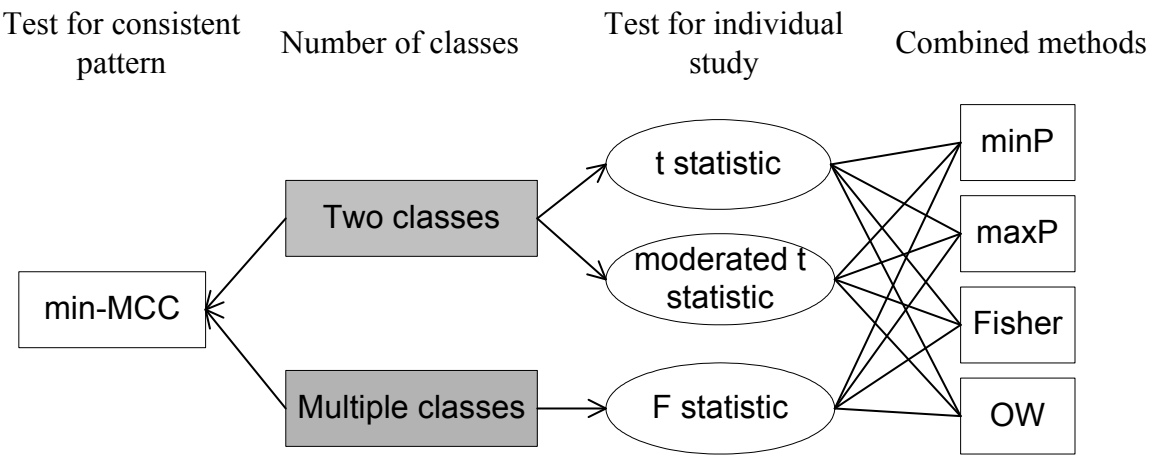

Figure 8: The methods implemented in *genomeMeta* package.

5.3 EXAMPLE

Following the example for the *GeneMeta* package, we split the breast tumor study reported by West et al. (2001) into two artificial studies and combine results from the two studies using different methods. As expected, according to the quantile-quantile plot of the observed vs. expected Cochran's Q values (Gentleman et al. (2008)), the two datasets are homogeneous. We fit the *moderated-t+OW* and *moderated-t+Fisher* methods to the artificial split. A plot of genes whose intensity levels are standardized within study is provided by the function $plot.Metagene$. If we are more interested in the number of genes that are below a given threshold for the FDR we can use $plot.fdr$.

$> dt < -list(data = list(d1, d2), label = list(l1, l2))$

$> library(genomeMeta)$

$> modt.ow < -genomeMeta(dt, nperm = 100, test =' modt', meta.method =' OW')$

$> plot.fdr(modt.ow)$

$> plot.Metagene(modt.ow, fdr.cut = 0.05)$

$> modt.maxp < -genomeMeta(dt, nperm = 100, test =' modt', meta.method =' Fisher')$

$> plot.fdr(modt.maxp)$

$> plot.Metagene(modt.maxp, fdr.cut = 0.05)$

The second example is the mouse energy metabolism data. We apply both the ANOVA-maxP and the min-MCC methods to the data. Plots are shown in Figure 10.

$> dt < -list(data = list(brown, liver, heart, skeletal),$

$label = list(colnames(brown), colnames(liver), colnames(heart), colnames(skeletal)))$

$> library(genomeMeta)$

$> F.maxp < -genomeMeta(dt, nperm = 100, test =' F', meta.method =' maxP')$

$> plot.fdr(F.maxp)$

$> plot.Metagene(F.maxp, fdr.cut = 0.05)$

$> minmcc < -genomeMeta(dt, nperm = 100, meta.method =' minMCC')$

$> plot.fdr(minmcc)$

$> plot.Metagene(minmcc, fdr.cut = 0.05)$

With 200 permutations and 2000 genes, the analysis takes approximately within one minute.

5.4 DISCUSSION

genomeMeta provides methods for analysis of multiple microarray studies. We organize the data and output in an object that can be queried for individual analysis results, combined analysis results and visualization of significant genes. More extensions are under way which include the min-MCC methods dealing with studies containing mismatched and different numbers of classes and other meta analysis methods as described in section 3 of chapter 2.

