ISSUES IN META-ANALYSIS OF CANCER MICROARRAY STUDIES: DATA DEPOSITORY IN R AND A META-ANALYSIS METHOD FOR MULTI-CLASS BIOMARKER DETECTION

by

Shu-Ya Lu

BS, National Taiwan University, Taiwan, 1998

MS, National Taiwan University, Taiwan, 2000

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

2009

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Shu-Ya Lu

It was defended on

July 15th 2009

and approved by

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

Lisa Weissfeld, Ph.D, Professor and Director of Academic Program, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

(Joyce) Chung-Chou H. Chang, Ph.D, Department of Medicine, School of Medicine, University of Pittsburgh

Gong Tang, Ph.D, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

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

ISSUES IN META-ANALYSIS OF CANCER MICROARRAY STUDIES: DATA DEPOSITORY IN R AND A META-ANALYSIS METHOD FOR MULTI-CLASS BIOMARKER DETECTION

Shu-Ya Lu, PhD

University of Pittsburgh, 2009

Systematic information integration of multiple related microarray studies has become an important issue as the technology has become significant mature and more prevalent in public health relevance over the past decade. The aggregated information provides more robust and accurate biomarker detection. So far, published meta-analysis methods for this purpose mostly consider two-class comparison. Methods for combining multiclass studies and expression 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 guarantee to reflect the concordant expression pattern information across studies, we propose a multi-class correlation measure (MCC) to specifically look for biomarkers of concordant inter-class patterns across a pair of studies. For both approaches, we focus on identifying biomarkers differentially expressed in all studies (i.e. ANOVA-maxP and min-MCC). The min-MCC method is further extended to identify biomarkers differentially expressed in partial studies using an optimally-weighted technique (OW-min-MCC). All methods are evaluated by simulation studies and by three meta-analysis applications to multi-tissue mouse metabolism data sets, multi-condition mouse trauma data sets and multi-malignant-condition human prostate cancer data sets.

The results show complementary strength of ANOVA-based and MCC-based approaches for different biological purposes. For detecting biomarkers with concordant inter-class patterns across studies, min-MCC has better power and performance. If biomarkers with discordant inter-class patterns across studies are expected and are of biological interests, ANOVA-maxP better serves this purpose.

PREFACE

I am indebted to many people for the successful completion of this dissertation. I am grateful for the generous support of my advisor, Dr.Chien-Cheng Tseng (George), who has been with me throughout the years as a mentor, colleague, editor, and friend. I also extend special thanks to Dr.Chung-Chou H. Chang (Joyce) who has been an unending source of advice.

I thank the other members of my committee, Dr.Lisa Weissfeld and Dr.Gong Tang for their valuable input. I would also like to give my hearty thanks to my GSR supervisor, Dr.Mary Ganguli. Without her support, I could not have gotten this.

There are many other people without whom I would never have made it to the end of a successful graduate career. Many thanks to my dear lab partner, Jia Li. Working with her is one of the greatest experience in my life. I also would like to give my deepest thanks to the bioinformatics and statistical learning group. Their hard work in reseach always encourages me.

I give my deepest appreciation to my Lord, Jesus Christ. His unlimited grace and support have been my power to keep going for the past couple years. Just like Psalm 23:1 says: "Jehovah is my Shepherd; I will lack nothing."

Finally, I owe my greatest debts to my family. I thank my parents for life and the strength and determination to live it. This dissertation is equally their achievement. For me, it is they who make all things possible. Most of all, I thank my wife, who shared my burdens and my joys during the past years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

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.

In this chapter, we will provide a brief introduction of the technology of microarray and meta-analysis. Then, we will describe the progress of microarray meta-analysis for the past decade.

1.1 THE PRINICIPLE AND TECHNOLOGY OF MICROARRAY

A microarry is a tool used to sift through and analyze the information contained within a genome or proteome. It consists of different nucleic acid or protein probes, called features, that are chemically attached to a substrate, which can be a microchip, a glass slide or a microsphere-sized bead. It is a multiplex technology used in molecular biology and in medicine. It can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. To summarize, we can say a microarray is a tool for analyzing gene expression

that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern.

Why are microarrays important? Because they may contain a very large number of genes and because of their small size. Therefore, microarrays are useful when an investigator wants to survey a large number of genes quickly or when the sample to be studied is small. In addition, a microarray can be used to examine the expression of hundreds or thousands of genes at once, which promises to revolutionize the way scientists examine gene expression. This technololgy is still considered to be in its infancy; therefore, many initial studies using microarrays have represented simple surveys of gene expression profiles in a variety of cell types.

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), in genotyping or in resequencing mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data.

1.1.1 Technology of microarray

In standard microarrays, the probes are attached to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are commonly known as gene chip or colloquially as an "Affy chip" when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of their detection system.

Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. The use of a collection of distinct DNAs in arrays for expression profiling was first described in 1987.

The arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (Saccharomyces cerevisiae) on a microarray was published in 1997.

Arrays of DNA can be spatially arranged, as in the commonly known gene chip (also called genome chip, DNA chip or gene array), or can be specific DNA sequences labelled such that they can be independently identified in solution. The traditional solid-phase array is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon biochip. The affixed DNA segments are known as probes (although some sources use different terms such as "reporters"). Thousands of them can be placed in known locations on a single DNA microarray.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.

Figure 1 shows the details of labeling and hybridization of microarrays. The left one is for cDNA arrays and the right one is for Affymetrix Gene Chip.

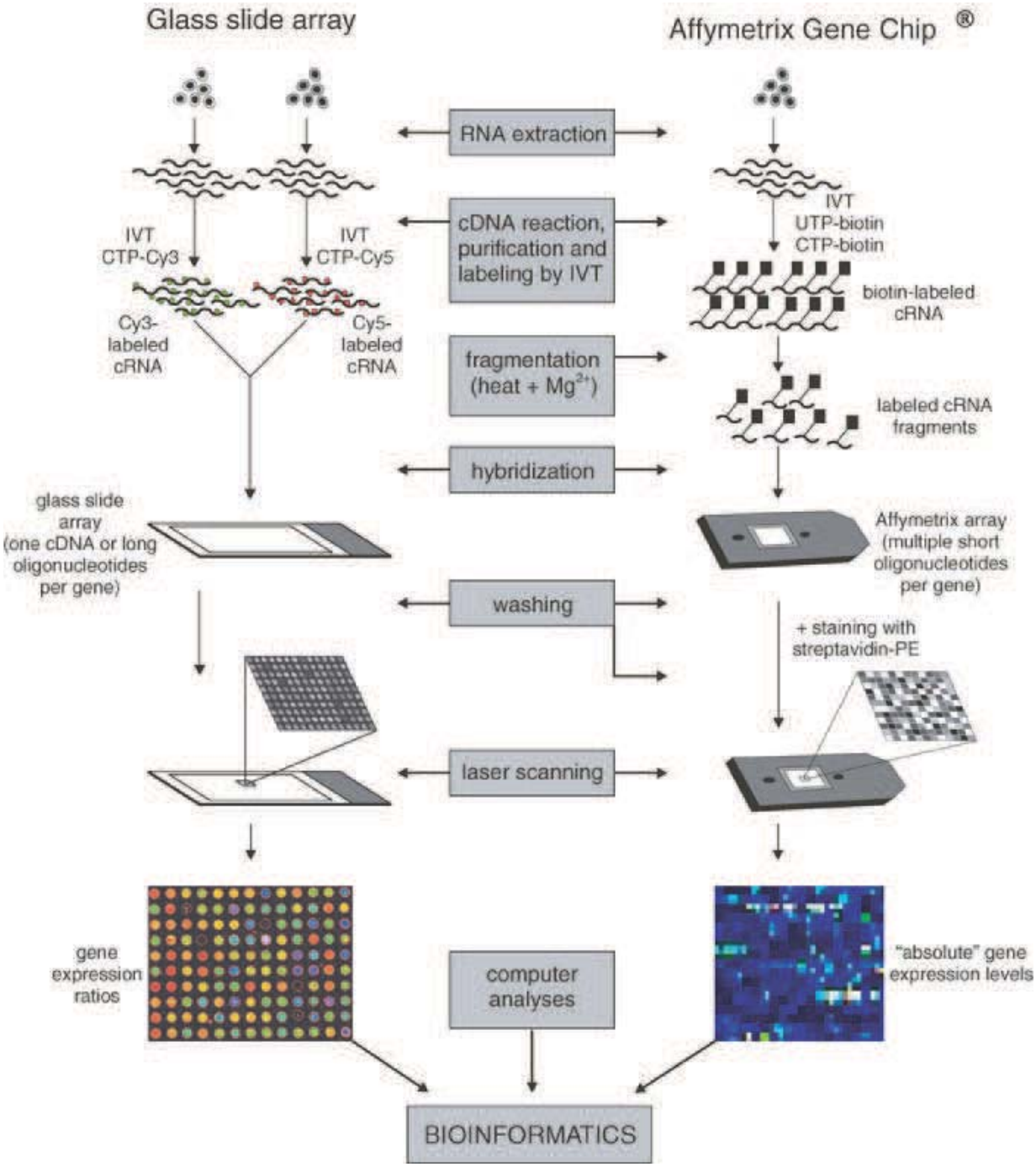

Figure 1: How the micorarray technology works.

1.1.2 Steps of a microarray experiment

The steps of a microarray experiment are as follows: experimental design, data standardization image acquisition and analysis, data pre-processing and normalization, and various statistical and data mining techniques to study data sets.

1. RNA is first isolated from different tissues via developmental stages, disease states or samples subjected to appropriated treatments or extracted from different experimental conditions.
2. The RNA is then labeled with two different fluorescent dyes and co-hybridized to a microarray that allows expression to be asseyed and compared between appropriate sample pairs.
3. The array is then scanned to acquire fluorescent images. Independent gray scale TIFF (Tagged Information File Format) images are generated for each pair of samples to be compared.
4. These images must then be analyzed to identify arrayed spots and to measure the relative fluorescence intensities for each element (dye). Low quality data is filtered out and the remaining high quality data is normalized.
5. Finally depending on the aim of the sudy, one can infer satistical significance or differential expression, perform various exploratory data analyses, classify samples according to their disease subtypes and carry pathway analysis.

Complete information from all the steps should be collected according to certain standards, such as the Minimum Information About a Microarray Experiment (MIAME) and archived properly. Figure 2 is a flow chart for the steps of microarray experiments. For more details, please refer to *www.ebi.ac.uk/arrayexpress/Standards/index.html*.

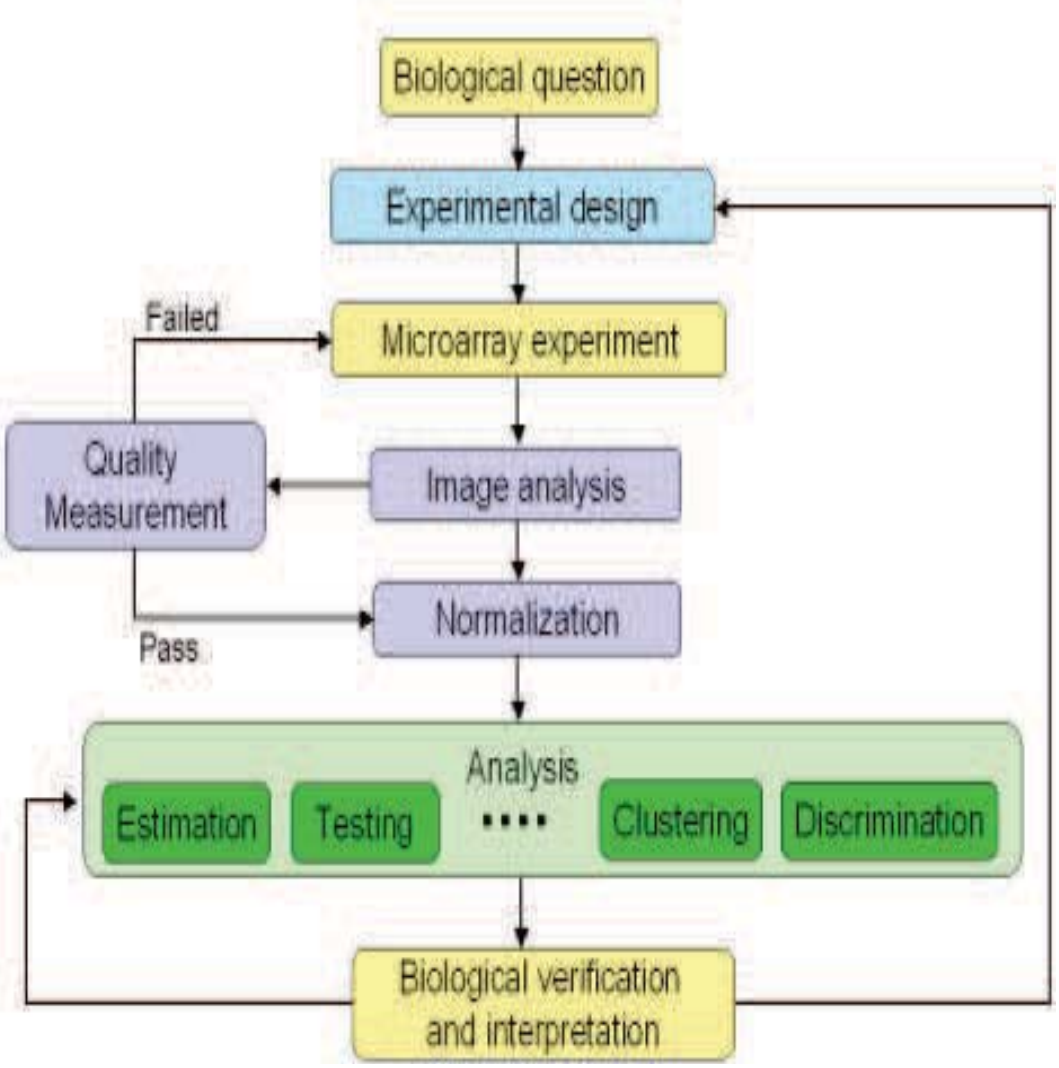

Figure 2: Steps of a microarray experiment.

1.1.3 Types of microarrays

Several competing technologies for microarray probe implementation have been developed over the past decade. We call them platforms. The existence of multiple technologies has raised the possiblity of cross-platform comparison and integration of data. Table 1 shows some comparisons among four major different platforms.

1.1.4 Steps of microarray analysis

Once we get data from a microarray experiment described in the previous section, the next step is to do data analysis. The procedure for microarray data analysis is as follows:

1. **Experimental design (Implementation):**
   Define the biological question and hypothesis clearly. Design the microarray experimental scheme carefully; include biological replication in experimental design.
2. **Data collection and archiving:**
   Compliance with microarray information collection standards (e.g. MIAME).
3. **Image acquisition:**
   Avoid photo-bleaching. Try to balance the overall intensities between the two dyes and scan images at appropriate resolution.
4. **Image analysis:**
   Inspect the gridding result manually and adjust the mask and flag poor-quality spots if necessary. Choose and apply appropriate aegmentation algorithm.
5. **Data pre-processing:**
   Remove poor-quality spots with intensity by lowering the background plus two standard deviations. Log-transform the intensity ratios.
6. **Data normalization:**
   Use diagnositc plots to evaluate the data and consider a proper method for normalization.
7. **Identifying differentially expressed genes:**
   Calculate a statistic based on the replicate array data for ranking genes. Select a cut-off value for rejecting the null-hypothesis. Do not use a fixed threshold (i.e. two-fold increase or decrease) to infer significance.

Table 1: Comparison of four different major microarray platforms

Platform	CodeLink	Affymetrix	Agilent	cDNA
Array format	30-mer	25-mer	60-mer	
Hybridization time	18h	16h	17h	17h
Hybridizaion temp	37^oC	45^oC	60^oC	60^oC
Colors	One-color	One-color	One/Two color	Two-color
Manufacturer	GE company	Affymetrix company	Agilent company	Individual Laboratory
Probe preparation	3-D aqueous gel matrix	Probes are short oligos synthesized through a photolighographic approach	Probes are printed by InKjet technology form HP	Probes are usually amplified by PCR and spotted by root
Density	∼57000 probes	423500 probes	22000 probes	Maximum of ∼15000 probes
Advantages	Sensitivity; 3D surface; Customization is possible;	Reproducibility; Mature platform; Customization;	Reproducibility; Mature platform; Customization	Inexpensive; Customization is possible
Disadvntages	Non-contract printing	Less sensitive	Two-color bias	Poor specificity

8. **Exploratory data analysis:**

   Use different analysis tools with different settings to "explore" the data. Validate the results by follow-up experiments.

9. **Other down-stream analyses:**

   Do other down-stream analyses based on the differentially expressed genes, like clustering, classfication, pathway analysis,...etc.

1.2 META-ANALYSIS

1.2.1 Basic ideas of meta-analysis

A meta-analysis is a statistical technique used to evaluate a pool of studies for systematic reviews. Most often it is used to assess the clinical effectiveness of healthcare interventions quantitatively and achieves this by combining data from two or more independent studies. Glass (1976) defines meta-analysis as: *The statistical analysis of a large collection of analysis results for the purpose of integrating the findings.*

The basic purpose of meta-analysis is to provide the same methodological rigor to a literature review that we require from experimental research. In other words, meta-anlyses are generally centered on the relationship between one explanatory and one response variable. This relationship, "the effect of X and Y", defines the analysis. So, meta-analysis provides an opportunity for *shared subjectivity* in reviews, rather than true objectivity.

Karl Pearson performed the first meta-analysis in 1904. Pearson wanted to overcome reduced statistical power in studies with small sample sizes. He analyzed the results from a group of studies and concluded that a new piece of research could be created to allow for more accuracy in analysis. Later, Tippett (1931), Fisher (1948), and Wilkinson (1951) also proposed methods for meta-analysis. Today, meta-analysis is widely used in epidemiology and evidence-based medicine.

By far the most common use of meta-analysis has been in *quantitative literature reviews*. These are review articles where the authors select a research finding that has been investigated in a certain research project under a large number of different circumstances. They then use meta-analysis to help them describe the overall strength of the effect, and under what circumstances it is stronger and weaker.

