TEST STATISTICS AND Q-VALUES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS

TEST STATISTICS AND Q-VALUES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS

By

Chang Ye, BSc

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2010) McMaster University
(Statistics) Hamilton, Ontario

TITLE: Test Statistics and q-values to identify differentially expressed genes in microarrays

AUTHOR: Chang Ye, B.Sc
(McMaster University, Canada)

SUPERVISOR: Professor Angelo Canty

NUMBER OF PAGES: xiii, 118

Acknowledgements

I am so grateful for all the wonderful people who have supported me throughout my academic career at McMaster. In particular, I would like to thank my supervisor Dr. Angelo Canty for giving an opportunity to work on this project, and his continuous support, valuable guidance and encouragement over the years. His enthusiasm, approachability, and patience were essential to the completion of this thesis.

I would like to acknowledge my thesis reviewers, Dr. Roman Viveros-Aguilera and Dr. Peter Macdonald for agreeing to be my thesis examiners and their valuable feedback. I especially thank two postdoctoral fellows who worked with my supervisor on the analysis of Microarray data, Dr. Amadou Sarr and Dr. Shaheena Bashir. Their passion to discuss any and all statistics related topics makes me think deeply about my research.

I am also grateful to the people from Dr. Jayne Danska's lab at the Hospital for Sick Children in Toronto for providing data. Special thanks also go to Dr. Tanya Prasolava and Dr. Evgueni Ivakine who produced the Affymetrix data sets used in the applications.

I wish to express deep gratitude to my husband and parents for their unconditional love and support. I would not have been able to succeed in my studies without them.

Contents

List of Tables

List of Figures

Abstract

A major goal of gene expression microarray studies is to identify differentially expressed genes. The q-value is widely used to control false positives among all genes called significant. It can be derived either by considering the false discovery rate (FDR) of all rejection regions containing a gene or by adjusting the usual p-value. Both methods use a modified t-statistic, referred to as d-statistic in this study. However there is no associated distribution theory for this statistic. We derive its distribution in the two-sample setting by numerical integration under normality assumption and taking the adjusting factor to be a constant. Because the distribution depends on the true population variance of each gene and this is unknown in practice, we propose various estimators to deal with this issue. We compare the three methods in terms of d-statistics and t-statistics respectively and assess their power to detect differentially expressed genes in simulated data sets and real microarray data. Methods based on the d-statistic perform much better as they can identify more significant genes with lower false discovery rate. When the distribution of gene expressions is close to normal distribution, the d distribution method works better, but when gene expressions are heavy-tailed, FDR and permutation methods are more powerful. Real data analysis indicates that the estimators proposed are unstable and need to be improved.

Chapter 1

Introduction

1.1 Introduction to the problem

In today's genomics, DNA microarray technology has emerged as a widely used platform for genomic studies. It has expanded the scale of biological research from studying single genes or proteins to studying all genes or proteins simultaneously. DNA microarrays powerfully provide a global view of changes in gene expression patterns in various disease conditions. In biomedical research, such an approach can determine biological behavior of both normal and diseased tissues, show insights into disease mechanisms and identify novel markers and candidates for diagnostic, prognostic and therapeutic intervention. Microarray data analysis raises numerous statistical issues such as multiple testing, experimental design, and discriminant analysis, etc.

This study describes statistical methods for the analysis of gene expression data from a study of Type 1 diabetes in mice. The goal of the DNA microarray experiments is to identify genes that are differentially expressed under two different strains of mice.

The biological question can be restated as a statistical problem in multiple hypothesis testing: the simultaneous tests for all genes to determine which ones are differentially expressed. There are two important statistical challenges associated with such microarray data analysis. One is that the number of biological samples is usually small compared to the huge number of genes tested. Since microarray experiments are still expensive, only relatively small sample sizes are possible. This limitation tends to result in less statistical power. The other challenge is that the large number of genes tested dramatically intensifies the multiple testing problem. Suppose there are m genes independently tested with level α, then the probability of at least one gene falsely called significant is $1-(1-\alpha)^m$. However, the number of genes in a microarray experiment is usually very large. This probability is increased sharply to be close to 1. Therefore (1) defining an appropriate error rate measure and (2) devising a powerful procedure to control this error rate have been paid tremendous attention in microarray studies (Dudoit *et al*, 2003).

The traditional measure in multiple testing is the family-wise error rate (FWER), which is the probability of at least one type I error occurring among all the hypotheses. Bonferroni corrections are the classical procedure most widely used to control it. It uses α/m as the threshold p-value for each test when m is the number of tests. It makes the FWER less than α. However, it is very stringent: If 10,000 genes are tested with α=0.05, the threshold p-value for each test is 5×10^{-6}. Such a stringent correction seriously increases the rate of false negatives and decreases the power to detect a significant test. Although there are other adjustments on the threshold from some authors, such as Holm (1979) and Hochberg (1988), they are still very conservative. Since the goal of a microarray experiment is to discover as many differentially expressed genes as possible, these multiple test corrections are not desirable. Benjamini and Hochberg (1995) introduced an error measure called the false discovery rate (FDR), which facilitates to know how many of the selected genes (i.e. those that have been called significant) could be false positives. Storey (2002) and Storey & Tibshirani (2003a) made a modification to Benjamini and Hochberg's FDR with a better estimate by considering the number of true nulls (non-differential expressions) among all the tests. This modification improves the FDR estimate and provides more power when a relatively large percentage of genes is selected.

In detecting the differentially expressed genes, some traditional parametric tests have been applied. The commonly used methods include various versions of the two-sample t-test. However, the strong normality assumption of the t-test can be violated in practice. The problem tends to be complicated by the fact that expressions may have non-identical and dependent distributions between genes. In this context, two nonparametric statistical methods are attractive: the Significance Analysis of Microarray (SAM) method of Tusher *et al.* (2001) and the empirical Bayes (EB) method of Efron *et al.* (2001). Both methods depend on constructing a test statistic similar to the t-test to estimate the null distribution. In addition, Smyth (2004) proposed the Linear Models for Microarray Analysis (LIMMA) method to address the issue that the variability of expressions differs between genes. In this study, we will give a brief description of the three methods with the aim to have a deeper understanding of the theoretical ideas behind them. We will discuss the performance in identification of significant expressions in terms of the two different test statistics, the regular t-statistic and the modified t-statistic established in SAM and EB. Because there is no associated distribution theory for the modified statistic, we will derive a computational method to find its cumulative distribution function.

The related performance will be demonstrated in our simulation studies and real microarray data analysis.

1.2 Mouse Model for Type 1 Diabetes

Diabetes is a complex disease that has been attracted more and more attention from health professionals. Public Health Agency of Canada 2005-2006 National Health Survey shows approximately 1.9 million Canadians had been diagnosed with diabetes. This represents about 1 in 17 Canadians – 5.5 % of all women and 6.2 % of all men. Moreover, 5 to 10% of people with diabetes have type 1 diabetes (T1D). T1D is an autoimmune disease and often develops in childhood or adolescence. It occurs when the cells of the pancreas are destroyed by the immune system and no longer produce insulin. Insulin is an important protein that the body needs to convert sugar, starches and other food into energy for daily life. Worth noting is that Canada has the third highest rate of T1D in the world and the incidence is rising. For patients, T1D greatly increases the chance of heart attack, stroke, blindness and limb amputation, as well as shortened life expectancy. As far as the nation is concerned, T1D cost the Canadian healthcare system $1.32 billion in 2002 and is estimated to rise to $1.6 billion by 2010. However, the exact cause of T1D is still not fully understood since it is affected not only by multiple genetic risk factors but also unknown environmental factors (Benkalhaa and Polychronakos, 2008). Therefore biologists and computational statisticians have devoted themselves to investigate complex genetic factors for T1D in a genome-wide study.

The data used in our study arise from the mouse model. Because mice breed quickly and share 99% genes with human, mouse models can provide a bridge to investigate human disease pathogenesis and to identify which genes control disease susceptibility. Biological researchers have found that allelic variation at many genetic regions (loci) contribute to susceptibility to T1D. There are at least 20 insulin-dependent diabetes (Idd) regions which have been identified in the diabetes-prone NOD mouse (Ivakine *et al*, 2005). In our study, we are interested in three regions among them: Idd4, Idd5 and Idd13. They display linkage to T1D, especially Idd4 exhibits sex-specific effect on T1D. We have two parental strains of mice Non-obese Diabetic (NOD) and Non-Obese Resistant (NOR). These two parental mouse strains are identical by descent in 88% of the genome. But the rate of NOD mice getting T1D is 82-85% much higher than NOR mice (only 3-5%) by the age of six months. One of the reasons is the difference in Idd4, Idd5 and Idd13 regions between the two strains. By mating NOD and NOR mice and selectively inbreeding multiple generations, two congenic strains NOD.NOR_Idd4 and NOR.NOD_Idd5/13 are derived. The NOD.NOR_Idd4 strain is identical to the parental NOD strain except for region Idd4 which inherits from the NOR mice. Similarly, the double congenic NOR.NOD_Idd5/13 strain is identical to the parental NOR strain except for two regions Idd5 and Idd13 which inherit from the NOD mice. In this research, we will study the two comparisons: NOR against NOR.NOD_Idd5/13 and NOD against NOD.NOR_Idd4.

1.3 Organization of the thesis

In the study we derive a new methodology and compare with the currently used methods to identify differentially expressed genes. Our analysis makes use of a simulation study and two real data sets from microarray experiments using Affymetrix GeneChip

MGU74Av2 for mice. The thesis is organized as follows: Chapter 2 gives genetics background and introduces microarray technology. We also review three widely used methods for finding differentially expressed genes such as SAM (Significant Analysis of Micorarrays), EB (Empirical Bayes) and LIMMA (Linear Models for Microarray Analysis). Chapter 3 describes three approaches to calculate the q-value for the selection of differentially expressed genes: methods based on the FDR, adjusted permutation p-values and adjusted p-values from test statistic null distribution. For the latter approach, we derive the distribution of the d-statistics based on the unknown parameter true population variance of each gene, and then propose some estimators to deal with it. In Chapter 4, we conduct a simulation study generating data from various symmetric distributions, either with a common variance across genes or allowing the variance to differ. From this simulation we assess the power of all of the methods to detect genes known to be differentially expressed. Two microarray data sets which are used to evaluate different methods are described, and the results of the comparison study are presented in Chapter 5. Finally, Chapter 6 summarizes our findings and discusses issues for future work.

Chapter 2

Genetics Background and Microarray Analysis

2.1 Basic genetics

Microarrays were designed in response to the need to analyze gene expression data. Consequently, a basic understanding of genetics becomes useful in understanding how the data are collected and how they should be handled. This section will introduce the main concepts of genetics and how they relate to microarrays.

A cell is the smallest unit of life. Almost every cell contains a complete copy of its genetic material in the form of DNA. The genetic information stored in DNA can be copied as mRNA to form functional proteins. Proteins are the most important determinants of the properties of the cells and organisms. Take T1D disease in our study for example. Insulin is a protein in the human body that plays a major role in decreasing the levels of sugar in the blood. Lack of it makes humans suspectable to

T1D and other complications. The production of proteins are controlled by genes which are encoded in DNA. The DNA is a double-stranded compound molecule composed of nucleotides. Each nucleotide comprises a phosphate, a sugar and a base. There are four different types of base: guanine (G), cytosine (C), adenine (A) and thymine (T). The bases on the two strands are paired according to the Watson-Crick pairing rule: G in one strand binds only to C in the other, and A binds only to T. Therefore, the two strands are each other's complement so that each strand stores the same sequence information. This biological fact is the basic property by which microarrays work. A gene is a segment of DNA that specifies a functional mRNA (Lee, 2004). The mRNA is a single-stranded molecule made up of G, C, A and U bases. The U base stands for uracil and replaces T when DNA is transcribed into mRNA. The mRNA delivers DNA's genetic message to the cell where proteins are made. Biologically speaking, the mRNA is similar to one single strand of DNA. The mRNA are targets applied in microarrays to find their complementary siblings.

The process by which genetic information from the DNA template of a gene is used in the synthesis of a functional gene product, mRNA and eventually protein, is called gene expression. For a gene, expression level is the amount of mRNA that is used to synthesize protein. The process from a gene to protein involves two essential stages, known as transcription and translation. During transcription, mRNA is produced using one DNA strand as the template. An mRNA molecule includes nucleotide sequences that correspond to amino acid sequences of its protein. During translation, three nucleotides codons in the mRNA sequence are read, and corresponding amino acids are assembled into protein with the help of corresponding tRNAs. Microarray technology utilizes these properties of binding of a single strand of DNA to mRNA to measure expression levels for many genes.

2.2 Microarray technology and Affymetrix GeneChip

Most cells in human body contain identical genes, but not all the genes are active in each cell. Some genes are turned on, or expressed, when needed. Studying which genes are active in different cell types can help biologists know how these cells function normally and how they are affected when related genes do not perform properly. When a gene is activated, cell begins to copy certain amount of mRNA of that gene. Due to Watson-Crick complementarity, the mRNA produced by the cell can bind to the original segment of one DNA strand from which the mRNA was transcribed. Therefore the expression for a gene can be obtained by measuring the amount of mRNA produced by the gene. In the past, biologists were limited to study a few genes per experiment. With the rapid development of DNA microarray technology, however, they can more easily examine tens of thousands of genes simultaneously and compare the expression of the same genes in different samples at the same time (McLachlan *et al*, 2004).

Typically, a microarray is a glass slide onto which known DNA sequences are immobilized in an orderly manner at specific spots. A microarray contains tens of thousands of spots and each spot contains a few million copies of identical DNA sequences. The DNA in a spot may either be genomic DNA or short stretch of oligonucleotide strands that correspond to parts of a gene. The spots are printed onto the glass slide by a robot or are synthesised by the process of photolithography. To explain the application of microarrays to measure gene expression levels, Figure 2.1 gives a general picture of the experimental steps involved. (1) mRNA Extraction. The mRNAs are extracted from sample because it is an indicator of which genes are being used in a specific cell. (2) Labeling and hybridization. These mRNA molecules are reverse transcribed into cDNA by using an enzyme and labeled with a fluorescent dye. In this step, any labeled cDNA sequence is hybridized to specific spots containing its complementary sequence according to base pairing property. (3) Scanning and detecting. After washing away all of the mRNA molecules which are not hybridized, the microarray slide is scanned by an optical detector device to get a fluorescent image. Researchers can look at the microarray image and see which RNA remains stuck to spots. Since we know which gene each spot represents and the mRNA only sticks to the gene that encoded it, we can determine which genes are expressed in samples. The amount of fluorescence corresponds to the amount of mRNA. If a gene is very active and producing many mRNA molecules, the corresponding spot will be very bright. In contrast, if a gene is less active with less mRNA molecules, the spot will be darker. A black fluorescent spot indicates none of mRNA molecules are produced and that gene is inactive.

Several technological platforms have been developed, differing in array design, manufacturing procedure (standardised printing or randomised microbeads), experimental design (absolute or relative expression level) and target oligonucleotide sequence length (Wilder *et al*, 2009). The most advanced and predominantly used platforms are Affymetrix GeneChip (Santa Clara, California) and Illumina Sentrix BeadChip (San Diego, California) (Wilder *et al*, 2009). In our study, our microarray data were from experiments using MGU74Av2 chip produced by Affymetrix. MGU74Av2 chip is a murine genome array chip containing 12,488 probe sets (Takahashi *et al*, 2005). Therefore, we will introduce here particularly Affymetrix array and its principle in measuring gene expression. GeneChip is a prefabricated oligonucleotide chip. Its probes are synthesized

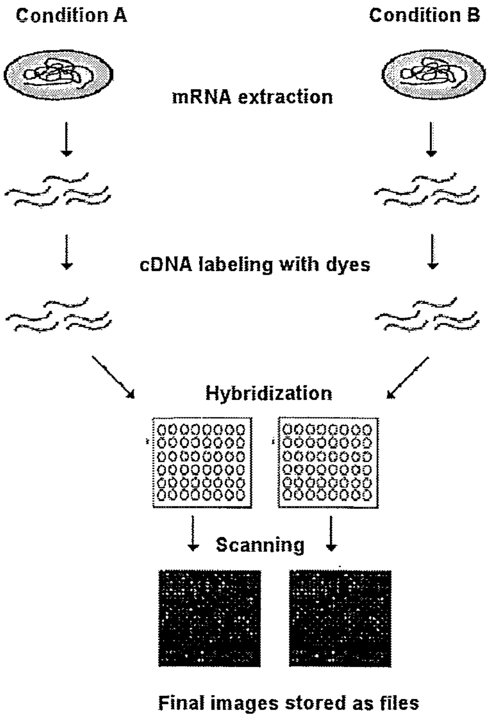


Figure 2.1: *An microarray experiment flowchart. Graphics from http://irfgc.irri.org. Image Courtesy: Dr. Madan Babu Mohan.*

in situ using a photolithographic method with marks directly on the surface of the chip. This method can produce many high-density arrays of the same design with consistent quality. The probes are 25 nucleotides long and organized as perfect match (PM) versus mismatch (MM) pairs. A probe set usually consists of 16-20 probe pairs. The PM probes are made perfectly complementary to the mRNA of target gene, while the MM probes are identical to the PM probes except for the central position. The design of PM-MM contrast in each probe set is intended to subtract out non-specific binding to the probes. After arrays are scanned and images generated, we can obtain an intensity value vector of two readings for each probe, one for PM and the other for MM. This intensity value represents how much hybridization occurred. In the end, all data about intensities and physical locations of probe sets are stored in .CEL files. To make biological sense of these .CEL files, the .CDF files are created to store the information about mapping probes to probe sets.

2.3 Robust Multi-array Average (RMA)

In the analysis of microarrays, our ultimate goal is to compare the expression values of each probe set from different chips. The measured intensity values in each chip are usually not directly used for comparison because some non-biological variation may have many different effects on them. The variation could be caused by optical noise, non-specific hybridization, probe-specific effects and measurement error, during all experimental steps (mRNA preparation, labeling, hybridization and scanning) (Irizarry *et al*, 2003b). Therefore it is necessary to minimize non-biological variations using a process called normalization so that biological differences can be more easily distinguished.

In our project, we normalize raw data stored in .CEL files using Robust Multi-array Average (RMA) method by rma function in **R** package **Affy** (Irizarry *et al*, 2003a; Gautier *et al*, 2004).

There are typically three steps to obtain an expression measure from raw data. (1) Background correction. Many expression measures are based on PM−MM (Affymetrix's AvDiff) or log(PM/MM) (Affymetrix's Average Log Ratio) with the assumption that the MM value represents the background. However, many MM values contain signals and in many cases the MM value is larger than the corresponding PM value (Irizarry *et al*, 2003b). So the RMA method completely ignores the MM values and only uses PM values. This method converts original PM probe intensity into an exponential signal and normal noise by maximum likelihood deconvolution (Irizarry *et al*, 2003a). After background correction, each PM probe intensity is transformed in log base 2 scale. (2) Normalization. These background corrected and log transformed PM intensities are normalized by quantile normalization method proposed by Bolstad *et al*(2003). This method makes the empirical distribution of these PM intensities the same for arrays $i = 1, 2, ..., I$. The normalization maps probe level data from all arrays so that an I-dimensional quantile-quantile plot follows the I-dimensional identity line (Bolstad *et al*, 2003). (3) Estimating expression. For a particular probe set, the normalized PM intensities follow a linear model,

$$Y_{ij} = \mu_i + \alpha_j + \epsilon_{ij}, \;\; i = 1, ..., I, \;\; j = 1, ..., J,$$

where I and J are number of arrays and probes within the probe set respectively. For the probe set, Y_{ij} denotes the normalized PM intensity for the ith array and the jth probe within the probe set, μ_i denotes the log-scale expression for ith array, α_j denotes probe effect for the jth probe, and ϵ_{ij} denotes random error with mean 0. The sum of α_j for all probe are assumed 0. To protect against outlier probes, a robust procedure such as median polish, is used to estimate parameters (Irizarry *et al*, 2003b). The estimated μ_i is referred to as the log scale expression corresponding to the probe set for the ith array.

2.4 Significance Analysis of Microarray (SAM)

After microarray raw data are normalized, the next task is data analysis to identify differentially expressed probe sets, which comes to the topic of multiple hypothesis testing. For each probe set, the null hypothesis that there is no difference in mean expression under two groups is tested against the alternative hypothesis that there is a difference in mean expressions. If there is enough evidence to show a difference, we reject the null hypothesis. For m probe sets we have m pairs of hypotheses:

$$\begin{cases} H_{j0} : \text{probe set } j \text{ is not differentially expressed} \\ H_{ja} : \text{probe set } j \text{ is differentially expressed,} \end{cases}$$

where $j = 1, 2, ..., m$.

Normally there are tens of thousands of probe sets ($m > 10000$) tested simultaneously for a microarray experiment. In this context, the situation becomes much more complicated. Multiple hypotheses testing concerns how we measure the probability that probe sets are falsely rejected (i.e. false positive) and the probability that probe sets are correctly rejected (i.e. power). A suitable overall error measure is required so that we can identify many differentially expressed probe sets without too many false positives.

Significance Analysis of Microarray (SAM) is a commonly used method for detecting significant features in DNA microarrays, first proposed by Tusher *et al.* (2001). SAM calculates a statistic d_j for each probe set on the basis of change in gene expression relative to the standard deviation. Probe sets with d_j value in a rejection region are declared significant. Asymmetrical rejection regions are chosen in SAM since nobody knows in advance how many differential expressions are in the positive and negative direction respectively. Particularly SAM includes some recently developed methodologies for estimating FDR and q-values.

Traditionally, a test statistic for assessing differential gene expression is the standard t-statistic:

$$t_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j}, j = 1, 2, ..., m \tag{2.1}$$

where s_j is the pooled standard error for probe set j, and $\bar{x}_{j2} - \bar{x}_{j1}$ is the difference of the average gene expression for probe set j under two groups. Let n_1 and n_2 be the number of arrays in group 1 and 2 respectively. The pooled standard error s_j is defined as:

$$s_j = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1 - 1)s_{j1}^2 + (n_2 - 1)s_{j2}^2}{n_1 + n_2 - 2}}, \tag{2.2}$$

where s_{j1}^2 and s_{j2}^2 are the sample variance in the two groups. However, the t-statistic is formed using only information from the probe set itself and might obtain large value when the difference in gene expression is near zero with very small s_j. Tusher *et al.* (2001) add a positive constant s_0 to increase the value of the denominator of t_j such that the variance of t_j is independent of the gene expressions. The modified t-statistic is defined as follows:

$$d_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s_0}, j = 1, 2, ..., m. \tag{2.3}$$

The extra term s_0 is the function of s_j, taken as a percentile of $\{s_1, ..., s_m\}$ which minimizes the coefficient of variation of d_j. Therefore, the modified t-statistic can not only borrow strength across the probe sets but also downweight the effect of probe sets with very low variances as compared to all genes. Chu *et al.* (2005) provide the procedure for computing s_0:

1. Let s^α be the percentile of s_j. The d_j^α is defined as

$$d_j^\alpha = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s^\alpha}.$$

2. Compute the percentiles of s_j, denoted by $\min\{s_j\} = q_0 < q_1 < q_2 < ... < q_{100} = \max\{s_j\}$.

3. For each $\alpha \in (0, 0.05, 0.10, ..., 1.0)$, compute v_i^α by

$$v_i = \text{mad}(d_j^\alpha | s_j \in [q_{i-1}, q_i)), i = 1, 2, ..., 99,$$

$$v_{100} = \text{mad}(d_j^\alpha | s_j \in [q_{99}, q_{100}]).$$

where mad is the median absolute deviation from the median, divided by 0.64. Then compute the coefficient of variation of the v_i by

$$cv(\alpha) = \frac{\text{sd}(v_i^\alpha)}{\text{mean}(v_i^\alpha)}$$

4. Choose $\hat{\alpha}$ as value which minimizes $cv(\alpha)$ and compute $s_0 = s^{\hat{\alpha}}$.

The algorithms for identifying significant probe sets are given as follows (Storey & Tibshirani, 2003a):

1. Compute the ordered statistics: $d_{(1)} \leq d_{(2)} ... \leq d_{(m)}$.

2. Make B sets of permutations of the group labels. For each permutation b, compute d_j^b and get corresponding order statistics $d_{(1)}^b \leq d_{(2)}^b ... \leq d_{(m)}^b$. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(j)} = (1/B) \sum_{b=1}^{B} d_{(j)}^b$.

3. Plot the expected order statistics from the permutations $\bar{d}_{(j)}$ against the observed statistics $d_{(j)}$, and form the pairs data $(\bar{d}_{(1)}, d_{(1)}), ..., (\bar{d}_{(m)}, d_{(m)})$. The point at median of expected null order statistics $\{\bar{d}_{(1)}, ..., \bar{d}_{(m)}\}$ is denoted by M which divides the data into two regions. The region with $\bar{d}_{(j)} \geq M$ is called upper region, the other is called lower region.

4. For a fixed threshold Δ, starting at this point M and moving up to the upper region, find the first probe set $j = j_1$ such that $(d_{(j_1)} - \bar{d}_{(j_1)}) \geq \Delta$ and $\bar{d}_{(j_1)} \geq M$. All probe sets with $d_{(j)} \geq d_{(j_1)}$ are declared to be positive. Similarly, moving down to the lower region, find the first probe set $j = j_2$ such that $(d_{(j_2)} - \bar{d}_{(j_2)}) \leq -\Delta$ and $\bar{d}_{(j_2)} \leq M$. All probe sets with $d_{(j)} \leq d_{(j_2)}$ are declared to be negative. If $j = j_1$ or $j = j_2$ do not exist, we declare that there are no positive or negative significant probe sets.

Thus, the threshold Δ can be adjusted to yield larger or smaller sets of significant probe sets. The larger Δ, the fewer the number of significant probe sets detected.