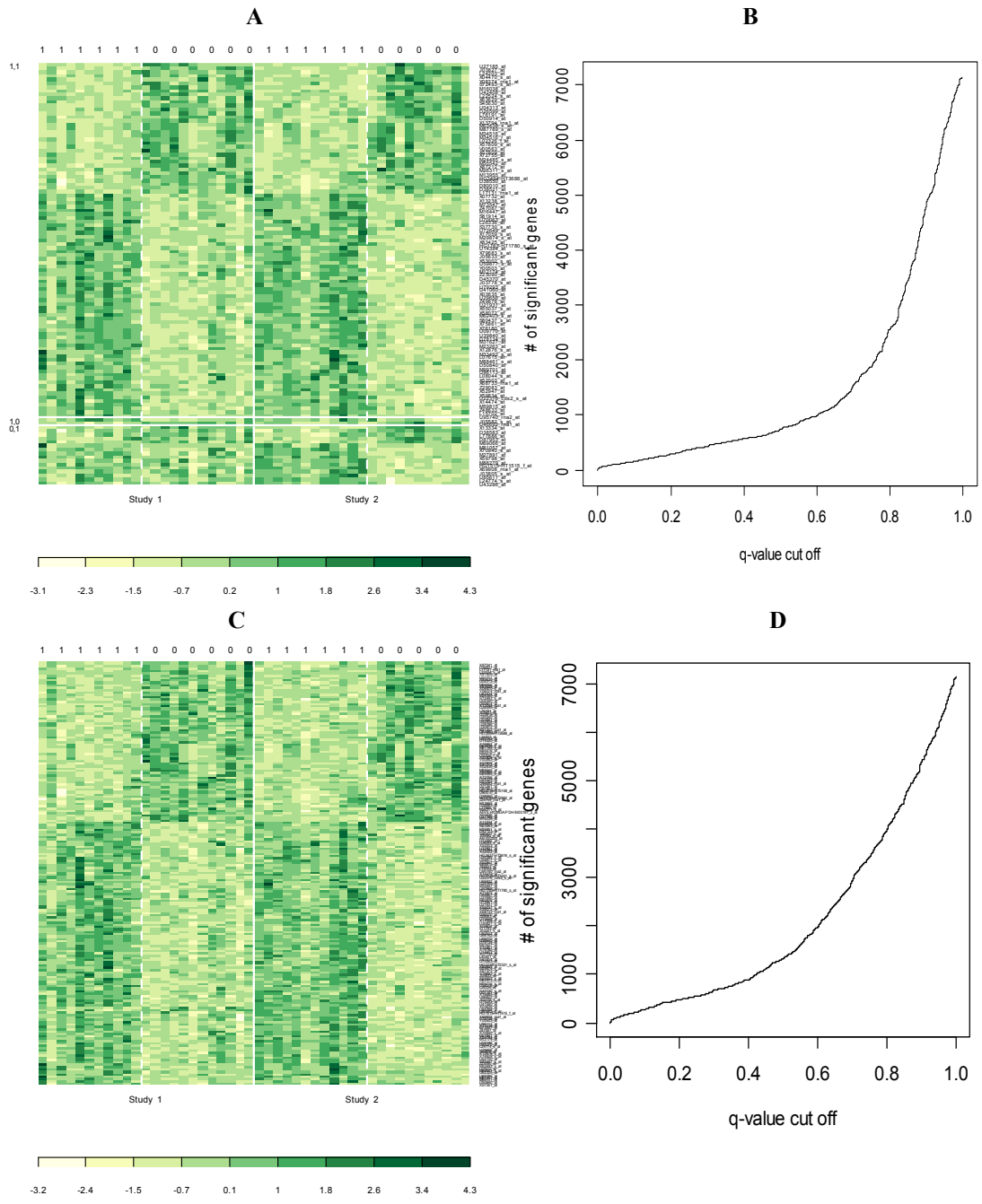

Figure 9: Genes identified using Meta-analysis for breast tumor studies. A. moderated t + OW method. Genes identified by OW with $FDR \leq 0.05$. OW classifies genes into several categories. B. moderated t + OW method. The number of genes that are below the threshold for FDR. C. moderated t + Fisher method. Genes identified by OW with $FDR \leq 0.05$. D. moderated t + Fisher method. The number of genes that are below the threshold for FDR.