Recently, as knowledge of meta-analytic techniques has become more widespread, researchers have begun to use *meta-analyitc summaries* within research papers. In this case, meta-analysis is used to provide information supporting a specific theoretical statement, usually about the overall strength or consistency of a relationship within the studies being conducted. As might be expected, calculating a meta-analytic summary is typically a much simpler procedure than performing a full quantitative literature review.

In general, two metrics are commonly combined in the meta-analysis. The first metric combines significance levels or their transformation scores. The famous Fisher's method belongs to this category that sums up the log-transformed p-values. Many other statistics including a trimmed version of Fisher's method (Olkin and Saner, 2001), minimum p-value (Tippett, 1931) and Wilkinson's (1951) rth smallest p-value have also been considered. The second metrix is to combine 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.

1.2.2 Traditional methods for combining significance levels

[1] Tippet's method

Tippet et al (1931) proposed a method to deal with meta-analysis. It is a technique to take the minimum p-value over different studies. The formula is

$$V^{minP} = \min_{1 \le i \le S} p_i \tag{1.1}$$

Null hypothesis is rejected if $V^{minP} \leq 1-(1-\alpha)^{1/S}$. Here α is the overall significance level and V^{minP} follow a $Beta(1,S)$ distribution under null hypothesis. This method is sensitive to outliers, so a variant uses the rth smallest p-value as an alternative (Wilkinson, 1951).

[2] Fisher's Method

Fisher et al (1948) proposed a method for meta-analysis. It is a technique for combining the results from a variety of independent tests bearing upon the same overall hypothesis as if in a single large test. If there are S independent experimental studies, Fisher's method combines the p-values of these studies and transfroms them into one statistic V having an χ^2 distribution using the formula

$$V = -2\sum_{i=1}^{S} log(p_i) \tag{1.2}$$

The degree of freedom of V is $2S$.

Fisher's method was shown 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).

[3] Wilkinson's method

Wilkison et al (1951) proposed a method for meta-analysis which generalized Tippett's procedure to a more robust rth smallest p-value.

$$V^W = p_{(r)} \tag{1.3}$$

It is obvious that Maximum p-value is a special case of V^W and the most frequently used. It is often refered to as Wilkinson's method. The formula is as follows:

$$V^{maxP} = \max_{1\leq i\leq S} p_i \tag{1.4}$$

V^{maxP} follow a $Beta(S,1)$ distribution under null hypothesis. One of our proposed methods, ANOVA-maxP, applies this idea for detecting genes with interesting intra-class patterns across multiple Microarray studies.

1.2.3 Traditional methods for combing effect size

Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect. Unlike significance tests, these indices are independent of sample size. ES measures are the common currency of meta-analysis studies that summarize the findings from a specific area of research. See, for example, the influential meta-analysis of psychological, educational, and behavioral treatments by Lipsey and Wilson (1993).

There is a wide array of formulas used to measure ES. For the occasional reader of meta-analysis studies, like myself, this diversity can be confusing. One of my objectives in putting together this set of lecture notes was to organize and summarize the various measures of ES.

In general, ES can be measured in two ways:
a) as the standardized difference between two means,
b) as the correlation between the independent variable classification and the individual scores on the dependent variable. This correlation is called the "effect size correlation" (Rosnow Rosenthal, 1996).

These notes begin with the presentation of the basic ES measures for studies with two independent groups. The issues involved when assessing ES for two dependent groups are then described.

For convenience sake we will assume that our contrast will be defined as (treatment group v.s. control group). When considering the role of this difference in the design of the study we will call the variable differentiating these groups as the "treatment factor".

The simplest effect size based on mean differences is Cohen's g, defined as

$$g = \frac{\bar{Y}_t - \bar{Y}_c}{s_p} \tag{1.5}$$

where $\bar{Y}_t$ is the mean of the treatment group, $\bar{Y}_c$ is the mean of the control group, and s_p is the pooled sample standard deviation.

While intuitive, the effect size g is actually a biased estimator of the population effect size

$$\delta = \frac{\mu_t - \mu_c}{\sigma} \tag{1.6}$$

Using g produces estimates that are too large, especially with small samples.

To correct g we multiply it by a correction term

$$J_m = 1 - \frac{3}{4m - 1} \tag{1.7}$$

where $m = n_t + n_c - 2$. The resulting statistic

$$d = g \cdot (1 - \frac{3}{4m - 1}) = g \cdot (1 - \frac{3}{4(n_t + n_c) - 9}) \tag{1.8}$$

is known as Hedges's d, and is an unbiased estimator of δ. It is generally best to record both g and d for each effect in our meta-analysis.

The variance of d, given relatively large samples, is

$$\sigma_d^2 = \frac{n_t + n_c}{n_t n_c} + \frac{d^2}{2(n_t + n_c)} \tag{1.9}$$

Using these statistics we can construct a level C confidence intervel for δ

$$d \pm z^*(\sigma_d) \tag{1.10}$$

Where z^* is the critical value from the normal distribution, such that the area between $-z^*$ and z^* is equal to C.

There are many developed formulas to calculate g from a number of different test statistics. One thing worthy to remember is that no matter which formula you choose to use, you should correct g for its sample size bias using the equation presented above.

1.2.4 Correlation of effect sizes

Correlations are widely used outside of meta-analysis as a measure of the linear relationship between town continuous variables. The correlation between two variables x and y may be calculated as

$$r_{xy} = \frac{\sum z_{xi} z_{yi}}{n} \tag{1.11}$$

where z_{xi} and z_{yi} are the standardized scores of the x and y for case i, n is the sample size.

Correlations can range between -1 and 1. Correlations near -1 indicate a strong negative relationship; correlations near 1 indicate a strong positive relationship, while correlations near 0 indicate a nolinear relationship.

The correlation coefficient r is a slightly biased estimator of ρ, the population correlation coefficient. An approximation of the population correlation may be obtained from the formula

$$G_{(r)} = r + \frac{r(1 - r^2)}{2(n - 3)} \tag{1.12}$$

The sampling distribution of a correlation coefficient is somewhat skewed, especially if the population correlaion is large. It is therefore conventional in meta-analysis to convert correlations to z scores using Fisher's r-to-z transfromation as

$$z_r = \frac{1}{2} ln(\frac{1 + r}{1 - r}) \tag{1.13}$$

where ln(x) is the natural logarithm function. All meta-analytic calculations are then performed using the transformed values. z_r has a nearly normal distribution with variance

$$s_z^2 = \frac{1}{n - 3} \tag{1.14}$$

Using these statistics we can construct a level C confidence interval for the population value

$$z_r \pm \frac{z^*}{\sqrt{n - 3}} \tag{1.15}$$

where z^* is the critical value from the normal distribution such that the area between $-z^*$ and z^* is equal to C.

If we wish to work with unbiased estimates of ρ, you should first calculate the correlation $G_(r)$ for each study and then transform the $G_(r)$ values to z-scores for analysis.

Once we have made the necessary computations, we can use Fisher's z-to-r transformation

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1} \tag{1.16}$$

where e is the base of the natural logarithm, to convert the results back into correlations.

1.3 MICROARRAY META-ANALYSIS

As we mentioned in the previous sections, microarray is a useful technology which provides an opportunity for global monitoring of the expression levels of thousands of genes simultaneously. This technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. The use of a collection of distinct DNAs in arrays for expression profiling was first described in 1987, and the arrayed DNAs were used to identify genes whose expression is modulated by interferon.

The use of miniaturized microarrays for gene expression profiling was first reported in 1995. 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.

Microarrays present new statistical problems because the data is very high dimensional with very little replication. Microarrays offer an exciting entry point for statisticians and computational scientists into the modern areas of computational biology and bioinformatics.

Among 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, 2004; 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 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, 2004; Hong et al, 2006) and others. We will introduce two methods for combining effect size in microarray meta-analysis below.

1.3.1 Modern methods for combining effect size in microarrary meta-analysis

When studies have a similar design and measure the outcome in a similar manner, combining estimates is usually preferred by some researchers to the omnibus methods in the previous section as suggested by some researchers. In this section we will introduce the two most used methods in this area, either the fixed or the random effects model (Hedges and Vevea, 1998).

Let T and C stand for two experimental conditions (treatment verses control), and let there be S independent studies and (n_{iT}, n_{iC}) replicates for the ith study. Briefly, a standardized mean difference was obtained as an effect size index for the measurement of differential expression of a gene in any given study.

$$d_i = \frac{\bar{T}_i - \bar{C}_i}{S_p}$$

Where $\bar{T}_i$ and $\bar{C}_i$ represent the means of treatment and control group in the ith study, and S_p indicates the estimated variation. Then we can model the effect size index d_i across studies as follows:

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

where μ denotes the parameter of interest (treatment effect), and s_i^2 and τ^2 represent the within-study and between-study variance.

1.3.2 Fixed effects model

Fixed effects models consider only within-study variability, s_i^2, and assume that all studies use identical methods, samples and measurements. They also assume there is a constant effect size μ for all studies. It means that $\theta_1 = \theta_2 = ... = \theta_S = \mu$. Thus, $d_i \sim N(\mu, s_i^2)$. The most efficient and unbiased estimator of μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{S} (s_i^2)^{-1} d_i}{\sum_{i=1}^{S} (s_i^2)^{-1}}$$

and the variance of $\hat{\mu}$ is

$$Var(\hat{\mu}) = \frac{1}{\sum_{i=1}^{S} (s_i^2)^{-1}}$$

A Z-score will be derived from $\hat{\mu}/\sqrt{Var(\hat{\mu})}$ to test the null hypothesis when $\sum_{i=1}^{S}(n_{iT} + n_{iC}) \to \infty$.

1.3.3 Random effects model

As an alternative approach, the random effects model allows the study outcomes to vary in a normal distribution among studies. That is, the true study effect size θ_i is no longer a constant effect and varies across studies. We represent the equations as follows:

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

The estimator of μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{S}(s_i^2 + \tau^2)^{-1} d_i}{\sum_{i=1}^{S}(s_i^2 + \tau^2)^{-1}}$$

and the variance of $\hat{\mu}$ is

$$Var(\hat{\mu}) = \frac{1}{\sum_{i=1}^{S}(s_i^2 + \tau^2)^{-1}}$$

A Z-score will be derived from $\hat{\mu}/\sqrt{Var(\hat{\mu})}$ to test the null hypothesis.

1.4 TWO COMPLEMENTARY HYPOTHESIS SETTINGS

Li and Tseng (2009) elucidated two statistical hypothesis settings behind two separate biological goals in combining multiple array studies. We consider meta-analysis of K gene expression profile studies: $D_1, D_2, ..., D_k$, x_{kgs} is the gene expression intensity of gene g and sample s, where sample $s = 1, ..., n_k$ belong to a normal group and $s = n_k + 1, ..., n_k + m_k$ belong to the disease group. Normally we consider a null hypothesis for each gene g:

$$H_0 : \theta_{g1} = \cdots = \theta_{gK} = 0,$$

where θ_{gk} represents the gene effect of gene g and study k. Following the convention of Birnbaum (1954), two complementary alternative hypotheses can be considered depending on the nature of the experimental situations in which the gene effects(θ_{gk}s) are obtained:

$$HS_A : \{ \quad H_0 \text{ versus } H_A : \theta_{gk} \neq 0, \ \forall 1 \leq k \leq K\}$$
$$HS_B : \{ \quad H_0 \text{ versus } H_B : \text{at least one } \theta_{gk} \neq 0, \ 1 \leq k \leq K\}$$

Under the two major alternative hypotheses categories, different subset or variations have been explicitly or implicitly targeted by different existing methods:

$$\begin{aligned}
HS_{A1} : &\quad \{H_0 \text{ versus } H_{A1} : \theta_g = \theta_{g1} = ... = \theta_{gk} \neq 0\} \\
HS_{A2} : &\quad \{H_0 \text{ versus } H_{A2} : \theta_g \neq 0, \ \theta_{gk} \sim N(\theta_g, \tau^2)\} \\
HS_{Bh} : &\quad \{H_0 \text{ versus } H_{Bh} : \textstyle\sum_{k=1}^{K} I(\theta_{gk} \neq 0) = h \ (1 \leq h \leq K)\} \\
HS_{Bh'} : &\quad \{H_0 \text{ versus } H_{Bh'} : \textstyle\sum_{k=1}^{K} I(\theta_{gk} \neq 0) = h \text{ and } \theta_{gk} = \theta_g \text{ if } \theta_{gk} \neq 0.(1 \leq h \leq K)\}
\end{aligned}$$

$I(\cdot)$ is an indicator function that equals 1 when statement true and 0 otherwise.

Without danger of confusion, we will use the notation of alternative hypothesis (e.g. H_A) to denote the paremeter space of the corresponding alternative hypothesis. It is clear that $H_A \subset H_B$. In H_A, gene g is identified only when it is differentially expressed in all studies. In H_B, it is selected if it is differentially expressed in one or more studies. We can easily note that $H_{A1} \subset H_A$ represents an equal fixed effect model. H_{A2} represents a random effect model for a similar purpose of H_A while $H_{A2} \not\subseteq H_A$ in general. We also can find that $H_B = \bigcup_{1 \leq h \leq K} H_{Bh}$, $H_{Bh'} \subseteq H_{Bh} \ (1 \leq h \leq K)$ and $H_{BK'} = H_{A1}$.

An optimally-weighted statistic was modified from the Fisher's score and was proposed for the former hypothesis setting (H_{A1}). The optimal weights provided natural categorization of the detected biomarkers for further biological investigation.

1.5 MOTIVATION

The current methods in Microarray meta-analysis, including those described previously, just focus on the detection of differentially expressed (DE) genes between two-class 'disease-versus-normal' or 'treatment-versus-control' setting. Methods for combining studies with

more than two classes are rarely discussed. This is the reason why we would like to propose a new method to deal with the issue of multiple-class studies in this dissertation. When more than two classes are considered, no single effect size can be computed, and instead, the inter-class patterns[1] become the concern. The former category is, however, extensible to data sets with more than two classes. The F-statistics (or equivalent ANOVA model) and its variants can be applied and p-values can be assessed and combined across studies.

In this dissertation, we explore the method of ANOVA-maxP, which detects biomarkers with a significant pattern (large between-class variance versus small within-class variation) in all studies. We also note that small p-values (equivalently large F-statistics) in all studies do not guarantee consistent inter-class patterns across studies. We can use real data as an illustration. Table 2 shows mouse metabolism data which has four tissues (brown fat, liver, heart and skeletal) and within each tissue there are three genotypes (Wilde Type, VLCAD, LCAD)[2]. The replicates of each tissue within each genotype are slightly different and the total number of arrays are 44.

Table 2: Mouse metabolism data

tissue type	brown fat			liver			heart			skeletal			Total
genotype	WT	V-	L-	WT	V-	L-	WT	V-	L-	WT	V-	L-	
n of arrays	4	4	4	4	4	4	3	4	4	3	3	3	44

WT: wilde type; V-:VLCAD; L-:LCAD

We can treat each tissue as a study and each genotype as a class, so now we have a Microarray meta-analysis data set with four studies and within each study has three classes. Then we applied ANOVA-maxP to do the biomarker detection. Figure 3 shows two genes in mouse metabolism data detected by ANOVA-maxP and the inter-calss patterns across

[1] In a loose definition, the "inter-class pattern" is defined as the vector of mean indensities in the classes within a study when the within-class variation is very small

[2] More details are described in section 1 of chapter 3

studies are concordant (the upper one) and discordant (the lower one). In Figure 3, we can see that Gene Amacr has a concordant inter-class pattern across these four tissues. It has high expression levels in two genotypes (WT, LCAD) and a low expression level in VLCAD, regardless of the kind of tissue. Gene Scd1 is also detected by ANOVA-maxP, however, it does not have a concordant inter-class pattern across these four tissues.

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 inter-class pattern biomarkers. The methodologies and procedures of these two methods are described in Chapter 3. We also included the performance of ANOVA-maxP and min-MCC assessed through simulation and applications to a multi-tissue mouse energy metabolism data set and a multi-platform mouse trauma data set. The result shows that ANOVA-maxP detects genes with both concordant and discordant inter-class patterns and min-MCC only detects genes with concordant inter-class genes. The two methods are complementary and serve different biological purposes.

In addition to the biomarker detection, we also develop a R depository package for containing the experimental information of these cancer microarray studies and some functions for data quality control, data operation and retrieval. The details are decribe in Chapter 2. In Chapter 4, we will make some conclusions and talk about the further works.

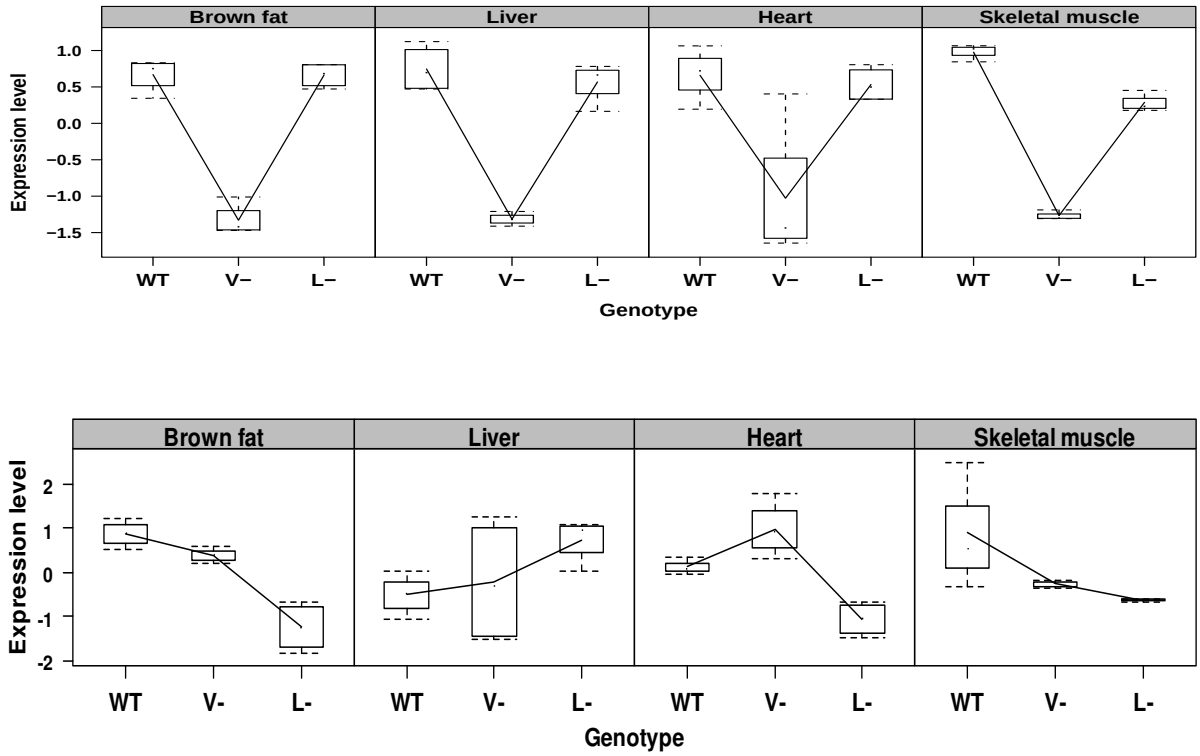

Figure 3: Two examples for inter-class patterns.