2.5 The false discovery rate (FDR) and q-values

When identifying significant probe sets, we also need to consider some suitable measures of error. In a microarray setting, Table 2.1 describes the possible error one can make while testing m probe sets. There are m_0 true null hypotheses and R probe sets identified

	Declared non-significant	Declared significant	Total
True H_0	$m_0 - V$	V	m_0
False H_0	$m - m_0 - S$	S	$m - m_0$
Total	$m - R$	R	m

Table 2.1: *Outcomes from m hypothesis tests.*

significantly differentially expressed. Among those probe sets, there are V false positives (Type I errors) and S true positives. The most traditional method is the familywise error rate (FWER) which focuses on controlling Type I errors, i.e. $FWER = \Pr(V \geq 1)$. However, the FWER is so conservative that a very limited number of probe sets can be called significant. Benjamini and Hochberg (1995) proposed a new error measure, the false discovery rate (FDR) which is the expected proportion of false positive findings among those differential expressions,

$$FDR = E[V/R|R > 0] \cdot \Pr(R > 0). \tag{2.4}$$

Microarray data analysis is an exploratory method to extract potential candidates for further investigation. Several false findings will not distort the conclusions during the investigation, as long as their proportion is small in comparison to the number of the rejected hypotheses (Reiner *et al*, 2002). Therefore the FDR control is more appropriate and practical.

The FDR aims to control the false positives among those probe sets called significant. If all null hypotheses H_0 are true, i.e. $m_0 = m$, the control of FDR is equivalent to the control of FWER. Otherwise, when some null hypotheses are true and some are false, i.e. $m_0 < m$, FDR is less strict than FWER and so is more powerful. Benjamini and

Hochberg (1995) derived a sequential p-value method to control the FDR: For a desired FDR level α, the ordered p-value $P_{(i)}$ is compared to threshold level $(i/m) \cdot \alpha$, then reject $H_{(1)}$,..., $H_{(k)}$ if $k = \max\{i : P_{(i)} \leq (i/m) \cdot \alpha\}$ exists. When the test statistics are independent, this procedure guarantees the FDR is controlled at a desired level α, i.e. FDR $\leq (m_0/m) \cdot \alpha \leq \alpha$. Benjamini and Yekutieli (2001) extended the same procedure to more general dependence structures, such as positive regression dependence. This procedure was shown to also control the FDR under the dependence situation.

Storey (2002) proposed a variant of the FDR, termed the positive false discovery rate (pFDR):

$$pFDR = E[V/R|R > 0] \tag{2.5}$$

provided that at least one positive finding has occurred so that $\Pr(R > 0) = 1$. The pFDR control is as liberal and powerful as the FDR control. Storey (2002) showed that pFDR and FDR are asymptotically equivalent for a fixed rejection region when data set is high-dimensional such as microarray data with large m. Instead of fixing the error rate and estimating its corresponding rejection region in Benjamini and Hochberg method, Storey (2002) proposed the opposite approach: fix the rejection region and then estimate its corresponding error rate. These two methods originally worked under the assumption of independent p-values. However Storey (2003) relaxed this assumption and found that the properties of the pFDR and FDR still hold approximately under weak dependence (e.g. dependence exists in finite block genes).

For a fixed rejection region threshold Δ, we may use a simple estimate of the FDR, which is the ratio of the average number of significant probe sets in B permutations and the number of significant probe sets. Since permutations make all probe sets non-differentially expressed, the average number of significant probe sets called in B permutations represents the number of false positives. However, this estimate tends to be biased upward. So Storey (2002) and Storey & Tibshirani (2003a) consider the proportion of non-differentially expressed probe sets (π_0) to improve the simple estimate of the FDR. Then the estimate of FDR is obtained by multiplying by an estimate of π_0,

$$\widehat{FDR}(\Delta) = \hat{\pi}_0 \frac{\text{the average number of significant probe sets from } B \text{ permutations}}{\text{the number of significant probe sets from observations}}$$

The estimate of π_0 can be obtained by either of these two methods.

Method A suggested by Chu *et al.* (2005): Take three-quarter $q_{.75}$ and one-quarter $q_{.25}$ quartiles of all permutation d-statistics. m is the total number of probe sets and B is number of permutations, then we get $m \times B$ such d-statistics. Then calculate the proportion of the original statistics fall in this interval divided by the proportion of permutation statistics in the interval, i.e.

$$\hat{\pi}_0 = \frac{\#\{d_j \in (q_{.25}, q_{.75})\}}{0.5 \cdot m}.$$

Method B suggested by Storey & Tibshirani (2003b): Set λ to a range of λ=0, 0.01, 0.02, ..., 0.95. Take 100 intervals in the form of $(\lambda/2, 1-\lambda/2)$. Using the above method for these intervals, calculate corresponding $\hat{\pi}_0$s. Then fit a natural cubic spline with three degrees of freedom. The spline evaluated at $\lambda = 1$ is the final $\hat{\pi}_0$.

The FDR gives a global measure of the overall accuracy of a set of significant probe sets. But it does not provide a specific measure of the significance of each probe set. For this purpose, Storey (2002) introduced a measure in terms of the FDR, which is called the q-value. The q-value of a observed statistic is the minimum FDR over all rejection regions containing that statistic. For probe set j, the largest rejection region threshold

Δ for which that probe set is called significant is denoted by Δ_j, the q-value is estimated by (Storey & Tibshirani, 2003a)

$$q_j^{FDR} = \min_{\Delta \leq \Delta_j} FDR(\Delta). \tag{2.6}$$

Storey (2002) also gave a simpler definition of q-value when statistics are the independent p-values. The rejection region takes the form $[0, \gamma]$ and the q-values in terms of p values can be rewritten as

$$q_j^P = \min_{\gamma \geq p} \frac{\pi_0 \gamma}{\Pr(p \leq \gamma)}. \tag{2.7}$$

Storey & Tibshirani (2003b) proposed the following algorithm for calculating q-value in terms of p-values in practice :

1. For m probe sets, compute and order the p-values $p_1, p_2, ..., p_m$. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ be the ordered p-values.

2. Estimate $\hat{\pi}_0$ from either method A or method B mentioned above.

3. Calculate the q-value for the largest p-value $p_{(m)}$ by the formula

$$\hat{q}(p_{(m)}) = \min_{\gamma \geq p_{(m)}} \frac{\hat{\pi}_0 m \cdot \gamma}{\#\{p_i \leq \gamma\}} = \hat{\pi}_0 \cdot p_{(m)}.$$

4. For $i = m-1, m-2, ..., 1$, calculate

$$\hat{q}(p_{(i)}) = \min_{\gamma \geq p_{(i)}} \frac{\hat{\pi}_0 m \cdot \gamma}{\#\{p_j \leq \gamma\}} = \min\left(\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right).$$

5. The estimated q-value for the jth most significant probe set is $\hat{q}(p_{(j)})$.

As we know, the p-value can provide a measure of significance for an individual probe set. Similarly, the q-value accomplishes the same goal with respect to FDR.

The smaller q-value, the more significant the differential expression of that probe set. But there are two major discrepancies between the q-values and p-values. Firstly, The q-value gives each probe set individual measure of significance in terms of the FDR, whereas the p-value in terms of the Type I error. Secondly, because the q-value is defined in terms of the FDR, the q-value automatically takes into account the fact that tens of thousands of probe sets are simultaneously being tested, while the p-values say little about the multiple comparison. Therefore, in microarray data analysis, a q-value threshold is more practical since it directly provides potential significant probe sets list for the future investigation in biology and also controls the FDR at a desirable level.

2.6 Empirical Bayes Analysis in Microarray

Efron *et al.* (2001) proposed an empirical Bayes (EB) method using a simple inference model and explained the application to comparative microarray experiments: a probe set is either differentially expressed or non-differential expressed by the condition of interest. Similar to SAM method, the EB method constructs the same modified t-statistic denoted by Z where s_0 is chosen as the 95th percentile of s_j values. The EB method also establishes a null statistic denoted by Z^* by permutation. The null statistic Z^* is used to provide an approximate null distribution for Z. Based on this test statistic Z, a mixture density model is established and resulting statistical inference is made. The methodology is implemented in the **R** package `EBarrays` (Yuan *et al*, 2007).

Let p_1 be the probability that a probe set is differentially expressed and $p_0 = 1 - p_1$ be the probability non-differentially expressed; $f_1(z)$ be the density of Z for differentially expressed probe sets and $f_0(z)$ be the density of Z for non-differentially expressed probe sets. Then the mixture density of the two populations can be constructed as:

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

Here $f(z)$ is estimated directly from all score z_j, $j = 1, ..., m$. The null density $f_0(Z)$ is approximated by the empirical distribution of the null scores $\{z_j^*\}$ obtained by permuting the condition labels.

According to Bayes' rule, posterior probabilities $p_1(z)$ and $p_0(z)$ from the mixture model can be calculated by the following two equations:

$$p_1(z) = 1 - p_0 \frac{f_0(z)}{f(z)}$$

$$p_0(z) = 1 - p_1(z) = p_0 \frac{f_0(z)}{f(z)}.$$

where $p_1(z)$ is the posterior probability for differentially expressed probe sets and $p_0(z)$ is the posterior probability for non-differentially expressed probe sets. Then $p_1(z)$ and $p_0(z)$ are the function of ratio $f_0(z)/f(z)$ and p_0. The estimate of $f_0(z)/f(z)$ is determined by relative densities from observed z statistics and permuted z^* statistics. We consider values of z statistics as "success" and values of permuted z^* statistics as "failure". In the logistic regression model, the probability $\pi(z)$ of a success at point z is defined by

$$\pi(z) = \frac{f(z)}{f(z) + Bf_0(z)},$$

where B is the number of permutations. Thus the ratio $f_0(z)/f(z)$ is rewritten as

$$\frac{f_0(z)}{f(z)} = \frac{1 - \pi(z)}{B\pi(z)}.$$

By logistic regression the $\pi(z)$ is estimated so that the estimate of ratio $f_0(z)/f(z)$ can be obtained. So for the determination of p_0, Efron *et al.* (2001) suggests taking the upper bound of $f_0(z)/f(z)$ for p_0 from the following relationships since the posterior probabilities are nonnegative for all z statistics,

$$p_1 \geq 1 - \min_z \left\{ \frac{f(z)}{f_0(z)} \right\}$$

and

$$p_0 \leq \min_z \left\{ \frac{f(d)}{f_0(z)} \right\}.$$

The EB method for microarray data can be summarized in the following algorithm (Efron *et al*, 2001):

1. Compute the statistics z_j for observed expressions, $j = 1, ..., m$.
2. Generate B sets of permutations of the condition labels. For each permutation, compute statistics z_j^*.
3. Estimate the ratio $f_0(z)/f(z)$ by using logistic regression. The ratio $f_0(z)/f(z)$ is based on the relative densities of Z and Z^*.
4. Take the upper bound of $f_0(z)/f(z)$ to estimate p_0.
5. Find the posterior probability $p_1(z)$ for each probe set by Bayes' rule.

Efron *et al.* (2001) points out that the empirical Bayes analysis is closely related to Benjamini and Hochberg's FDR criterion. Benjamini and Hochberg's FDR is a global definition. In empirical Bayes analysis, a false discovery rate for a gene is defined as

$$\text{fdr}(z) = p_0 \frac{f_0(z)}{f(z)}$$

This quantity is called the local false discovery rate (fdr). The fdr(z) is a posterior probability $p_0(z)$ that a probe set with statistic z is non-differentially expressed. To estimate global FDR for a rejection region λ, we need to permute the observed z statistics. Then the estimated FDR for a rejection region λ is obtained by

$$\widehat{FDR}_\lambda = \hat{p}_0 \cdot \frac{\sum_{b=1}^{B} \#\{z_i^{*b} \in \lambda\}}{B \cdot \#\{z_i \in \lambda\}}.$$

Therefore, we establish a connection between the estimated posterior probabilities and the FDR, thereby allowing for the analyst to handle multiple testing issues that arise when dealing with a large number of simultaneous tests.

2.7 Linear Models for Microarray Data (LIMMA)

Smyth (2004) developed a linear model approach for general microarray experiments with two or more groups, based on the hierarchical model of Lönnstedt and Speed (2002). A single linear model proposed by Kerr *et al.* (2000) was used to fit an entire microarray experiment and assumed the equal variance for probe sets. However, LIMMA is designed to fit a linear model for each gene and also assume different variances across probe sets. The methodology is implemented in the **R** package `limma` (Smyth, 2005).

Suppose we have a set of n microarrays yielding a response vector $\mathbf{y}_j^T = (y_{j1}, ..., y_{ji}, ..., y_{jn})$ for the jth probe set. Each component y_{ji} represents the expression of probe set j in array i $(j = 1, ..., m; i = 1, ..., n)$. In our case, n is the total number of arrays under two different conditions. Then we obtain a linear model for the jth probe set:

$$\mathrm{E}(\mathbf{y}_j) = X\boldsymbol{\alpha}_j, \ \mathrm{var}(\mathbf{y}_j) = W_j\sigma_j^2$$

where X is a design matrix, $\boldsymbol{\alpha}_j$ is a coefficient vector, and W_j is a known nonnegative definite weight matrix. The contrasts of interest are defined by $\boldsymbol{\beta}_j = C^T\boldsymbol{\alpha}_j$ where C is contrast matrix. We can fit this model for each probe set to obtain coefficient estimators

$\hat{\alpha}_j$, estimators s_j^2 of σ_j^2, estimated covariance matrices $\mathrm{var}(\hat{\alpha}_j) = V_j s_j^2$ where V_j is a positive definite matrix, and contrast estimators $\hat{\beta}_j = C^T \hat{\alpha}_j$ with covariance matrices $\mathrm{var}(\hat{\beta}_j) = C^T V_j C s_j^2$. In our study of two-sample comparison, the contrast estimator $\hat{\beta}_j$ is scalar.

Given the large number of gene-wise linear model fits arising from a microarray experiment, there is need to take advantage of the parallel structure where the same model is fitted to each probe set. Lonnstedt and Speed (2002) define a hierarchical Bayesian model for this purpose, that describes how the unknown coefficients β_j and unknown variance σ_j^2 vary across probe sets. This is done by assuming prior distribution for these sets of parameters. Prior information on β_j is assumed to be normally distributed and σ_j^2 is assumed to follow approximately a scaled chisquare distribution. The posterior mean of σ_j^2 given s_j^2 is obtained by

$$\tilde{s}_j^2 = E(\sigma_j^2 | s_j^2) = \frac{d_0 s_0^2 + d_j s_j^2}{d_0 + d_j}.$$

The posterior values shrink the observed variances s_j^2 towards the prior values s_0^2 with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom. The posterior variance is used to replace the usual sample variance in the regular t-statistic. The modified statistic is called moderated t-statistic, defined by

$$\tilde{t}_j = \frac{\hat{\beta}_j}{\tilde{s}_j}.$$

The moderated t statistic is shown to follow a t distribution with degrees of freedom $d_j + d_0$. The added d_0 degrees of freedom reflect the extra information which is borrowed across probe sets. The moderated t has the advantage over the regular t-statistic that large statistics are less likely to arise only from underestimated sample variances. The posterior variance $\tilde{s}_j^2$ offsets the small sample variances heavily in a relative sense while larger sample variances are moderated to a lesser relative degree. The p-values from the $\tilde{t}_j$ will be calculated and then used to identify differentially expressed genes.

The moderated t-statistic is established on the basis of the given d_0 and s_0^2 values. For an microoarray experiment, d_0 and s_0^2 can be estimated directly from observed gene expression in an empirical Bayes manner. Smyth (2004) proposed a computationally intensive method to estimate these two parameters. The $\log s_j^2$ instead of s_j^2 is used because the moments of $\log s_j^2$ are finite for any degrees of freedom and the distribution of $\log s_j^2$ is more nearly normal so that moment estimate is more efficient. Estimate of d_0 is obtained by solving

$$\psi'(d_0/2) = \text{mean}\{(e_j - \bar{e})n/(n-1) - \psi'(d_j/2)\}$$

where $e_j = \log s_j^2 - \psi(d_j/2) + \log(d_j/2)$, $\bar{e}$ is mean of e_j for all n number of array, and $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively. So s_0^2 can be estimated by

$$s_0^2 = \exp\{\bar{e} + \psi(d_0/2) - \log(d_0/2)\}.$$

This estimate for s_0^2 is usually somewhat smaller than the mean of the s_j^2. Therefore a set of sample variance s_j^2 leads to the estimates of d_0 and s_0^2 which are the important quantity in the moderated t-statistics. Then an empirical Bayes log posterior odds statistic called B can be obtained by equation (7) of Lönnstedt & Speed (2002), which is proportional to a function of the moderated t-statistic, i.e.

$$B \propto \frac{1 + \tilde{t}_j^2}{1 + \frac{\tilde{t}_j^2}{1+nc}}$$

where n is sample size of each condition and c is a hyperparameter in the normal prior of the nonzero means. The B statistic is an increasing function of the square of the modified

t-statistic $\tilde{t}_j$ and used to rank genes in order of evidence for differential expression. Hence large absolute values of $\tilde{t}_j$ lead to large values of B. Note that the adjustment factor s_0^2 in the modified t-statistic $\tilde{t}_j$ is made to the sample variance whereas in the case of SAM it is made to the standard error.

In summary, the procedure for general microarray experiments with two or more groups essentially involves three steps. The first step is to reset data in the context of general linear models with appropriate design matrix. The second step is to derive estimators for parameters d_0 and s_0^2. The third step is to calculate B statistic in terms of the moderated t-statistic in which posterior standard deviations are used in place of regular standard deviations.

Chapter 3

Test statistics and q-values to identify differential expressions

In our study, we are interested in determining which probe sets show a statistically significant difference in expression between different strains. For simplicity, we limit the discussion of methodology to the case where there are two strains and all samples are independent. Therefore, the null hypothesis for each probe set is that there is no difference in mean expression between two strains, i.e. $H_{j0} : \mu_{j1} = \mu_{j2}$, $j = 1, ..., m$, where μ_{j1} and μ_{j2} denote the population mean expression in strain 1 and 2 for probe set j, respectively. There are often tens of thousands of probe sets tested simultaneously in a microarray experiment. In such experiments, gene expression data have three properties: (i) the dimension of data (m) is much larger than sample size, (ii) some probe sets are correlated, and (iii) a large proportion of the null hypotheses are expected to be true (Pollard & Van der Laan, 2004). So gene expression studies have motivated us to better understand multiple testing issues such as forming test statistic, calculating the null distribution for the test statistic, choosing rejection region and controlling the number of false positives.

3.1 Test statistics

A test statistic quantifies the evidence in the data against the null hypothesis being tested. We must make sure that the statistic used is appropriate. Because different statistics may give different p-values and q-values for each probe set, and therefore they can lead to different conclusions. A commonly used statistic for testing difference in the means of two strains is the well-known two-sample t-statistic which was applied to identify differential expressions by Dudoit *et al.* (2002).

Tusher *et al.* (2001), Efron *et al.* (2001) and Smyth (2004) modified this regular t-statistic with an offset standard deviation s_0. Since the number of RNA samples measured for each strain is always small, the variability for each probe set may not be stable. For example, if standard error (s_j) from one probe set is small, by chance, the t value becomes larger even when the corresponding difference between two average expressions $(\bar{x}_{j2} - \bar{x}_{j1})$ is small (Cui & Churchill, 2003). So adding the extra term s_0 offsets this instability. Tusher *et al.* (2001) estimated s_0 by minimizing the coefficient of variation of test statistics. In Chapter 2 we have introduced the procedure to get s_0, which was proposed by Chu *et al.* (2005). Efron *et al.* (2001) chose the 95th percentile of the distribution of all sample standard deviations which ensures less information loss. Smyth (2004) estimated s_0 by solving equations with respect to sample standard deviations and degrees of freedom of the distribution of modified t-statistics. In our study, we focus on the statistic proposed by Tusher *et al.* (2001), defined by

$$D_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_0 + s_j} \tag{3.1}$$

We will refer to this as the d-statistic. As introduced in Chapter 2, it borrows information across all probe sets, while the regular t-statistic only uses information from one probe set at a time.

3.2 Null distribution and p-values

A key feature of hypothesis testing is the null distribution of the test statistic. The choice of a proper null distribution is crucial in order to ensure a desirable control of FDR. In our study, we consider two choices of null distribution for both d-statistic and t-statistic: (1) numeric integration method to derive a null distribution; (2) permutation method to estimate a null distribution. In this section we focus on the t-statistic whose null distribution was derived under some assumptions (Welch, 1937; Best and Rayner,1987). We also introduce permutation method for t-statistic, which has been discussed widely (Dudoit and van der Laan, 2008).

3.2.1 t distribution and the calculation of p-value

If gene expressions for each probe set are normally distributed, the null distribution of T_j follows Student t-distribution, $T_j \sim t_\nu$, (Student, 1908). In our study we make the assumption that variances under two strains are equal for a given probe set. Under this assumption, t-statistic follows a t distribution with degrees of freedom $n_1 + n_2 - 2$. The p-value is hence easily calculated by the cumulative distribution function of the t

distribution. For two-sided alternative hypothesis, p-value for the t-statistic is:

$$p\text{-value for probe set } j = 2 \times P(t_{n_1+n_2-2} \geq |t_j|).$$

Here we expect both overexpression and underexpression probe sets so hypothesis testing is two-sided.

Note that, although the variability of the raw expression values increases with the mean, the effect of normalization and the log transformation will tend to attenuate this relationship and so the assumption of equal variances seems justified. We have not, however, examined the effect of unequal variances by strain in our study.

3.2.2 Permutation method and the calculation of p-value

In practice it may not be valid to assume that the null distribution of T_j is a t distribution, especially as small sample sizes are very common in a microarray data and samples may not come from normal distributions. So we apply the permutation method to estimate the null distribution without limitation on sample size and any parametric assumption.

By permutation, the estimated null distribution can be obtained by calculating all possible values of the test statistic under rearrangements of strain labels in a manner of sampling without replacement. This method requires that all observations are exchangeable under null hypothesis H_0 where two strains have identical expression patterns. Each arrangement may be viewed as a permutation of the $n_1 + n_2$ expression values with n_1 values assigned to strain 1 and the remainder assigned to strain 2. There are B such permutations in two-sample case, where

$$B = \frac{(n_1 + n_2)!}{n_1!n_2!}. \tag{3.2}$$

B	Observed expression values										t
	x_{j11}	x_{j12}	x_{j13}	x_{j14}	x_{j15}	x_{j21}	x_{j22}	x_{j23}	x_{j24}	x_{j25}	statistic
b_1	1	1	1	1	1	2	2	2	2	2	t_j^1
b_2	1	1	1	1	2	1	2	2	2	2	t_j^2
...						...					...
b_{251}	2	2	2	2	1	2	1	1	1	1	t_j^{251}
b_{252}	2	2	2	2	2	1	1	1	1	1	t_j^{252}

Table 3.1: *Simple demonstration of a permutation test for probe set j.*

For example, in our simulation study where $n_1 = n_2 = 5$ then B=(5+5)!/(5!5!)=252. When H_0 is true, the p-value for probe set j is the fraction of the B permuted t-statistics denoted by t_j^b that are greater or equal to the observed t-statistics denoted by t_j^{obs} in absolute value, i.e.

$$p\text{-value for probe set } j = \frac{\#\{b : |t_j^b| \geq |t_j^{obs}|\}}{B}.$$

Table 3.1 shows a simple illustration of a permutation test for probe set j in our simulation study. The strain vector is (1 1 1 1 1 2 2 2 2 2). The observed expression values are denoted by x_{jik}, where $j = 1, ..., 10000$ denotes probe set, $i = 1, 2$ denotes strain and $k = 1, ..., 5$ denotes the number of sample under each strain. Permutation matrix of each probe set is a 252 × 10 matrix in terms of strain labels (Table 3.1). Then each probe set has an observed test statistic as well as a set of 252 permuted test statistics. When we combine all probe sets together, we get a 10000 × 252 matrix in terms of permuted test statistics shown in Table 3.2. Because we permute the entire vectors of expression values, the correlation between expression values can be maintained. By pooling over

	b_1	b_2	$\cdots$	b_{252}
probe set1	t_1^1	t_1^2	$\cdots$	t_1^{252}
probe set2	t_2^1	t_2^2	$\cdots$	t_2^{252}
...	...	...	...	...
probe set$j$	t_j^1	t_j^2	$\cdots$	t_j^{252}
...	...	...	...	...
probe set10000	t_{10000}^1	t_{10000}^2	$\cdots$	t_{10000}^{252}

Table 3.2: *Simple demonstration of permutation test by pooling all probe sets.*

all m probe sets, the p-value for probe set j is given by

$$p\text{-value for probe set } j = \sum_{b=1}^{B} \frac{\#\{i : |t_i^b| \geq |t_j^{obs}|, i = 1, .., m\}}{mB}. \tag{3.3}$$

Here we assume that the test statistics have the same null distribution across probe sets so that all $m \times B$ permuted values are used in the calculation of the p-value. We also assume independence of the test statistics which is unlikely to be true in practice but is unlikely to badly affect the p-value calculation.

In summary, the calculation of p-value using the permutation distribution of the test statistics T_j, $j = 1, ..., m$, is carried on by the following algorithm.

1. Permute the elements in strain vector (1 1 1 1 1 2 2 2 2 2) as in Table 3.1. The number of permutation B is calculated by equation (3.2).

2. For each probe set, compute observed t-statistics t_j^{obs} and a set of B permuted t-statistic: $t_j^1, ..., t_j^B$, $j = 1, ..., m$.

3. Establish a $m \times B$ permutation matrix in terms of permuted t-statistic as in Table 3.2.

4. Compare the absolute values of all $m \times B$ permuted t-statistics with the absolute value of the observed statistic $|t_j^{obs}|$ for probe set j. Calculate the proportion of the absolute values of permuted t-statistics which are greater or equal to $|t_j^{obs}|$. Thus the proportion is the p-value for probe set j defined by equation (3.3).