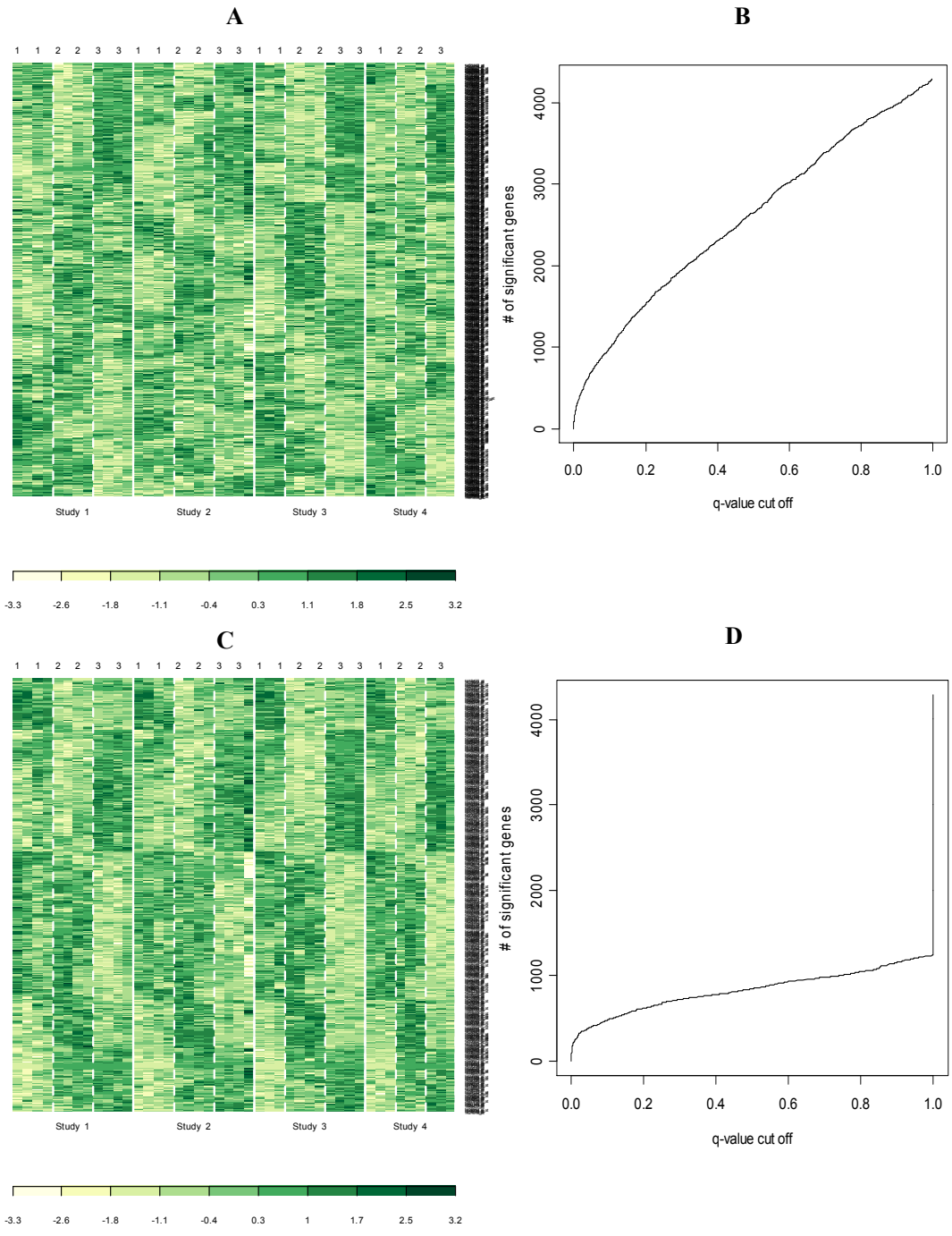

Figure 10: The heatmaps of genes identified using Meta-analysis for mouse energy metabolism studies. A. F + maxP method. F+maxP identifies genes ($FDR \leq 0.05$) whose expressions significantly vary across genotypes in all tissues. B. F + maxP method. The number of genes that are below the threshold for FDR. C. min-MCC method. Only genes with consistent variation across genotypes in all tissues are detected ($FDR \leq 0.05$). D. min-MCC method. The number of genes that are below the threshold for FDR.