Two examples for showing gene expression indensities with and without concordant inter-class patterns across these four tissues. Box-plots of each genotype in each tissue are plotted and the mean expression levels are connected. Upper: Gene Amacr, involved in metabolic process, has a concordant inter-class pattern across these four tissues. Lower: Gene Scd1, involved in fatty acid synthesis pathway, does not have a concordant inter-class pattern across these four tissues.

2.0 DATA DEPOSITORY FOR CANCER MICROARRAY STUDIES

In this chapter, we are going to introduce a depository package for cancer microarray studies. By using this package, users can very easily get the experimental information of microarray studies. A function of retrieving gene expression infromation for a special gene which users are interested in is also developed. Because these functions are based on two free pieces of softwares, R and Bioconductor, a brief introduction of these two free software packages is provided as well.

2.1 AN INTRODUCTION TO R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and his colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has

been with the defaults for the minor design choices in graphics, but the user retains full control. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

2.1.1 The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy,
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution, and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy. More can be learned about specific functions from the R help system; either `help(name)` or `?name` will yield the help files.

2.1.2 Strengths and weaknesses

Strengths

- free and open source, supported by a strong user community
- highly extensible and flexible
- implementation of modern statistical methods
- moderately flexible graphics with intelligent defaults

Weaknesses

- slow or impossible with large data sets
- non-standard programming paradigms

2.1.3 Basic operations

[1] Vectors and assignment

To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
R> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

This is an *assignment* statement using the *function* `c()` which in this context can take an arbitrary number of vector *arguments* and whose value is a vector got by concatenating

its arguments end to end.

```
R> 1/x
```

The reciprocals of the five values would be printed at the terminal (and the value of x, of course,unchanged). The further assignment

```
R> y <- c(x, 0, x)
```

would create a vector y with 11 entries consisting of two copies of x with a zero in the middle place.

[2] Vector arithmetic

With the above assignments the command

```
R> v <- 2*x + y + 1
```

generates a new vector v of length 11 constructed by adding together, element by element, `2*x` repeated x, 2 times, y repeated just once, and 1 repeated 11 times.

The elementary arithmetic operators are the usual `+, -, *, /` and $\wedge$ for raising to a power. In addition, all of the common arithmetic functions are available. `log, exp, sin, cos, tan, sqrt`, and so on, all have their usual meaning. Max and min select the largest and smallest elements of a vector respectively. `range` is a function whose value is a vector of length two, namely `c(min(x),max(x))`. `length(x)` is the number of elements in x, `sum(x)` gives the total of the elements in x, and `prod(x)` their product.

[3] Missing values and Index vectors

The function `is.na(x)` gives a logical vector of the same size as x with value `TRUE` if and only if the corresponding element in `x` is `NA` or `NaN`.

```
R> z <- c(1:3,NA); ind <- is.na(z)
```

and another command like

```
R> y <- x[!is.na(x)]
```

create (or re-create) an object `y` which will contain the non-missing values of `x`, in the same order. Note that if x has missing values, `y` will be shorter than `x`. Also,

```
R> (x+1)[(!is.na(x)) & x>0] -> z
```

creates an object `z` and places in it the values of the vector `x+1` for which the corresponding value in `x` was both non-missing and positive. Another example for missing values is

```
R> x[is.na(x)] <- 0
```

This replaces any missing values in x by zeros and

```
R> y[y < 0] <- -y[y < 0]
```

has the same effect as

```
> y <- abs(y)
```

For more details of R language, please refer to *www.r-project.org*.

2.2 AN INTRODUCTION TO BIOCONDUCTOR

Bioconductor is an open source and open development software project that provides tools for the analysis and comprehension of genomic data. Bioconductor is based primarily on the R programming language, but does contain contributions in other programming languages. It has two releases each year that follow the biannual releases of R. At any one time there is a release version, which corresponds to the released version of R, and a development version, which corresponds to the development version of R. Most users will find the release version appropriate for their needs. In addition, there are a large number of meta-data packages available that are mainly, but not solely, oriented towards different types of microarrays.

The Bioconductor project was started in the Fall of 2001 and is overseen by the Bioconductor core team, based primarily at the Fred Hutchinson Cancer Research Center with other members coming from various US and international institutions. It gained widespread exposure in the groundbreaking Genome Biology 2004 paper Bioconductor: open software development for computational biology and bioinformatics. More project details are available online in the Bioconductor annual reports.

2.2.1 Bioconductor Packages

Most Bioconductor components are distributed as R packages, which are add-on modules for R. Initially most of the Bioconductor software packages focused primarily on DNA microarray data analysis. As the project has matured, the functional scope of the software packages has broadened to include the analysis of all types of genomic data, such as SAGE, sequence, or SNP data.

2.2.2 Goals of the Bioconductor project

The broad goals of the Bioconductor project are:

- To provide widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data.
- To facilitate the inclusion of biological meta-data in the analysis of genomic data, e.g. literature data from PubMed, annotated data from LocusLink.
- To provide a common software platform that enables the rapid development and deployment of extensible, scalable, and inter-operable software.
- To further scientific understanding by producing high-quality documentation and reproducible research.
- To train researchers on computational and statistical methods for the analysis of genomic data.

2.2.3 Main features of the Bioconductor project

- **The R Project for Statistical Computing.** R and the R package system provide a broad range of advantages to the Bioconductor project.

- **Documentation and reproducible research.** Each Bioconductor package contains at least one vignette, which is a document that provides a textual, task-oriented description of the package's functionality. These vignettes come in several forms. Many are simple "HowTo"s that are designed to demonstrate how a particular task can be accomplished

with that package's software. Others provide a more thorough overview of the package or might even discuss general issues related to the package. In the future, we are looking towards providing vignettes that are not specifically tied to a package, but rather are demonstrating more complex concepts. As with all aspects of the Bioconductor project, users are encouraged to participate in this effort.

- **Statistical and graphical methods.** The Bioconductor project aims to provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data. Analysis packages are available for: pre-processing Affymetrix and cDNA array data, identifying differentially expressed genes, graph theoretical analyses, and plotting genomic data. In addition, the R package system itself provides implementations for a broad range of state-of-the-art statistical and graphical techniques, including linear and non-linear modeling, cluster analysis, prediction, resampling, survival analysis, and time-series analysis.

- **Annotation.** The Bioconductor project provides software for associating microarray and other genomic data in real time to biological metadata from web databases such as GenBank, LocusLink and PubMed (annotate package). Functions are also provided for incorporating the results of statistical analysis in HTML reports with links to annotation WWW resources. Software tools are available for assembling and processing genomic annotation data, from databases such as GenBank, the Gene Ontology Consortium, LocusLink, UniGene, the UCSC Human Genome Project (AnnotationDbi package). Data packages are distributed to provide mappings between different probe identifiers (e.g. Affy IDs, LocusLink, PubMed). Customized annotation libraries can also be assembled.

- **Bioconductor short courses.** The Bioconductor project has developed a program of short courses on software and statistical methods for the analysis of genomic data. Courses have been given for audiences with backgrounds in either biology or statistics. All course materials (lectures and computer labs) are available on the WWW. Customized short courses may also be designed for interested parties.

- **Open source.** The Bioconductor project has a commitment to full open source discipline, with distribution via a SourceForge-like platform. All contributions are expected to exist under an open source license such as Artistic 2.0, GPL2, or BSD. There are many different reasons why open–source software is beneficial to the analysis of microarray data and to computational biology in general. The reasons include:
  - To provide full access to algorithms and their implementation
  - To facilitate software improvements through bug fixing and software extension
  - To encourage good scientific computing and statistical practice by providing appropriate tools and instruction
  - To provide a workbench of tools that allow researchers to explore and expand the methods used to analyze biological data
  - To ensure that the international scientific community is the owner of the software tools needed to carry out research
  - To lead and encourage commercial support and development of those tools that are successful
  - To promote reproducible research by providing open and accessible tools with which to carry out that research (reproducible research is distinct from independent verification)

2.2.4 A simple example: AnnotationDbi

AnnotationDbi is used primarily to create maps that allow easy access from R to underlying annotation databases. AnnotationDbi introduces a new future for the Bioconductor annotation data packages by changing the paradigm that is used for exchanging annotations. We will use a database called *hgu95av2.db* to be an illustration. This database is for a Affymetrix array whose name is HgU95AV2.

First of all, we need to call this database into R

```
R> library("hgu95av2.db")
```

The same basic set of objects is provided with this databse package:

```
R> ls("package:hgu95av2.db")
```

```
 [1] "hgu95av2"                 "hgu95av2_dbconn"
 [3] "hgu95av2_dbfile"          "hgu95av2_dbInfo"
 [5] "hgu95av2_dbschema"        "hgu95av2ACCNUM"
 [7] "hgu95av2ALIAS2PROBE"      "hgu95av2CHR"
 [9] "hgu95av2CHRLENGTHS"       "hgu95av2CHRLOC"
[11] "hgu95av2CHRLOCEND"        "hgu95av2ENSEMBL"
[13] "hgu95av2ENSEMBL2PROBE"    "hgu95av2ENTREZID"
[15] "hgu95av2ENZYME"           "hgu95av2ENZYME2PROBE"
[17] "hgu95av2GENENAME"         "hgu95av2GO"
[19] "hgu95av2GO2ALLPROBES"     "hgu95av2GO2PROBE"
[21] "hgu95av2MAP"              "hgu95av2MAPCOUNTS"
[23] "hgu95av2OMIM"             "hgu95av2ORGANISM"
[25] "hgu95av2PATH"             "hgu95av2PATH2PROBE"
[27] "hgu95av2PFAM"             "hgu95av2PMID"
[29] "hgu95av2PMID2PROBE"       "hgu95av2PROSITE"
[31] "hgu95av2REFSEQ"           "hgu95av2SYMBOL"
[33] "hgu95av2UNIGENE"          "hgu95av2UNIPROT"
```

To demonstrate the steps for converting probe set IDs of HgU95AV2 to EntrezID, we'll start with a random sample of probe set IDs.

```
R> all_probes <- ls(hgu95av2ENTREZID)
R> length(all_probes)
[1] 12625
```

There are 12625 probe set IDs in HgU95AV2.

```
R> set.seed(0xa1beef)
R> probes <- sample(all_probes, 5)
R> probes
[1] "31882_at" "38780_at" "37033_s_at" "1702_at" "31610_at"
```

These are the probe set IDs which are randomly selected from HgU95AV2. The usual ways of accessing annotation data are by using the package: `hgu95av2ENTREZID`. Suppose we would like to know the EntrezID of "31882_at", there two ways to do it: using the order number of "31882_at" in the set of `probes` or the probe set ID directly. Below are the demostrations.

```
R> hgu95av2ENTREZID[[probes[1]]]
[1] "9136"
```

```
R> hgu95av2ENTREZID$"31882_at"
```

[1] "9136"

The EntrezID of "31882_at" is 9136.

If we would like to know the symbols (gene names) of `probes`, we can use the function: `mget`.

```
R> syms <- unlist(mget(probes, hgu95av2SYMBOL))
R> syms
```

31882_at	38780_at	37033_s_at	1702_at	31610_at
"RRP9"	"AKR1A1"	"GPX1"	"IL2RA"	"PDZK1IP1"

The symbols of these five probe set IDs are annotated successfully. For more details about Bioconductor, please refer to *www.bioconductor.org*.

2.3 R PACKAGES FOR CANCER STUDY DEPOSITORY

Before we do any kind of microarray meta-analysis, the first step is collecting enough data sets and containing the information of the experiment as much as possible. Thus, creating a convenient and useful structure of depository for the data sets is a very important work. For fulfilling this purpose, we developed a data depository based on the idea of **eSet** in Bioconductor and can be executed in R. This kind of structure has the following features:

- Simplified data content.
- Structured class hierarchy.
- Alternative storage modes.
- More validity checking.

This depository package is constituted of four components, **assayData, phenoData, featureData, experimentData**. The details of these four components are as follows:

- **assayData**:

  This is a matrix containing the gene expression values of the cancer study. Rows are **features**, e.g., gene IDs. And columns are samples represented on each study.

- **phenoData**:

  (1)**pData**: Rows are sample identifiers and columns are measured covariates, like age, sex, tumor type, survival status, or other experimental information of samples.

  (2)**pheno_varDescrptn**: Rows are measured covariate labels and columns are covariate descriptors.

- **featureData**:

  (1)**fData**: Rows are feature identifiers. These match row names of assayData. Columns are measured covariates.

  (2)**feature_varDescrptn**: Rows are measured covariate labels. Columns are covariate descriptors.

- **experimentData**:

  This is an information structure based on MIAME (Minimum Information About a Microarray Experiment) protocol, and includes

  - **name**: the last name of the first author on the paper
  - **lab**: the last name of the last author on the paper
  - **contact**: the contact information of responding author
  - **title**: the title of the paper
  - **abstract**: the abstract of the paper
  - **url**: the link of the paper
  - **pubMedIds**: the pubMed ID
  - **citation**: the citation of this paper
  - **organism**: the organism which the paper used
  - **data_url**: the data link of this study
  - **GEO**: GEO number of this data set
  - **journal**: the journal in which this paper was published
  - **year**: the year in which this paper was published
  - **geneIDType**: what kind of gene ID in this data set
  - **annotation**: what kind of array type did this study use

We are going to illustrate the functions of this package by the Magee study, a prostate cancer study published in "Cancer Research", 2001. Its array type is Affymetrix HG6800. We have

already prepared the four components and saved them in a depository package which is named "Magee". So, if we type `Magee` in R, like

```
R> Magee
```

You will see the following information:

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 7129 features, 15 samples
element names: exprs
phenoData
sampleNames: Benign.1, Benign.2, ..., Met3 (15 total)
varLabels and varMetadata description:
Cell_type: Benign: Benign, Primary: Tumor, Met: Metastasis
Tumor_type: Normal: Benign, Tumor:Primary, Other: Metastasis
featureData
featureNames: AFFX-BioB-5_at, AFFX-BioB-M_at, ..., Z78285_f_at (7129 total)
fvarLabels and fvarMetadata description:
geneID: AFFY.HU6800
experimentData: use 'experimentData(object)'
pubMedIds: 11479199
Annotation: HG6800
```

From the messages above, we can see that there are 15 samples in this study and that within each sample there are 7129 probe set IDs. The array type for this study is HG6800. If you would like to know more details about this experiment, type in `experimentData(Magee)`, then you will get all the experimental information.

We have collected about 66 cancer studies in our database and the total number is growing continually. Table 3 shows a brief summary of the cancer types and array types in the database.

Table 3: Summary of the studies in the database

Cancer Type						
Cancer type	Breast	Colon	Leukemia	Lung	Prostate	Total
Number of studies	16	11	13	10	16	66

Array Type				
Array type	cDNA	Affymetrix	Aligent	Total
Number of studies	15	50	1	66

More details of these studies about the lab, year, and array types, are in Appendix A.

2.4 FUNCTIONS TO OPERATE THE PACKAGES

2.4.1 Function for data quality evaluation

For microarray meta-analysis, data collection is the first step. Once we get enough numbers of data sets, the next step is evaluating the quality of these data sets, because if we include the samples with bad quality in our down-stream analysis, it could give us wrong information and cause bias of the biomarker detection. For this purpose, we developed a package which is named `dataQE` to do some qualtiy checking of microarray studies before doing the data analysis.

2.4.1.1 Principles and steps The principles of `dataQE` is based on Pearson correlation and the idea of mean and standard deviation. There are two parts in this package. The first one is for evaluating sample quality and the second one is for evaluating the sample normalization. The steps are as follows:

Sample Quality Evaluation

Suppose we have a study S and there are K samples in this study.

[1] Firstly, dataQE calculates pair-wise correlations of these K samples. We will get a symmetric matrix, C , containing these pair-wise correlations.
[2] Secondly, taking the mean of each row of C, we will get a set of numbers, M, with K values of mean of the pair-wise correlations.
[3] Thirdly, calculating the mean and standard deviation of M separately, denoted by corMean and corSD. If $\texttt{M}_\texttt{i} \leq \texttt{corMEAN} - 3 * \texttt{corSD}$, then sample i will be considered as a sample with bad quality, because the average correlation with other samples is much lower than others.

Sample Normalization Evaluation

The second part of dataQE is for evaluating the sample normalization. The steps are as follows:
[1] Firstly, dataQE calculates the mean of sample i , denoted by $\texttt{mean}_\texttt{i}$, where $(1 \leq i \leq K)$.
[2] Secondly, taking the mean and standard deviation of these $\texttt{mean}_\texttt{i}$s , denoted by Sample_mean and Sample_meanSD. If $|\texttt{mean}_\texttt{i}| \geq \texttt{Sample\_mean} + 3 * \texttt{Sample\_meanSD}$, then sample i will be considered as a sample with bad normalization in mean values, because the mean is much lower than others.
Similarly, dataQE will calculate the standard deviation of each sample, denoted by $\texttt{sd}_\texttt{i}$, and calculate the mean and standard deviation of $\texttt{sd}_\texttt{i}$s, denoted by Sample_SDmean and Sample_sdSD separately.
If $|\texttt{sd}_\texttt{i}| \geq \texttt{Sample\_SDmean} + 3 * \texttt{Samle\_SDsd}$, then sample i will be considered as a sample with bad normalization via standard deviation, because the standard devation is much lower than others.

2.4.1.2 An example of real data

Now we will use real data to demonstrate how dataQE works. It is prostate caner data with 148 samples, and tje data name in R is pnew. If we type dataQE(pnew) in R, it will show the following warning messages,