3.3 The distribution of d-statistics under null hypothesis

The d-statistic in the equation (3.1) has been widely used in practice, such as in SAM method of Tusher *et al.* (2001) and empirical Bayes of Efron *et al.* (2001). These methods employ the above permutation technique to estimate the null distribution, however, do not have an associated distribution theory. In this section, we use numeric integration method to derive the distribution of d-statistic when H_0 is true.

3.3.1 Derivation of d distribution and calculation of p-values

Suppose Z is standard normal variable and W has a chi-square distribution with r degrees of freedom that is independent of Z, then the variable

$$T = \frac{Z}{\sqrt{W/r}}$$

is t distributed with r degrees of freedom. When a positive constant c is added in the denominator, a new variable is defined by D as:

$$D = \frac{Z}{c + \sqrt{W/r}}$$

T variable is a special case of D variable when c is zero. In our study, we only derive the cumulative distribution function of D since we only need to calculate p-values. Let $X = \sqrt{W/r}$, then $W = X^2 r \sim \chi^2_{(r)}$ with probability density function

$$f_W(w) = \frac{1}{2^{r/2}\Gamma(r/2)} w^{r/2-1} e^{-w/2}.$$

And,

$$F_D(d) = P(D \leq d) = P\left(\frac{Z}{c+X} \leq d\right) = P(Z \leq d(c+X)).$$

To compute the probability of $Z \leq d(c+X)$, we need to get the joint distribution of X and Z. Since Z and W are independent, X and Z are independent. Given Z being standard normal, the marginal density function of X can be attained by changing variable,

$$\begin{aligned} f_X(x) &= \frac{1}{2^{r/2}\Gamma(r/2)} (x^2 r)^{r/2-1} e^{-x^2 r/2} \cdot 2xr \\ &= \frac{1}{2^{r/2-1}\Gamma(r/2)} r^{r/2} x^{r-1} e^{-x^2 r/2}. \end{aligned}$$

Thus the joint distribution of X and Z is

$$\begin{aligned} f_{X,Z}(x,z) &= f_X(x) \cdot f_Z(z) \\ &= \frac{1}{2^{r/2-1}\Gamma(r/2)} r^{r/2} x^{r-1} e^{-x^2 r/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \\ &= \frac{r^{r/2}}{\sqrt{2\pi} 2^{r/2-1}\Gamma(r/2)} x^{r-1} e^{-(x^2 r + z^2)/2}. \end{aligned}$$

Then the cumulative distribution function of D is obtained by integrating $f_{X,Z}(x,z)$ over $z \in (-\infty, d(x+c))$ and $x \in (0, +\infty)$,

$$\begin{aligned} F_D(d) &= P(Z \leq d(c+X)) \\ &= \int_0^{+\infty} \int_{-\infty}^{d(x+c)} f_X(x) \cdot f_Z(z) dz dx \end{aligned}$$

$$
\begin{aligned}
&= \int_0^{+\infty} \frac{1}{2^{r/2-1}\Gamma(r/2)} r^{r/2} x^{r-1} e^{-x^2 r/2} \cdot \int_{-\infty}^{d(x+c)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz dx \\
&= \int_0^{+\infty} \frac{1}{2^{r/2-1}\Gamma(r/2)} r^{r/2} x^{r-1} e^{-x^2 r/2} \cdot \Phi(d(x+c)) dx \\
&= \int_0^{+\infty} 2rx \frac{1}{2^{r/2}\Gamma(r/2)} (rx^2)^{r/2-1} e^{-x^2 r/2} \cdot \Phi(d(x+c)) dx.
\end{aligned}
$$

Therefore cdf $F_D(d)$ is denoted by

$$F_D(d) = \int_0^{+\infty} 2rx \frac{1}{2^{r/2}\Gamma(r/2)} (rx^2)^{r/2-1} e^{-x^2 r/2} \cdot \Phi(d(x+c)) dx. \tag{3.4}$$

where $2rx\frac{1}{2^{r/2}\Gamma(r/2)}(rx^2)^{r/2-1}e^{-x^2r/2}$ is probability density function of the random variable X where $W = X^2 r \sim \chi^2_{(r)}$, and $\Phi(\cdot)$ is standard normal cumulative distribution function. Under the assumption of common variance for each probe set, the degrees of freedom $r = n_1 + n_2 - 2$.

Because the numerator of D is standard normal variable whose distribution is symmetric about 0, the distribution of D is symmetric about 0 too. For two-sided alternative hypothesis, the p-value for d-statistic is

$$p\text{-value for probe set } j = 2(1 - F_D(|d_j|)), \text{for all } d_j.$$

In our simulation study, sample size for each strain is 5 so $r = 5+5-2 = 8$, and c is defined in the following section. Since it is cumbersome to know the true underlying d distribution, we collected d-statistics (d.null) for those truly non-differentially expressed probe sets in Simulation 1 and drew empirical cdf plot using `plot(ecdf(d.null))` in R. In Figure 3.1, the thin curve of $F_D(d)$ from numeric integration fits empirical plot (thick curve) perfectly, indicating that $F_D(d)$ is the true cumulative distribution function of d.

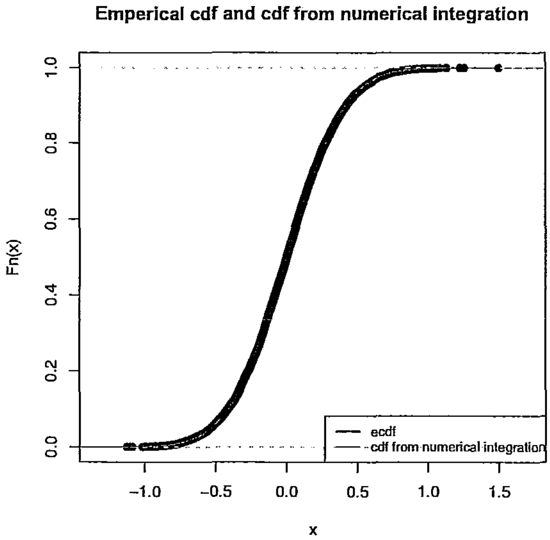


Figure 3.1: *The fitting of empirical cdf and $F_D(d)$ from numeric integration.*

3.3.2 Determination of c in D

Suppose we have two independent random samples from strain 1 and 2 for probe set j, $X_{j11},...,X_{j1n_1} \sim N(\mu_{j1},\sigma_j^2)$ and $X_{j21},...,X_{j2n_2} \sim N(\mu_{j2},\sigma_j^2)$, then $\bar{x}_{j1}-\bar{x}_{j2} \sim N(\mu_{j1}-\mu_{j2},\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2}))$. Under null hypothesis, d-statistic for probe set j in equation (3.1) is defined as:

$$\begin{aligned} D_j &= \frac{\bar{x}_{j1}-\bar{x}_{j2}}{s_0+s_j} \\ &= \frac{(\bar{x}_{j1}-\bar{x}_{j2}-0)/\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}{(s_0+s_j)/\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}} \\ &= \frac{(\bar{x}_{j1}-\bar{x}_{j2})/\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}{\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}+\sqrt{\frac{s_j^2}{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}}. \end{aligned}$$

Since $\frac{(n_1-1)s_{j1}^2}{\sigma_j^2}$ and $\frac{(n_2-1)s_{j2}^2}{\sigma_j^2}$ have Chi-squared distribution with degrees of freedom $n_1 - 1$ and $n_2 - 1$ respectively, it can be shown by moment generating function that

$$\frac{(n_1-1)s_{j1}^2 + (n_2-1)s_{j2}^2}{\sigma_j^2} \sim \chi^2_{n_1+n_2-2}.$$

By plugging in equation 2.2, thus

$$\begin{aligned}
D_j &= \frac{(\bar{x}_{j1} - \bar{x}_{j2})/\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}{\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}} + \sqrt{\frac{s_j^2}{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}} \\
&= \frac{(\bar{x}_{j1} - \bar{x}_{j2})/\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}{\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}} + \sqrt{\frac{(\frac{1}{n_1}+\frac{1}{n_2})\frac{(n_1-1)s_{j1}^2+(n_2-1)s_{j2}^2}{n_1+n_2-2}}{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}} \\
&\sim \frac{N(0,1)}{\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}} + \sqrt{\frac{(n_1-1)s_{j1}^2+(n_2-1)s_{j2}^2}{(n_1+n_2-2)\sigma_j^2}}} \\
&\sim \frac{N(0,1)}{\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}} + \sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1+n_2-2}}}.
\end{aligned}$$

The first term in the denominator of d-statistic $\frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}$ is denoted by

$$c_j = \frac{s_0}{\sqrt{\sigma_j^2(\frac{1}{n_1}+\frac{1}{n_2})}}, \tag{3.5}$$

which is a function of s_0 and sample sizes when common variance σ_j^2 is known. In practice, however, σ_j^2 is unknown and needs to be estimated. Here we consider the following estimates of σ_j^2 and corresponding estimates of c_j.

The pooled estimate of σ_j^2: $s_{j;pool}^2$. Since both strains for probe set j are assumed to have the same population variance, we naturally consider the pooled variance $s_{j;pool}^2$ defined by

$$s_{j;pool}^2 = \frac{(n_1-1)s_{j1}^2 + (n_2-1)s_{j2}^2}{n_1+n_2-2}. \tag{3.6}$$

Correspondingly, the estimate of c_j is defined with respect to the pooled variance $s^2_{j;pool}$ by

$$\hat{c}^1_j = \frac{s_0}{\sqrt{\hat{\sigma}^2_j\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{s_0}{\sqrt{s^2_{j;pool}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}. \tag{3.7}$$

Note that the denominator of $\hat{c}^1_j$ is s_j from equation (2.2). The denominator of the d-statistic moderates this using $s_j + s_0$ so we considered using the same here to get

$$\hat{c}^2_j = \frac{s_0}{s_0 + \sqrt{s^2_{j;pool}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{1}{\frac{1}{\hat{c}^1_j}+1}, \tag{3.8}$$

which is the function of $\hat{c}^1_j$. Since s_0 is always positive, then $\hat{c}^2_j$ is always smaller than $\hat{c}^1_j$.

The null estimate of σ^2_j: $s^2_{j;null}$. Since there is no difference in expressions under two strains when null hypothesis is true, all samples come from the same population. We propose an estimator called null sample variance, $s^2_{j;null}$ for probe set j, treating all samples $\{X_{j11}, ..., X_{j1n_1}, X_{j21}, ..., X_{j2n_2}\}$ as a random sample, then the sample variance can be computed by:

$$s^2_{j;null} = \frac{1}{n_1+n_2-1}\sum_{i=1}^{n_1+n_2}(x_{ji}-\bar{x}_j)^2, \tag{3.9}$$

where $\bar{x}_j$ is the average expression over two strains for probe set j. As a result, an estimate of c_j is denoted by $\hat{c}^3_j$, as

$$\hat{c}^3_j = \frac{s_0}{\sqrt{s^2_{j;null}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}. \tag{3.10}$$

The pooled estimate in (3.6) and the null estimate in (3.9) assume that each probe set has a different variance. We also propose two more estimates by considering some cases where all probe sets may share a common variance. By taking the average of $s^2_{j;pool}$ across all probe sets, we construct a common pooled variance for each probe set,

$$\bar{s}^2_{pool} = \frac{1}{m}\sum_{j=1}^{m} s^2_{j;pool}, \tag{3.11}$$

the corresponding estimate of c_j is denoted by $\hat{c}_j^4$ as

$$\hat{c}_j^4 = \frac{s_0}{\sqrt{\bar{s}^2_{pool}(\frac{1}{n_1} + \frac{1}{n_2})}}. \tag{3.12}$$

Similarly,

$$\bar{s}^2_{null} = \frac{1}{m}\sum_{j=1}^{m} s^2_{j;null}, \tag{3.13}$$

and

$$\hat{c}_j^5 = \frac{s_0}{\sqrt{\bar{s}^2_{null}(\frac{1}{n_1} + \frac{1}{n_2})}}. \tag{3.14}$$

3.4 Selection of probe sets via q-values

As we know, the p-value can be used to assign to each probe set a level of significance in terms of the false positive rate. A p-value threshold of 5% yields a false positive rate of 5% among all null hypotheses being tested. However, it dose not provide a measure of the errors among the probe sets called significant. Information on this is provided by the FDR, which gives a measure of the proportion of false positives among significant probe sets. Moreover, the q-value with respect to FDR provides an individual measure of the significance of each probe set. If a q-value threshold α is chosen, then we can call all probe sets with $q \leq \alpha$ significant. Then a set of significant probe sets can be produced. For the large number of probe sets being tested, the $FDR \leq \alpha$ so that the proportion of false positives among all significant probe sets is controlled at level α. Therefore q-value not only gives individual measure of significance for a probe set but also control the FDR at the prechosen level α.

In our study we calculate the q-value for a probe set by two methods. One method is based on equation (2.6) in terms of FDR. In this method we need to find the largest rejection region threshold Δ_j where a particular probe set is called significant and estimate FDRs over all larger rejection regions containing that probe set. The method is implemented in `SAM.R` written by Dr. Angelo Canty (2005). It cycles through all probe sets starting from the median M of expected order statistics $\{\bar{d}_{(1)}, ..., \bar{d}_{(m)}\}$ and moving outwards to find the values of Δ_j for each probe set, it then finds the FDR for each such value of $\Delta_j, j = 1, ..., m$ and uses the minimum as the q-value by

$$q_j^{FDR} = \min_{\Delta \leq \Delta_j} FDR(\Delta),$$

where Δ_j is the largest rejection region threshold for which probe set j is called significant.

The other method is based on equation (2.7) in terms of p-value. In this method, we need to first compute p-values for all probe sets. If the null distribution of test statistics is very clear (e.g., the null distribution of a t test is the t distribution when gene expressions are normally distributed), then p-values can be computed directly from that distribution. Otherwise we apply permutation method mentioned in previous section to estimate null distribution. After ordering p-values as $p_{(1)} \leq p_{(2)} ... \leq p_{(m)}$, q-value for the probe set with the largest p-value $p_{(m)}$ is first estimated by

$$\hat{q}(p_{(m)}) = \hat{\pi}_0 \cdot p_{(m)},$$

and then q-value for $(m-1)$th most significant probe set is estimated by

$$\hat{q}(p_{(i)}) = \min\left\{\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right\}, \text{for } i = m-1, ..., 1.$$

The ordering of probe sets in terms of their q-values is the same as that in terms of their p-values. The algorithm for estimating q-value in terms of p-value (Storey & Tibshirani, 2003b) has also been presented in Chapter 2 and contained in `qvalue` package written by Dabney and Storey (2009). We wrote an R function `qvalue.pvall` based on this software to implement the algorithm.

3.5 Summary of methods used in our study

The classical approach to hypothesis testing is to calculate a statistic and associated p-value. When the test statistic has a known null distribution, then we can easily get p-value from the null distribution. Otherwise we use permutations of replicated measurements to estimate a null distribution and then calculate the permutation p-value. Whether p-value is based on a known null distribution or estimated null distribution, p-value provides a good approach to achieving q-value for each probe set. We denote this approach as q-value in terms of p-value. According to different method to get p-values, we denote the approach using distribution to obtain p-values by **DIS** and the approach using permutation by **PERM**. On the other hand, the SAM method proposes a different approach by using original test statistic instead of its associated p-value to estimate the q-value for each probe set. We denote this approach as q-value in terms of FDR by **FDR**. The biggest advantage of **FDR** method is that we directly use original test statistic regardless of its null distribution.

As discussed in previous sections, we consider two types of test statistics, regular t-statistic and d-statistic, and would like to compare their performance when calculating q-values from FDR method and p-value method. Therefore we add **.t** and **.d** in the end of **FDR**, **DIS** and **PERM** in order to distinguish which test statistic is used. Table 3.3 lists six methods we discuss in our simulation study and real data analysis. When we

Methods	d-statistic	t-statistic
FDR method	**FDR.d**	**FDR.t**
p-values from permutation	**PERM.d**	**PERM.t**
p-values from distribution	**DIS.d**	**DIS.t**

Table 3.3: *Six methods discussed in simulation study and application.*

calculate d-statistics for all probe sets, we implement three methods to get q-values which are **FDR.d**, **PERM.d** and **DIS.d**. Similarly, the three methods **FDR.t**, **PERM.t** and **DIS.t** are calculated based on regular t-statistics.

We will pay a closer attention to the distribution method based on d-statistic, **DIS.d**. The p-value is calculated from the cdf of d distribution which is based on true population variance σ_j^2. In practice, however, it is unknown and need to be estimated. From the previous discussion, we consider four estimates: $s_{j;pool}^2$ and $s_{j;null}^2$, and their average quantities $\bar{s}_{pool}^2$ and $\bar{s}_{null}^2$. Based on $s_{j;pool}^2$ and $\bar{s}_{pool}^2$, we propose three estimates of c_j: $\hat{c}_j^1$, $\hat{c}_j^2$ and $\hat{c}_j^4$. We also have other two estimates $\hat{c}_j^3$ and $\hat{c}_j^5$ associated with $s_{j;null}^2$ and $\bar{s}_{null}^2$. Therefore we discuss five methods in terms of d distribution: **DIS.d.d1**, **DIS.d.d2**, **DIS.d.d3**, **DIS.d.d4** and **DIS.d.d5**. Table 3.4 lists these methods and associated estimates of σ_j^2 and c_j.

Methods	$\hat{\sigma}_j^2$	$\hat{c}_j$
DIS.d.d1	$s_{j;pool}^2 = \frac{1}{n_1+n_2-2}[(n_1-1)s_{j1}^2 + (n_2-1)s_{j2}^2]$	$\hat{c}_j^1 = \frac{s_0}{\sqrt{s_{j;pool}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$
DIS.d.d2	$s_{j;pool}^2 = \frac{1}{n_1+n_2-2}[(n_1-1)s_{j1}^2 + (n_2-1)s_{j2}^2]$	$\hat{c}_j^2 = \frac{1}{\frac{1}{\hat{c}_j^1}+1}$
DIS.d.d3	$s_{j;null}^2 = \frac{1}{n_1+n_2-1}\sum_{i=1}^{n_1+n_2}(x_{ji}-\bar{x}_j)^2$	$\hat{c}_j^3 = \frac{s_0}{\sqrt{s_{j;null}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$
DIS.d.d4	$\bar{s}_{pool}^2 = \frac{1}{m}\sum_{j=1}^{m} s_{j;pool}^2$	$\hat{c}_j^4 = \frac{s_0}{\sqrt{\bar{s}_{pool}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$
DIS.d.d5	$\bar{s}_{null}^2 = \frac{1}{m}\sum_{j=1}^{m} s_{j;null}^2$	$\hat{c}_j^5 = \frac{s_0}{\sqrt{\bar{s}_{null}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$

Table 3.4: *Distribution method based on d-statistics when using different estimates of* σ_j^2 and c_j.

Chapter 4

Simulation Studies

We implemented and evaluated the six methods proposed in Chapter 3 for identification of differentially expressed probe sets in microarray data: FDR method, permutation method and distribution method in terms of d-statistic and t-statistic, respectively. The evaluation includes application to both simulated data and real microarray data. In this chapter we investigate the performance of the six methods using simulated data sets. Since we know which probe sets are truly differentially expressed during data generation, we can compute the real FDR to compare which method can get the most accurate results. In addition, we conducted a simulation study to observe the power for detecting true differences from each method.

4.1 Simulation study design

We generated 100 simulated data sets, of the expressions of 10,000 probe sets from 2 strains of 5 samples each. Each data set is denoted as $Y_{10,000\times10}$ data matrix. For convenience, we did not consider day effect. The first five columns in data matrix Y are viewed as samples from strain 1; the second five are the samples from strain 2. We denote by entry y_{ji} the expression level for probe set j from the i-th sample, j=1,2,...,10,000 and i=1,2,...,10. In order to examine the FDR behaviors of all methods, we specifically set the first 10% probe sets/rows in Y matrix as differentially expressed and the remaining as non-differentially expressed. Thus the proportion of true null hypotheses π_0 is $(100 - 10)\% = 90\%$.

In Simulation 1, we consider an extremely simple case where expressions are normally distributed and all probe sets have the same variance 1. The expression for probe set j from the ith sample can be modeled as

$$y_{ji} = \mu_{ji} + d_{ji} + \epsilon_{ji}, \tag{4.1}$$

where μ_{ji}=2,

$$d_{ji} = \begin{cases} -3 + \frac{2}{499}(j-1) & j = 1, ..., 500; i = 6, ..., 10 \\ 1 + \frac{2}{499}(j-501) & j = 501, ..., 1000; i = 6, ..., 10 \\ 0 & \text{otherwise.} \end{cases}$$

and $\epsilon_{ji} \sim N(0,1)$ $(i = 1, ..., 10, \text{iid})$. Thus all probe sets share the common variance 1. The model (4.1) makes the first 500 probe sets negatively differentially expressed with different amounts of differential expressions between -3 and -1, the second 500 probe sets positively differentially expressed with different amounts of differential expressions between 1 and 3, and the remaining 9000 probe sets equally expressed.

In Simulation 2, we assume that the data are heavy-tailed and the error term ϵ_{ji} is distributed as $\epsilon_{ji} \sim t_{(3)}/\sqrt{3}$, where $\sqrt{3}$ is the standard deviation of t distribution with 3 degrees of freedom. In practice the sources of variability may vary across probe sets,

Error term distribution	$\sigma_j = 1$	$\sigma_j \sim \chi^2_{(20)}/20$
N(0,1)	simulation 1	simulation 3
$t(3)/\sqrt{3}$	simulation 4	simulation 2
$t(5)/\sqrt{5/3}$	simulation 5	
$t(10)/\sqrt{10/8}$	simulation 6	
$Laplace(0,1)/\sqrt{2}$	simulation 7	

Table 4.1: *Description of seven simulations by distribution and variability of variance*

so we assume different variances σ_j^2 for probe sets, which are generated from $\chi^2_{(20)}/20$ distribution. Therefore the expression for probe set j from the ith sample conditional on σ_j can be modeled as

$$y_{ji} = \mu_{ji} + \sigma_j * d_{ji} + \sigma_j * \epsilon_{ji}, \tag{4.2}$$

where $\epsilon_{ji} \sim t_{(3)}/\sqrt{3}$ $(i = 1, ..., 10, \text{iid})$, μ_{ji} and d_{ji} are the same as in Simulation 1.

Then we consider two more cases between Simulation 1 and 2, normally distributed data with different variances (Simulation 3) and heavy-tailed distributed data with equal variance (Simulation 4). Simulation 3 is the same as Simulation 2 in model (4.2) except that we generate ϵ_{ji} from $N(0,1)$. Simulation 4 is the same as Simulation 1 in model (4.1) except that we generate ϵ_{ji} from $t_{(3)}/\sqrt{3}$ distribution.

Furthermore we consider the bigger degrees of freedom for t distribution error terms when assuming equal variance 1 across all probe sets. We perform Simulation 5 using model (4.1) where $\epsilon_{ji} \sim t_{(5)}/\sqrt{5/3}$, and Simulation 6 where $\epsilon_{ji} \sim t_{(10)}/\sqrt{10/8}$. In Simulation 7, we assume error terms follow symmetric Laplace distribution with mean 0 and parameter $b = 1$ divided by its standard deviation $\sqrt{2}$, i.e. $\epsilon_{ji} \sim$ Laplace(0,1)$/\sqrt{2}$.

In summary, Table 4.1 lists the simulations we performed by error term distribution and variability of probe sets. When all probe sets have the common variance 1, we consider gene expressions following normal distribution N(0,1), t distribution with different degrees of freedom and Laplace(2,1) separately. When variances across all probe sets are varying, we only consider gene expression following normal distribution N(0,1) and t distribution with 3 degrees of freedom. Each simulation can be implemented as follows:

- Step 1: Generate a simulated data set using an appropriate model.
- Step 2: Apply methods listed in Tables 3.3 and 3.4 to the data sets generated in Step 1, record the number of probe sets called significant and the number of probe sets falsely called significant, and calculate true FDR, for each method.
- Step 3: Repeat 100 times Step 1 and Step 2. For each method, compute the average number of probe sets called significant, the average number of probe sets falsely called significant and the average of true FDRs.

4.2 Power simulation

To investigate the powers of the methods presented in Tables 3.3 and 3.4 for detecting true differences between strains, we conducted power simulation where we generated 100 simulated data sets of 13200 genes from 2 strains of 5 samples each. Similar to Simulation 1, gene expression values came from normal distribution and variances σ_j^2 were fixed at 1. There are 3200 truly differentially expressed probe sets so the proportion of true null hypotheses is $(1 - 3200/13200) \times 100\% = 75.76\%$. True negative differences are the set of $A_1 = \{$-4, -3.5, -3, -2.5, -2, -1.5, -1, -0.5$\}$ and true positive differences are the set of $A_2 = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$. Then all true differences are the union of sets A_1, A_2 and $\{0\}$. Each true difference point is denoted by δ_k where $k = 1, 2, ..., 17$ and $\delta_k \in A_1 \cup A_2 \cup \{0\}$. The median of true differences δ_9=0 indicates probe sets are non-differentially expressed. Increasing true differences allowed us to test robustness of the methods. For the first 1600 truly differentially expressed probe sets, an amount in set A_1 was added to the expression of samples. For the second 1600 probe sets, an amount in set A_2 was added to the expression of samples. There are 200 probe sets at every true difference point δ_k excluding $k = 9$ and 10000 probe sets at δ_9=0. Therefore the expression for probe set j from the ith sample can be modeled as

$$y_{ji}^k = \mu_{ji} + d_{ji}^k + \epsilon_{ji}, \tag{4.3}$$

where μ_{ji}=2, k=1,2,...,17,

$$d_{ji}^k = \begin{cases} 0 & j = 1, ..., 13200; i = 1, ..., 5 \\ \delta_{ji}^1 & j = 1, ..., 200, i = 6, ..., 10 \\ \cdots & \\ \delta_{ji}^8 & j = 1401, ..., 1600 \text{ and } i = 6, ..., 10 \\ \delta_{ji}^{10} & j = 1601, ..., 1800 \text{ and } i = 6, ..., 10 \\ \cdots & \\ \delta_{ji}^{17} & j = 3001, ..., 3200 \text{ and } i = 6, ..., 10 \\ 0 & j = 3201, ..., 13200 \text{ and } i = 6, ..., 10 \end{cases}$$

and $\epsilon_{ji} \sim N(0, 1)$ $(i = 1, ..., 10, \text{iid})$. For the first 3200 probe sets, the model (4.3) makes 200 probe sets differentially expressed at each true difference point with the same amount of differential expressions. The remaining 10000 probe sets are equally expressed.