6.0 CONCLUSION

Large-scale meta-analysis of genomic studies is becoming increasingly common, because it is now more feasible and there are a greater volume of data sets available. In this dissertation, I have addressed these issues as follows:

In Chapter 3, I clarified that some existing meta-analytic methods are for different biological purposes. The need to distinguish them is more obvious in genomic studies. Two complementary hypothesis settings are outlined. And all meta-analytic methods targeted for their respective hypothesis settings are clarified. Focusing on one of the hypotheses, we proposed a new meta-analytic method to facilitate combination of similar microarray studies. This method is named as the optimally-weighted statistic, which is based on the weighting scheme of the famous Fisher's method. The statistic is constructed as the minimum of the p-values of the weighted Fisher's statistic for all possible weights. We have shown that this method is admissible under Birnbaum's theorem. By comparing it with other popular methods, we showed that it had better power for detecting genes differentially expressed in part of the data sets with a tradeoff of slightly lesser power than Fisher's method in detected genes differentially expressed in all data sets. In general our method is powerful and superior to the others. Moreover, the optimal weights of each study for each gene produced by the method can lead to further categorization of differentially expressed genes, showing in which studies the identified genes are differentially expressed. The application to the multi-tissue mouse energy metabolism study and multiple prostate cancer studies have shown that the optimally-weighted statistic is very well suited to the setting where we expect tests to vary across data sets. We conducted extensive simulations to assess the performance of our proposed method and some existing methods with regard to the number of genes detected in different settings and the estimate of the false discovery rate under different scenarios.

Optimally-weighted method is more robust to varying scenarios.

In contrast to the combining estimates approach, the optimally-weighted method combines p-values. The principal objection to the combining p-value approaches is failure in producing direction and magnitude. However, data structures and statistical hypotheses in multiple microarray studies may differ, making a direct combination of estimates inappropriate. Moreover, there are two ways to incorporate direction information for a p-value based method. One is to incorporate signs according to the effect estimates, the other is to consider a one-sided p-value. We utilized the first way to further filter out identified genes with inconsistent regulations across data sets. However, this strategy is not feasible where there are multiple classes in each study.

To deal with studies with more than two classes, we proposed two methods in chapter 4. The first is a natural extension of Wilkinson's maximum p-value method. The ANOVA model is first used to test the significance of variation in gene expressions across multiple classes in each study. The corresponding p-values from the F-test are then combined by taking the maximum. This approach is denoted as the ANOVA-maxP. Obviously, the ANOVA-maxP identifies genes significantly vary across classes in all studies. It is unable to filter genes that differ in any pattern between classes across studies. Here, we proposed a second test statistic, the multi-class correlation statistic (MCC), based on an equal mixture of bivariate normal distributions. We assume that the gene expressions from the same class of two studies follow a bivariate distribution. We further assume that there is an equal probability of sampling each class. This assumption allows us to conveniently construct a statistic to assess the correlation between two studies. Moreover, a large positive MCC indicates a similar pattern between two studies. It also overcomes the infeasibility of Pearson's correlation when the studies have unequal sample sizes. The MCC measure is further extended to identify genes with a consistent pattern across more than two studies. We considered minimizing all pairwise multi-class correlations of all studies (min-MCC). The min-MCC was compared with the ANOVA-maxP in a simulation study and in applications to multi-tissue mouse energy metabolism and mouse trauma data sets. In the simulation, the min-MCC shows much higher power in identifying genes with consistent patterns across studies than ANOVA-maxP. It is not surprising that the ANOVA-maxP also identifies plenty of genes with inconsistent

patterns across studies. Both methods have been shown to be indispensable in the applications, and the min-MCC helps to detect genes have same pattern across studies. On the other hand, genes that are differentially expressed with different patterns across studies are also of interest to researchers. In this case, the combination of the ANOVA-maxP and the min-MCC can be implemented.