```
Warning:These samples may not have good quality:  7 87 88 89
Warning:These samplee may not have good normalization:  7 87 88 94 95
```

It means that sample 7, 87, 88, 89 have lower average correlation than other samples, and sample 7, 87, 88, 94, 95 have worse normalization than others. Users can refer to these warning messages and consider if it is necessary to remove these samples before doing any microarray data analysis. Figure 4 shows the heatmaps without and with removing these samples. It looks like we can get better and robust data after removing these samples.

There is another option in `dataQE: plot=TRUE`. If you type `dataQE(data,plot=TRUE)`, it will give you four plots: `(1) histogram of sample correlations, (2) histogram of sample means, (3) histogram of sample standard deviations, (4) a list of questionable samples`. For example, if we type `dataQE(pnew,plot=TRUE)` in R, then we will get a figure with four plots as shown in Figure 5.

2.4.2 Functions for data operation and retrieval

What we have discussed above just focused on a single study. We collected data sets one by one and saved them in a depository package in R. Now we can do some quality evaluation before doing the down-stream analysis. However, our purpose is not only talking about the matters in a single microarray study. What we would like to do is microarray meta-analysis. So, it is necessary for us to merge these data sets before doing analysis. Therefore, we developed two functions for this purpose.

2.4.2.1 Function for gene ID convertion As we mentioned in 1.1.3, there are many different microarray types and all of their gene identifiers (gene IDs) are also different. When we want to merge these different studies, we will encounter the problem of gene ID matching among these studies. For solving this problem, we developed a function which is named `gene.match`. In `gene.match`, we choose EntrezID as our medium to merge data sets. It means any kind of gene IDs of microarray platforms will be converted to EntrezID first and then thees EntrezIDs will be used to merge studies.

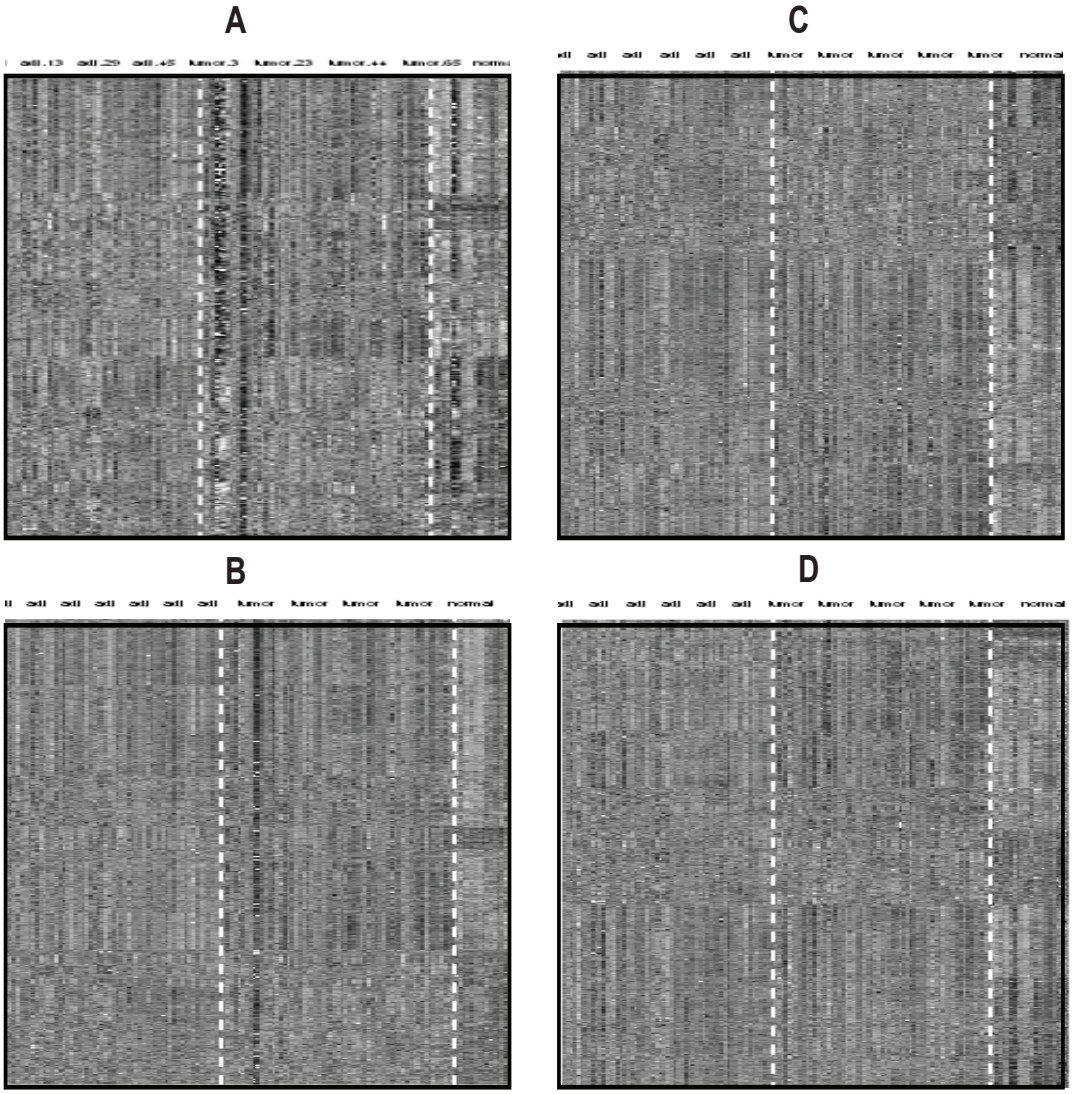

Figure 4: Heatmaps of dataQE results.

Heatmaps without and with removing the questionable samples which were detected by dataQE. A: All samples are included. B: Remove 3 lowest correlation samples (7,87,88). C: Remove 4 lowest correlation samples (7,87,88,89). D: Remove 4 lowest correlation samples (7,87,88,89),and 2 lowest column mean samples (94,95).

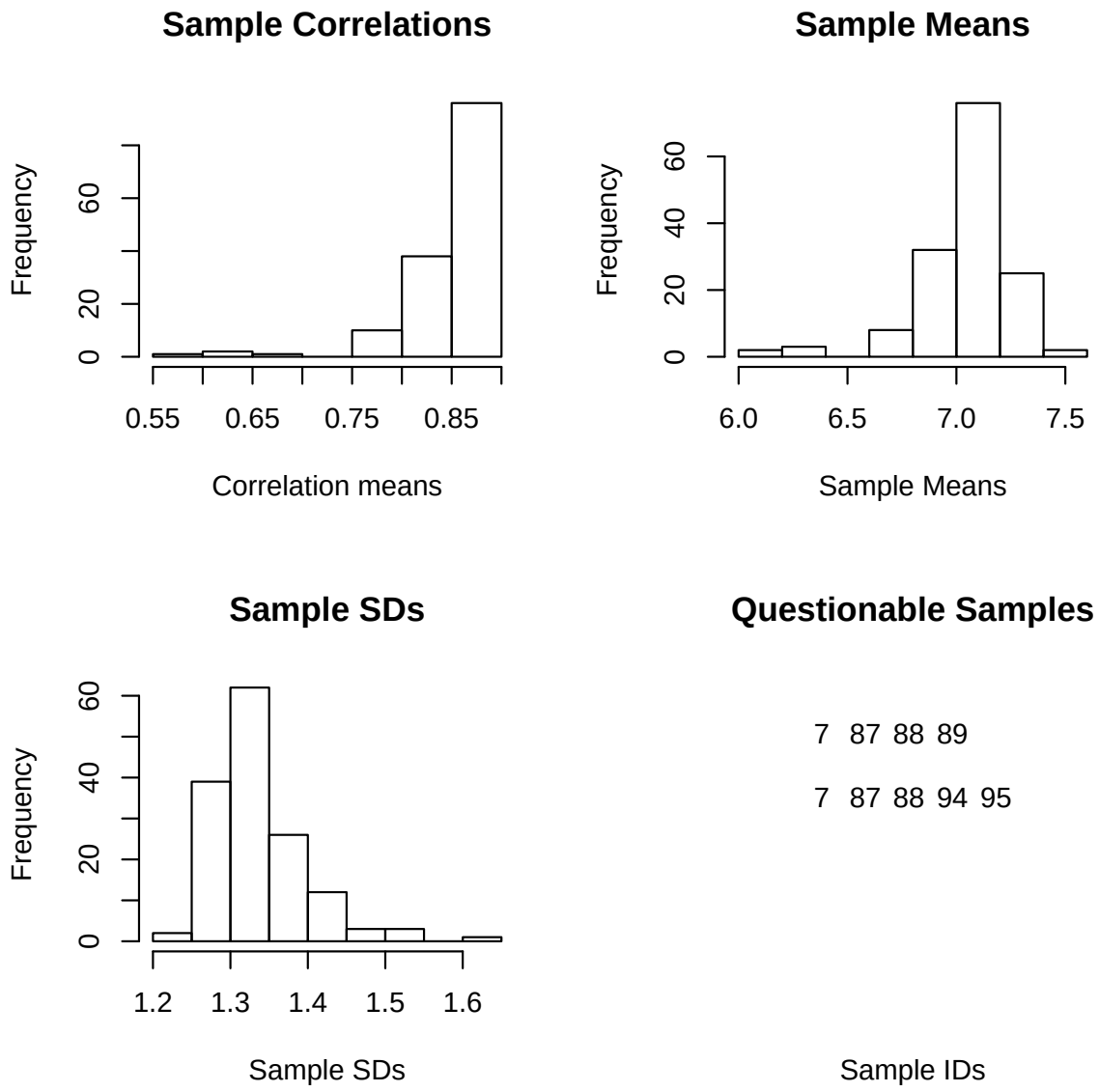

Figure 5: dataQE outputs if plot=TRUE.

dataQE outputs if plot=TRUE. Upper left: Histogram of average pair-wise correlations. Lower left: Histogram of Sample standard deviations. Upper right: Histogram of sample means. Lower right: Questionable samples.

There is another problem when we do gene-matching by EntrezID. It could happen that some gene IDs may convert to the same EntrezID, therefore we will have duplicated EntrezIDs after conversion. There are two options in `gene.match` to solve this problem. One is `"average"` and another one is `"IQR"`[1].

For Affymetrix array, we have prepared the functions for `HU6800, HGU133, HGU133A, HGU133A2, HGU133B, HGU133PLUS2, HGU95A, HGU95Av2`. For cDNA array, we have `UniGene ID, GenBank Accession Number, RefSeq Identifiers, Official Gene Symbol, Gene Symbol`.

2.4.2.2 Function for correlation of correlations In addition to retrieving the single gene information among multiple studies, someone could be interested in the genomic correlations of these cancer studies. For this purpose, we developed another function, `CorCor`, to calculate the genomic meta-correlations. We call it "Correlations of Correlations". The steps issue are as follows. Suppose we have S studies.
[1] Calculate the pair-wise correlations of whole genes in each study.
[2] Applying the concepts in 1.2.4 to do a modification and transformation to get unbiased estimators and $r - to - z$ score for meta-analysis correlations.
[3] Calculate the pair-wise correlations between any two of the S studies and show the results as a symmetric correlation matrix.

For example, if we type the following code in R,

```
da <- list(brown,liver,heart,ske)
na <- c("brown","liver","heart","ske")
CorCor(da,names=na)
```

CorCor will give you the following correlation matrix,

[1] IQR is the abbreviation of Interquartile range. More details are described in Appendix B

```
            brown         liver         heart           ske
brown  1.00000000  0.0948003492  0.0976056777  0.0006159800
liver  0.09480035  1.0000000000  0.0510178781  0.0008724284
heart  0.09760568  0.0510178781  1.0000000000  0.0007053562
ske    0.00061598  0.0008724284  0.0007053562  1.0000000000
```

2.4.2.3 Function for gene retrieving of clinical features Before any biomarker detection, based on some knowledge, the biologists could be roughly interested in the expression behavior of a certain gene among different data sets or cancer studies. For this purpose, we developed a function, `meta.gene.retrieve`. There are five parts in this function.

1. `data`: a list of data sets which the user would like to investigate.
2. `gene`: the name of the gene that the user is interested in. It should be included in double quotes, "".
3. `ID.Type`: the ID type of gene. For Affymetrix array, the options are `HU6800, HGU133, HGU133A, HGU133A2, HGU133B, HGU133PLUS2, HGU95A, HGU95Av2`. For cDNA array, the options are `UniGene ID, GenBank Accession Number, RefSeq Identifiers, Official Gene Symbol, Gene Symbol`.
4. `plot.type`: plot types of the results. There are two types, "boxplot" and "CI" (95%confidence interval), which users can choose.
5. `study.names`: a list of the names of data sets in `data`.
6. `pool.replicate`: the pool methods for duplicated genes in gene convertion. There are two options, "average" and "IQR", which users can choose.

For example, MT1H is an important gene in metabolism processes and we would like to know the expression behavior among some interesting cancer data sets, prostate cancer: `Magee, Welsh, Luo, Dhanasekaran, Singh, Varambally`, and lung cancer: `Bhattacharjee, Garber, Beer, Wachi, Gordon`. This function is only for the comparison of " Normal vs. Tumor". If we would like to use box-plot as the plot type and "IQR" as the pool.replicated method, the following commands are used,