To study the effect of true difference in expression on power, we concentrated on the three specific effect sizes, denoted by $E_1 = \pm 4$, $E_2 = \pm 3$ and $E_3 = \pm 2$. The signs +

and - represent the effect size in positive and negative directions respectively. Here we only generated one data set for each E_l, $l = 1, 2, 3$, and assigned 500 probe sets at the positive and negative effect sizes respectively. We refer to RUN1 for $E_1 = \pm 4$, RUN2 for $E_2 = \pm 3$ and RUN3 for $E_3 = \pm 2$. The expression for probe set j from the ith sample in the lth run can be modeled as

$$y_{ji} = \mu_{ji} + d_{ji} + \epsilon_{ji}, \tag{4.4}$$

where μ_{ji}=2,

$$d_{ji} = \begin{cases} -|E_l| & j = 1, ..., 500; i = 6, ..., 10 \\ |E_l| & j = 501, ..., 1000; i = 6, ..., 10 \\ 0 & \text{otherwise.} \end{cases}$$

$\epsilon_{ji} \sim N(0, 1)$ $(i = 1, ..., 10; j = 1, ..., m, \text{iid})$. Model (4.4) makes the first 500 probe sets negatively differentially expressed with the same amount of differential expressions, the second 500 probe sets positively differentially expressed with the same amount of differential expressions and the remaining 9000 probe sets equally expressed.

4.3 Results

4.3.1 Simulations 1-4

Simulation 1. Here gene expressions are normally distributed and share a common variance across probe sets. From Table 4.2, we observe that the methods based on d-statistic detect many more significant probe sets with less false positives so that they control true FDRs at the relatively lower level. Therefore the methods based on d-statistic appear to be more powerful in this case. For any specific statistic, FDR method and permutation method come up with fairly similar results in detection of significant probe sets and controlling of FDR; the distribution methods can detect more significant probe sets and control the true FDR close to the nominal level 5%.

When employing d distribution method to calculate q-value, we proposed four estimates $s^2_{j;pool}$, $\bar{s}^2_{pool}$, $s^2_{j;null}$ and $\bar{s}^2_{null}$ of unknown σ^2_j. In this simulation, true σ^2_j was fixed at 1 and hence all related quantities based on different estimates reported in Table 4.3 must be close to ones from true σ^2_j to show a good performance. It is clear from Table 4.3 that when cutoff level is at 0.05, DIS.d.d1 yields much greater number of significant probe sets than DIS.d. Among those probe sets, however, quite a lot of them are truly non-differentially expressed, leading to a rather high FDR of about 20%. Methods DIS.d.d2 and DIS.d.d3 are too conservative and detect very few significant probe sets. Since $\bar{s}^2_{pool}$ and $\bar{s}^2_{null}$ assume the common variance across probe sets, their performances are quite close to DIS.d. In particular, DIS.d.d4 based on $\bar{s}^2_{pool}$ is the best in this case.

Here we will take a closer look at the performance of methods for a single simulated data set.

In Figure 4.1(a)-(c), we make comparison on the q-value performance across three methods based on d-statistics. We can see that FDR.d and PERM.d display the similar performance. In Figure 4.1(a) we see the q-values from both methods are very close and fall in the diagonal. In Figure 4.1 (b) and (c) we also see that the q-values from DIS.d method are generally a little smaller than those from the other two methods. Methods based on t-statistics have the same pattern of q-values.

Figure 4.2(a)-(c) presents the relationships of q-values among methods based on different statistics. Vertical line in each panel represents the cutoff level. We further observe that q-values based on t-statistic are bigger than d-statistic, implying that meth-

Simulations		d-statistic			t-statistic		
		FDR.d	PERM.d	DIS.d	FDR.t	PERM.t	DIS.t
Simulation 1	#sig	461.70	459.94	557.50	166.69	162.24	189.06
	V	11.27	11.01	26.43	7.46	7.14	9.28
	FDR	0.0244	0.0239	0.0473	0.0438	0.0428	0.0487
Simulation 2	#sig	648.63	647.49	527.63	532.28	530.87	445.82
	V	22.54	22.19	5.84	23.09	22.80	12.50
	FDR	0.0347	0.0342	0.0110	0.0433	0.0429	0.0280
Simulation 3	#sig	404.36	402.30	559.70	166.30	161.78	188.63
	V	10.13	9.89	26.55	7.44	7.14	9.24
	FDR	0.0250	0.0245	0.0474	0.0437	0.0428	0.0486
Simulation 4	#sig	660.49	659.32	535.46	532.10	530.71	445.55
	V	22.12	21.70	5.44	23.07	22.79	12.46
	FDR	0.0334	0.0329	0.0101	0.0433	0.0429	0.0279

Table 4.2: *Comparison of 6 methods between Simulations 1-4 at q-value cutoff 0.05. Simulation 1:* $y_{ji} \sim N(0,1)$ *and* $\sigma_j^2 = 1$*; Simulation 2:* $y_{ji} \sim t(3)/\sqrt{3}$ *and* $\sigma_j^2 \sim \chi^2_{(20)}/20$*; Simulation 3:* $y_{ji} \sim N(0,1)$ *and* $\sigma_j^2 \sim \chi^2_{(20)}/20$*; Simulation 4:* $y_{ji} \sim t(3)/\sqrt{3}$ *and* $\sigma_j^2 = 1$.

		true σ^2	s^2_{jpool}, $\hat{c}^1_j$	s^2_{jpool}, $\hat{c}^2_j$	s^2_{jnull}, $\hat{c}^3_j$	$\bar{s}^2_{pool}$, $\hat{c}^4_j$	$\bar{s}^2_{null}$, $\hat{c}^5_j$
Simulations		DIS.d	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
Simulation 1	#sig	557.50	732.84	0.02	0.05	557.84	502.37
	V	26.43	165.78	0	0	26.47	16.35
	FDR	0.0473	0.2260	0	0	0.0474	0.0325
Simulation 2	#sig	527.63	656.93	508.57	455.27	519.16	500.44
	V	5.84	64.72	12.73	12.28	5.76	4.59
	FDR	0.0110	0.0984	0.0250	0.0269	0.0110	0.0091
Simulation 3	#sig	559.70	737.20	0.26	1.03	543.70	484.23
	V	26.55	168.01	0	0	32.87	20.89
	FDR	0.0474	0.2276	0	0	0.0603	0.0431
Simulation 4	#sig	534.57	673.88	496.32	452.62	536.05	514.32
	V	5.41	70.70	9.89	11.60	5.63	4.05
	FDR	0.0101	0.1047	0.0199	0.0256	0.0104	0.0078

Table 4.3: *Comparison of estimates of σ_j^2 and c_j at q-value cutoff 0.05 between Simulations 1-4. Simulation 1: $y_{ji} \sim N(0,1)$ and $\sigma_j^2 = 1$; Simulation 2: $y_{ji} \sim t(3)/\sqrt{3}$ and $\sigma_j^2 \sim \chi^2_{(20)}/20$; Simulation 3: $y_{ji} \sim N(0,1)$ and $\sigma_j^2 \sim \chi^2_{(20)}/20$; Simulation 4: $y_{ji} \sim t(3)/\sqrt{3}$ and $\sigma_j^2 = 1$.*

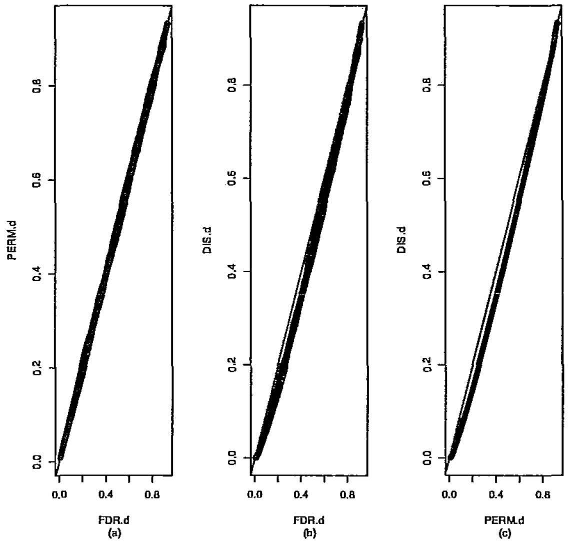


Figure 4.1: *Plots of pairwise relationship of q-values among three methods based on d-statistics for one data set in Simulation 1.*

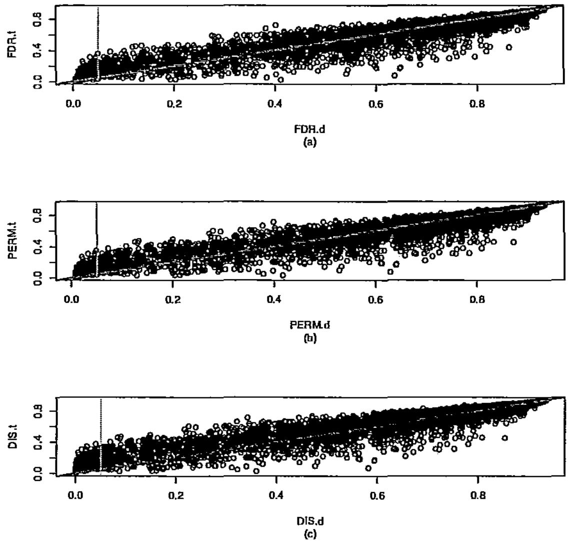


Figure 4.2: *The plots of relationships of q-values among FDR method, permutation method and distribution method based on different statistics for one data set.*

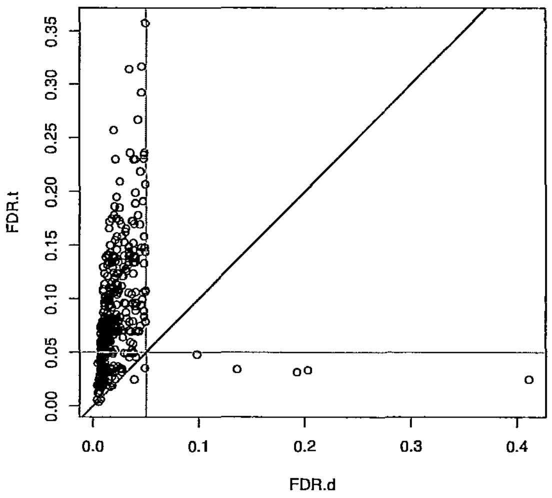


Figure 4.3: *Plot of the relationship of q-values from FDR method using different statistic for one data set.*

ods based on t-statistic will detect fewer significant probe sets at cutoff level 0.05. This justifies the finding in Table 4.1 where there are only a half and even less number of probe sets called significant from t-statistic. In order to figure out more reasons why t-statistic gives fewer significant genes than d-statistic, we will take FDR method for example.

Figure 4.3 shows the relationship of the q-values based on two different statistics. This plot is separated into four regions by cutoff levels. Some probe sets with small q-values ($<= 0.05$) based on both two statistics fall in the lowest left region. Among them q-values based on t-statistic are uniformly bigger than based on d-statistic except two probe sets. For the probe sets with small q-value ($<= 0.05$) based on t-statistic, there are only five probe sets with large q-values (> 0.05) based on d-statistic, whereas for the probe sets with small q-values based on d-statistic, there are lots of probe sets with large q-values based on t-statistic. Under the same cutoff level, methods based on t-statistic tend to generate a larger q-value for the same probe set so that they can't detect as many as possible number of significant probe sets. Within the upper left region, there are 302 probe sets with small q-values based on d-statistic but large q-values based on t-statistic. If we choose the method based on t-statistic, there are 296 truly differentially expressed probe sets which are unable to be detected. This feature explains why t-statistic gives fewer differentially expressed probe sets than d-statistic.

In Figure 4.3 the five probe sets with small q-value based on t-statistic but large q-value based on d-statistic are ID 1535, 5197, 5493, 7291 and 7848, which fall in the lower right region. If we choose a method based on t-statistic and set the same cutoff level, then these five probe sets must be called significant. However, these probe sets are actually non-differentially expressed since only the first 1000 genes are differentially expressed. Take probe set ID 1535 for example. It is extremely different between the values of d-statistic (0.6457) and t-statistic (6.2815) and the corresponding q-values as well (q-value based on d is 0.6457 and one based on t is 0.02474). Given the d and t statistics and $s_0 = 1.3421$, we can get the difference expression $\bar{x}_{j1} - \bar{x}_{j2} = 0.96$ and standard error $s_j = 0.153$ where $j = 1535$. Such a small standard error makes the t-statistic quite large which would mislead us to call this probe set significant. However adding extra term s_0 can damp down this effect and the associated d-statistic can be sharply decreased so that we can correctly call this probe set non-significant. Since probe set 1535 is actually non-differentially expressed, the conclusion based on d-statistic is correct, while conclusion based on t-statistic is incorrect. It indicates that method based on t-statistics is more likely detect some false positives.

Take another probe set ID 903 for example, which falls in the upper left region in figure 4.3. The two test statistics for this probe set are slightly different ($d =$ -1.1036, $t =$ -2.5932) but the q-values based on them are quite different (q-value based on d is 0.0455 and one based on t is 0.3166). Then we can get $\bar{x}_{j1} - \bar{x}_{j2} =$ -2.85 and $s_j = 1.1$. Since probe set 903 is actually differentially expressed, the conclusion based on d-statistic that it is significant is correct, whereas the conclusion based on t-statistic is wrong. This indicates that methods based on t-statistic will miss some truly differentially expressed probe sets. Therefore above two examples give the good evidences to explain why it is necessary to add s_0 in the denominator of regular t-statistic.

Simulation 2. We generated gene expressions from heavy-tailed t-distribution and allowed variance to differ for each probe set. From Table 4.2, we observe that methods based on d-statistics are more powerful and they can identify more significant probe sets and control the FDRs at the relative lower levels, and that FDR method and permutation method are almost the same in performance. These findings are similar to ones in Simulation 1. However, distribution methods (DIS.d and DIS.t) become conservative in this simulation because they identify less significant probe sets and control the FDRs at quite low level compared to the other two methods based on the same test statistic. We also observe from Table 4.3 that DIS.d.d1 method is still liberal. DIS.d.d2 and DIS.d.d3 methods can work here, but have false discovery rate larger than that of DIS.d. DIS.d.d4 and DIS.d.d5 methods assuming a common variance across probe sets are still close to DIS.d method, even when probe sets do not have common variance.

Comparison between Simulation 1 and Simulation 2. From Table 4.2, we observe that in simulation 2 both FDR.d and PERM.d methods identify more significant probe sets with more false positives, and that the FDRs are closer to nominal FDR

5%. When gene expressions follow normal distribution in simulation 1, DIS.d method identifies most number of significant probe sets among all methods and controls FDR closest to nominal. However, when gene expressions follow a heavy-tailed distribution in simulation 2, it becomes conservative and controls FDR at fairly low level (only 1%). The methods based on t-statistic exhibit quite different performance between two simulations. There are around 500 probe sets called significant in simulation 2 while only about one-third as many probe sets called significant in simulation 1. FDR.t and PERM.t methods control FDRs at almost the same level (about 4%) in two simulations. DIS.t method seems more accurate in simulation 2 because it controls FDR at much lower level (2.8%) when identifying more significant probe sets.

From the above comparison, those methods (except DIS.d) in Simulation 2 appear more liberal. Table 4.4 and Figures 4.4-4.7 can give some further explanations. Table 4.4 shows the distributions of estimated proportion of non-differential expressions $\hat{\pi}_0$, fudge factor s_0 and associated percentile. Compared to Simulation 2, we find out that $\hat{\pi}_0$ in simulation 1 is more overestimated which might produce bigger q-value for each probe set so that smaller number of probe sets can be identified. We also find out that s_0 in simulation 1 was taken by the maximum among standard errors s_j, while s_0 in simulation 2 was chosen at less than the 5^{th} percentile of s_j. Such a large value of s_0 tends to shrink down d-statistic in equation (3.1) and results in smaller values of d-statistic. Therefore fewer number of probe sets are called significant in simulation 1.

Next we will take a closer look at the distributions of four quantities for a particular data set in each simulation: standard error s_j, difference in expression(diff_j), t-statistic and d-statistic, for all probe sets. It can be seen from Figure 4.4 that the distribution of s_j in simulation 2 are skewed with very long tail, while distribution in simulation 1

Quantity	Simulation	Min	1stQu	Median	Mean	3rdQu	Max
$\hat{\pi}_0$	Simulation 1	0.9004	0.9185	0.9228	0.9241	0.9306	0.9518
	Simulation 2	0.8862	0.9033	0.9107	0.9104	0.9154	0.9330
	Simulation 3	0.8974	0.9192	0.9249	0.9251	0.9298	0.9538
	Simulation 4	0.8892	0.9052	0.9104	0.9115	0.9166	0.9388
s0	Simulation 1	1.200	1.254	1.292	1.290	1.320	1.422
	Simulation 2	0.1698	0.1753	0.1776	0.1791	0.1793	0.2366
	Simulation 3	0.8237	1.1160	1.4980	1.4120	1.5640	1.7880
	Simulation 4	0.1895	0.1976	0.2164	0.2150	0.2233	0.2825
s0.percentile	Simulation 1	100	100	100	100	100	100
	Simulation 2	1	1	1	1.12	1	5
	Simulation 3	88	99	100	99.51	100	100
	Simulation 4	1	1	2	1.96	2	8

Table 4.4: *The comparison of the distributions of* π_0, s_0 *and* s_0*.percentile between Simulations 1-4. Simulation 1:* $y_{ji} \sim N(0,1)$ *and* $\sigma_j^2 = 1$*; Simulation 2:* $y_{ji} \sim t(3)/\sqrt{3}$ *and* $\sigma_j^2 \sim \chi^2_{(20)}/20$*; Simulation 3:* $y_{ji} \sim N(0,1)$ *and* $\sigma_j^2 \sim \chi^2_{(20)}/20$*; Simulation 4:* $y_{ji} \sim t(3)/\sqrt{3}$ *and* $\sigma_j^2 = 1$.

are symmetrical, and from Figure 4.5 that there are most of probe sets concentrated within ±2 in both simulations but simulation 2 has longer tail in negative direction. In figure 4.6 and 4.7, we can see that when s_0 was taken 0, the distribution of t-statistics in simulation 2 in Figure 4.6 has longer tails; When adding s_0, the distribution of d-statistics in simulation 2 spread out with longer tails, while distribution in simulation 1 is more concentrated around 0. These figures altogether explain why simulation 2 can detect many more significant probe sets.

Comparison across Simulations 1-4. From Table 4.2 and Table 4.3 we observe that simulation 1 and 3 where gene expressions follow normal distribution are similar in performance, and that simulation 2 and 4 where gene expressions follow a heavy-tailed distribution are similar in performance. We also observe that whether or not each probe set has different variance does not have much effect on the performance of methods. The similarities can also be seen from the performance of $\hat{\pi}_0$, s_0 and associated percentiles in Table 4.4, the distributions and differences in expression, standard errors, t-statistics and d-statistics in Figure 4.4, 4.5, 4.6 and 4.7.

4.3.2 Simulations 5-7

In this section we only consider the cases where gene expressions are generated from t distribution with different degrees of freedom, and assume probe sets share a common variance. We also examine the performance of methods in Simulation 7 where gene expressions were generated from a symmetric Laplace distribution with mean 0 and scale parameter 1. Table 4.5 shows the comparison across Simulation 5,6 and 7 with respect to the number of significant probe sets, false positives and true FDR. Comparing

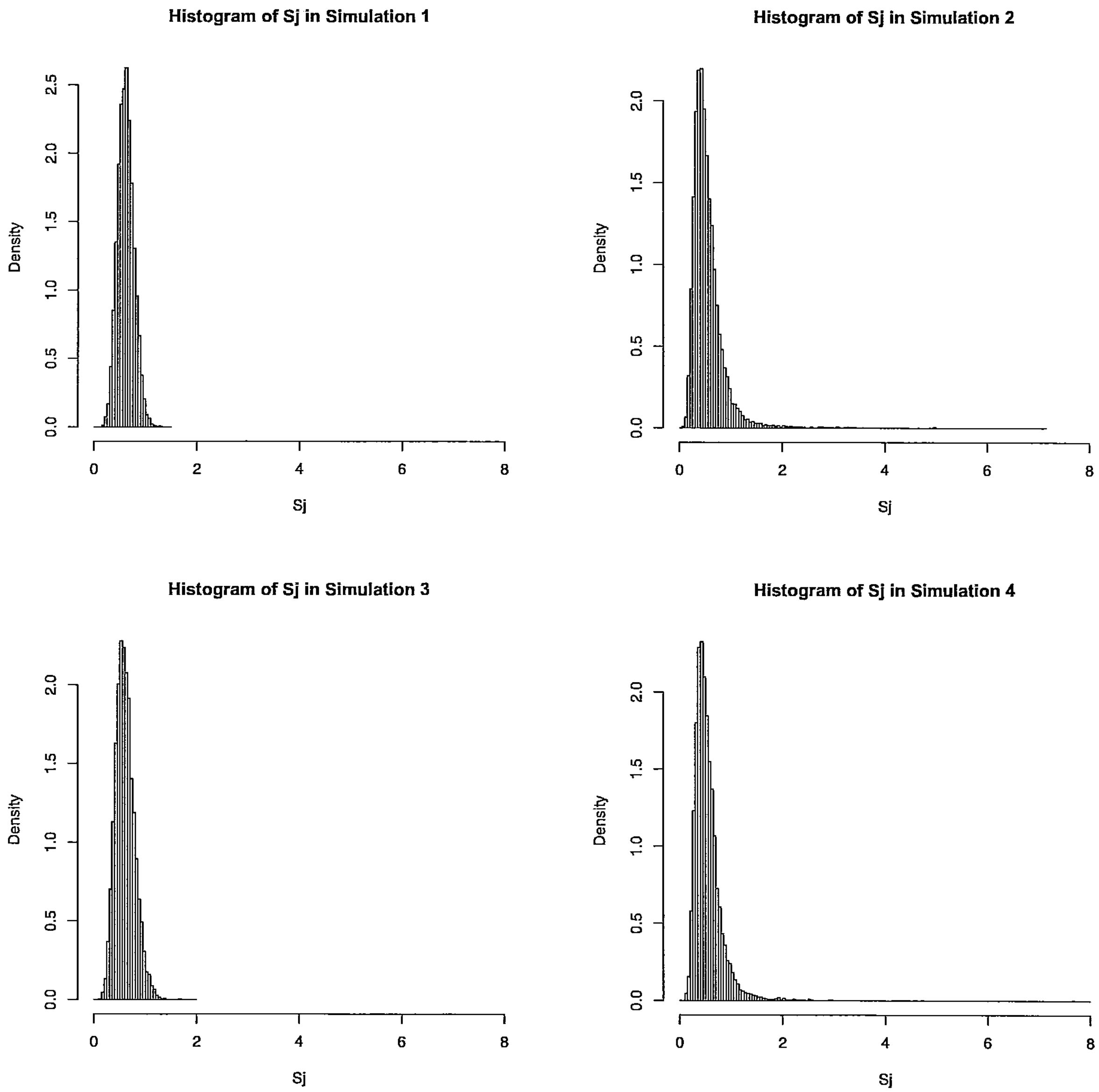


Figure 4.4: *The histogram of the distribution of standard errors of all probe sets for the first data set in Simulations 1-4.*