There are a few possible extensions and future directions. The two proposed methods utilize the most conservative and extreme statistic (maximum of ANOVA p-values and minimum of pairwise MCC). This may be too stringent and sensitive to outliers especially when K is large. A quick modification may be to use the rth ranked statistic instead. Currently we consider all studies to have identical K classes. Both the ANOVA-maxP and the min-MCC can be extended for studies containing mismatched and different number of classes. For the min-MCC, the pairwise MCC can be defined using only the overlapping classes across a pair of studies.

In applications, we only consider matched genes from all studies. In practice, studies to be combined usually come from different microarray platforms. Genes that are missing in part of the studies are not considered. As the number of studies increases, much fewer genes can be matched. An extension allowing non-matched genes is necessary. This dissertation only considered integrative analysis of microarray data sets, the statistical methods we proposed are flexible and can be easily applied to other similar data sets such as SNP arrays, genome arrays, methylation arrays, proteomic experiments and ChIP-on-chip experiments.

In the end, we developed a R package - *genomeMeta* which implemented the statistical tools we described in this dissertation. More extensions as we described above are under process.

This work attempts to improve the identification of biologically meaningful genes and could improve significant findings by providing more accurate answers and providing a brand new direction to a rarely discussed statistical issue yet important in practice.

BIBLIOGRAPHY

[1] R.L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:29–300, 1982.

[2] A. Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49:559–574, 1954.

[3] A. Birnbaum. Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *The Annals of Mathematical Statistics*, 26:21–36, 1955.

[4] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H.D. Rosas, S.M. Hersch, P. Hogarth, B. Bouzou, R.V. Jensen, and D. Krainc. Genome-wide expression profiling of human blood reveals biomarkers for huntingtons disease. *Proceedings of the National Academy of Sciences*, 102:11023–11028, 2005.

[5] R. Breitling and P. 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*, 3:1171–1189, 2005.

[6] R. Breitling, Armengaud. P., A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 573:83–92, 2004.

[7] J. Cardoso, J.H. Boer, H. Morreau, and R. Fodde. Expression and genomic profiling of colorectal cancer. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1775:103–137, 2007.

[8] J.K. Choi, U. Yu, S. Kim, and O.J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19:84–90, 2003.

[9] J. Cohen. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, 2 edition, 1988.

[10] J. Cohen. A power primer. *Psychological Bulletin*, 112:155–159, 1992.

[11] E.M. Conlon, J.J. Song, and J.S. Liu. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, 7:247–2590, 2006.

[12] R.D. Cousins. Annotated bibliography of some papers on combining significances or p-values. *arXiv:0705.2209v1*, 2007.

[13] M. Dai, P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, W.E. Bunney, R.M. Myers, T.P. Speed, H. Akil, S.J. Watson., and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res*, 33, 2005. available online: e175. 10.1093/nar/gni179.

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

[15] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin, and A.M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412:822–826, 2001.

[16] R. M. J. Donahue. A note on information seldom reported via the p value. *The American Statistician*, 53:303–3066, 1999.

[17] R. Edgar, M. Domrachev, and A.E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, 2002.

[18] B. Efron, J. D. Tibshirani, R. andStorey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[19] R.A. Fisher. *Statistical Methods for Research Workers*. Edinburgh, Oliver & Boyd, 4 edition, 1932.

[20] R.A. Fisher. *The Design of Experiment*. New York: Hafner Press, 1935.

[21] E.O. George. *Combining independent one-sided and two-sided statistical tests-Some theory and applications*. PhD thesis, University of Rocheser, 1977.

[22] D. Ghosh, T.R. Barrette, D. Rhodes, and A.M. Chinnaiyan. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomic*, 3:180–188, 2003.

[23] G.V. Glass. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5:3–8, 1976.