```
data <- list(Magee,Welsh,Luo,Singh,Varambally,Bhattacharjee,Beer,Wachi,Gordon)
sname <- c("Magee","Welsh","Luo","Singh","Varambally","Bhattacharjee","Beer",
"Wachi","Gordon")
meta.gene.retrieve(data=data, gene="MT1H", ID.Type="SYMBOL", plot.type=c("boxplot"),
study.names=sname, pool.replicate="IQR")
```

We will get the results as shown in Figure 6. On the bottom of each box-plot, `meta.gene.retrieve` will show the sample size and provide the p-value of a simple t-test.

If there is not such a comparison type in this dataset, `meta.gene.retrieve` will show `"NO SUCH TYPE"` in the middle of the plot. If there is not a gene in this dataset, `meta.gene.retrieve` will show `"NO SUCH GENE"`.

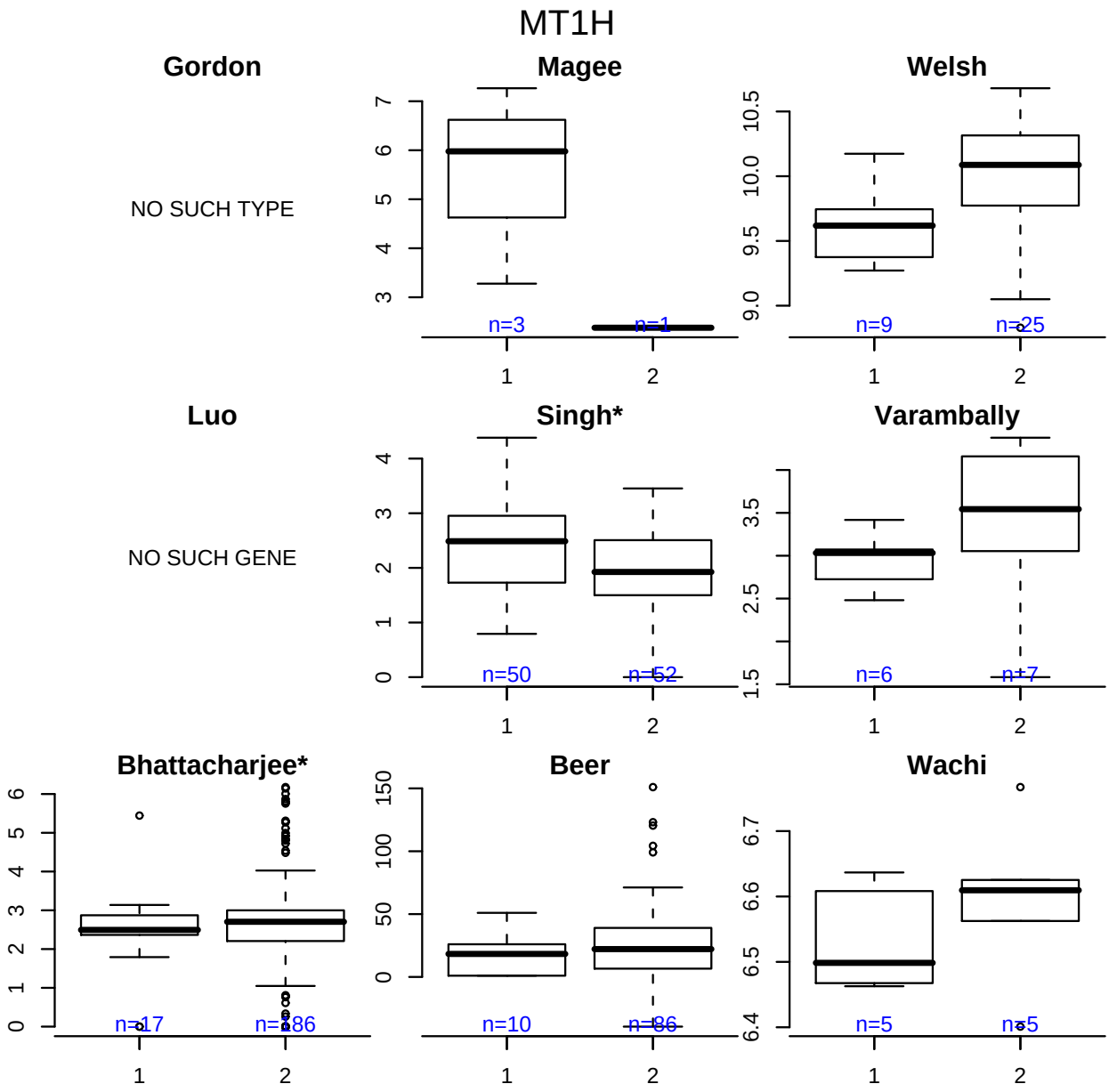

Figure 6: MT1H Retrieve Plots.

Box plots of the expression levels of MT1H among different cancer data sets. On the bottom of each plot, `meta.gene.retrieve` provides the p-values of simple t-test

3.0 BIOMAKER DETECTION METHODS FOR MULTIPLE MULTI-CLASS STUDIES

As the microarray technology becomes mature and prevalent in biomedical research, an increasing number of data sets have been accumulated in the public domain. Systematic information integration of multiple related studies should improve biomarker detection. So far, published meta-analysis methods for this purpose mostly consider two-class comparison. Methods for combining multi-class studies and pattern concordance have been rarely explored.

We first consider a natural extension of combining p-values from the traditional ANOVA model. Since p-values from ANOVA do not guarantee to reflect the concordant expression pattern information across studies, we propose a multi-class correlation measure (MCC) to specifically look for biomarkers of concordant inter-class patterns across a pair of studies. For both approaches, we focus on identifying biomarkers differentially expressed in all studies (i.e. ANOVA-maxP and min-MCC). The min-MCC method is further extended to identify biomarkers differentially expressed in partial studies using an optimally-weighted technique (OW-min-MCC). All methods are evaluated by simulation studies and by three meta-analysis applications to multi-tissue mouse metabolism data sets, multi-condition mouse trauma data sets and multi-malignant-condition human prostate cancer data sets.

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 classes of samples are measured with n_{sk} replicates for study s and class k. Denote by x_{sgki} the expression intensity of gene g (1

$\leq g \leq$ G), study s ($1 \leq s \leq$ S), class k ($1 \leq k \leq$ K) and replicated sample i ($1 \leq i \leq n_{sk}$). In this proposal, we particularly consider the situation when K > 2.

3.1 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, ANOVA model is first used to test the significance of variation in gene expressions across phenotype classes in each study. Corresponding p-values from F-test are then combined by taking the maximum.

3.1.1 Procedure for ANOVA-maxP

1. Compute F-statistics, F_{gs}, for gene g in the sth study.

2. Permute group labels in each study for B times, and similarly calculate 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 p-value of F_{gs} as

$$p_{gs} = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(F_{gs}^{(b)} \geq F_{gs})}{(B \cdot G)} \tag{3.1}$$

   where $I(\cdot)$ is the indicator function that takes values one when the statement is true and zero otherwise.

   Similarly given $F_{gs}^{(b)}$, compute p-value of $F_{gs}^{(b)}$ as

$$p_{gs}^{b} = \frac{\sum_{b'=1}^{B} \sum_{g'=1}^{G} I(F_{gs}^{(b')} \geq F_{gs})}{(B \cdot G)} \tag{3.2}$$

4. The maximum P-value statistic is defined as

$$V_g = \max_{1 \le s \le S} p_{gs}$$

. Similarly define

$$V_g^{(b)} = \max_{1 \le s \le S} p_{gs}^{(b)}$$

.

5. Estimate p-value of V_g by

$$p(V_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(V_{g'}^{(b)} \le V_g)}{(B \cdot G)} \tag{3.3}$$

6. Estimate q-value for each gene as

$$q(V_g) = \hat{\pi_0} \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(V_{g'}^{(b)} \le V_g)}{(B \sum_{g'=1}^{G} I(V_g' \le V_g))} \tag{3.4}$$

where $\hat{\pi_0}$ is the estimate of proportion of null genes. A conservative suggestion is to set $\hat{\pi_0}$ as 1. Genes with q-values smaller than 0.05 are detected as biomarkers.

3.2 MULTI-CLASS CORRELATION(MCC) FOR A PAIR OF STUDEIS

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 $x_{kj}(1 \le k \le K, 1 \le j \le n_k)$ to represent expression intensity of class k, sample j for the first study and $y_{kj}(1 \le k \le K, 1 \le j \le m_k)$ 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 $(x_{11}, ..., x_{1n_1}, ..., x_{k1}, ..., x_{kn_k})$ and $(y_{11}, ..., y_{1m_1}, ..., y_{k1}, ..., y_{km_k})$ if $n_k = m_k, \forall k$. However, since $n_k \ne m_k$ in general and this naive definition ignores the exchangeability within $(x_{11}, ..., x_{1n_k})$ and $(y_{11}, ..., y_{1n_k})$ for a given $1 \le k \le K$, 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. $E(X_k) = \mu_{X_k}, E(Y_k) = \mu_{Y_k}, Var(X_k) = \sigma^2_{X_k}, Var(Y_k) = \sigma^2_{Y_k}$. Aslo 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

$$F_{(X,Y)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X_k,Y_k)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X_k)}(s)F_{(Y_k)}(t) \tag{3.5}$$

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

$$MCC = cor((X_1, ..., X_K), (Y_1, ..., Y_K)) = \frac{E(XY) - EX \cdot EY}{\sqrt{Var(X) \cdot Var(Y)}} \tag{3.6}$$

$$\frac{(\frac{1}{K}\sum_{k=1}^{K}\mu_{X_k}\mu_{Y_k}) - (\frac{1}{K}\sum_{k=1}^{K}\mu_{X_k})(\frac{1}{K}\sum_{k=1}^{K}\mu_{Y_k})}{\sqrt{[\frac{1}{K}\sum_{k=1}^{K}\sigma^2_{X_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{X_k} - \overline{\mu}_X)^2][\frac{1}{K}\sum_{k=1}^{K}\sigma^2_{Y_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{Y_k} - \overline{\mu}_Y)^2]}} \tag{3.7}$$

where

$$\overline{\mu}_X = \frac{1}{K}\sum_{k=1}^{K}\mu_{X_k} \tag{3.8}$$

and

$$\overline{\mu}_Y = \frac{1}{K}\sum_{k=1}^{K}\mu_{Y_k} \tag{3.9}$$

This correlation measure takes values between -1 and 1. A large positive correlation indicates similar patterns between two studies for a given gene.

When $n_1 = n_2 = ... = n_K = n$ and $m_1 = m_2 = ... = m_K = m$, MCC can be rewritten as below:

$$MCC = \frac{r_{\bar{X}\bar{Y}}}{\sqrt{\frac{1}{F_X} + 1}\sqrt{\frac{1}{F_Y} + 1}} \tag{3.10}$$

Where

$$r_{\bar{X}\bar{Y}} = \frac{(\frac{1}{K}\sum\mu_{X_k}\mu_{X_k}) - (\frac{1}{K}\sum\mu_{X_k})(\frac{1}{K}\sum\mu_{Y_k})}{\sqrt{\frac{1}{K}\sum(\mu_{X_k} - \overline{\mu}_X)^2}\sqrt{\frac{1}{K}\sum(\mu_{Y_k} - \overline{\mu}_Y)^2}} \tag{3.11}$$

and

$$F_X = \frac{\sum(\mu_{X_k} - \overline{\mu}_X)^2/K}{\sum\sum(x_{kj} - \mu_{X_k})^2/(n \cdot K)}$$

,

$$F_Y = \frac{\sum(\mu_{Y_k} - \overline{\mu}_Y)^2/K}{\sum\sum(y_{kj} - \mu_{Y_k})^2/(m \cdot K)} \tag{3.12}$$

From the equation (3.3), F_X and F_Y are related to F-statistic in ANOVA. $r_{\bar{X}\bar{Y}}$ is the sample correlation of $(\mu_{X_1}, ..., \mu_{X_K})$ and $(\mu_{Y_1}, ..., \mu_{Y_K})$. When the within-class variation is much smaller than the between-class variation, F_X and F_Y become large. MCC converges to $r_{\bar{X}\bar{Y}}$ as expected.

In practice, distributions of (X_k, Y_k) are unknown. The means (μ_{X_k} and μ_{Y_k}) and the variances (σ_{X_k} and σ_{Y_k}) are not available. 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 $F_{X'_k}(t) = \sum_{j=1}^{n_k} I(x_{kj} \le t)$ and similarly $F_{Y'_k}(t) = \sum_{j=1}^{n_k} I(y_{kj} \le t)$. Define (X', Y') to be equal mixture of bivariate distribution (X'_k, Y'_k) such that

$$F_{(X',Y')}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X'_k,Y'_k)}(s,t) = \frac{1}{K}\sum_{k=1}^{K} F_{(X'_k)}(s)F_{(Y'_k)}(t) \tag{3.13}$$

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

$$MCC = cor(\tilde{x}, \tilde{y})) = \frac{E(X'Y') - EX' \cdot EY'}{\sqrt{Var(X') \cdot Var(Y')}} \tag{3.14}$$

$$= \frac{(\frac{1}{K}\sum_{k=1}^{K} \mu_{X'_k}\mu_{Y'_k}) - (\frac{1}{K}\sum_{k=1}^{K} \mu_{X'_k})(\frac{1}{K}\sum_{k=1}^{K} \mu_{Y'_k})}{\sqrt{[\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{X'_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{X'_k} - \overline{\mu}_{X'})^2][\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{Y'_k} + \frac{1}{K}\sum_{k=1}^{K}(\mu_{Y'_k} - \overline{\mu}_{Y'})^2]}} \tag{3.15}$$

where

$$\overline{\mu}_{X'} = \frac{1}{K}\sum_{k=1}^{K} \mu_{X'_k}, \ \overline{\mu}_{Y'} = \frac{1}{K}\sum_{k=1}^{K} \mu_{Y'_k} \tag{3.16}$$

,

$$\mu_{X'_k} = \frac{\sum_{j=1}^{n_k} x_{kj}}{n_k}, \ \mu_{Y'_k} = \frac{\sum_{j=1}^{m_k} y_{kj}}{m_k} \tag{3.17}$$

and

$$\sigma^2_{X'_k} = \frac{\sum_{j=1}^{n_k}(x_{kj} - \mu_{X'_k})^2}{n_k}, \;\; \sigma^2_{Y'_k} = \frac{\sum_{j=1}^{m_k}(y_{kj} - \mu_{Y'_k})^2}{m_k} \tag{3.18}$$

When $n_1 = n_2 = ... = n_K = n$ and $m_1 = m_2 = ... = m_K = m$, MCC will have another form as below which is related to F statistic from ANOVA.

$$MCC = \frac{r_{\bar{X}'\bar{Y}'}}{\sqrt{\frac{1}{F_{X'}} \cdot \frac{K-1}{K} + 1}\sqrt{\frac{1}{F_{Y'}} \cdot \frac{K-1}{K} + 1}} \tag{3.19}$$

where

$$r_{\bar{X}'\bar{Y}'} = \frac{\sum(\bar{x}_k - \bar{x}..)(\bar{y}_k - \bar{y}..)}{\sqrt{\sum(\bar{x}_k - \bar{x}..)^2}\sqrt{\sum(\bar{y}_k - \bar{y}..)^2}} \tag{3.20}$$

and

$$F_{X'} = \frac{\sum(\bar{x}_k - \bar{x}_{..})^2/(K-1)}{\sum\sum(x_{ki} - \bar{x}_k)^2/(n-1)K} \tag{3.21}$$

,

$$F_{Y'} = \frac{\sum(\bar{y}_k - \bar{y}_{..})^2/(K-1)}{\sum\sum(y_{ki} - \bar{y}_k)^2/(n-1)K} \tag{3.22}$$

From the equation (3.6), $F_{X'}$ and $F_{Y'}$ are exactly the F-statistics in ANOVA for $\tilde{x}$ and $\tilde{Y}$. $r_{\bar{X}\bar{Y}}$ is the sample correlation of $(\mu_{X'_1}, ..., \mu_{X'_K})$ and $(\mu_{Y'_1}, ..., \mu_{Y'_K})$. When the within-class variation is much smaller than the between-class variation, $F_{X'}$ and $F_{Y'}$ become large. MCC converges to $r_{\bar{X}'\bar{Y}'}$ as expected.

3.2.1 Procedure of MCC for combining two studies

1. Compute MCC statistic, MCC_g, for each gene g.

2. Permute group labels in each study for B times, and similarly calculated the permuted statistics, $MCC_g^{(b)}$ where $1 \leq g \leq G, 1 \leq s \leq S$ and $1 \leq b \leq B$.

3. Estimate p-value of MCC_g as

$$p(MCC_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(MCC_{g'}^{(b)} \geq MCC_g)}{(B \cdot G)} \tag{3.23}$$

where $I(\cdot)$ is the indicator function that takes values one when the statement is true and zero otherwise.

4. Estimate q-value for each gene of MCC_g as

$$q(MCC_g) = \hat{\pi_0} \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(MCC_{g'}^{(b)} \geq MCC_g)}{(B \sum_{g'=1}^{G} I(MCC'_g \geq MCC_g))} \tag{3.24}$$

where $\hat{\pi_0}$ is the estimate of proportion of null genes. A conservative suggestion is to set $\hat{\pi_0}$ as 1. Genes with q-values smaller than 0.05 are detected as biomarkers.

3.2.2 Minimum MCC (min-MCC) for more than two studies

The MCC measure described above measures the 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. The procedures are described as follows:

1. Compute MCC statistic, $MCC_{g(u)(v)}$, for each gene g and for a pair of studies u and v.

2. Permute group labels in each study for B times, and similarly calculate the permuted statistics, $MCC_{g(u)(v)}^{(b)}$ where $1 \leq g \leq G, 1 \leq s \leq S$ and $1 \leq b \leq B$.

3. Calculate

$$min - MCC_g = \min_{1 \le u \le v \le S} MCC_{g(u)(v)} \tag{3.25}$$

and

$$min - MCC_g^{(b)} = \min_{1 \le u \le v \le S} MCC_{g(u)(v)}^{(b)} \tag{3.26}$$

4. Estimate p-value of $min - MCC_g$ as

$$p(min - MCC_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(min - MCC_{g'}^{(b)} \ge min - MCC_g)}{(B \cdot G)} \tag{3.27}$$

where $I(\cdot)$ is the indicator function that takes values one when the statement is true and zero otherwise.

5. Estimate q-value for each gene of $min - MCC_g$ as

$$q(min - MCC_g) = \hat{\pi_0} \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(min - MCC_{g'}^{(b)} \ge min - MCC_g)}{(B \cdot \ sum_{g'=1}^{G} I(min - MCC_{g'} \ge min - MCC_g))} \tag{3.28}$$

where $\hat{\pi_0}$ is the estimate of proportion of null genes. A conservative suggestion is to set $\hat{\pi_0}$ as 1. Genes with q-values smaller than 0.05 are detected as biomarkers.

3.3 OPTIMALLY-WEIGHTED STATISTICS FOR MINIMUM MCC (MIN-MCC)

Li and Tseng (2009) proposed an optimally weighted (OW) statistic was modified from the Fisher's score and was proposed for the former hypothesis setting HS_B. The optimal weights provided natural categorization of the detected biomarkers for further biological investigation. They compared this method to the classical Fisher's equally weighted statistic (EW), Tippett's minimum p-value statistic (minP) and Pearson's statistic (PR).

In general, there exists no uniformly most powerful test. All of the four methods compared are admissible under a simplified Gaussian scenario. Nevertheless, the proposed OW statistic consistently has the best or near ot the best power in a wide variety of alternative

hypotheses, especially when EW and minP perform poorly in two extreme alternative hypotheses respectively.

3.3.1 An introduction to Optimally-Weighted statistic

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

$$U_g(w_g) = -\sum_{s=1}^{S} w_{gs} log(p_{gs}) \tag{3.29}$$

where p_{gs} is the p-value of gene g in study s, w_s is the weight assigned to the sth study and $w_g = (w_{g1}, \ldots, w_{gs})$. Under the null hypothesis that $\theta_{gs} = 0 \ \forall \ s$, the p-value of the observed weighted statistic, $p(u_g(w_g))$, can be obtained for a given gene g and weight w_g (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:

$$V_g^{OW} = \min_{w_g \in W} p(u_g(w_g)) \tag{3.30}$$

where $u_g(w)$ is the observed statistic for $U_g(w)$ and W is a pre-specified searching space. The choice of searching space in this paper is $W = \{w | w_i \in \{0, 1\}\}$, which results in an affordabl 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.

The vector of optimal-weight, $w_g^* = \arg\min_{wg \in W} p(u_g(w_g))$ actually serves as a convenient basis fogene categorization in the follow-up of biological interpretation and exploration. The optimal weights obtained from OW have the additional advantages of filtering discordant biomarkers and providing natural categorization of the detected genes for further biological investigation.

3.3.2 Procedure for OW-min-MCC

We also applied OW concept in min-MCC for multiple multi-class studies to detected biomarkers with conditional inter-class patterns across these studies. An algorithm applying the idea of OW statistics was developed as follows:

Suppose there are S studies, $W = \{w = (w_1, \ldots, w_s | w_i \in \{0, 1\},\ 2 \leq \sum_{i=1}^{S} w_i \leq S\})$

1. Given a weight $w = (w_i, \ldots, w_s)$, then min-MCC based on w is defined as

$$M_g(w) = \min_{w_u=1, w_v=1, 1 \leq u \neq v \leq S} (MCC_{g(u)(v)}) \tag{3.31}$$

for study u and v , $1 \leq g \leq G$.

2. Randomly permute class labels in each study for B times, and similarly calculate the permuted min-MCC which is defined as

$$M_g^{(b)}(w) = \min_{w_u=1, w_v=1, 1 \leq u \neq v \leq S} (MCC_{g(u)(v)}^{(b)}) \tag{3.32}$$

$1 \leq g \leq G$, $1 \leq b \leq B$.

3. Estimate the p-value of $M_g(w)$ and $M^(b)_g(w)$ as

$$p(M_g(w)) = \frac{\sum_{b=1}^{B} \sum_{g=1}^{G} I(M_g^{(b)} \geq M_g(w))}{B \cdot G} \tag{3.33}$$

$$p(M_g^{(b)}(w)) = \frac{\sum_{b'=1}^{B} \sum_{g'=1}^{G} I(M_{g'}^{(b')} \geq M_g^{(b)}(w))}{B \cdot G} \tag{3.34}$$

4. Define H_g as the optimally weighted(OW) statistic of $p(M_g(w))$, such that

$$H_g = \min_{w \in W} p(M_g(w))$$

Similarly,

$$H_g^{(b)} = \min_{w \in W} p(M_g^{(b)}(w))$$

5. Assess p-value of H_g as

$$p(H_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(H_{g'}^{(b)} \leq H_g)}{B \cdot G} \tag{3.35}$$

6. Assess q-value of H_g as

$$q(H_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(H_{g'}^{(b)} \leq H_g)}{B \cdot \sum_{g'=1}^{G} I(H_{g'} \leq H_g)} \tag{3.36}$$

7. Calculate optimal weight as

$$w_g^* = \text{argmin}_{w \in W} p(M_g(w)) \tag{3.37}$$

3.4 RESULTS

3.4.1 Simulation study

We conducted a simulation scenario for combining three genomic studies to assess the performance of our proposed ANOVA-maxP, min-MCC method, and OW-min-MCC. Denote by x_{sgki} the expression intensity of study s ($1 \leq s$), 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 simulation scenario, we simulated three studies (S = 3). Each study had three classes (K = 3). The numbers of replicates, $n_{sk}(1 \leq s \leq S, 1 \leq k \leq K)$, were different among each class of each study.

A total of G = 2000 genes in each study were simulated. Among these 2000 genes, 300 genes showed concordant inter-class patterns across all studies (category I), 100 genes were with discordant inter-class pattern (category II), and 100 genes with concordant pattern in study 1 and study 2 but no pattern in study 3 (category III). The expression intensities were simulated from $x_{sgki} \sim N(\mu_{sk}, \sigma_s^2)$. For genes with concordant inter-class pattern, study 1: $\mu_1 = (1, 3, 5)$, study 2: $\mu_2 = (2, 4, 6)$, and study 3: $\mu_3 = (1, 4, 7)$. So, mean vectors

$\mu_s = (\mu_{s1}, ..., \mu_{sk})$ across studies had pair-wise correlation one. For genes with discordant inter-class patterns, that study 1 is $\mu_1 = (1, 3, 5)$, study 2 is $\mu_2 = (6, 4, 2)$, and study 3: $\mu_3 = (1, 7, 4)$, pair-wise correlations of mean vectors were low or negative. For the genes in category III, mean vector for study 1 is $(1, 3, 5)$, for study 2 is $(2, 4, 6)$, and $(0, 0, 0)$ for study 3. For the rest of the 1500 genes are null genes form $(\mu_{s1}, ..., \mu_{sk}) = (0, ..., 0)$.

The effect size was defined as the ratio of the standard deviation of the mean vectors to the within class standard deviation, σ_s. We chose effect sizes to be 0.5, 0.6, and 0.7. False discovery rate (FDR) was controlled at 0.05 for each method, and each simulation was repeated 200 times. The details of simulation settings are described in Table 4.

Table 4: Settings of simulation scenario

		Study1	Study 2	Study 3
Effect size	N	$(n_{11}, n_{12}, n_{13}) = (10, 5, 8)$	$(n_{21}, n_{22}, n_{23}) = (5, 8, 10)$	$(n_{31}, n_{32}, n_{33}) = (8, 10, 5)$
		$(\mu_{11}, \mu_{12}, \mu_{13}), \sigma_1$	$(\mu_{21}, \mu_{22}, \mu_{23}), \sigma_2$	$(\mu_{31}, \mu_{32}, \mu_{33}), \sigma_3$
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	(1,7,1),5.9
	III	(1,3,5),3.5	(6,4,2),3.1	(0,0,0),5.3
	Null	(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	(1,7,1),4.8
	III	(1,3,5),2.9	(6,4,2),2.6	(0,0,0),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	(1,7,1),4.3
	III	(1,3,5),2.5	(6,4,2),2.2	(0,0,0),3.8
	Null	(0,0,0),2.5	(0,0,0),2.2	(0,0,0),3.2

I: Genes with concordant patterns across studies. The mean vector in Study 1 is (1,3,5) which has the same pattern (trend) as the mean vector, (2,4,6), in Study 2 and (1,4,7) in Study 3.

II: Genes with discordant patterns across studies. The mean vector in Study 1 is (1,3,5) which has the different pattern (trend) from the mean vector, (6,4,2), in Study 2 and (1,7,1) in Study 3.

III: Genes with concordant patterns in study 1 and study 2 but no pattern in study 3. The mean vector for Study 1 is (1,3,5), (6,4,2) for Study 2 and (0,0,0) for Study 3.

Null: Null genes.Do not have any pattern in the studies.

The average numbers of genes identified in each category by the three methods (ANOVA-maxP, min-MCC and OW-min-MCC) under each simulation scenario is presented in Table 2. We note that each of the three methods has specific target of biomarkers: ANOVA-maxP detects concordant and discordant biomarkers across all three studies (i.e. category I and II); min-MCC identifies concordant biomarkers across all three studies (i.e. category I); OW-min-MCC finds concordant biomarkers across two or more studies (i.e. category I and III). Thus, we define three types of false discov-ery rates for the corresponding biological purposes: FDR1=(III+null)/(I+II+III+null) for ANOVA-maxP, FDR2=(II+III+null)/(I+II+III+null) for min-MCC and FDR3=(II+null)/(I+II+III+null) for OW-min-MCC.

As expected, ANOVA-maxP detects both category I and cate-gory II genes because the p-values in ANOVA do not reflect the inter-class pattern information for each individual study. On the other hand, min-MCC method detects almost only concordant inter-class genes (category I). For example, when effect size equals 0.6, min-MCC detects an average of 260.52 (out of 300) genes of true concordant inter-class pattern genes, while ANOVA-maxP only identifies 193.11 concordant genes together with 73.5 discordant biomarkers. Both of ANOVA-maxP and min-MCC detect few bio-markers of category III because genes in category III are differen-tially expressed only in study 1 and 2 but not in study 3. OW-min-MCC detects biomarkers mostly in category I and III but not in II since it is designed to detect concordant biomarkers in two or more studies. The simulation result clearly confirms that selection among the three methods depend on the ultimate biological purpose.

Result of the FDR calculation reveals an interesting issue of our proposed algorithm. FDR3 for OW-min-MCC is controlled near the nominal FDR=5%. FDR1 and FDR2 are, however, anti-conservative. This is an issue of HSA that the null hypothesis is essentially a composite null hypothesis, instead of a simple null hypothesis. We will discuss further in the "Conclusion Section".

Table 5: Results of simulation scenario

Effect size	Methods	I	II	III	Null	FDR1(%)	FDR2(%)	FDR(%)
0.5	ANOVA-maxP	107.08	44.23	7.56	5.22	7.8[a]	34.7	30.1
	min-MCC	197.42	0.22	10.05	7.95	8.3	8.4[b]	3.8
	OW-min-MCC	163.67	3.76	29.74	5.23	17.3	19.1	4.4[c]
0.6	ANOVA-maxP	193.11	73.50	13.47	9.45	7.9[a]	33.3	28.6
	min-MCC	260.52	0.05	15.56	10.85	9.2	9.2[b]	3.8
	OW-min-MCC	253.19	7.10	57.98	8.73	20.4	22.6	4.8[c]
0.7	ANOVA-maxP	250.77	88.98	16.26	11.63	7.6[a]	31.8	27.4
	min-MCC	287.97	0.03	19.24	11.96	9.8	9.8[b]	3.8
	OW-min-MCC	288.54	9.74	80.41	10.15	23.3	25.8	5.1[c]

I: genes with concordant patterns across studies.

II: genes with discordant patterns across studies.

III: genes with concordant patterns in study 1 and study 2 but no pattern in study 3.

Null: Null genes.

[a]: ANOVA-maxP detects both concordant and discordant genes (category I and II). FDR1 ($=\frac{III+Null}{I+II+III+Null}$) is a better measure for false discoveries.

[b]: Min-MCC detects only concordant genes (category I). FDR2 ($=\frac{II+III+Null}{I+II+III+Null}$) is a better measure for false discoveries.

[c]: OW-min-MCC detects category I and III genes only. FDR3 ($=\frac{II+Null}{I+II+III+Null}$) is a better measure for false discoveries.

3.4.2 Data description

Mouse Metabolism Data

Three 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 the microarray experiment separately to study the expression changes across genotypes. For duplicate spots, mean of them were 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 quality defect and were deleted from further analysis. The detailed sample information is described in Table 6.

Table 6: Mouse metabolism data

tissue type	brown fat			liver			heart			skeletal			Total
genotype	WT	V-	L-	WT	V-	L-	WT	V-	L-	WT	V-	L-	
n of arrays	4	4	4	4	4	4	3	4	4	3	3	3	44

WT: wilde type; V-:VLCAD -/-; L-:LCAD -/-

Mouse Trauma Data

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 performed 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 had problematic quality and were removed from further analysis. Table 7 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.

Table 7: Mouse trauma data

array platform	Codlink					Affymetrix					Total
experimental conditions	I	II	III	IV	V	I	II	III	IV	V	
number of arrays	4	3	4	4	4	4	3	4	4	4	38

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.

Prostate Cancer Data

The third data is prostate cancer data from three different data sets, Dha-nasekaran et al (2001), Lapointe et al (2004), and Varambally et al (2005). These three data sets all have three tumor types, normal, tumor, and meta-static. Data from the three studies were log2 transformed and converted to Entrez ID for combining with other studies. Genes with large missing percentage (means > 40 %) or little information content (average log2-scaled means < 7 or average log2-scaled standard deviations < 0.4) are filtered out. KNN imputation method was used for the missing values before applying proposed methods. The basic experimental information is in Table 8.

Table 8: Sample description of three prostate cancer studies

	Dhanasekaran			Lapointe			Varambally		
Array type	cDNA			cDNA			Affy HU133Plus2		
Tissue type	N	T	M	N	T	M	N	T	M
N of arrays	14	13	20	41	62	9	6	7	6

N: Normal; T: Tumor; M: Metastasis

The first data set contains a multi-tissue design. The application of HS_A helps to detect tissue-specific biomarkers and HS_B 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. HS_B is of interest to generate highly confident biomarkers confirmed by both platforms while HS_A becomes of less biological interests. The detected platform-specific biomarkers from HS_A may help to identify technical issues across the two platforms.

3.4.3 Application to mouse metabolism data

We first applied ANOVA-maxP and min-MCC methods to the mouse metabolism data. For the first biological goal, we were interested in genes with clear inter-class patterns but may be discordant across tissues. These tissue-dependent biomarkers reflected tissue-specific biological changes under VLCAD and LCAD mutations.

ANOVA-maxP identified 637 genes and the heatmap of detected biomarkers is shown in the left plot of Figure 7A. Figure 7B shows a histogram of min-MCC (minimum of pair-wise multi-class correlation measures across four tissues) of the 637 detected biomarkers. 408 of the 637 genes (64.05%) had negative min-MCC (i.e. with discordant inter-class patterns in at least a pair of tissues). The right panel of Figure 7A shows the heatmap of these 408 discordant genes and the heatmap of 229 genes with positive min-MCC. The genes with discordant patterns are potential targets to identify tissue-specific regulators in the mutations.

The second biological goal was to identify biomarkers that have consistent inter-class pattern across all four tissues (i.e. reliable tissue-invariant biomarkers). To achieve this goal, min-MCC served better for this purpose. A total of 387 genes were identified and are displayed in Figure 7C. It is clearly seen that these genes have clear concordant inter-class pat-terns across all four tissues. A simple cluster analysis can further group them into six major patterns for further biological investigation. A manuscript of this on-going project containing detailed bio-logical discoveries is under preparation.

Finally, we applied OW-min-MCC method to allow detection of concordant biomarkers in partial (two or more) studies. Table 8 shows all 11 possible categories of the weights and the number and percentage of detected biomarkers. Among the 1,209 biomarkers detected by OW-min-MCC, we found that the number of significant biomarkers is relatively small (398) when the optimal weights included skeletal while heart involved the highest number of bio-markers (960). Heatmap of the 1,209 biomarkers is shown in Figure 8.