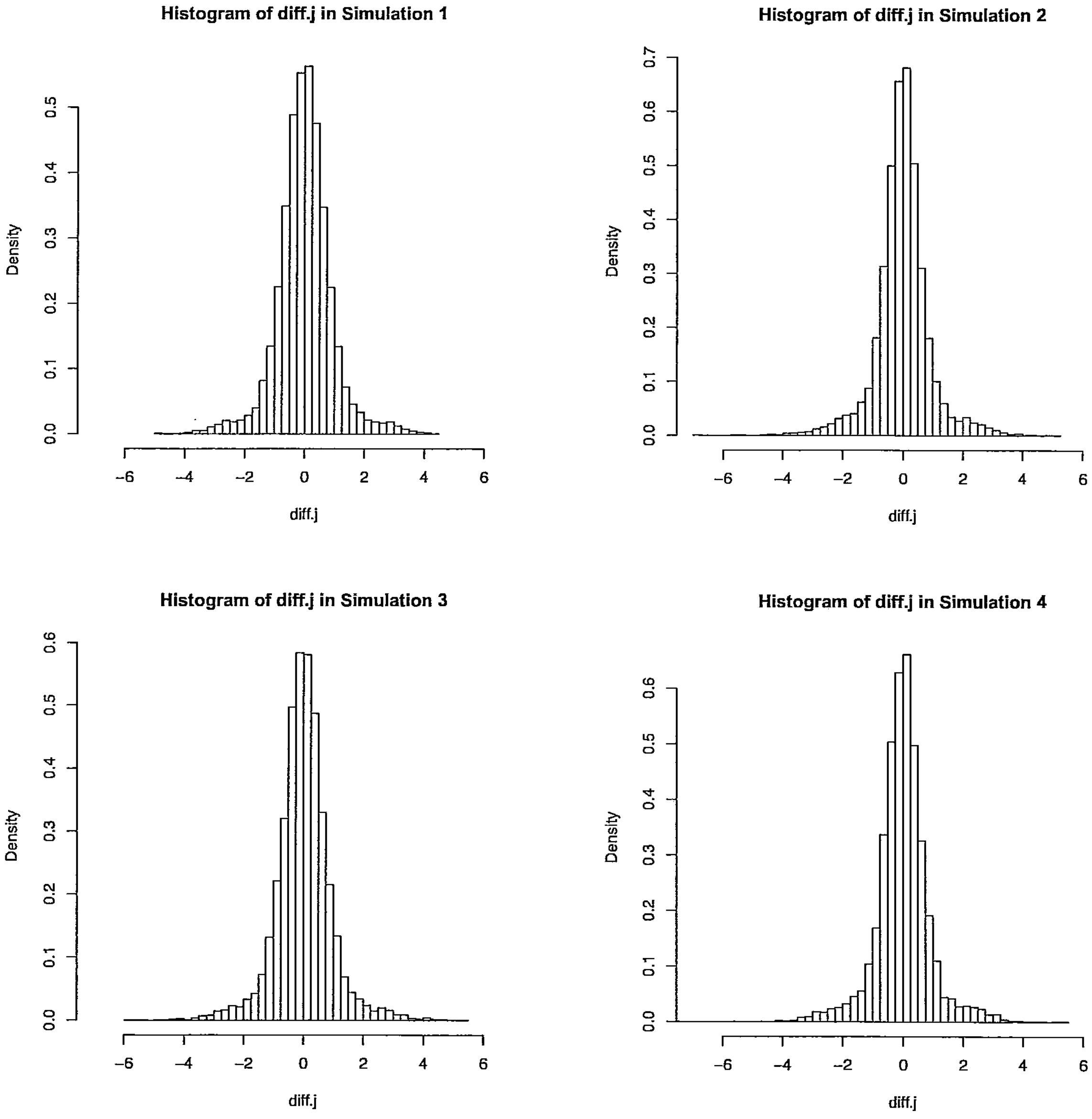
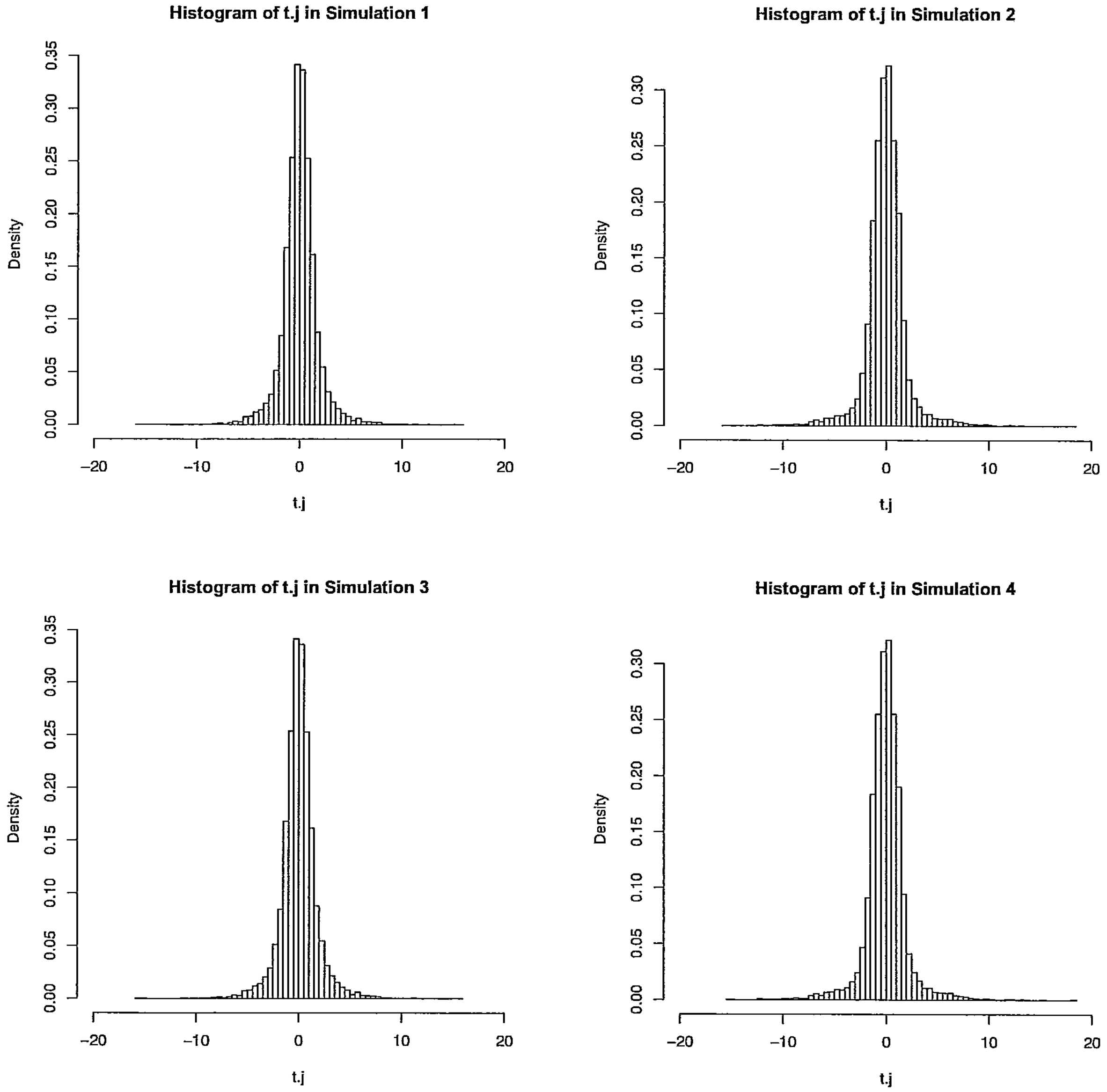


Figure 4.5: *The histogram of the distribution of differences in expression of all probe sets for the first data set in Simulations 1-4.*



Histogram of t.j in Simulation 1

Histogram of t.j in Simulation 2

Histogram of t.j in Simulation 3

Histogram of t.j in Simulation 4

Figure 4.6: *The histogram of the distribution of t-statistics of all probe sets for the first data set in Simulations 1-4.*

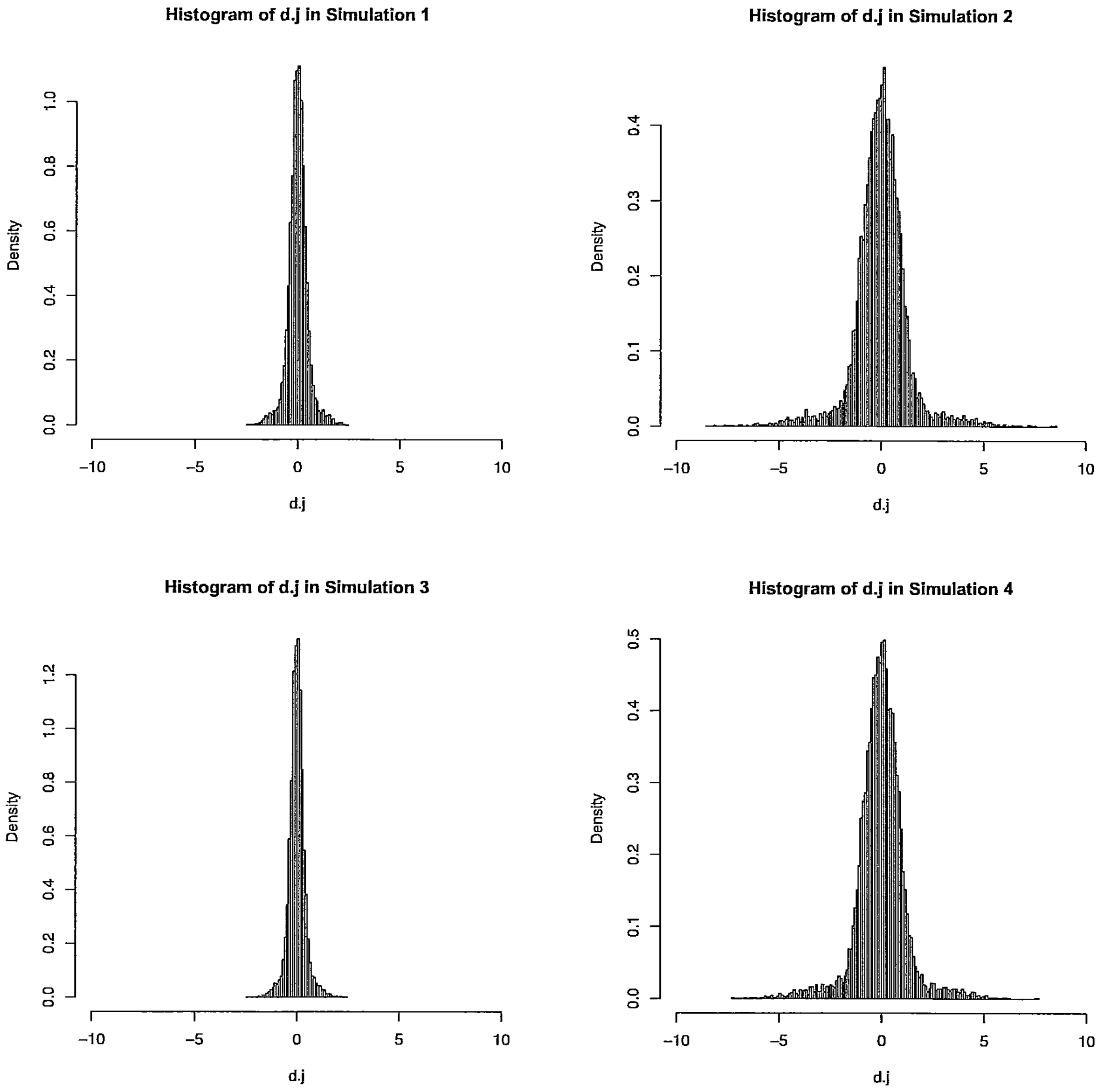


Figure 4.7: *The histogram of the distribution of d-statistics of all probe sets for the first data set in Simulations 1-4.*

Simulation 5 and 6 in Table 4.5 with Simulation 4 in Table 4.2, we observe that when degrees of freedom are increased, number of significant probe sets and false positives and true FDR level are monotonically decreased in FDR.d, PERM.d, FDR.t and PERM.t methods. In DIS.d method, the number of false positives and FDR level have positive relationship with degrees of freedom, but the number of significant probe sets does not display any clear relationship. In DIS.t method, the number of significant probe sets and false positives have negative relationship with degrees of freedom yet FDR level has positive relationship. We also see that when degrees of freedom is as large as 10 (Simulation 6), the performances of all six methods are quite close to Simulation 1 where gene expressions are normally distributed. This is due to the fact that t-distribution approaches the normal distribution as degrees of freedom are increased. In addition, Simulation 7 where data was generated from Laplace distribution is between Simulation 4 and Simulation 5 in terms of the number of significant probe sets and false positives and true FDR.

Table 4.6 presents the impact of gene expression distribution on the performance of the estimates when applying d-distribution method to calculate q-value. We observe that DIS.d.d1 is always liberal no matter which distribution gene expressions follow. DIS.d.d2 and DIS.d.d3 method can not work when expressions approximately follow normal distribution, but can identify significant probe sets when gene expression distribution have heavy tails. Since DIS.d.d4 and DIS.d.d5 assume a common variance across probe sets, they are very close to DIS.d in various data set. As degrees of freedom are decreasing, false discovery rates in both methods become lower and lower.

Table 4.7 shows the performances of distributions of estimated proportion of non-differential expressions $\hat{\pi}_0$, fudge factor s_0 and associated percentile among Simulation

Simulations	Test statistic	d-statistic			t-statistic		
		FDR.d	PERM.d	DIST.d	FDR.t	PERM.t	DIS.t
Simulation 5	#sig	540.66	539.23	526.99	338.97	337.25	288.58
	V	15.04	14.71	12.83	14.46	14.02	9.63
	FDR	0.0277	0.0272	0.0243	0.0426	0.0415	0.0333
Simulation 6	#sig	472.13	471.02	565.62	238.22	234.29	218.98
	V	9.37	9.33	24.49	9.97	9.52	8.18
	FDR	0.0197	0.0197	0.0432	0.0416	0.0403	0.0370
Simulation 7	#sig	551.72	550.42	502.76	407.60	405.93	296.43
	V	15.87	15.68	9.37	16.16	15.92	6.27
	FDR	0.0288	0.0285	0.0186	0.0395	0.0391	0.0210

Table 4.5: *Comparison of 6 methods between Simulation 5-7 at q-value cutoff 0.05. Simulation 5: $y_{ji} \sim t(5)/\sqrt{5/3}$ and $\sigma_j^2 = 1$; Simulation 5: $y_{ji} \sim t(10)/\sqrt{10/8}$ and $\sigma_j^2 = 1$; Simulation 7: $y_{ji} \sim Laplace(0,1)/\sqrt{2}$ and $\sigma_j^2 = 1$.*

		true σ^2	$s^2_{jpool}, \hat{c}^1_j$	$s^2_{jpool}, \hat{c}^2_j$	$s^2_{jnull}, \hat{c}^3_j$	$\bar{s}^2_{pool}, \hat{c}^4_j$	$\bar{s}^2_{null}, \hat{c}^5_j$
Simulations		DIS.d	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
Simulation 5	#sig	526.99	682.96	149.49	234.10	526.54	487.00
	V	12.83	111.45	0.20	3.80	12.77	8.05
	FDR	0.0243	0.1628	0.0012	0.0159	0.0242	0.0165
Simulation 6	#sig	565.62	744.45	0.02	0.67	566.21	509.17
	V	24.49	158.15	0	0	24.65	14.32
	FDR	0.0432	0.2121	0	0	0.0434	0.0281
Simulation 7	#sig	502.76	624.71	256.01	274.96	502.67	469.87
	V	9.37	79.18	0.92	3.35	9.38	6.29
	FDR	0.0186	0.1264	0.0033	0.0121	0.0186	0.01334

Table 4.6: *Comparison of estimates of* σ^2 *and* c_j *at q-value cutoff 0.05 between Simulation 5-7. Simulation 5:* $y_{ji} \sim t(5)/\sqrt{5/3}$ *and* $\sigma_j^2 = 1$*; Simulation 6:* $y_{ji} \sim t(10)/\sqrt{10/8}$ *and* $\sigma_j^2 = 1$*; Simulation 7:* $y_{ji} \sim Laplace(0,1)/\sqrt{2}$ *and* $\sigma_j^2 = 1$.

Quantity	Simulation	Min	1stQu	Median	Mean	3rdQu	Max
$\hat{\pi}_0$	Simulation 5	0.8948	0.9127	0.9190	0.9189	0.9246	0.9436
	Simulation 6	0.8910	0.9179	0.9234	0.9236	0.9298	0.9442
	Simulation 7	0.8930	0.9100	0.9159	0.9154	0.9232	0.9382
s0	Simulation 5	0.4075	0.5502	0.5950	0.5919	0.6368	0.7529
	Simulation 6	1.076	1.558	1.681	1.660	1.812	2.793
	Simulation 7	0.2998	0.3984	0.4395	0.4448	0.4966	0.5816
s0.percentile	Simulation 5	16	46	56	54.67	64.25	81
	Simulation 6	99	100	100	99.89	100	100
	Simulation 7	5	17	24	25.84	35	52

Table 4.7: *The comparison of the distributions of* π_0, s_0 *and* s_0*.percentile between between Simulations 5-7. Simulation 5:* $y_{ji} \sim t(5)/\sqrt{5/3}$ *and* $\sigma_j^2 = 1$*; Simulation 5:* $y_{ji} \sim t(10)/\sqrt{10/8}$ *and* $\sigma_j^2 = 1$*; Simulation 7:* $y_{ji} \sim Laplace(0,1)/\sqrt{2}$ *and* $\sigma_j^2 = 1$.

5,6 and 7. We observe that as degrees of freedom increasing, $\hat{\pi}_0$ is more overestimated, s_0 is increased and taken the larger percentile of s_j. From Table 4.2 and 4.7, we also observe that Simulation 6 is close to Simulation 1 and Simulation 7 is between Simulation 4 and 5 in terms of the $\hat{\pi}_0$ and s_0, which are consistent with the findings in Table 4.6.

4.3.3 Power simulation

Since we know which probe sets are truly differentially expressed and how many probe sets are truly differentially expressed at each true difference point, we can calculate the power for detecting true positives at each point. In this simulation, we define the power

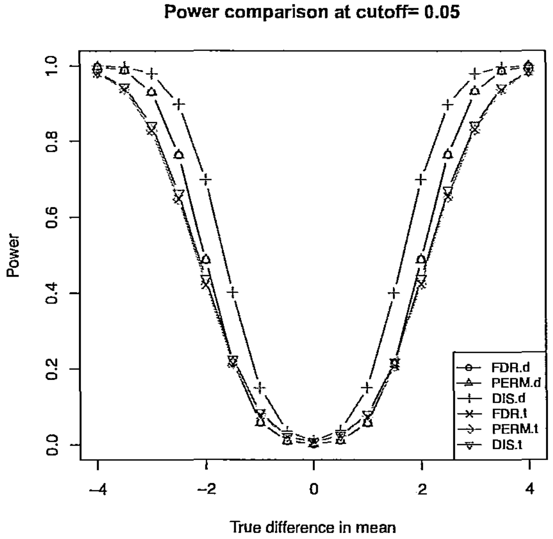


Figure 4.8: *Comparison of the powers for detecting differentially expressed genes at all true difference points across six methods when setting cutoff level 5%.*

as the ratio of the number of detected significant probe sets over 200 truly differential expressions. For non-differential expressions, we define the power as the proportion of probe sets rejected among them. Then the equation of power is defined by

$$\beta_k = \begin{cases} \frac{\#\{\text{detected significant probe sets}\}}{200} & \text{for every true difference point } \delta_k, \text{i.e.,} k \neq 9; \\ \frac{\#\{\text{rejected probe sets}\}}{13200-3200} & \text{for non-differential expressions, i.e.,} k = 9. \end{cases} \tag{4.5}$$

Figure 4.8 exhibits the "U" shape concave-down curves of the behavior of the power for detecting true differences from six methods. The X-axis represents the true difference points, and the Y-axis shows the proportion of such probe sets that were called significant with cutoff 0.05. For all methods, powers go up as absolute values of true differences increase. When absolute value of true difference exceed 4, the corresponding powers approach to 1. At the true difference -2.5 or 2.5, the powers sharply increase

k	Diff	FDR.d	PERM.d	DIS.d	FDR.t	PERM.t	DIS.t
1	-4.0	0.9983	0.9983	0.9998	0.9815	0.9816	0.9841
2	-3.5	0.9889	0.9889	0.9978	0.9385	0.9382	0.9454
3	-3.0	0.9302	0.9299	0.9801	0.8286	0.8290	0.8425
4	-2.5	0.7641	0.7637	0.8975	0.6486	0.6472	0.6643
5	-2.0	0.4870	0.4865	0.6977	0.4205	0.4205	0.4373
6	-1.5	0.2162	0.2160	0.4016	0.2132	0.2131	0.2251
7	-1.0	0.0581	0.0576	0.1499	0.0790	0.0790	0.0848
8	-0.5	0.0098	0.0097	0.0353	0.0199	0.0199	0.0220
10	0.5	0.0107	0.0107	0.0374	0.0230	0.0227	0.0244
11	1.0	0.0576	0.0571	0.1505	0.0751	0.0749	0.0804
12	1.5	0.2137	0.2130	0.4001	0.2077	0.2073	0.2183
13	2.0	0.4894	0.4879	0.6994	0.4238	0.4230	0.4393
14	2.5	0.7654	0.7648	0.8967	0.6550	0.6535	0.6719
15	3.0	0.9323	0.9319	0.9800	0.8316	0.8311	0.8434
16	3.5	0.9877	0.9874	0.9972	0.9369	0.9366	0.9429
17	4.0	0.9990	0.9990	0.9999	0.9848	0.9847	0.9871

Table 4.8: *Comparison of the powers for detecting differentially expressed probe sets at all true difference points across six methods with cutoff level 5%.*

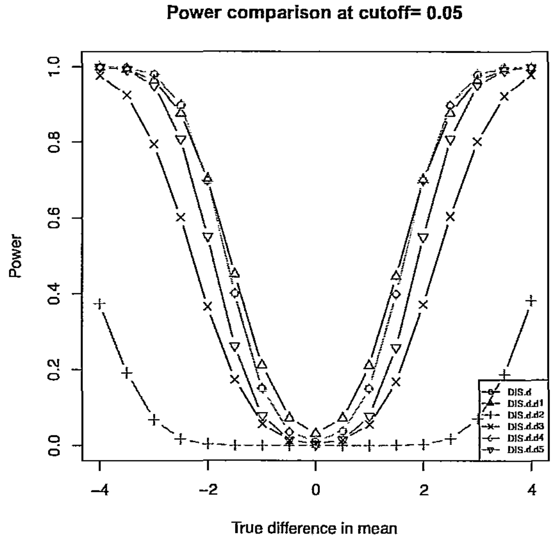


Figure 4.9: *Comparison of the powers for detecting differentially expressed genes at all true difference points across DIS.d, DIS.d.d1, DIS.d.d2, DIS.d.d3, DIS.d.d4 and DIS.d.d5 methods when setting cutoff level 5%.*

up to over 70% for methods based on d-statistic and over 60% for methods based on t-statistic, which can be verified by Table 4.8 listing the power at each true difference point for all methods. In Figure 4.8, we can also see that methods based on d-statistic are generally more powerful at every true difference point than methods based on t statistics. Particularly, DIS.d method is most powerful if population variance for each probe set is known. FDR.d and PERM.d methods have almost the same power behavior since their power plots are overlap. Similarly, FDR.t and PERM.t methods have the same power, and DIS.t method is slightly more powerful but not clearly.

To test the powers of methods based on estimates of unknown variance σ_j^2, we calculated the power at each true difference point (Table 4.9) and drew associated plots

k	Diff	DIS.d	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
1	-4.0000	0.9998	0.9989	0.3744	0.9771	0.9998	0.9990
2	-3.5000	0.9978	0.9932	0.1918	0.9239	0.9978	0.9930
3	-3.0000	0.9801	0.9649	0.0680	0.7955	0.9801	0.9498
4	-2.5000	0.8975	0.8764	0.0179	0.6023	0.8974	0.8091
5	-2.0000	0.6977	0.7027	0.0030	0.3659	0.6982	0.5512
6	-1.5000	0.4016	0.4519	0.0002	0.1733	0.4017	0.2623
7	-1.0000	0.1499	0.2119	0.0001	0.0571	0.1502	0.0789
8	-0.5000	0.0353	0.0721	0.0000	0.0138	0.0355	0.0145
10	0.5000	0.0374	0.0723	0.0000	0.0149	0.0375	0.0156
11	1.0000	0.1505	0.2112	0.0000	0.0556	0.1505	0.0779
12	1.5000	0.4001	0.4465	0.0003	0.1688	0.4001	0.2591
13	2.0000	0.6994	0.7010	0.0030	0.3722	0.6992	0.5506
14	2.5000	0.8967	0.8762	0.0186	0.6043	0.8967	0.8083
15	3.0000	0.9800	0.9645	0.0708	0.8020	0.9801	0.9512
16	3.5000	0.9972	0.9921	0.1877	0.9215	0.9972	0.9912
17	4.0000	0.9999	0.9991	0.3817	0.9804	0.9999	0.9994

Table 4.9: *Comparison of the powers for detecting differentially expressed genes at all true difference points across DIS.d, DIS.d.d1, DIS.d.d2, DIS.d.d3, DIS.d.d4 and DIS.d.d5 methods with cutoff level 5%.*

(Figure 4.9) at q-value cutoff 0.05. Since we know the true variance σ_j^2, we may compare the power of each method to DIS.d at each point. If a method is closest to DIS.d, we may conclude that the method is most powerful. In Figure 4.9, we see that the power plot of DIS.d.d4 is identical to the power plot of DIS.d. It is due to the fact that we assume all gene expressions share a common variance in power simulation design. The power plot DIS.d.d1 is quite close to DIS.d when the absolute value of true difference is larger than 2. However it includes many false positives when true difference is extremely small. Both the power plots of DIS.d.d3 and DIS.d.d5 fall under DIS.d plot. Compared to DIS.d.d3, DIS.d.d5 is closer to DIS.d since it assumes a common variance across probe sets. DIS.d.d2 is least powerful because it is far down away DIS.d.

For the 10,000 non-differentially expressed probe sets ($\delta_9 = 0$), we estimated the number of probe sets falsely declared significant among these. This is essentially an estimate of the Type I error after adjustment for multiple testing. A Bonferroni family-wise error rate correction would use a level of 0.05/10000 for this so we would expect the power curve to approach 0.000005. In our study, however, we are still using the more liberal FDR control and so the power curves do not get this low. For the FDR and permutation methods of q-value calculation based on the d-statistic we get an estimated Type I error rate of 0.0019 and for that based on the distribution of the d-statistic using the known variance it is 0.0099. For the other methods in Table 4.8 based on the t-statistic the estimate are 0.0074 for the FDR and permutation based methods and 0.008 for the method based on the t distribution. When looking at the various methods for estimating the common variance when using the distribution of the d-statistic the method using the average pooled standard deviation (DIS.d.d4) gives the same estimate as using the true variance (0.0099). DIS.d.d1 has the highest estimate of Type I error

(0.0317) whereas DIS.d.d2 failed to declare any of these significant. Using DIS.d.d3 and DIS.d.d5 gave estimates of 0.0044 and 0.0031 respectively. These are in keeping with the results we found for the power to detect truly differentially expressed probe sets in that those methods with lower power also tended to make fewer Type I errors.