[24] I.J. Good. On the weighted combination of significance tests. *Journal of the Royal Statistical Society*, 17:264–26, 1955.

[25] L.V. Hedges. Meta-analysis. *Journal of Educational Statistics*, 17:279–296, 1992.

[26] L.V. Hedges and I. Olkin. *Statistical methods for meta-analysis*. New York:Academic, 1985.

[27] L.V. Hedges and J.L. Vevea. Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3:486–504, 1998.

[28] J.P.T. Higgins and S.G. Thompson. Quantifying heterogeneity in a metaanalysis. *Statistics in Medicine*, 21:1539–1558, 2002.

[29] F. Hong, Breitling R., C.W. McEntee, B.S. Wittner, J.L. Nemhauser, and J. Chory. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22:2825C2827, 2006.

[30] Y. Jung, M. Oh, D.W. Shin, S. Kang, and H.S. Oh. Identifying differentially genes in meta-analysis via bayesian model-based clustering. *Biometrical Journal*, 48:435–450, 2006.

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

[32] R.C. Littell and J. L. Folks. Asymptotic optimality of fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66:802–806, 1971.

[33] J. Luo, D.J. Duggan, Y. Chen, J. Sauvageot, C.M. Ewing, M.L. Bittner, J.M. Trent, and W.B. Isaacs. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Research*, 61:4683–4688, 2001.

[34] Y. Moreau, S. Aerts, B. De Moor, B. De Strooper, and M. Dabrowski. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, 19:570–577, 2003.

[35] M.A. Newton, A. Noueiry, D. Sarkar, and P. Ahlqui st. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176, 2004.

[36] I. Olkin and H. Saner. Approximationsfor trimmed fisher proceduresin research synthesis. *Statistical Methods in Medical Research*, 10:267–276, 2001.

[37] A.B. Owen. Pearsons test in a large scale multiple meta-analysis. Technical report, Stanford University, 2007.

[38] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G.G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. Arrayexpressa public repository for microarray gene expression data at the ebi. *Nucleic Acids Research*, 33:553–555, 2005.

[39] E.S. Pearson. The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30:134–148, 1938.

[40] M. Pirooznia, V. Nagarajan, and Y. Deng. Gene venn - a web application for comparing gene lists using venn diagram. *Binformation*, 1:420–422, 2007.

[41] S.W. Raudenbush and A.S. Bryk. Empirical bayes meta-analysis. *Journal of Educational Statistic*, 10:75–98, 1985.

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

[43] S. N. Roy. On a heuristic method of testconstruction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24:220–238, 1953.

[44] D.B. Rubin. Using empirical bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, 75:801–816, 1980.

[45] D.B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6:337–401, 1981.

[46] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 3:1090–1098, 2004.

[47] G. Sherlock, T. Hernandez-Boussard, G. Kasarskis, A.and Binkley, J.C. Matese, S.S. Dwight, Kaloper, S. M., Weng, Ball Jin, H., M.B. C.A., Eisen, P.T. Spellman, P.O. Brown, D. Botstein, and J.M. Cherry. Combining affymetrix microarray results. *Nucleic Acids Research*, 29:152–155, 2001.

[48] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:3, 2004.

[49] J. Stevens and R.W. Doerge. Combining affymetrix microarray results. *BioMed Central Bioinformatics*, 6:57, 2005.

[50] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Ser. B*, 64:479–495, 2002.

[51] S. Stouffer, E. Suchman, L. DeVinnery, S. Star, and R. W. Jr. *The American Soldier,volumn I: Adjustment during Army Life.* Princeton University Press, 1949.

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

[53] G.C. Tseng, M.K. Oh, L. Rohlin, J.C. Liao, and W.H. Wong. Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*, 29:2549–2557, 2001.

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

[55] J.B. Welsh, L.M. Sapinoso, A.I. Su, S.G. Kern, J. Wang-Rodriguez, C.A. Moskaluk, H.F. Frierson, and G.M. Jr Hampton. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, 61:5974–5978, 2001.

[56] B. Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48:156–157, 1951.

[57] D.V. Zaykin, Lev A. Zhivotovsky, P.H. Westfall, and B.S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22:170–185, 2002.