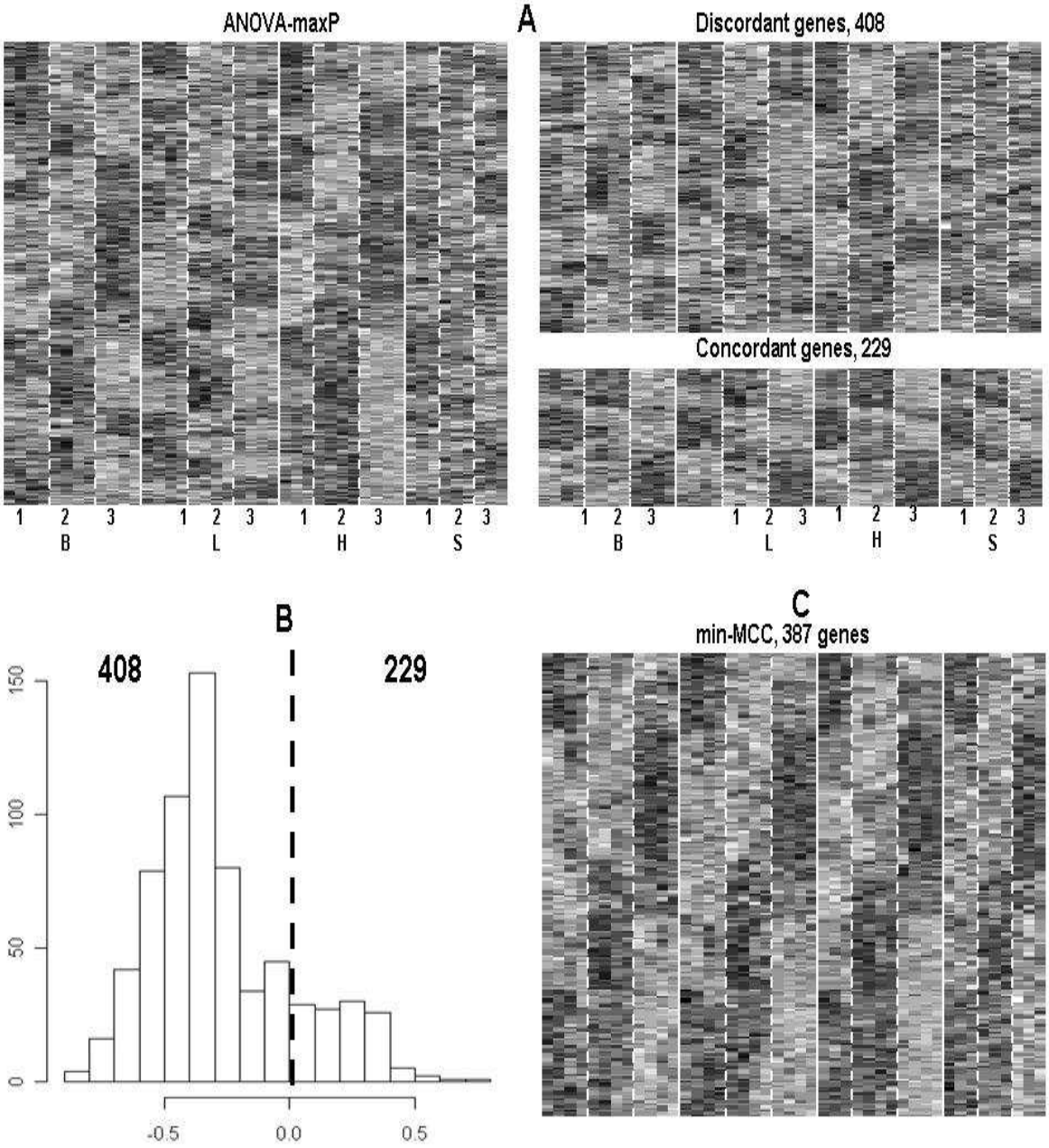

Figure 7: Heatmap of significant genes detected in mouse metabolism data.
A(Left): heatmap of 637 ANOVA-maxP genes, A(Right): heatmap of discordant ANOVA-maxP genes and concordant ANOVA-maxP genes. B: histogram of min-MCC of the 637 ANOVA-maxP genes, C: heatmap of 387 min-MCC genes (B: Brown fat; L: Liver; H: Heart; S: Skeletal muscle. 1: Wild type; 2: VLCAD -/-; 3: LCAD -/-.)

Table 9: OW-min-MCC results for mouse metabolism data

Category	Weight	Tissue	number of sig. genes	%
1	(1,1,0,0)	Brown, Liver	126	10.42
2	(1,0,1,0)	Brown, Heart	258	21.34
3	(1,0,0,1)	Brown, Skeletal	31	2.56
4	(0,1,1,0)	Liver, Heart	240	19.85
5	(0,1,0,1)	Liver, Skeletal	54	4.47
6	(0,0,1,1)	Heart, Skeletal	98	8.11
7	(1,1,1,0)	Brown, Liver, Heart	187	15.47
8	(1,1,0,1)	Brown, Liver, Skeletal	38	3.41
9	(1,0,1,1)	Brown, Heart, Skeletal	75	6.20
10	(0,1,1,1)	Liver, Heart, Skeletal	53	4.38
11	(1,1,1,1)	Brown, Heart, Live, Skeletal	49	4.05
Total			1209	100

Brown: Categories including: $1, 2, 3, 7, 8, 9, 11$; Total number of biomarkers: 764.

Liver: Categories including: $1, 4, 5, 7, 10, 11$; Total number of biomarkers: 747.

Heart: Categories including: $2, 4, 6, 7, 9, 10, 11$; Total number of biomarkers: 960.

Skeletal: Categories including: $3, 5, 6, 8, 9, 10, 11$; Total number of biomarkers: 398.

The number of significant biomarkers is relatively small when the weight included Skeletal. It could give us some information that genes could have different expression levels in Skeletal.

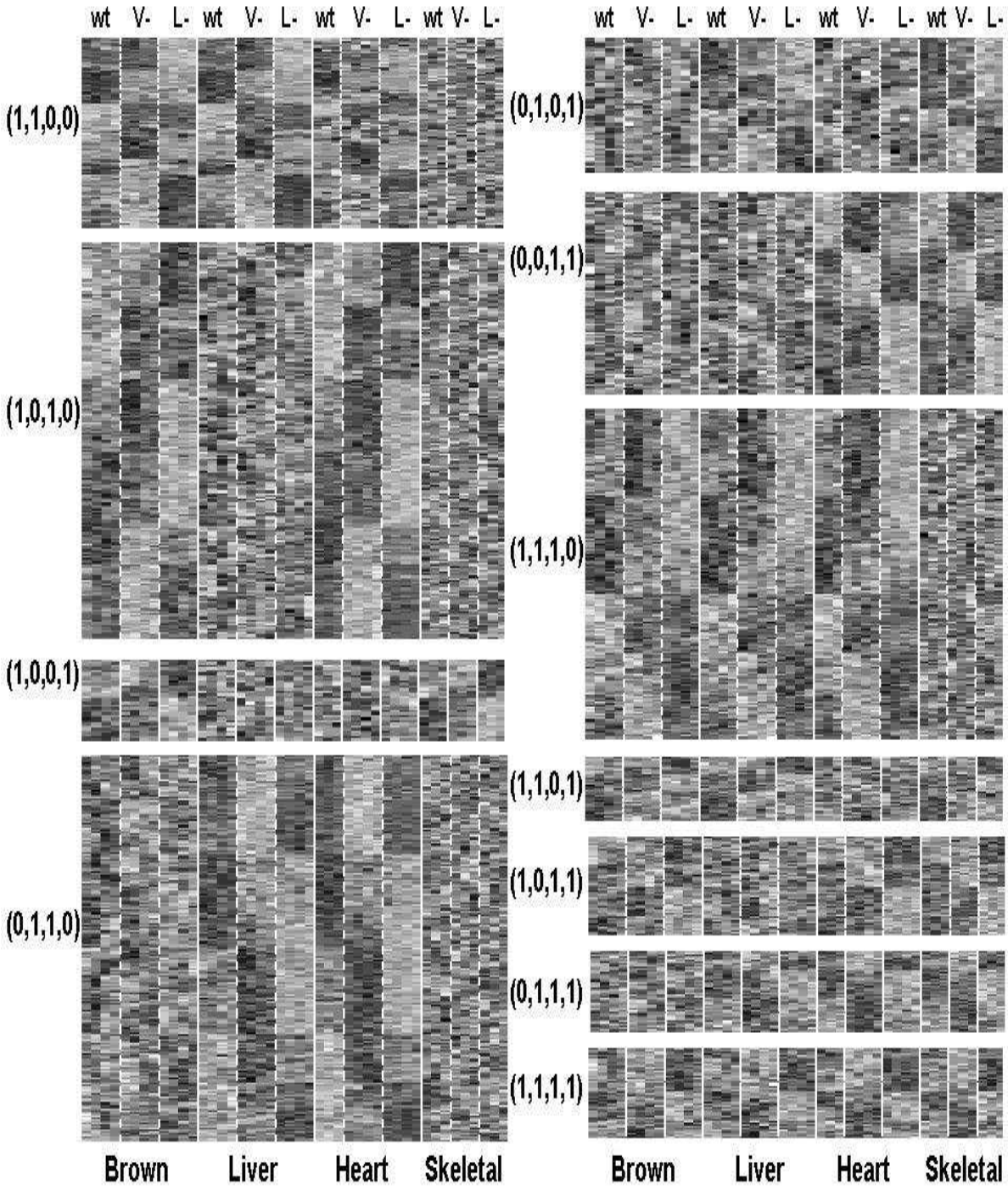

Figure 8: Heatmap for OW-min-MCC.

Heatmap of significant biomarkers for these 11 categories of weights. (wt: Wild Type, V-: VLCAD-/-, L-: LCAD-/-)

3.4.4 Application to mouse trauma data

We similarly applied ANOVA-maxP and min-MCC to the mouse trauma data. We did not apply OW-min-MCC since there were only two studies (platforms). 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 exist no discordant biomarkers. Combining the two data sets should increase statistical power and detect more concordant inter-class pattern genes.

Indeed, by controlling FDR at 0.05, ANOVA-maxP detected 3388 genes (heatmap shown in the left plot of Figure 9A) and 179 (5.28%) genes showed discordant inter-class patterns of negative MCC across the two platforms (right plot of Figure 9A). Figure 9B shows the histogram of min-MCC of the 3388 ANOVA-maxP genes. On the other hand, 3633 genes were identified using min-MCC (figure 9C). These highlyreliable biomarkers confirmed by both platforms were used for further cluster analysis and pathway analysis to understand the biological changes under different severity of trauma (manuscript in preparation).

The higher proportion of genes with concordant inter-class patterns confirmed that the two array platforms are highly reproducible. The 179 discordant inter-class pattern genes are, however, technical errors and need further investigation. The discordances are possibly due to mistaken gene annotation, differential hybridization efficiencies caused by different probe selection criteria or non-specific cross-hybridization.

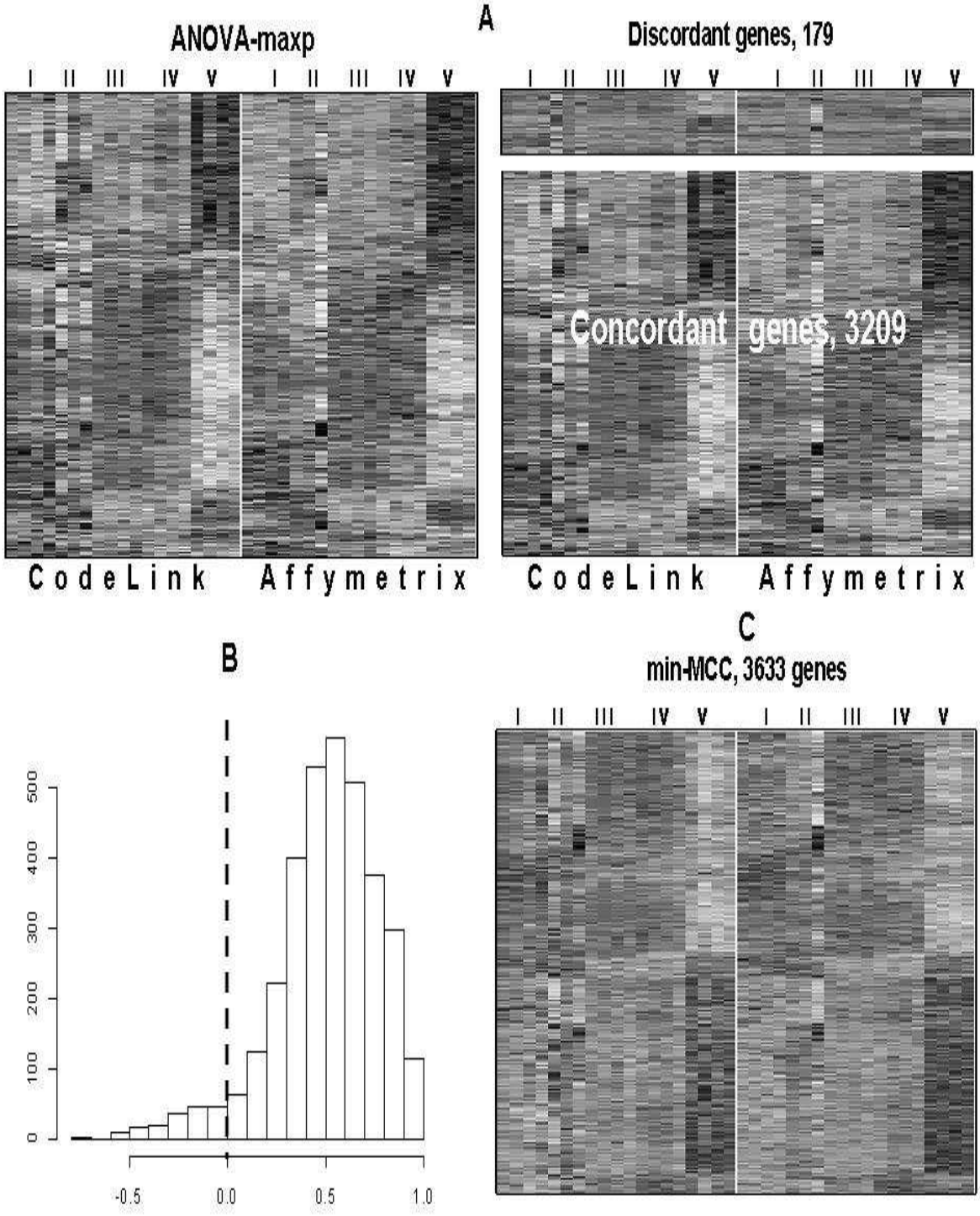

Figure 9: Heatmap of significant genes in mouse trauma data.

A(Left): heatmap of 3388 ANOVA-maxP genes, A(Right): heatmap of 3209 concordant ANOVA-maxP genes and 179 discordant ANOVA-maxP genes. B: histogram of min-MCC of the 3388 ANOVA-maxP genes, C: heatmap of 3633 min-MCC genes.

3.4.5 Application to prostate cancer data

We applied ANOVA-maxP and min-MCC to three prostate cancer studies (Dhanasekaran, Lapointe, and Varambally) that investigated three classes of normal, localized tumor and metastatic tumor tissues. After filtering and merging these three data sets, we obtained 1,004 genes for downstream analysis.

Among these 1,004 genes, ANOVA-maxP detected 206 biomarkers, and min-MCC identified 120 biomarkers. We performed pathway analysis using Ingenuity Pathway Analysis (IPA) software to investigate which biological functions are related (enriched) in the biomarker lists. The top p-values of enrichment analysis without multiple comparison correction are presented in Table 10. In these 56 cancer-related biological functions, p-values of min-MCC are always much more significant than those from ANOVA-maxP.

In this application, ANOVA-maxP clearly adds in biomarkers of discordant patterns that are likely to be false positives. The result of min-MCC provides a better biomarker list to investigate the biological process from normal to localized tumors and to further malignant metastatic tumors.

Table 10: IPA biological functions results for prostate cancers.

Category	min-MCC*	ANOVA-maxP*	N1	N2	N3
Cardiovascular Disease	0.000885	0.029	5	9	1
Cell Death	0.000885	0.0419	4	2	0
Dermatological Diseases	0.000885	0.0419	10	2	0
Cellular Development	0.00167	0.0419	6	2	1
Skeletal and Muscular System	0.00167	0.0419	3	2	0
Tissue Development	0.00167	0.0419	17	6	0
Nervous System Development	0.00167	>0.05	12	0	0
Respiratory Disease	0.00167	>0.05	5	0	0
Cellular Movement	0.00266	>0.05	10	0	0

*: enrichment p-value.

N1: number of biomarkers for min-MCC;

N2: number of biomarkers for ANOVA-maxP;

N3: number of intersection.

4.0 CONCLUSIONS AND FUTURE WORK

4.1 CONCLUSIONS

Meta-analysis or information integration of multiple genomic studies helps to increase statistical power of biomarker detection. The evaluation of performance and choice of the best method depend on the ultimate biological goal. Many meta-analysis methods have been proposed for combining two-class genomic studies. In this dissertation, we proposed and compared ANOVA-maxP, min-MCC and OW-min-MCC methods for combining multi-class microarray studies.

To our knowledge, this is the first systematic investigation in this area. These three methods provide complementary utilities to identify both concordant and discordant biomarkers across all studies (ANOVA-maxP), concordant biomarkers across all studies (min-MCC) and concordant biomarkers across partial studies (OW-min-MCC).

Our simulation result demonstrates the complementary advantages of the three methods for different types of biomarkers. The three applications to real data further elucidate the advantage and timing of using each of the methods. For the mouse metabolism data, study-specific biomarkers are expected and are of biological interests.

In the analysis of mouse-metabolism data, ANOVA-maxP can help to identify tissue-specific biomarkers under different tissue physiology. Applications of min-MCC contribute another aspect of biological interest if researchers would like to find biomarkers with consistent expressions across these four tissues. However, if researchers are just interested in a

part of these four tissues, like, Brown and Liver, then OW-min-MCC will be a solution to fit this purpose. So, for different biological interests, investigators can choose different methods.

In the analysis of mouse-trauma data, analyses of ANOVA-maxP and min-MCC produced very similar results, confirming reproducibility of the two commercial array platforms. For the prostate cancer data, pathway analysis on the detected biomarkers concluded that min-MCC generates results of better biological contents.

In conclusion, we showed when study variations are expected and both study-invariant (concordant inter-class pattern) and study-specific (discordant inter-class pattern) genes are of biological interests, ANOVA-maxP serves for this purpose. When detecting highly reliable biomarkers (i.e. only study-invariant or con-cordant inter-class pattern genes) is the goal, min-MCC and OW-min-MCC are better choices and provide better statistical power.

In the simulation result, the FDR control of ANOVA-maxP and min-MCC (that focus on HS_A) is anti-conservative. This is because the permutation test implicitly assumes no differential expression in all studies for null genes while, in our simulation and in real applications, biomarkers differentially expressed in partial studies exist.

In other words, the null hypothesis in the simulation and real applications is a composite null hypothesis but the permutation test assumes a simple null hypothesis. More research is needed to adequately model the mixture of composite null hypothesis for an accurate FDR control.

4.2 FUTURE WORK

There are some works we can put effort on in the future.

1. Missing values pose another issue in merging multiple studies. For instance, if we just select the genes which exist among all the studies, a small set of genes could be gained and a lot of information lost which in turn has an adverse impact on ensuing analyses. On the other hand, if we target a certain percentage of existence of the genes, say 80% which means 20% of the values are allowed to be missing, this can give us more information for the meta-analysis. We are searching for a reasonable percentage of missing values and are investigating the influence on the analysis.

2. In addition to the issue of combining multiple studies, there are a few possible extensions and future directions. Currently we consider all studies to have identical K classes. Both ANOVA-maxP and min-MCC can be extended to studies containing mismatched classes of different sizes. For example, for min-MCC, the pairwise MCC can be defined using only the overlapping classes across a pair of studies. If we have three studies as following:

Study	1				2					3		
class	A	B	C	D	A	B	C	D	E	B	C	D

Let MCC_1 be the pairwise MCC of Study 1 and Study 2 for the overlapping classes (A, B, C, D), MCC_2 be the pairwise MCC of Study 1 and Study 3 for the overlapping classes (B, C, D) and MCC_3 be the pairwise MCC of Study 2 and Study 3 for the overlapping classes (B, C, D). Then, the min-MCC for these three studies is $min_{1\leq i\leq 3}MCC_i$.

3. These three 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 the class, K, is large. A quick modification may be to use the rth ranked statistic instead.

4. The biological meanings and functions of these detected biomarkers are of interest. Gene Ontology (GO) enrichment analyses can be conducted to investigate the biological pathways and knowledge of these detected biomarkers. Gene cluster analysis is another potentially effective strategy to further investigate the deteced biomarkers.

5. Developing a hierarchical meta-analysis method, as 4.1 shows, to investigate the genes' behavior not only across different studies, but also across different types of cancers is also of interest.

$$\overbrace{\overbrace{S_1 \dots S_{n_1}}^{Breast}\ \overbrace{S_1 \dots S_{n_2}}^{Leukemia}\ \overbrace{S_1 \dots S_{n_3}}^{Lung}\ \overbrace{S_1 \dots S_{n_4}}^{Prostate}}^{Cancer\ Types} \tag{4.1}$$

APPENDIX

A.1 CANCER STUDIES

Cancer Type	Author	Year	Array Type
Prostate cancer	Dhanasekaran	2001	cDNA
	Luo	2001	cDNA
	Magee	2001	Affy HG6800
	Welsh	2001	Affy HU95A
	Ernst	2002	Affy HU95A
	Luo	2002	Affy HU95A
	Singh	2002	Affy HU95Av2
	Henshall	2003	Affy Eos Hu03
	Chen	2004	Affy HU95Av2
	Lapointe	2004	cDNA
	Stuart	2004	Affy HU95Av2
	Yu	2004	Affy HU95A
	Best	2005	Affy HU133A
	Varambally	2005	Affy HU133
	Nanni	2006	Affy Hu133A
	Tomlins	2007	cDNA
Lung cancer	Bhattacharjee	2001 (Harvard)	Affy HU95A
	Garber	2001 (Stanford)	cDNA
	Beer	2002 (Michigan)	Affy HG6800
	Gordon	2002	Affy HU95A
	Wigle	2002	cDNA
	Michael	2004	cDNA
	Magda	2005	Affy HU133A
	Washi	2005	Affy HU133A
	Gemma	2006	Affy HU133A

Cancer Type	Author	Year	Array Type
Breast cancer	Perou	1999	cDNA
	Perou	2000	cDNA
	Sorlie	2001	cDNA
	vant Veer	2002	cDNA
	Chang	2003	Affy HU95Av2
	Sorlie	2003	cDNA
	Sotiriou	2003	cDNA
	Acevedo	2004	Affy HU133A
	Frasor	2004	Affy HU95A
	Ma	2004	Agilent
	Troester	2004	cDNA
	Farmer	2005	Affy HU133A
	Moggs	2005	Affy HU133A
	Rouzie	2005	Affy HU133A
	Yang	2005	Affy HU133A
	Bild	2006	Affy HU95A
Colon cancer	Alon	1999	Affy HU6800
	Agrawal	2002	Affy HU95A
	Fleet	2003	Affy HU95A
	Anderle	2004	Affy HU133A
	Bertucci	2004	Affy
	Bertucci	2004	Affy
	Bandres	2005	Affy
	Barrier	2005	Affy HU133A
	Dommels	2005	Affy
	West	2005	Affy HU133Plus2
	Koinuma	2006	Affy HU133A
	Whitney	2006	Affy HU133

Cancer Type	Author	Year	Array Type
Leukemia cancer	Cheok	2003	Affy HU95A
	Yagi	2003	Affy HU95A
	Addya	2004	Affy HU133A
	Haslinger	2004	Affy HU95A
	Scandura	2004	Affy HU133A
	Stegmaier	2004	Affy HU133A
	Crossman	2005	Affy HU95A
	Depitta	2005	cDNA
	Dik	2005	Affy HU133A
	Falt	2005	Affy HU95
	Gutierrez	2005	Affy HU133A
	Neumann	2005	Affy HU-Focus
	Raetz	2006	Affy HU133A

A.2 INTERQUARTILE RANGE, IQR

In descriptive statistics, the interquartile range (IQR), also called the midspread, middle fifty and middle of the number of observations, is a measure of statistical dispersion, being equal to the difference between the third and first quartiles.

Unlike the (total) range, the interquartile range is a robust statistic, having a breakdown point of 25%, and is thus often preferred to the total range. The IQR is used to build box plots, simple graphical representations of a probability distribution.

For a symmetric distribution (so the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD). The median is the corresponding measure of central tendency.

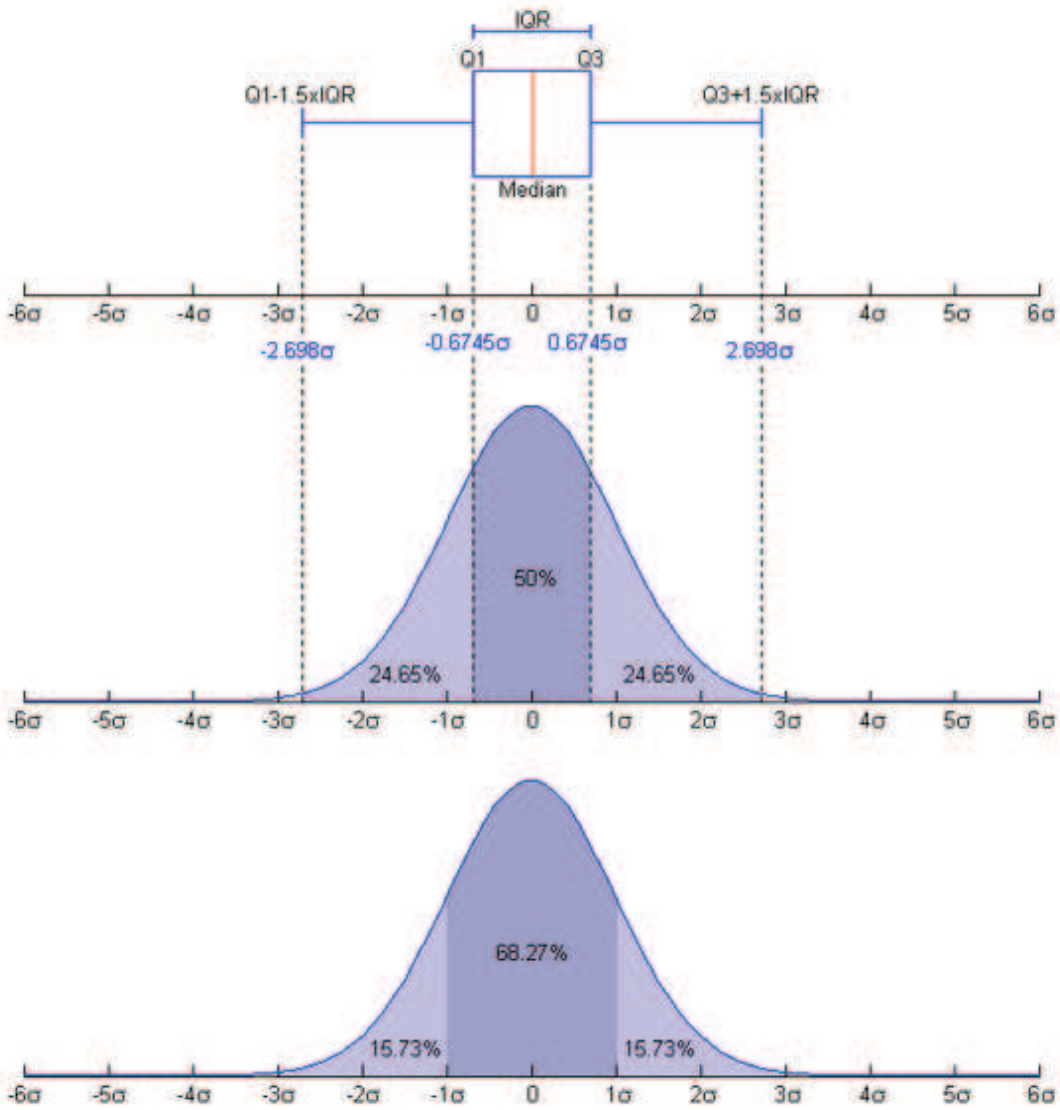

Figure 10: Interquartile range (IQR).

Boxplot (with an interquartile range) and a probability density function (pdf) of a Normal $N(0,\sigma^2)$ Population

BIBLIOGRAPHY

[1] Best CJ, Gillespie JW, Yi Y, Chandramouli GV, Perlmutter MA, Gathright Y, Erickson HS, Georgevich L, Tangrea MA, Duray PH, Gonzlez S, Velasco A, Linehan WM, Matusik RJ, Price DK, Figg WD, Emmert-Buck MR, Chuaqui RF. Molecular alterations in primary prostate cancer after androgen ablation therapy. *Clin Cancer Res.* 2005; 11(19 Pt 1):6823-34.