Table 4.10 and Figure 4.10 give the distributions of permuted d-statistics and t-statistics when true differences are ±4, ±3 and ±2 for 9,000 non-differential expressions. We observe that the distributions of permuted test statistics for non-differential probe sets are identical no matter how big effect size is. Furthermore we observe that given the same effect size, permuted d-statistics are more concentrated around 0, while permuted t-statistics spread out very much with longer tails. Such long tails could make us incorrectly call some probe set significant. Table 4.11 and Figure 4.11 show the distributions of permuted d-statistics and t-statistics for all probe sets. Adding differentially expressed probe sets, we observe that the tails of the distributions become longer as effect size is increased for both permuted test statistics. This indicates that more significant probe sets can be detected when effect size is larger.

Table 4.12 shows the comparison of six methods. At effect size of 4 or -4, methods based on d-statistic detect as many as possible significant probe sets with very few false positives and false negatives. DIS.d method not only detect all true positives but also contains some false positives. Methods based on t-statistics falsely call some probe sets significant and miss some truly differential expressions as well, although they detect more significant probe sets than methods based on d-statistics. At effect size of 3 or -3, however, methods based on d-statistic can detect more significant probe sets and control FDR at much lower level. At effect size of 2 or -2, methods based on t-statistic work badly that only about 100 probe sets are called significant and FDR is unable to

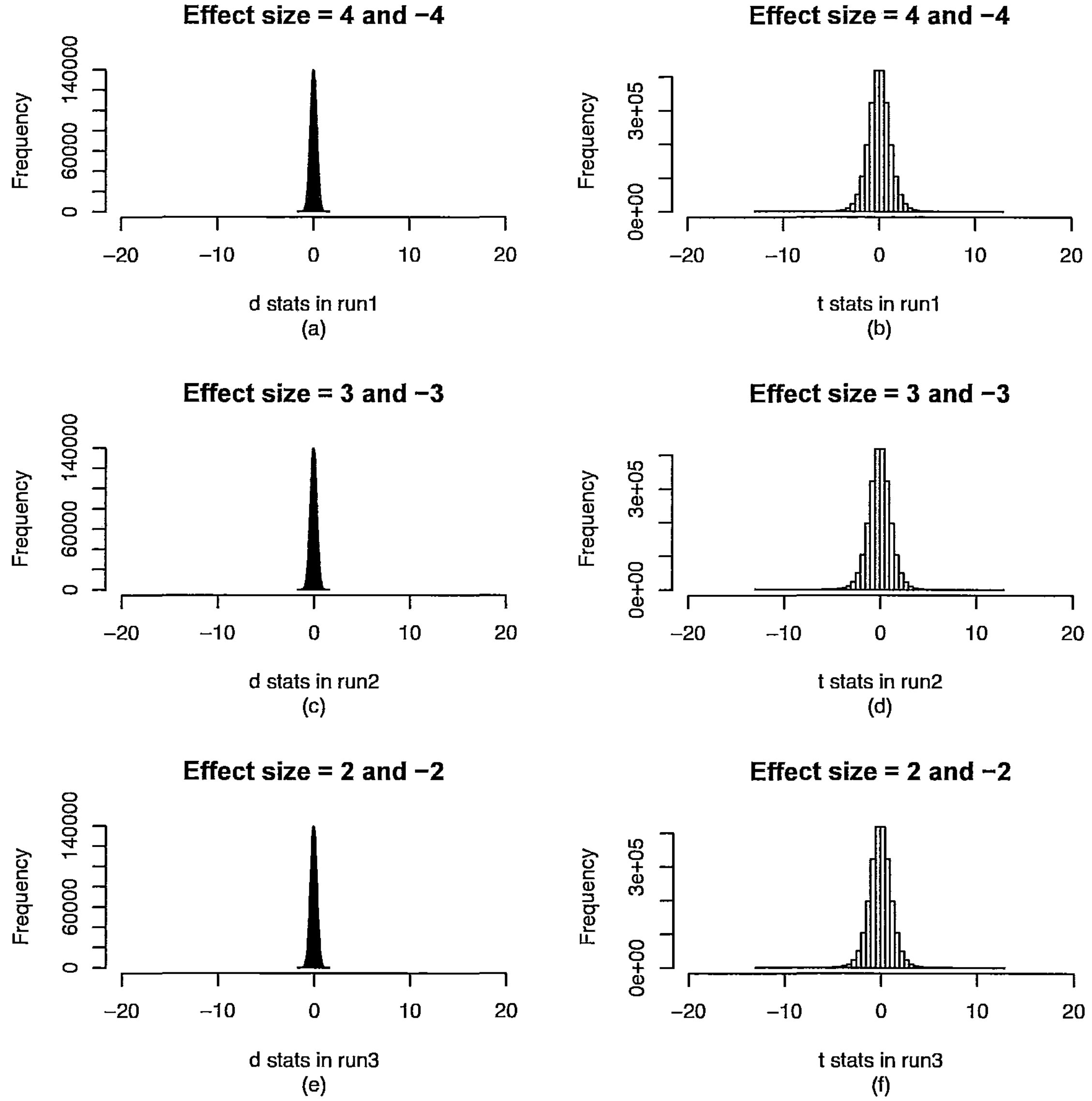


Figure 4.10: *Comparison of distributions of permuted d-statistics and t-statistics in runs 1-3 for 9,000 non-differentially expressed probe sets.*

	Run	Min	1stQu	Median	3rdQu	Max
d-statistic	Run 1	-1.6610	-0.2188	0	0.2188	1.6610
	Run 2	-1.6610	-0.2188	0	0.2188	1.6610
	Run 3	-1.6610	-0.2188	0	0.2188	1.6610
t-statistic	Run 1	-12.67	-0.7068	0	0.7068	12.67
	Run 2	-12.67	-0.7068	0	0.7068	12.67
	Run 3	-12.67	-0.7068	0	0.7068	12.67

Table 4.10: *Comparison of distributions of permuted d-statistics and t-statistics among runs 1-3 for 9,000 non-differentially expressed probe sets.*

	Run	Min	1stQu	Median	3rdQu	Max
d-statistic	Run 1	-3.449	-0.2295	0	0.2295	3.449
	Run 2	-2.814	-0.2267	0	0.2267	2.814
	Run 3	-2.178	-0.2238	0	0.2238	2.178
t-statistic	Run 1	-23.38	-0.7029	0	0.7029	23.38
	Run 2	-19.07	-0.7045	0	0.7045	19.07
	Run 3	-14.77	-0.7057	0	0.7057	14.77

Table 4.11: *Comparison of distributions of permuted d-statistics and t-statistics among runs 1-3 for all 10,000 probe sets*

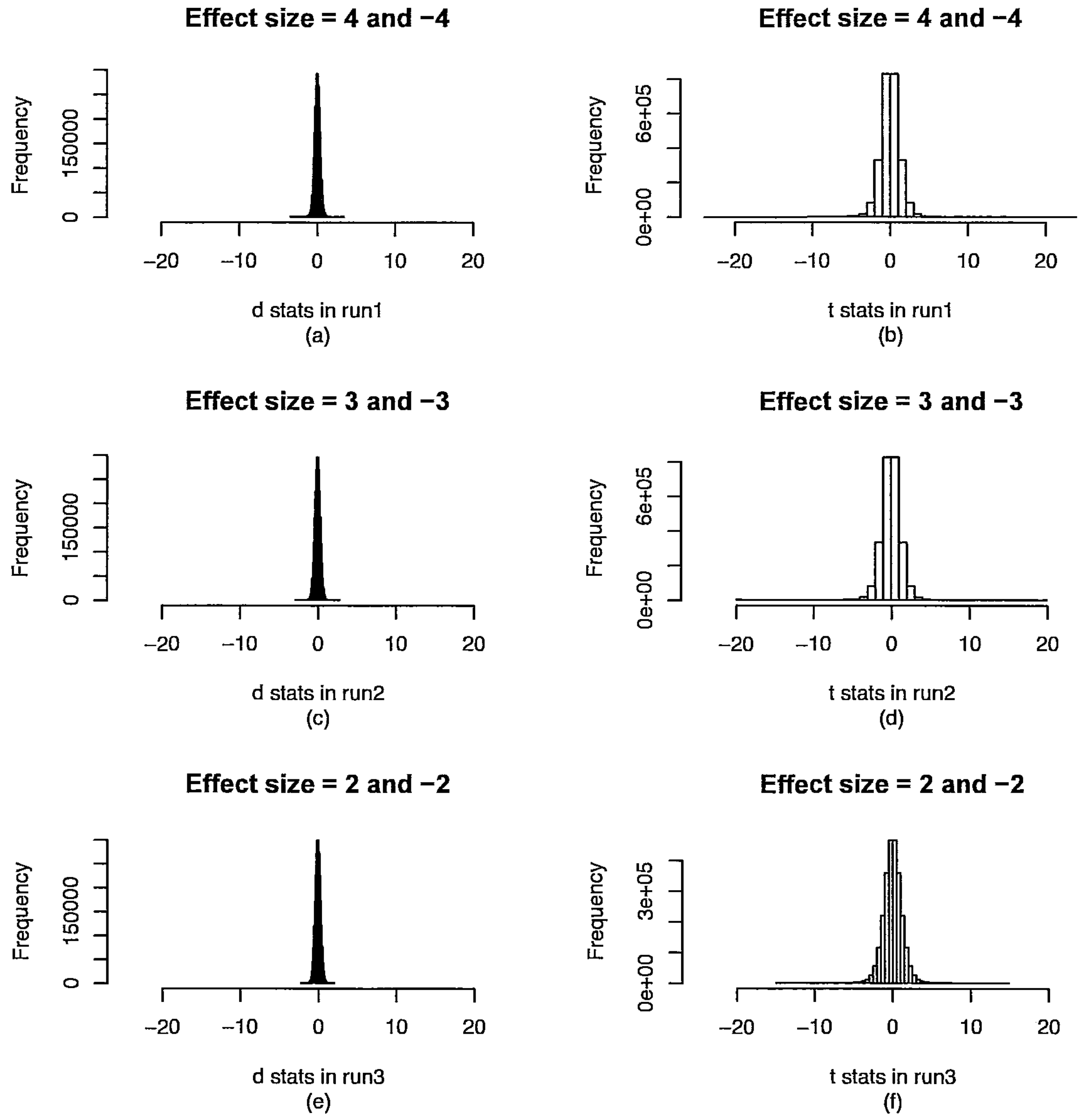


Figure 4.11: *Comparison of distributions of permuted d-statistics and t-statistics in runs 1-3 for all 10,000 probe sets.*

Effect size		d-statistic			t-statistic		
		FDR.d	PERM.d	DIST.d	FDR.t	PERM.t	DIST.t
±4	#sig	1002	999	1043	1019	1018	1023
	V	5	3	43	48	48	52
	$m_0 - (R - V)$	3	4	0	29	30	29
	FDR	0.0050	0.0030	0.0412	0.0471	0.0472	0.0508
±3	#sig	660.49	659.32	535.46	532.10	530.71	445.55
	V	12	10	43	34	33	38
	$m_0 - (R - V)$	77	80	38	264	261	248
	FDR	0.0128	0.0108	0.0428	0.0442	0.0427	0.0481
±2	#sig	474	477	612	129	111	111
	V	10	10	27	7	6	6
	$m_0 - (R - V)$	536	533	415	878	895	895
	FDR	0.0211	0.0210	0.0441	0.0543	0.0541	0.0541

Table 4.12: *Comparison of 6 methods among different effect sides when setting cutoff 5%. $m_0 - (R - V)$: the number of false negatives.*

be controlled. In contrast, methods based on *d*-statistic perform well even though effect size is as small as ±2, they identify about 400 significant probe sets and also control FDR at a desirable level. Overall, methods based on *d*-statistic are more powerful for detecting significant probe sets at each effect size and even small effect size.

4.4 Conclusions

In general, the methods based on *d*-statistic outperform the methods based on *t*-statistic because they can identify more significant probe sets with lower false discovery rate. When the absolute value of difference in expression is as large as 4, methods based on both statistics can identify the large amount of significant probe sets but *t*-statistic methods produce much more false positives. When the absolute value of difference in expression is as small as 2, methods based on *t*-statistic do not work well that very few probe sets can be identified and FDR exceeds threshold. Therefore methods based on *d*-statistic are more accurate and stable. Power plots in Figure 4.8 also indicate that methods based on *d*-statistic are more powerful because their associated power plots are generally above plots from methods based on *t*-statistic at true differences greater than 2 or less than -2.

For each test statistic, FDR and permutation methods are essentially the same in that they identify almost the same number of significant probe sets and control the false discovery rate at almost the same level. However distribution method has different behavior depending on the distribution of gene expressions. If gene expressions are normally or approximately normally distributed, distribution method can identify more significant probe sets than FDR and permutation methods (Simulation 1 and 3). But if gene expressions are heavy tailed, it identifies less significant probe sets and controls FDR at quite lower level (Simulation 2, 4, 5, 6 and 7). It is due to the fact that distribution method is established under the normality assumption. Once this assumption is violated, distribution method becomes conservative.

According to comparison among Simulation 1, 2, 3 and 4, we find out that the distribution of gene expressions would affect performance of methods. For instance, both Simulation 1 and Simulation 3 where gene expressions were generated from normal distribution have similar performance, although Simulation 1 assumed a common variance across probe sets and Simulation 2 allowed variances to differ. Similarly, both Simulation 2 and Simulation 4 have similar performance since their expression data were generated from the same t distribution. Moreover, for t distribution data, we consider different degrees of freedom. From the comparison among Simulation 4,5 and 6, we find out that when decreasing degrees of freedom, the larger number of significant probe sets are identified, and that the FDRs are still less than nominal 5%. In addition, the study on the different type symmetric distributed data sets consisting of normal distribution data (Simulation 1 and 3), t distribution data (Simulation 2,4,5 and 6) and Laplace distribution data (Simulation 7), indicates that methods based on d-statistic are superior to methods based on t-statistic if gene expressions are symmetrically distributed.

Compared to the t distribution method, the d distribution method works better in all situations we discussed, when true variance of each probe set is known. When gene expression distribution has heavy tail, the d distribution method can not only identify more significant probe sets but also control FDR at quite lower level. When data approach normal distribution, the d distribution method can identify much more significant probe sets although control FDR at almost the same level as the t distribution. Therefore, no matter what distribution gene expressions follow, the d distribution method always outperforms the t distribution method if true variance is known.

However, a big issue faced when employing the d distribution method is that true variance σ_j^2 is usually unknown in practice. Under this circumstance, we proposed four estimates: the pooled variance ($s_{j;pool}^2$), the null variance ($s_{j;null}^2$), the average of pooled variances $\bar{s}_{pool}^2$ and the average of null variances $\bar{s}_{null}^2$, and applied them to the various data sets. We find out that the method using the pooled variance (DIS.d.d1) is overly liberal in all cases. It can be verified by power plot (Figure 4.9) where the number of probe sets called significant from DIS.d.d1 is much greater than DIS.d, when true differences in expression are small ($<$ 2 or $>$ -2). So DIS.d.d1 must falsely call some probe sets significant which are truly equally expressed. The methods using the pooled variance and transformed $\hat{c}_j$ (DIS.d.d2) and using the null variance for each probe set (DIS.d.d3) are distribution-dependent. They only work in the case where data have heavy tails. The methods using the same variance across all probe sets (DIS.d.d4 and DIS.d.d5) work very well in both normal distribution and t distribution data even when probe set do not have a common variance. Particularly the method using a common pooled variance (DIS.d.d4) is closest to the method using the true variances across probe sets and the method using a common null variance (DIS.d.d5) is somewhat conservative.

Chapter 5

Application to real Microarray data

To further test the performance of all methods introduced in Chapter 3, we consider here their applications to two real microarray data sets, which were generated from Affymetric MGU74Av2 Gene chips. Each gene chip has 12,488 probe sets. We will compare the performances of PERM.d, FDR.t, PERM.t and DIS.t with FDR.d, in terms of the number of differential expressions declared and the ordering of the differential expressions. When applying d distribution method, we will also compare the performance of proposed estimates of population variance σ_j^2 with FDR.d.

5.1 Description of real data sets

As introduced in Chapter 1, Idd4, Idd5 and Idd13 are of interest among those regions identified by recent biological research in mouse model. By inbreeding two strains of NOD mice and NOR mice by multiple generations, we can obtain two congenic strains NOD.NOR Idd4 and NOR.NOD_Idd5/13. We firstly consider the data set with gene expression levels between NOR and NOR.NOD_Id5/13. The entire data set comprise I=10 tissue samples and J=12,488 probe sets. The first five samples were taken from NOR mice and the last five samples from NOR.NOD_Id5/13 mice. The hybridizations were completed in two days. The first three samples of each strain were measured in Day 1 and the last two samples in Day 2. After reading and normalizing raw gene expression values in **ReadAffy** and **rma** functions from **Affy** package in **R** (Gautier *et al*, 2004), a data matrix can be defined as Y_{ji}, where i=1,2,...,10 and j=1,2,..., 12,488. The corresponding strain vector is (1 1 1 1 1 2 2 2 2 2) where 1 and 2 denote NOR and NOR.NOD_Id5/13 strains respectively, and day effect vector is (1 1 1 2 2 1 1 1 2 2) where 1 and 2 denote Day 1 and Day 2, respectively.

Since region Idd4 exhibits sex-specific effect on T1D, we focus here on the comparison of gene expression data between female NOD mice and female NOD.NOR_Idd4 mice. This data set contains I=8 tissue samples and J=12,488 probe sets. The first four samples were taken from female NOD mice, two of which were hybridized in Day 1 and Day 2 respectively. Similarly, the last four samples were taken from female NOD.NOR_Idd4 mice and also hybridized on these two days. Then the corresponding strain vector is (1 1 1 1 2 2 2 2), and day effect vector is (1 1 2 2 1 1 2 2).

5.2 Discussion on methods when day effect present

The FDR and Permutation methods in terms of d and t statistics introduced in Chapter 3 involve permutation tests, which assume that expression values are exchangeable under null hypothesis. In each real data set, however, expression values were generated from two different days. The day on which hybridization is done can change the distribution of expression values and thus expression values might not be exchangeable between the two days. Therefore it is necessary to consider day effect as a factor as it produces non-biological variation. The linear model for a given probe set j is

$$y_{ji} = \beta_{j0} + \beta_{j1} D_i + \beta_{j2} S_i + \varepsilon_{ji}, i = 1, ..., n_1 + n_2 \tag{5.1}$$

where D_i is the day indicator for the ith sample, S_i is the strain indicator for the ith sample, β_{j1} and β_{j2} are corresponding regression coefficients, and error term $\varepsilon_{ji} \sim N(0, \sigma_j^2)$.

Permutation within day. For the first application, samples 1,2,3,6,7 and 8 came from Day 1, and samples 4,5,9 and 10 came from Day 2. We allow permutation of the samples within two days separately, but not across the two days. Within Day 1, samples 1,2 and 3 and samples 6,7 and 8 were taken from two different strains, so there are

$$B_{\text{day1}} = \frac{(3+3)!}{3! \times 3!} = 20$$

permutations by equation (3.4). Similarly, there are

$$B_{\text{day2}} = \frac{(2+2)!}{2! \times 2!} = 6$$

permutations within Day 2. Hence the total number of possible permutations is

$$B_{\text{total}} = B_{\text{day1}} \times B_{\text{day2}} = 120.$$

For each probe set, we obtained a 120×10 permutation matrix in terms of strain labels and then calculated a set of 120 corresponding permuted test statistics. When pooling all probe sets, we got a $12,488 \times 120$ matrix of permuted test statistics.

For the second application, samples 1,2 and samples 5,6 came from two different strains within Day 1, so there are

$$B_{\text{day1}} = \frac{(2+2)!}{2! \times 2!} = 6$$

permutations. Similarly, there are the same number of permutations within Day 2. Thus the total number of permutation is

$$B_{\text{total}} = B_{\text{day1}} \times B_{\text{day2}} = 6 \times 6 = 36.$$

For each probe set, we obtained a 36 × 8 permutation matrix in terms of strain labels and a set of 36 corresponding permuted test statistics. When pooling all probe sets, we got a $12,488 \times 36$ matrix of permuted test statistics.

Test statistics. The t-statistic is now defined as the estimated $\hat{\beta}_{j2}$ for the strain effect in model (5.1) divided by the standard error of $\hat{\beta}_{j2}$, i.e.

$$t_j = \frac{\hat{\beta}_{j2}}{se(\hat{\beta}_{j2})}. \tag{5.2}$$

Under the normality and constant variance assumptions, t_j follows a t distribution with $(n_1 - 1) + (n_2 - 1) - (N_{\text{day}} - 1)$ where N_{day} is the number of levels of D_i. Similarly, the d-statistic is defined by

$$d_j = \frac{\hat{\beta}_{j2}}{se(\hat{\beta}_{j2}) + s_0}. \tag{5.3}$$

where s_0 is found using the algorithm in section 2.4 with $\bar{x}_{j2} - \bar{x}_{j1}$ replaced by $\hat{\beta}_{j2}$ and s_j replaced by $se(\hat{\beta}_{j2})$. Thus d_j follows the d distribution with $(n_1-1)+(n_2-1)-(N_{\text{day}}-1)$ degrees of freedom.

The estimates of σ_j^2. Since day effect is present in applications, the non-biological variance sources are not only from strain effect not also from day effect. The equivalent of the pooled variance estimate $s_{j;pool}^2$ is the mean square error from the model (5.1) with these two effects. But the equivalent of the null variance estimate $s_{j;null}^2$ is the mean square error from the model with only day effect, i.e.

$$y_{ji} = \beta_{j0} + \beta_{j1} D_i + \varepsilon_{ji}, i = 1, ..., n_1 + n_2, \tag{5.4}$$

where $\varepsilon_{ji} \sim N(0, \sigma_j^2)$. Because the null variance estimate is established under null hypothesis where all samples are regarded as the same population and then strain effect is removed. Accordingly, the estimates in DIS.d.d4 and DIS.d.d5 are simply the average of them over 12488 probe sets.

5.3 Comparison of methods in microarray data

Worth noting here is that **DIS.d** method is not applicable in real data sets since we are unable to observe true variance σ_j^2. We test FDR.d, PERM.d, FDR.t, PERM.t, DIS.t and *d* distribution methods using various estimates. Since the SAM methodology which FDR.d method is based on has been widely applied to identify differential expressions in microarray analysis, we will compare other methods with it in finding lists of significant differential expressions and examining behaviors of *q*-values.

5.3.1 NOR V NOR.NOD_Idd5/13

Firstly, we test the performance of the methods on gene expressions of NOR and NOR.NOD_Idd5/13 mice. In this application, we obtained s_0=0.0308 and the estimated proportion of differentially expressions $1 - \hat{\pi}_0 = 1 - 0.66 = 0.34$. From Figure 5.1 we can see that the pooled standard deviations across all probe sets are right skewed with variance 0.0033, indicating that all probe sets are not equally variant. Figure 5.2 exhibits the distributions of *d*-statistics and *t*-statistics in this application. We can see that two distributions are symmetric about 0, but the distribution of *d*-statistics is narrower with shorter tails and that of *t*-statistics spreads out more with longer tails.

Table 5.1 presents the number of significant expressions declared by various methods.

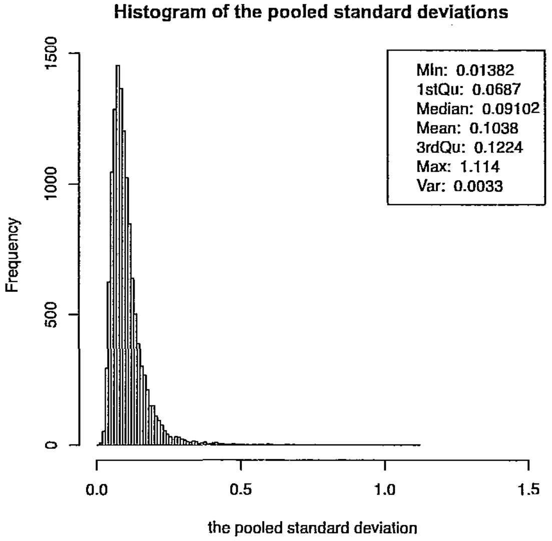


Figure 5.1: *Plot of the pooled standard deviations for NOR V NOR.NOD_Idd5/13.*

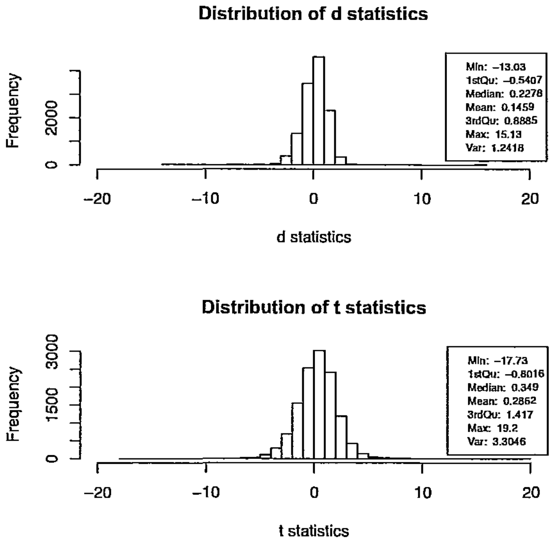


Figure 5.2: *Histograms of different test statistics for NOR V NOR.NOD_Idd5/13.*

FDR.d	PERM.d	FDR.t	PERM.t	DIS.t	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
88	80	63	57	62	1024	76	45	86	69

Table 5.1: *Number of significant probe sets identified at q-value=0.05 level using different methods and estimates for NOR V NOR.NOD_Idd5/13.*

From this table, we can see that PERM.d method is closest to FDR.d method, methods based on t-statistics produce smaller number of significant probe sets. We can also see that DIS.d.d4 method with the pooled variance estimate assuming a common variance across probe sets is closest to FDR.d method among d distribution methods. DIS.d.d1 method with the same estimate but allowing variance to differ is very liberal.

Table 5.2 and 5.3 show the top 20 probe sets as ranked by q-values from FDR.d method. Each table includes Probe set ID, test statistic and q-value calculated by various methods. From Table 5.2, we can see that q-values from PERM.d give almost the same ordering as FDR.d although the former are slightly bigger, whereas methods based on t-statistics give a different ordering and also several q-values exceeds cutoff level 0.05. From Table 5.3, we can see that the q-values from DIS.d.d4 and DIS.d.d5 are generally smaller ones from FDR.d and do not change ordering of q-values much. We can also see that the top 20 probe sets declared by FDR.d are also declared by DIS.d.d1 with much smaller q-values, whereas a few of the 20 probe sets get larger q-values when testing DIS.d.d2 and DIS.d.d3 methods.

Furthermore, Figure 5.3 and 5.4 give the behaviors of q-values of probe sets declared significant from FDR.d method at cutoff 0.05 against other methods. In each panel of

			q-values				
Probeset ID	d.stats	t.stats	FDR.d	PERM.d	FDR.t	PERM.t	DIS.t
97206_at	15.1273	19.2036	0.0055	0.0109	0.0109	0.0164	0.0018
101845_s_at	-7.8347	-9.5681	0.0073	0.0109	0.0201	0.0226	0.0156
96336_at	-13.0317	-17.7337	0.0073	0.0109	0.0137	0.0164	0.0018
100606_at	-6.9533	-10.8032	0.0077	0.0109	0.0170	0.0208	0.0125
100908_at	-7.0921	-9.9785	0.0077	0.0109	0.0191	0.0219	0.0148
104671_at	-6.3664	-8.5011	0.0079	0.0109	0.0246	0.0279	0.0244
96562_at	-6.2092	-14.9188	0.0079	0.0109	0.0146	0.0197	0.0040
97916_at	-6.1728	-9.7750	0.0079	0.0109	0.0191	0.0226	0.0156
99580_s_at	-5.6518	-6.9936	0.0088	0.0131	0.0383	0.0444	0.0423
102965_at	-5.5352	-10.5737	0.0090	0.0135	0.0170	0.0208	0.0125
104183_at	-5.2498	-7.2883	0.0091	0.0135	0.0371	0.0421	0.0396
95571_at	7.1063	12.6885	0.0091	0.0109	0.0164	0.0197	0.0072
98015_at	-5.0918	-10.5309	0.0097	0.0148	0.0170	0.0208	0.0125
102348_at	5.3585	6.8765	0.0133	0.0135	0.0452	0.0444	0.0423
92397_at	-4.8303	-8.3123	0.0140	0.0164	0.0257	0.0292	0.0244
96606_at	-4.7999	-5.8879	0.0140	0.0164	0.0580	0.0587	0.0603
103346_at	-4.5488	-7.9412	0.0144	0.0171	0.0286	0.0312	0.0290
95520_at	-4.6260	-8.5573	0.0144	0.0171	0.0246	0.0279	0.0244
99146_at	-4.6014	-9.1994	0.0144	0.0171	0.0201	0.0226	0.0189
102851_s_at	-4.4038	-8.3749	0.0153	0.0186	0.0250	0.0281	0.0244

Table 5.2: *q-values of top 20 probe sets detected from FDR.d method against other methods (PERM.d, FDR.t, PERM.t and DIS.t) for NOR V NOR.NOD_Idd5/13.*

		q-values					
Probeset ID	d.stats	FDR.d	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
97206_at	15.1273	0.0055	0.0000	0.0000	0.0017	0.0000	0.0000
101845_s_at	-7.8347	0.0073	0.0001	0.0009	0.0187	0.0000	0.0000
96336_at	-13.0317	0.0073	0.0000	0.0000	0.0017	0.0000	0.0000
100606_at	-6.9533	0.0077	0.0000	0.0001	0.0178	0.0000	0.0000
100908_at	-7.0921	0.0077	0.0000	0.0002	0.0187	0.0000	0.0000
104671_at	-6.3664	0.0079	0.0001	0.0015	0.0293	0.0001	0.0001
96562_at	-6.2092	0.0079	0.0000	0.0000	0.0082	0.0001	0.0002
97916_at	-6.1728	0.0079	0.0000	0.0002	0.0187	0.0001	0.0002
99580_s_at	-5.6518	0.0088	0.0009	0.0075	0.0469	0.0003	0.0006
102965_at	-5.5352	0.0090	0.0000	0.0001	0.0187	0.0004	0.0008
104183_at	-5.2498	0.0091	0.0002	0.0051	0.0410	0.0007	0.0014
95571_at	7.1063	0.0091	0.0000	0.0000	0.0106	0.0000	0.0000
98015_at	-5.0918	0.0097	0.0000	0.0002	0.0187	0.0011	0.0020
102348_at	5.3585	0.0133	0.0008	0.0075	0.0469	0.0006	0.0012
92397_at	-4.8303	0.0140	0.0000	0.0016	0.0334	0.0020	0.0035
96606_at	-4.7999	0.0140	0.0034	0.0247	0.0669	0.0020	0.0035
103346_at	-4.5488	0.0144	0.0000	0.0028	0.0379	0.0032	0.0055
95520_at	-4.6260	0.0144	0.0000	0.0015	0.0317	0.0029	0.0049
99146_at	-4.6014	0.0144	0.0000	0.0009	0.0285	0.0029	0.0050
102851_s_at	-4.4038	0.0153	0.0000	0.0021	0.0337	0.0045	0.0074

Table 5.3: *q-values of top 20 probe sets detected from FDR.d method against d distribution when applying different estimates for NOR V NOR.NOD_Idd5/13.*

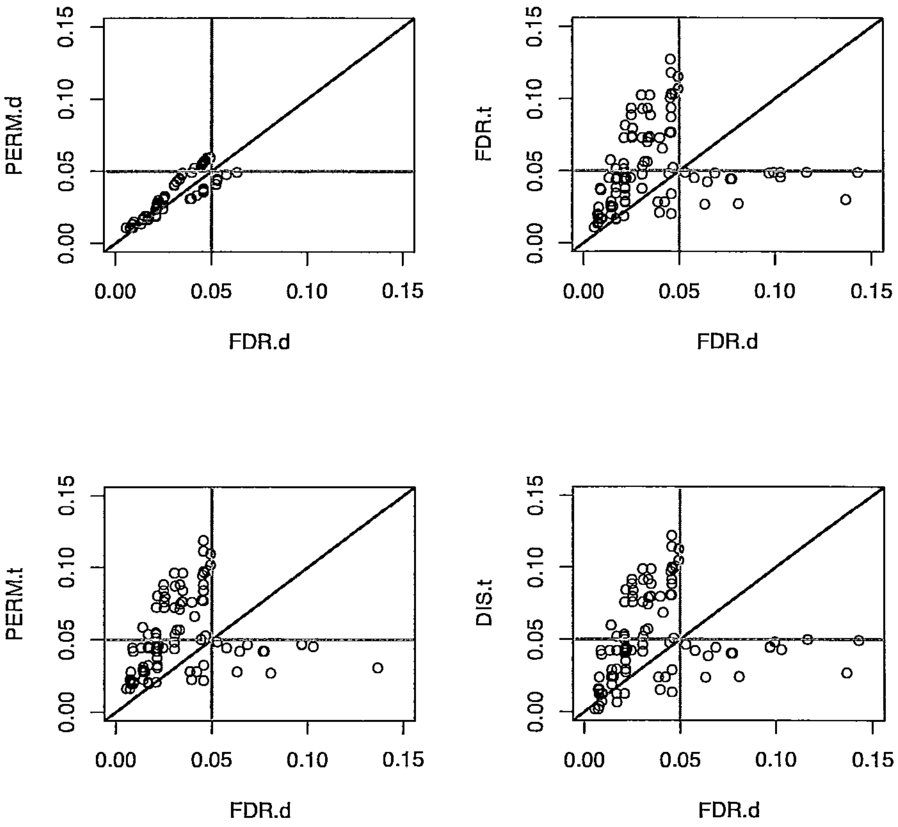


Figure 5.3: *Plots of the q-values of significant probe sets from FDR.d against those from other methods at q-value cutoff=0.05 for NOR V NOR.NOD_Idd5/13.*

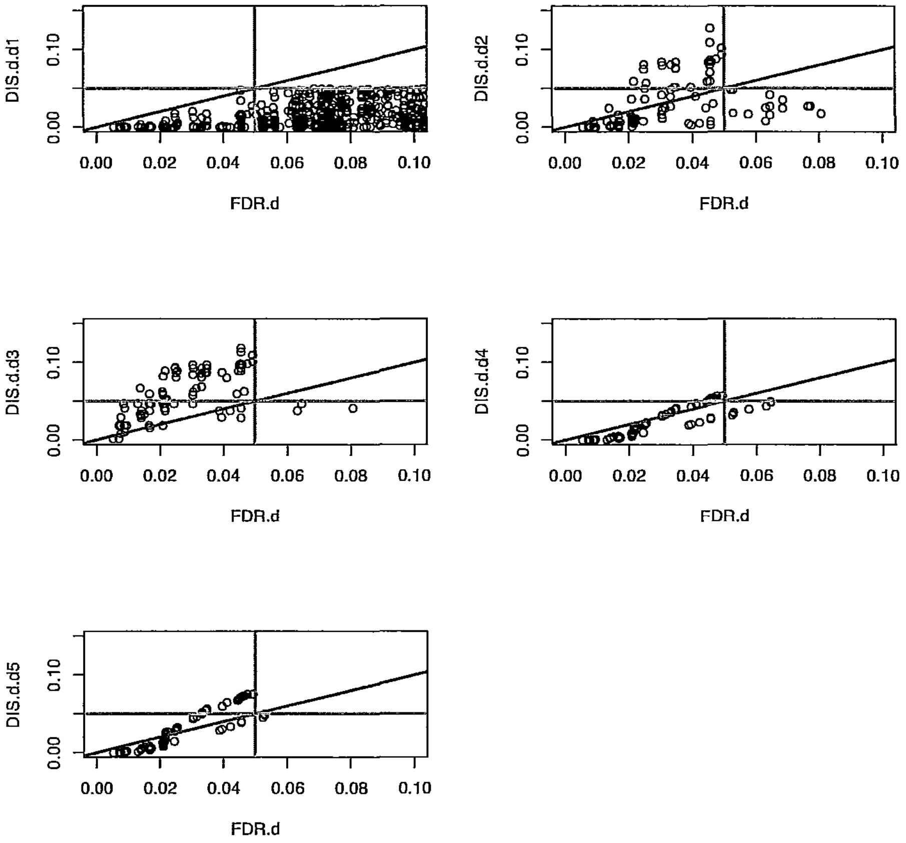


Figure 5.4: *Plots of the q-values of significant probes from FDR.d against those from other estimate methods at q-value cutoff=0.05 for NOR V NOR.NOD_Idd5/13.*

both figures, X-axis represents the q-values calculated from FDR.d method and Y-axis represents the q-values calculated from other methods. Points in panel are the probe sets which are declared significant from either FDR.d method or other methods when cutoff is 0.05, therefore points in left bottom corner of each panel between vertical line ($x=0.05$) and horizontal line ($y=0.05$) are probe sets declared significant by both FDR.d and any one of other methods. In Figure 5.3, we can observe from the first panel that the behaviors of q-values from FDR.d and PERM.d are quite similar and probe sets declared are concentrated in the left bottom corner, implying that two methods can produce similar differential expression lists. From the bottom panels and the right top panel in this figure, we find that although some probe sets are simultaneously called significant by FDR.d and methods based on t-statistics, quite a few probe sets detected by FDR.d cannot be called significant by others, and conversely some probe sets not detected by FDR.d are called significant by others. In Figure 5.4, the left top panel indicates that DIS.d.d1 not only detects the majority of probe sets produced by FDR.d but also calls more probe sets significant which are not detected by FDR.d. The right top panel and the left middle panel in this figure show that there are quite a few probe sets called significant by FDR.d which cannot be identified by DIS.d.d2 and DIS.d.d3 methods. The rest of panels in this figure exhibit that the majority of probe sets called significant by FDR.d are also declared by DIS.d.d4 and DIS.d.d5. So these two methods appear closer to FDR.d.

FDR.d	Perm.d	FDR.t	Perm.t	DIS.t
56	38	10	7	14

DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
1290	20	8	380	266

Table 5.4: *Number of significant probe sets identified at q-value=0.05 level using different methods and estimates for female NOD V NOD.NOR_Idd4.*

5.3.2 Female NOD V NOD.NOR_Idd4

Secondly, We test the performance of the methods on gene expressions of NOD and NOD.NOR_Idd4 female mice. In this application we obtain s_0=0.0680 and the estimated proportion of differentially expressions $1 - \hat{\pi}_0 = 1 - 0.80 = 0.20$. Compared to the first application, the pooled standard deviations in this application are more concentrated and distributed with the mean 0.0656 and variance 0.001 (Figure 5.5). The distribution of test statistics are more varying because they have much longer tails (Figure 5.6). From Table 5.4, we can see that PERM.D method still works closest to FDR.d method with relatively less probe sets declared, methods based on t-statistics produce much smaller number of probe sets. We can also see that DIS.d.d1 declares extremely large number of probe sets, while DIS.d.d2 and DIS.d.d3 only yields small number of probe sets. DIS.d.d4 and DIS.d.d5 using common variance estimate declare 200-400 signals out of 12488 probe sets which seems too high. But we are not sure the reason why these two methods call so many probe sets significant in this application where the pooled standard deviations are relatively concentrated.

Table 5.5 and 5.6 show the top 20 probe sets as ranked by q-values from FDR.d

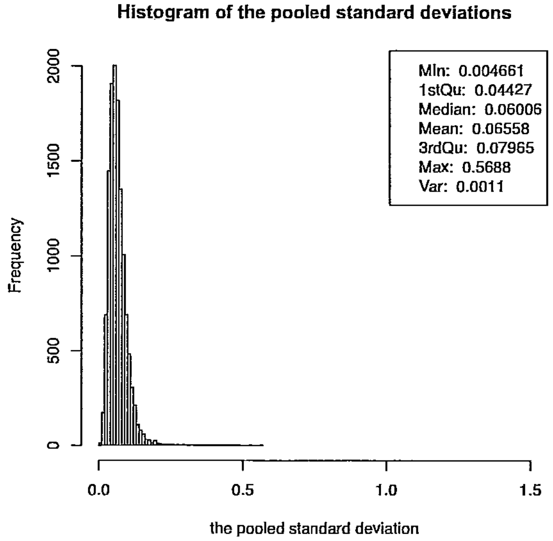


Figure 5.5: *Plot of the pooled standard deviations for female NOD V NOD.NOR_Idd4.*

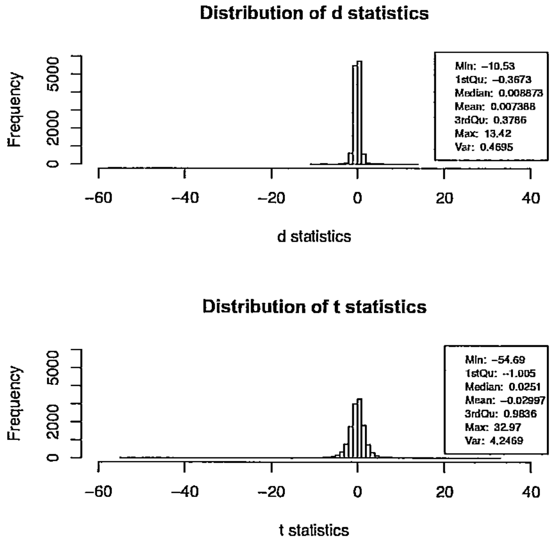


Figure 5.6: *Histograms of different test statistics for female NOD V NOD.NOR_Idd4.*

			q-values				
Probeset ID	d.stats	t.stats	FDR.d	PERM.d	FDR.t	PERM.t	DIS.t
101992_at	13.4183	32.9731	0.0222	0.0334	0.0371	0.0371	0.0019
93427_at	11.1455	27.7614	0.0222	0.0334	0.0389	0.0389	0.0028
100277_at	4.5916	11.2561	0.0304	0.0430	0.0677	0.0667	0.0509
101851_at	3.9798	8.1601	0.0304	0.0430	0.0947	0.0950	0.0736
102424_at	4.3053	10.3468	0.0304	0.0430	0.0767	0.0739	0.0544
102736_at	4.1741	7.5962	0.0304	0.0430	0.0965	0.0976	0.0766
103235_at	3.8043	9.7141	0.0304	0.0430	0.0776	0.0773	0.0544
104094_at	5.9220	16.7526	0.0304	0.0420	0.0519	0.0544	0.0173
104100_at	3.3610	7.3959	0.0304	0.0445	0.1004	0.1017	0.0798
160679_at	3.4886	6.7208	0.0304	0.0433	0.1040	0.1067	0.0863
92642_at	5.7207	9.7825	0.0304	0.0420	0.0776	0.0772	0.0544
93037_i_at	3.6906	7.1768	0.0304	0.0431	0.1004	0.1017	0.0825
93871_at	3.9755	10.1345	0.0304	0.0430	0.0767	0.0739	0.0544
94429_at	3.4493	7.7955	0.0304	0.0433	0.0956	0.0976	0.0737
96047_at	6.9835	12.6312	0.0304	0.0408	0.0653	0.0649	0.0444
100946_at	3.3343	7.9029	0.0311	0.0456	0.0956	0.0976	0.0737
101676_at	-7.6618	-21.3755	0.0311	0.0408	0.0350	0.0408	0.0069
103935_at	-10.5311	-54.6905	0.0311	0.0408	0.0222	0.0371	0.0004
93403_at	-7.4484	-15.4338	0.0311	0.0408	0.0417	0.0544	0.0230
101506_at	3.2235	9.7298	0.0313	0.0465	0.0776	0.0773	0.0544

Table 5.5: *q-values of top 20 probe sets detected from FDR.d method against other methods (PERM.d, FDR.t, PERM.t and DIS.t) for female NOD V NOD.NOR_Idd4.*

		q-values					
Probset ID	d.stats	FDR.d	DIS.d.d1	DIS.d.d2	DIS.d.d3	DIS.d.d4	DIS.d.d5
101992_at	13.4183	0.0222	0.0000	0.0000	0.0038	0.0000	0.0000
93427_at	11.1455	0.0222	0.0000	0.0000	0.0061	0.0000	0.0000
100277_at	4.5916	0.0304	0.0000	0.0015	0.1560	0.0000	0.0000
101851_at	3.9798	0.0304	0.0000	0.0248	0.1858	0.0000	0.0000
102424_at	4.3053	0.0304	0.0000	0.0038	0.1681	0.0000	0.0000
102736_at	4.1741	0.0304	0.0000	0.0324	0.1858	0.0000	0.0000
103235_at	3.8043	0.0304	0.0000	0.0120	0.1858	0.0000	0.0000
104094_at	5.9220	0.0304	0.0000	0.0000	0.0455	0.0000	0.0000
104100_at	3.3610	0.0304	0.0000	0.0564	0.1858	0.0000	0.0000
160679_at	3.4886	0.0304	0.0000	0.0770	0.1858	0.0000	0.0000
92642_at	5.7207	0.0304	0.0000	0.0019	0.1681	0.0000	0.0000
93037_i_at	3.6906	0.0304	0.0000	0.0511	0.1858	0.0000	0.0000
93871_at	3.9755	0.0304	0.0000	0.0073	0.1856	0.0000	0.0000
94429_at	3.4493	0.0304	0.0000	0.0486	0.1858	0.0000	0.0000
96047_at	6.9835	0.0304	0.0000	0.0000	0.0712	0.0000	0.0000
100946_at	3.3343	0.0311	0.0000	0.0511	0.1858	0.0000	0.0000
101676_at	-7.6618	0.0311	0.0000	0.0000	0.0205	0.0000	0.0000
103935_at	-10.5311	0.0311	0.0000	0.0000	0.0038	0.0000	0.0000
93403_at	-7.4484	0.0311	0.0000	0.0000	0.0455	0.0000	0.0000
101506_at	3.2235	0.0313	0.0000	0.0341	0.1858	0.0000	0.0000

Table 5.6: *q-values of top 20 probe sets detected from FDR.d method against d distribution when applying different estimates for female NOD V NOD.NOR_Idd4.*

method. From Table 5.5 we can see that q-values from PERM.d still give almost the same ordering as FDR.d, similar to the first application. Methods based on t-statistics not only give the different ordering but also quite a few q-values exceed 0.05. In table 5.6 the top probe sets declared by FDR.d are all declared by DIS.d.d1, DIS.d.d4 and DIS.d.d5 which yield extremely small q-values. DIS.d.d2 and DIS.d.d3 methods detect quite small number of probe sets.

In Figure 5.7, we observe that the behaviors of q-values from FDR.d and PERM.d are similar, which is consistent with the first application. From the rest of panels, we find that only a small portion of probe sets declared by FDR.d can be declared by the methods based on t-statistics. Figure 5.8 presents the behaviors of q-values of probe sets which show different patterns from the first application when testing various estimates. We observe that quite a lot of probe sets declared by DIS.d.d1, DIS.d.4 and DIS.d.d5 are not declared by FDR.d so the three methods are very liberal and might include many false positives. In contrast DIS.d.d2 and DIS.d.d3 are so conservative that they are missing some signals and only detect few probe sets.

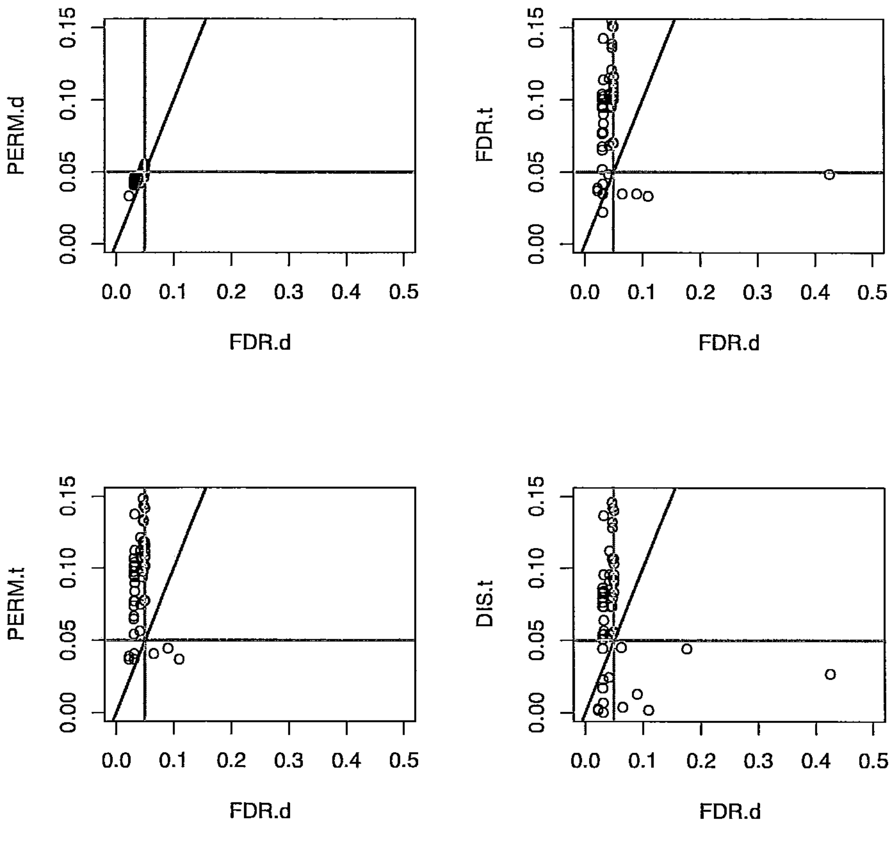


Figure 5.7: *Plots of the q-values of significant probe sets from FDR.d against the ones from other methods at q-value cutoff=0.05 for female NOD V NOD.NOR_Idd4.*

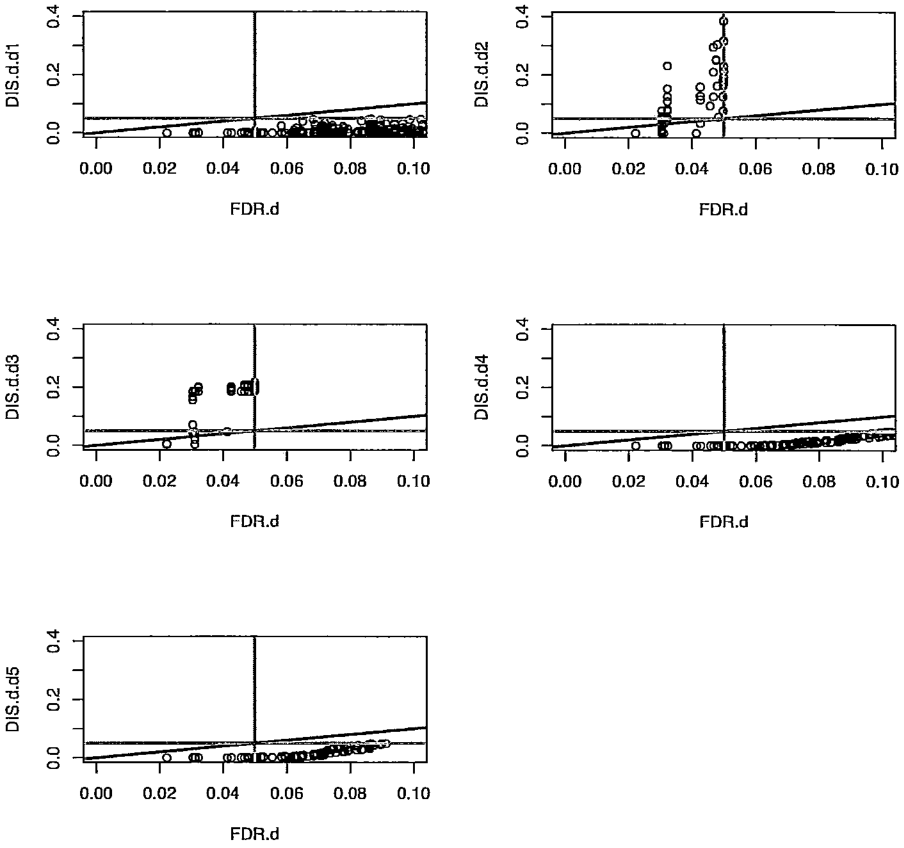


Figure 5.8: *Plots of the q-values of significant probe sets from FDR.d against the ones from other estimate methods at q-value cutoff=0.05 for female NOD V NOD.NOR_Idd4.*

Chapter 6

Discussion and Future Work

In this study, we have compared the performances of the FDR, permutation and analytical distribution methods based on d and t statistics respectively for identifying the differential expressions using microarray data sets. The methods were compared using both real gene expression data sets and the simulated data sets. These six methods can be classified as nonparametric and parametric approaches. The nonparametric approaches include FDR.d, PERM.d, FDR.t and PERM.t. The basic idea of methods is to estimate the null distribution of the test statistic from permutations rather than assuming a specific form of null distribution. The parametric approaches include DIS.d and DIS.t which assume the test statistics follow a specific distribution under null hypothesis. DIS.d method depends on the true population variance of each probe set which is generally unknown in practice. We proposed five estimators to deal with the unknown variance and evaluated their performance using simulated and real data sets.