[2] Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 2002: 8, 816 - 824

[3] Bhattacharjee, A. and Richards, W. G. and Staunton, J. and Li, C. and Monti, S. and Vasa, P. and Ladd, C. and Beheshti, J. and Bueno, R. and Gillette, M. and Loda, M.and Weber, G. and et al. (2001), Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad SciUSA*, 98, 13790-13795.

[4] Bertucci F, Salas S, Eysteries S, Nasser V, Finetti P, Ginestier C, Charafe-Jauffret E, Loriod B, Bachelart L, Montfort J, Victorero G, Viret F, Ollendorff V, Fert V, Giovaninni M, Delpero JR, Nguyen C, Viens P, Monges G, Birnbaum D, Houlgatte R.Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene.* 2004, 23(7):1377-91

[5] Bild, A., G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, Jr., J. R. Marks, M. West, H. K. Dressman, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature.* 2006 Jan 19;439(7074):353-7

[6] Birnbaum, A. (1954), Combining independent tests of significance. *Journal of the American Statistical Association*, 49, 559-574.

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

[8] Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H.D., Hersch, S.M.,Hogarth, P., Bouzou, B., Jensen, R.V., and Krainc, D. (2005), Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proceedings of the National Academy of Sciences*, 102, 11023-110

[9] Breitling, R. et al. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*,2004, 573, 83.92.

[10] Breitling, R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *J. Bioinf. Comp. Biol.*, 3, 1171.1189.

[11] Cardoso J., Boer J., H. Morreau H. , Fodde R. (2007), Expression and genomic profiling of colorectal cancer. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1775 ,103137.

[12] Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, 2003 Aug 2;362(9381):362-9.

[13] Chen CD, Welsbie DS, Tran C, Baek SH, Chen R, Vessella R, Rosenfeld MG, Sawyers CL. Molecular determinants of resistance to antiandrogen therapy. Nat Med. 2004;10(1):33-9. *Epub* 2003 Dec 21.

[14] Cheok MH, Yang W, Pui CH, Downing JR, Cheng C, Naeve CW, Relling MV, Evans WE.Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet*. 2003 May;34(1):85-90.

[15] Choi, J.K. et al. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19, 84.90.

[16] Choi, H. and Shen, R. and Chinnaiyan, A. M. and Ghosh, D. (2007), A Latent Variable Approach for Meta-analysis of gene expression data from Multiple microarray experiments. *BMC Bioinformatics*, 8, 364-383.

[17] Crossman LC, Mori M, Hsieh YC, Lange T, Paschka P, Harrington CA, Krohn K, Niederwieser DW, Hehlmann R, Hochhaus A, Druker BJ, Deininger MW.In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures. *Haematologica*. 2005 Apr;90(4):459-64.

[18] Conlon, E.M., Song, J.J., Liu, J.S. (2006), Bayesian models for pooling microarray studies with multiple sources of replications, *BMC Bioinformatics*, 7, 247-259.

[19] Conlon, E.M., Song, J.J., Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8, 80.

[20] Cousins, R.D. (2007), Annotated Bibliography of Some Papers on Combining Significances or p-values. arXiv:0705.2209v1.

[21] Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005), Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33, e175.10.1093/nar/gni17

[22] De Pitta C, Tombolan L, Campo Dell'Orto M, Accordi B, te Kronnie G, Romualdi C, Vitulo N, Basso G, Lanfranchi G.A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. *Haematologica*. 2005 Jul;90(7):890-8.

[23] Dhanasekaran S.M., Barrette T.R., Ghosh D., Shah R., Varambally S., Kurachi K.,Pienta K.J., Rubin M.A., Chinnaiyan A.M. (2001), Delineation of prognostic biomarkers in prostate cancer, *Nature*, 412, 822-82

[24] Dik WA, Brahim W, Braun C, Asnafi V, Dastugue N, Bernard OA, van Dongen JJ, Langerak AW, Macintyre EA, Delabesse E.CALM-AF10+ T-ALL expression profiles are characterized by overexpression of HOXA and BMI1 oncogenes. *Leukemia*. 2005 Nov;19(11):1948-57

[25] Efron, B. et al. (2001) Empirical Bayes analysis of a microarray experiment.*J. Am. Stat. Assoc.*, 96, 1151.1160.

[26] Edgar R, Domrachev M, Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*; 30(1), 207-10.

[27] Ernst T, Hergenhahn M, Kenzelmann M, Cohen CD, Bonrouhi M, Weninger A, Klren R, Grene EF, Wiesel M, Gdemann C, Kster J, Schott W, Staehler G, Kretzler M, Hollstein M and Grene H. Decrease and Gain of Gene Expression Are Equally Discriminatory Markers for Prostate Carcinoma. *American Journal of Pathology.* 2002. 160(6): 2169-2180

[28] Fleet JC, Wang L, Vitek O, Craig BA, Edenberg HJ.Gene expression profiling of Caco-2 BBe cells suggests a role for specific signaling pathways during intestinal differentiation. *Physiol Genomics*. 2003, 13(1):57-68

[29] Fisher, R.A. (1925) *Statistical Methods for Research Worker*. Oliver and Boyd, Edinburg and London.

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

[31] Garber, M. E. and Troyanskaya, O. G. and Schluens, K. and Petersen, S. and Thaesler, Z. and Pacyna Gengelbach, M. and van de Rijn, M. and Rosen, G. D. and Perou, C. M.

andWhyte, R.I. and Altman, R.B. and Brown, P.O. and et al. (2001), *Diversity of gene expression in adenocarcinoma of the lung*. Proc Natl Acad Sci USA, 98, 13784-13789.

[32] George, E.O. (1977), Combining independent one-sided and two-sided statistical tests-Some theory and applications. Unpublished doctoral dissertation, University of Rochese

[33] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R.Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*. 2002, 62(17):4963-7

[34] Good, I.J. (1955), On the weighted combination of significance tests. *Journal of the Royal Statistical Society*, Ser. B, 17, 264-265

[35] Ghosh,D. et al. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics*, 3, 180.188.

[36] Gutierrez NC, Lopez-Perez R, Hernandez JM, Isidro I, Gonzalez B, Delgado M, Ferminan E, Garcia JL, Vazquez L, Gonzalez M, San Miguel JF.Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. *Leukemia*. 2005 Mar;19(3):402-9.

[37] Haslinger C, Schweifer N, Stilgenbauer S, Dohner H, Lichter P, Kraut N, Stratowa C, Abseher R. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol*. 2004 Oct 1;22(19):3937-49.

[38] Henshall SM, Afar DE, Hiller J, Horvath LG, Quinn DI, Rasiah KK, Gish K, Willhite D, Kench JG, Gardiner-Garden M, Stricker PD, Scher HI, Grygiel JJ, Agus DB, Mack DH, Sutherland RL. Survival analysis of genome-wide gene expression profiles of prostate cancers identifies new prognostic targets of disease relapse. *Cancer Res*. 2003, 63(14):4196-203.

[39] Hong,F. et al. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22, 2825.27.

[40] Hu, P., Greenwood, C. MT. and Beyene, J. (2005), Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, 6, 128-138.

[41] Jones MH, Virtanen C, Honjoh D, Miyoshi T, Satoh Y, Okumura S, Nakagawa K, Nomura H, Ishikawa Y.Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*. 2004 Mar 6;363(9411):775-81

[42] Koinuma K, Yamashita Y, Liu W, Hatanaka H, Kurashina K, Wada T, Takada S, Kaneda R, Choi YL, Fujiwara SI, Miyakura Y, Nagai H, Mano H.Epigenetic silenc-

ing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene*. 2006, 25(1):139-46.

[43] Lancaster, H. (1961), The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3, 20-3

[44] Lapointe J, Li C, Higgins JP, van de Rijn M et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA*. 2004 Jan 20;101(3):811-6. PMID: 14711987

[45] Li J and Tseng GC. (2008) Optimally weighted statistic for combining multiple genomic studies. (submitted)

[46] Littell, R.C., Folks, J. L. (1971), Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66, 802-8

[47] Littell, R.C., Folks, J. L. (1973), Asymptotic optimality of Fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, 68, 193-1

[48] Loughin, T. M. (2004), A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics Data Analysis*, 47, 467-48

[49] Luo J., Duggan D.J., Chen Y., Sauvageot J., Ewing C.M., Bittner M.L., Trent J.M.,Isaacs W.B. (2001), Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Research*, 61, 4683-46

[50] Magee JA, Araki T, Patil S, Ehrig T, True L, Humphrey PA, Catalona WJ, Watson MA, and Milbrandt J. Expression Profiling Reveals Hepsin Overexpression in Prostate Cancer. *Cancer Res* 2001 61: 5692-5696

[51] Moreau, Y., Aerts, S., De Moor, B., De Strooper, B. and Dabrowski, M. (2003), Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, 19, 570-57

[52] Nanni S, Priolo C, Grasselli A, D'Eletto M, Merola R, Moretti F, Gallucci M, De Carli P, Sentinelli S, Cianciulli AM, Mottolese M, Carlini P, Arcelli D, Helmer-Citterich M, Gaetano C, Loda M, Pontecorvi A, Bacchetti S, Sacchi A, Farsetti A. Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer. *Mol Cancer Res*. 2006, 4(2):79-92

[53] Newton, M.A. et al. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5, 155.176.

[54] Neumann F, Teutsch N, Kliszewski S, Bork S, Steidl U, Brors B, Schimkus N, Roes N, Germing U, Hildebrandt B, Royer-Pokora B, Eils R, Gattermann N, Haas R, Kronenwett R.Gene expression profiling of Philadelphia chromosome (Ph)-negative CD34+

hematopoietic stem and progenitor cells of patients with Ph-positive CML in major molecular remission during therapy with imatinib. *Leukemia*. 2005 Mar;19(3):458-60.

[55] Olkin I and Saner H. (2001) Approximations for trimmed Fisher procedures in research synthesis. *Statistical Methods in Medical Research*, 10:267-276.

[56] Owen, A.B. (2007), Pearson's test in a large scale multiple meta-analysis. Technical report. Stanford University.

[57] Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, et al. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2005;33(Database issue):D553-5.

[58] Pearson, E.S. (1938), The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30, 134-148

[59] Pirooznia M., Nagarajan V., Deng Y., (2007), Gene Venn - A web application for comparing gene lists using venn diagram. *Binformation*, 1(10), 420-42

[60] Rhodes,D.R. et al. (2002) Meta-analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, 62, 4427-4433

[61] Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Keith Anderson3, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN and Pusztai L. Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy. *Clinical Cancer Research*, 2005; 11, 5678-5685.

[62] Roy, S. N. (1953), On a Heuristic Method of TestConstruction and its Use in Multivariate Analysis. *The Annals of Mathematical Statistics*, 24, 220-238.

[63] Segal, E., Friedman, N., Koller, D., Regev, A. (2004), A module map showing conditional activity of expression modules in cancer. *Nature Genetics* 36, 1090-1098

[64] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, et al. The Stanford Microarray Database.*Nucleic Acids Res* 2001;29(1):152-5.

[65] Shen, R., Ghosh, D. and Chinnaiyan, A.M. (2004), Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*,5, 94-10

[66] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. *CANCER CELL*, 2002, 1: 203-209

[67] Smyth, G. K. (2004), Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3.3.

[68] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lnning P, Brresen-Dale AL.Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001 Sep 11;98(19):10869-74

[69] Stevens, J. and Doerge, R.W. (2005). Combining Affymetrix microarray results. *BioMed Central Bioinformatics*. 6:57.

[70] Stevens, J. and Doerge, 2005. Meta-Analysis combines Affymetrix microarray results across laboratories. *Comparative and Functional Genomics*. 6:116-122.

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

[72] Storey, J.D. (2002), A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, Ser. B, 64, 479-49

[73] Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, Masys D, Arden K, Goodison S, McClelland M, Wang Y, Sawyers A, Kalcheva I, Tarin D and Mercola D. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2004, 101(2): 615-620

[74] Tippett, LHC. The Methods in Statistics, First edition. Williams and Norgate, Ltd.; 1931.

[75] Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet*. 2007, 39(1):41-51

[76] Tseng, George C. Min-Kyu Oh, Lars Rohlin, James C. Liao, and Wing Hung Wong. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*. 29: 2549-2557.

[77] Tusher,V.G. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc.*Natl Acad*. Sci, 98, 5116-5121

[78] Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, Wei JT, Pienta KJ, Ghosh D, Rubin MA, Chinnaiyan AM. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*. 2005, 8(5):393-406.

[79] Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 2005 Dec 1;21(23):4205-8

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

[81] West JD, Marnett LJ. Alterations in gene expression induced by the lipid peroxidation product, 4-hydroxy-2-nonenal. *Chem Res Toxicol.* 2005, 18(11):1642-53.

[82] Whitney EM, Ghaleb AM, Chen X, Yang VW. Transcriptional profiling of the cell cycle checkpoint gene krppel-like factor 4 reveals a global inhibitory function in macromolecular biosynthesis. *Gene Expr*, 2006;13(2):85-96

[83] Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, Breitkreutz BJ, Jorgenson P, Tyers M, Shepherd FA, Tsao MS. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* 2002, 62(11):3005-8

[84] Wilkinson,B.(1951) A statistical consideration in psychological research. *Psychological Bulletin*, 48, 156-157.

[85] Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo JH. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of Clinical Oncology*, 2004, 22(14):2790-2799