Overall, methods based on the d-statistic perform better than methods based on t-statistic because they can identify more significant differential expressions with lower false discovery rate. We find in simulation studies that for a variety of gene expression distributions, some with quite heavy tails, methods based on d-statistic can identify more significant differential expressions and control false discovery rate very well. In contrast, methods based on t-statistic appear to be weak in that they identify far less significant probe sets and control false positive rates very close to nominal FDR. It can be clearly shown by our power simulation analysis (say Figure 4.8) that the powers of methods based on t-statistic are lower than methods based on d-statistic and also get more Type I errors when true differences approach 0. So we can draw a conclusion that methods based on d-statistic are more accurate and powerful for identifying differential expressions. This has been confirmed by the results in two real data analyses (Table 5.1 and 5.4) where FRD.d and PERM.d always identify more significant probe sets than FDR.t and PERM.t. Moreover, we notice that methods based on t-statistic surprisingly identify quite small number of differential expressions even when the true gene expression distribution is Gaussian. This would not be in accordance with what we expected because t-statistic only consider information from one probe set at a time and most likely falsely call significant those expressions with small fold changes and smaller variance. So we may need more work to examine the performance of the methods based on t-statistic in future.

FDR and permutation methods usually give similar results whether d-statistic or t-statistic is applied. FDR method considers the FDR of all rejection regions containing a probe set and takes the minimum of FDR as the q-value for that probe set, while permutation method permutes array labels and derives p-value as the fraction of the permuted test statistics greater or equal to the observed statistics. The performance similarity of these two methods implies that q-values calculated from the original statistics are approximately equivalent to ones from their p-values. In simulation studies, we suppose gene expressions are independent and consequently their statistics and p-values are independent too. Therefore the two methods are almost identical in calling significant probe sets and controlling FDR. However, the results of two real microarray data sets indicate that permutation methods are little conservative with a smaller number of probe sets declared and the slightly larger q-values. One primary reason could be explained by the fact that some probe sets are dependent on each other in real microarray data. We may relax the independence assumption in our simulation study in future to explore the effect of dependence on the performance of these two methods.

The d distribution method combines the strengths of other methods based on d-statistics and the t distribution method. Similar to FDR.d and PERM.d, the d distribution methods construct test statistics by adding a small positive constant (s_0) in the denominator of regular t-statistic, which ensures that those probe sets with small fold changes will not be incorrectly selected as significant. In simulation study and real data analysis, we found that the distribution of d-statistics tends to have shorter tails and is more concentrated compared to t-statistics. We can conclude that the d distribution method could better control false positives and avoid missing true positives. In addition, the d distribution method similar to the t distribution method can directly calculate p-values without any permutation of test statistics. So it calculates q-values very fast compared to other nonparametric methods which are often computationally intensive. However, the d distribution method has some weakness. First, the d distribution method depends on population variances which are generally unknown in practice, although it appears to be the most powerful in our power simulation. Therefore its performance is determined to a large extent by the estimate of variance. Second, the *d* distribution method is sensitive to violation of normality assumption. In our simulation study analysis, we found that it is conservative and identifies less significant probe sets when gene expression distribution is heavy-tailed. In contrast, it identifies more significant probe sets and control false discovery rate close to nominal one (5%) when gene expression data follow a normal distribution. Therefore the estimate of population variance and the distribution of gene expression are two main factors influencing the performance of *d* distribution methods. Sample size might be another factor since large sample size ensures that normality assumption is satisfied by central limit theorem. So how big sample size can improve the performance of the *d* distribution methods could be examined in future.

In our study we also proposed five estimators for population variances across probe sets. DIS.d.d1, DIS.d.d2 and DIS.d.d3 methods allow different variances across all probe sets, while DIS.d.d4 and DIS.d.d5 methods assume a common variance. The results of simulation studies indicate that DIS.d.d4 method using the average of pooled sample variances is closet to DIS.d method using the known population variances in all cases, followed by DIS.d.d5 method, while DIS.d.d1 method is always overly liberal and falsely calls many probe sets significant which have small fold changes actually. DIS.d.d2 and DIS.d.d3 methods work poorly and identify very few significant probe sets when gene expressions follow a normal distribution, while they become workable when gene expressions are heavy-tailed. In real data analysis, the performances of DIS.d.d1, DIS.d.d2 and DIS.d.d3 methods are consistent with simulation. However, DIS.d.d4 and DIS.d.d5 appear to be unstable. In the first application, they are quite close to FDR.d in terms of the number of significant probe sets, which is nicely consistent with the finding in the simulations. Because we use the same sample size in simulation and this application.

In the second application, however, these two methods unexpectedly identify over 5-fold significant probe sets than FDR.d. The reason might be the fact that the sample size for each strain in this application is smaller and corresponding degrees of freedom become smaller when applying *d* distribution methods. Such an instability of the performances of DIS.d.d4 and DIS.d.d5 within applications needs more study in future work. Furthermore, the distributions of the pooled standard deviations across all probe sets in two applications indicate that all probe sets are not equally variant (Figure 5.1 and 5.3). Therefore DIS.d.d2 and DIS.d.d3 methods taking into probe set variability account are supposed to be more applicable than DIS.d.d4 and DIS.d.d5. But they appear to be quite conservative and produce very limited number of significant probe sets. It forces us in the future to find a better estimator of population variance so that *d* distribution method will be closer to FDR.d method and more applicable in practice.

Appendix A

Partial R codes

A.1 R codes for calculating p-values from permutation

```
pvalues <- function(perm.t.stats,obs.t.stats){

 # pvalues function is used to calculate p-values by permutation
 # method. This function implements the formula (3.3).

 # The following values should be input into pvalues function.
 # perm.t.stats = permuted test statistics
 # obs.t.stats = observed test statistics

 # The output of this function is a vector with the length of m.

  m <- length(obs.t.stats)
  # calculate the number of permutations for each probe set
  b <- dim(perm.t.stats)[2]
  perm.t.stats <- abs(perm.t.stats)
  obs.t.stats <- abs(obs.t.stats)
  p <- rep(NA,m)
  for(i in 1:m){
    p[i]<-mean(perm.t.stats >= obs.t.stats[i])
  }
  return(p)
}
```

A.2 R codes for calculating p-values from d-distribution method

```
fx <- function (x,k,r,c){

 # fx function is used to calculate cumulative distribution function of
 # d-statistics under null hypothesis in the formula (3.4).
 # x is a random variable whose distribution is of our interest;
 # k is a point of distribution and r is degrees of freedom.
 # c is a positive constant number.

    2*r*x*dchisq(r*x^2,r)*pnorm(k*(x+c))
}

 # When calculating p-values from the cdf of d-statistics, we
 # need to input following values into integrate function so
 # that we can obtain p-values.
 # obs.t.stats = observed d statistics
 # k = negative absolute value of observed d-statistics
 # n1 = sample size in strain 1
 # n2 = sample size in strain 2
 # c = a function of s0 in the formula (3.5)
 # sd = standard deviation of probe sets

   d <- obs.t.stats
   p <- rep(NA,length(d))
   for (j in 1:length(d))
       p[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
       c = s0/(sd[j] * sqrt(1/n1 + 1/n2)))$value) * 2
```

A.3 R codes for calculating q-values in terms of p-values

```
qvalue.pval <- function(p,pi0,genes){

 # qvalue.pva function is used to calculate q-values in terms
 # p-values. This function is applicable to the p-values either
 # from permutation method or analytical distribution method.
 # This function implements the algorithm in Storey and Tibshirani
 # (2003b) and borrows some codes in qvalue package wrote by Dabney
 # and Storey (2009).
 # The function performs PERM.d, DIS.d, DIS.d.d1, DIS.d.d2, DIS.d.d3
 # DIS.d.d4, DIS.d.d5, PERM.t, DIS.t methods.

 # The following values should be input into qvalue.pval function.
 # p = p-values either from permutation or from distribution
 # pi0 = the estimated proportion of non-differentially expressed
 #       genes
# genes = probe set ID or name

# The output of this function is a four-dimension data frame.

  u <- order(p)
  m=length(p)
  qvalue.rank <- function(x) {
       idx <- sort.list(x)
       fc <- factor(x)
       nl <- length(levels(fc))
       bin <- as.integer(fc)
       tbl <- tabulate(bin)
       cs <- cumsum(tbl)
       tbl <- rep(cs, tbl)
       tbl[idx] <- tbl
       return(tbl)
   }

  v <- qvalue.rank(p)
  qvalue <- pi0 * m * p/v
  qvalue[u[m]] <- min(qvalue[u[m]], 1)
  for (i in (m - 1):1) {
       qvalue[u[i]] <- min(qvalue[u[i]], qvalue[u[i + 1]], 1)
   }

  data.frame (pvalues=p, qvalues=qvalue, pi0 = pi0, genes = genes)

}
```

A.4 R codes for calculating p-values from d distribution when applying various estimators of population variance in simulations

```
# Since d distribution method depends on the true population
# variances across all probe sets which are generally unknown in
# reality, we propose five estimators and associated estimated c.

# following values need be prepared
# sd1 = standard deviation of expressions for each probe set in strain 1
# sd2 = standard deviation of expressions for each probe set in strain 2
# n1 = sample size in strain 1
# n2 = sample size in strain 2
# obs.t.stats = observed d statistics
# y = data matrix of gene expression under two strains
# m = the number of probe sets

  d <- obs.t.stats
  p1 <- p2 <- p3 <- p4 <- p5 <- rep(NA,length(d))
```

```
# 1. The pooled estimate: s2p in formula (3.8)
# associated estimated c in formula (3.9)

  s2p <- ((n1-1)*sd1^2 + (n2-1)*sd2^2)/(n1+n2-2)
  for (j in 1:length(d))
      p.d1[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
      c = s0/sqrt(s2p[j] * (1/n1 +1/n2)))$value)*2

# 2. The null estimate: s2p in formula (3.8)
# associated estimated c in formula (3.10)

  for (j in 1:length(d))
      p.d2[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
      c = s0/(s0 + sqrt(s2p[j] * (1/n1 +1/n2))))$value)*2

# 3. The pooled estimate: ss in formula (3.11)
# associated estimated c in formula (3.12)

  ss <- apply(y,1,var)
  for (j in 1:length(d))
      p.d3[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
      c = s0/sqrt(ss[j] * (1/n1 +1/n2)))$value)*2

# 4. The average of the pooled estimates across all probe sets
# in formula (3.13)
# associated estimated c in formula (3.14)

   ave.s2p <- sum(s2p)/m
   for (j in 1:length(d))
      p.d4[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
      c = s0/sqrt(ave.s2p * (1/n1 +1/n2)))$value)*2

# 5. The average of the null estimates across all probe sets
# in formula (3.15)
# associated estimated c in formula (3.16)

   ave.ss <- sum(ss)/m
   for (j in 1:length(d))
      p.d5[j] <- (integrate(fx, 0, Inf, k=-abs(d[j]), r=(n1 + n2 - 2),
      c = s0/sqrt(ave.ss * (1/n1 +1/n2)))$value)*2
```

A.5 R codes for simulations

```
# We use the following functions to generate m*(n1+n2)
# simulated data set y.
# m = the number of probe sets
# m1 = number of differentially expressed probe set
# mu = mean of normal distribution
# sd = standard deviation of normal distribution
# n1 = sample size of strain 1
# n2 = sample size of strain 2

# According to different error terms assumption and variance
# variability across probe set, we generate data by the following
# different function.

# 1. Generate data from normal distribution N(0,1) with a common
# variance across probe sets. The following function generate data
# for simulation 1. This function is applicable to simulation
# 4,5 and 6 replacing rnorm() as rt().

   genes.data.1 <- function(m, n1, n2, m1=m*0.1, mu, sd){

      diff1 <- c(seq(-3, -1, length=m1/2), seq(1,3,length=m1/2))
      diff <- c(diff1,rep(0, m-m1))

      # sd = 1 in this simulation
      strain1 <- mu + matrix(rnorm(m * n, 0, sd), nrow = m)*sd
      strain2 <- mu + diff*sd + matrix(rnorm(m * n, 0, sd), nrow = m)*sd
      cbind(strain1, strain2)
   }

# 2. Generate data from t distribution t(3)/sqrt(3) allowing
# different variance from chi-square distribution with df=20
# divided by 20. The following function generate data
# for simulation 1.

    sdi <- sqrt(rchisq(m,20)/20)

    genes.data.2 <- function(m, n1, n2, m1=m*0.1, mu, sdi){

       diff1 <- c(seq(-3, -1, length=m1/2), seq(1, 3, length=m1/2))
       diff <- c(diff1, rep(0,m-m1))

       z1 <- matrix(rt(m*n1,3)/sqrt(3), nrow=m)
       strain1 <- mu + sdi*z1
       z2 <- matrix(rt(m*n2,3)/sqrt(3), nrow=m)
       strain2 <- mu + sdi*(diff+z2)
       cbind(strain1, strain2)
    }

# 3. Generate data from normal distribution N(0,1) allowing
# different variance from chi-square distribution with df=20
# divided by 20. The following function generate data
# for simulation 3.

    sdi <- sqrt(rchisq(m,20)/20)
```

```
genes.data.3 <- function(m, n1, n2, m1=m*0.1, mu, sdi){

   set.seed(seed)
   diff1 <- c(seq(-3,-1,length=m1/2),seq(1,3,length=m1/2))
   diff <- c(diff1, rep(0,m-m1))

   z1 <- matrix(rnorm(m*n, 0, 1),nrow=m)
   z2 <- matrix(rnorm(m*n, 0, 1),nrow=m)
   strain1 <- mu+sdi*z1
   strain2 <- mu+sdi*(z2+diff)
   cbind(strain1, strain2)
}
```

```
# 4. Generate data from other various symmetric distributions with
# a common variance in each data set. The following function implement
# simulation 7.

    genes.data.4 <-function (m, n1, n2, m1 = m * 0.1, mu, sd) {

       diff1 <- c(seq(-3, -1, length = m1/2), seq(1, 3, length = m1/2))
       diff <- c(diff1, rep(0, m - m1))
       z1 <- matrix(rexp(m * n1, 1) * (2 * rbinom(m * n1, 1, 0.5)
                    - 1)/sqrt(2), nrow = m)
       strain1 <- mu + sd * z1
       z2 <- matrix(rexp(m * n2, 1) * (2 * rbinom(m * n2, 1, 0.5)
                    - 1)/sqrt(2), nrow = m)
       strain2 <- mu + sd * (diff + z2)
       cbind(strain1, strain2)
    }
```

A.6 R codes for applications

```
# When testing methods in applications, we use linear model (5.1)
# to calculate test statistics which take into account day effect.
# Accordingly the degrees of freedom of d and t distribution methods
# become (n1-1)+(n2-1)-(Nday-1) where Nday is the number of levels of
# day effect. Worth noting is that the estimates of variance when day
# effect is present are the mean square errors. The pooled variance
# estimate is the mean square error from model (5.1) and the null variance
# estimate is the mean square error from model (5.4).

# y = the full gene expression matrix with each row representing a probe set
# strain = strain effect vector
# day = day effect vector
# n = sample size of each strain
# 1. Calculate test statistics, s0 and pi0
     # Calculate observed test statistics
     blockstats <- block.stats(y, strain, day)
     dstats <- blockstats$d.stat  # d statistics
     tstats <- blockstats$t.stat  # t statistics
     sd1 <- blockstats$sd1
     sd2 <- blockstats$sd2
     s0 <- blockstats$s0[1]

     # Calculate permuted test statistics
     perm.matrix <- perms.block(strain, day)
     d.stats.perm <- function(strain, day, exprs, s0)
         block.stats(exprs, strain, day, s0, return.all=F)
     permuted.d.stats <- apply(perm.matrix, 1, d.stats.perm,
         exprs=y, s0=s0, day=day)

     t.stats.perm <- function(strain, day, exprs, s0=0)
         block.stats(exprs,strain,day,s0=0,return.all=F)
     permuted.t.stats <- apply(perm.matrix, 1, t.stats.perm,
         exprs=y, s0=0, day=day)

     # Calculate pi0
     pi0 <- calc.pi0(dstats, permuted.d.stats)

 # 2. Test methods FDR.d, PERM.d, FDR.t, PERM.t and DIS.t
     g.name <- row.names(y)   # take out probe set ID
     # A. Calculate q-values from FDR.d
     observed.ordered.d <- sort(dstats)
     ordered.permuted.d <- apply(permuted.d.stats, 2, sort)
     expected.permuted.d <- rowMeans(ordered.permuted.d)
     result1 <- qvalue1.new(dstats, expected.permuted.d,
          permuted.d.stats, pi0, g.name)

     # B. Calculate q-values from PERM.d
     p.d <- pvalues1(permuted.d.stats, dstats)
     result2 <- qvalue.pval1(p=p.d, pi0=pi0, genes=g.name)

     # C. Calculate q-values from FDR.t
     ordered.permuted.t <- apply(permuted.t.stats, 2, sort)
     expected.permuted.t <- rowMeans(ordered.permuted.t)
     result3 <- qvalue1.new(tstats, expected.permuted.t,
           permuted.t.stats, pi0, g.name)

     # D. Calculate q-values from PERM.t
     pvalues <- pvalues1(permuted.t.stats, tstats)
     result4 <- qvalue.pval1(p=pvalues, pi0=pi0, genes=g.name)

     # E. Calculate q-values from DIS.t
     t <- tstats
     p.t <- 2*pt(-abs(t),df=2*n-2-1)  # day effect would affect linear model
```

```
result5 <- qvalue.pval1(p=p.t, pi0=pi0, genes=g.name)

# F. d distribution method
d <- dstats
p.d <- p.d1 <- p.d2 <-p.d3 <- p.d4 <- p.d5 <- rep(NA,length(d))

sigma2 <- matrix(NA, nrow=nrow(norandidd), ncol=2)
for (i in 1:nrow(sigma2)) {
   # take out Residual standard error and then get "pooled estimate"
   sigma2[i,1] <- summary(lm(norandidd[i,]~day+strain))$sigma^2
   # take out Residual standard error and then get "null estimate"
   sigma2[i,2] <- summary(lm(norandidd[i,]~day))$sigma^2
}

# method 1 - pooled variance
s2p <- sigma2[,1]
for (j in 1:length(d)) p.d1[j]=(integrate(fx, 0, Inf, k=-abs(d[j]),
   r= (2*n-2-1), c=s0/sqrt(2*s2p[j]/n))$value)*2
result.d1 <- qvalue.pval1(p=p.d1,pi0=pi0,genes=g.name)

# method 2 - pooled variance and transformation of c
for (j in 1:length(d)) p.d2[j]=(integrate(fx, 0, Inf, k=-abs(d[j]),
   r=(2*n-2-1), c=s0/(sqrt(2*s2p[j]/n)+s0))$value)*2
result.d2 <- qvalue.pval1(p=p.d2,pi0=pi0,genes=g.name)

# method 3 - null variance
ss <- sigma2[,2]
for (j in 1:length(d)) p.d3[j]=(integrate(fx, 0, Inf, k=-abs(d[j]),
   r=(2*n-2-1),c=s0/sqrt(2*ss[j]/n))$value)*2
result.d3<- qvalue.pval1(p=p.d3, pi0=pi0, genes=g.name)

# method 4 - the average of pooled variance
ave.s2p <- mean(s2p)
for (j in 1:length(d)) p.d4[j]=(integrate(fx, 0, Inf, k=-abs(d[j]),
   r= (2*n-2-1),c=s0/sqrt(2*ave.s2p/n))$value)*2
result.d4 <- qvalue.pval1(p=p.d4, pi0=pi0, genes=g.name)

# method 5 - the average of null variance
ave.ss <- mean(ss)
for (j in 1:length(d)) p.d5[j]=(integrate(fx, 0, Inf, k=-abs(d[j]),
   r= (2*n-2-1),c=s0/sqrt(2*ave.ss/n))$value)*2
result.d5 <- qvalue.pval1(p=p.d5, pi0=pi0, genes=g.name)
```

Bibliography

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**: 289-300.

[2] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**: 1165-1188.

[3] Benkalhaa, H.O. and Polychronakos, C. (2008). The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. *Trends in Molecular Medicine* **14**: 268-275.

[4] Best, D.I. and Rayner, C.W. (1987). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics* **29**: 205 -220.

[5] Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19(2)**: 185-93.

[6] Canty, A. (2005). Analysis of Affymetrix Microarrays in R.

[7] Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V. (2005). SAM, Significance Analysis of Microarrays: Users Guide and Technical Document.

[8] Cui, X. and Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**: 210.

[9] Dabney, A. and Storey, J. with assistance from Warnes, G.R. (2009). qvalue: Q-value estimation for false discovery rate control. R package version 1.20.0. http://CRAN.R-project.org/package=qvalue.

[10] Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002). Statistical methods for identfying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**: 111-139.

[11] Dudoit, S., Shaffer, P.J. and Boldrick, C.J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**: 71-103.

[12] Dudoit, S. and van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics.* Springer. New York.

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

[14] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 3 (Feb. 2004): 307-315.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance . *Biometrika* **75**: 800-803.

[16] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**: 65-70.

[17] Irizarry, R. A., Gautier, L., and Cope, L.M. (2003a). An R package for analyses of affymetrix oligonucleotide arrays. In *The Analysis of Gene Expression Data: Methods and Software.* (Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L. editors). Springer.

[18] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**: 249-264.

[19] Ivakine, E.A., Fox, C.J., Mortin-Toth, S.M., Canty, A., Walton, D.S., Paterson, A.D., Aleksa, K., Ito, S. and Danska, J.S. (2005). Sex-Specific Effect of Insulin-Dependent Diabetes 4 on Regulation of Diabetes Pathogenesis in the Nonobese Diabetic Mouse. *The Journal of Immunology.* **174**: 7129-7140.

[20] Kerr, M.K., Martin, M., and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**: 819-837.

[21] Lee, M.T. (2004). *Analysis of Microarray Gene Expression Data.* Kluwer Academic Publishers. USA.

[22] Lönnstedt, I., and Speed, T.P. (2002). Replicated microarray data. *Statistica Sinica* **12**:13-46.

[23] McLachlan, G.J., Do, K.A. and Ambroise, C. (2004). *Analzing Microarray Gene Expression Data.* John Wiley and Sons. New Jersey.

[24] Pollard, K.S. and Van der Laan, M.J (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* **125**: 85100.

[25] Public Health Agency of Canada 2005-2006 National Health Survey, http://www.phac-aspc.gc.ca/cd-mc/diabetes-diabete/face-eng.php.

[26] Reiner, A., Yekutieli, D. and Benjamini, Y. (2002). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**: 368-375.

[27] Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differentially expression in Microarray experiments. *Statistical Applications in Genetics and Molecular Biology.* Vol. 3 : Iss. 1, Article 3.

[28] Smyth, G. K. (2005). Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pages 397–420.

[29] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**: 479-498.

[30] Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**: 2013-2035.

[31] Storey, J. and Tibshirani, R. (2003a). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software*, (Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L. editors). Springer.

[32] Storey, J. and Tibshirani, R. (2003b). Statistical significance for genomewide studies. *Proceedings of the National Academy of Science of the United States of America* **100**: 9440-9445.

[33] Student [William Sealy Gosset]. (1908). The probable error of a mean. *Biometrika.* **6**(1): 1-25.

[34] Takahashi, N., Rieneck, K., Van der Kraan, P.M., Van Beuningen, H.M., Vitters, E.L., Bendtzen, K. and Van den Berg, W.B. (2005). Elucidation of IL-1/TGF-beta interactions in mouse chondrocyte cell line by genome-wide gene expression. *Osteoarthritis Cartilage* **13**: 426-38.

[35] Tusher, V., Tibshirani, R. and Chu, C. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Preceedings of the National Academy of Sciences* **98**: 5116-5121.

[36] Welch, B.L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29**: 21-52.

[37] Wilder, S.P., Kaisaki, P.J., Argoud, K., Salhan, A., Ragoussis, J., Bihoreau, M.T. and Gauguier,. D. (2009). Comparative analysis of methods for gene transcription profiling data derived from different microarray technologies in rat and mouse models of diabetes. *BMC Genomics* **10**: 63.

[38] Yuan, M., Kendziorski, C., Newton,M. and Sarkar, M. (2007). EBarrays: Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification. R package version 2.8.0.