This is to certify that the
dissertation entitled

SHRINKAGE PROCEDURES FOR MIXED MODEL ANALYSES OF MICROARRAY EXPERIMENTS

presented by

Lan Xiao

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Department of Animal Science

Robert J. Tempelman

Major Professor's Signature

August 14, 2007

Date

SHRINKAGE PROCEDURES FOR MIXED MODEL ANALYSES OF MICROARRAY EXPERIMENTS

By

Lan Xiao

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Animal Science

2007

ABSTRACT

SHRINKAGE PROCEDURES FOR MIXED MODEL ANALYSES OF MICROARRAY EXPERIMENTS

By

Lan Xiao

Two color microarray systems are amongst the currently most popular functional genomics tools that have permeated animal science research. This novel technology facilitates the simultaneous profiling of the behavior of tens of thousands of genes under various experimental conditions.

Data generated by microarray experiments are typically influenced by a number of complex sources of systematic and random experimental variation. Mixed models provide a powerful means to account for multiple sources of variation in very general and efficient experimental designs. Now the number of hypotheses tests are a linear function of the number of genes, each test limited by generally few replicates per treatment condition due to the substantial costs of a microarray experiment. Although several Bayesian methods have been deemed effective for borrowing information across genes for the analysis of microarray data, there remains unresolved issues for more elaborate design structures characterized by differing levels of replication.

Two alternative Bayesian approaches to mixed model inference of microarray experiments are presented in this dissertation. These methods facilitate more reliable inferences on gene effects by borrowing information from the whole ensemble of genes on not just one but several layers of variability. A proposed empirical Bayes mixed model (EB-ANOVA) pools information on ANOVA mean squares for random and residual effects across genes, thereby improving sensitivity for detecting differential expression

while providing adequate control of the false discovery rate (FDR). A second model (BAYESRATIO) was subsequently constructed to generalize the common correlation assumption for microarrays having two or more spots per gene, as currently implemented in the popular R software package LIMMA. The BAYESRATIO model was shown to have better performance on ROC curves and FDR control, where LIMMA was found to be too liberal for controlling FDR. A third chapter compares different image analysis software combined with the statistical methods we proposed in previous two chapters. The significantly different data features from different image software result in dissimilar statistical inferences. The findings from this work support the contention that the background adjustment may substantially reduce the precision and increase the variability of intensity estimation.

ACKNOWLEDGMENTS

My first acknowledgement must go to my advisor, Dr. Robert Tempelman. Without his encouragement and patience, I would not have finished this work. I greatly appreciated his guidance not only for my academic and professional growth but also skill improvement, which is so helpful in my current career.

I also wish to express many thanks to my guidance committee members. Dr. Ernst devoted her precious time to help me with understanding genetics. Dr. Burton offered me collaboration opportunity and also provided me great assistance for getting laboratory experience. I'd like to thank Dr. Rosa, who was my second reader and always gave me great suggestions about my study and my current job, Dr. Huebner for her time and efforts in statistical genetic and genomics, where I could benefit from, and Dr. Cui for his help to be my additional examiner within a short notice.

I also owe great gratitude to my colleagues, Juan Pedro, Fernando, Pablo, Dave, Nora, LingChu, XiaoNing, Sally, Patty, Kelly and Sue, who helped me in different ways.

My heartfelt appreciation definitely goes to my family and my friends, whose support, encouragement, and companionship made my long journey as a graduate student here more enjoyable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

1. Statistical Challenges in Gene Selection for Microarray Data Analysis

Microarray technology has empowered biomedical researchers to study the simultaneous expression of thousands of genes as influenced by various experimental conditions. Of course, the large volume of gene expression data also poses many statistical inference challenges. One important goal in microarray studies is to identify a subset of genes which is differentially expressed between two or more treatments. A main limitation in this exercise is the limited availability of numbers of samples (i.e., less than 10 samples per treatment) and the extremely large number of genes (i.e., thousands of genes). Small sample sizes may substantially compromise statistical power for analyses conducted separately for each gene; compounding this problem further is the multiple testing considerations that must be considered when conducting thousands of hypothesis tests. Moreover, microarray experiments involve multi-step processes such as extraction of mRNA, reverse transcription, labeling, hybridization, scanning, and image analysis. Each step represents a potential source of variation and error, thereby affecting the measurement of gene expressions and further exacerbating the detection of differential expression.

The most common experimental design for two color arrays (cDNA microarray) is the reference design, where each experimental sample is hybridized against a common reference sample. A less common design is the loop design where each sample is hybridized to each of two different samples in two different dye orientations and can be

connected as a loop (Kerr & Churchill 2001). For more complex factorial treatment structures (Glonek & Solomon 2004), some contrasts might be considered to be more important than others, thereby influencing design choices. That is, an efficient design typically utilizes more direct comparisons, i.e. within arrays, for the most interesting contrasts and indirect comparisons for others. Furthermore, a typical microarray experiment has many sources of variation which can be attributed to biological and technical levels of replication (Zakharkin et al. 2005). Confusing technical duplicates with biological replicates will lead to misconceptions in conducting and interpreting statistical tests (Peng et al. 2003).

Image analysis is an important stage of microarray experiments and can have a potentially large impact on subsequent analyses, such as the identification of differentially expressed genes (Yang et al. 2002). The primary purpose of image analysis step is to convert TIFF images which contain both Red and Green channels to numeric data, specifically foreground and background intensity information for each spot and wavelength for each of the two different dyes (e.g. Cy3 and Cy5). The process of image analysis after scanning the array includes locating each spot on the slide, partitioning the pixels within each grid box into foreground and the background, and quantifying the intensity values and some quality control measures for the Cy3 and Cy5 channel for each spot on the microarray. This data set generated by such image analysis is the object used for statistical analysis. The choice of segmentation method in image analysis software forms a crucial preliminary step in microarray analysis as any errors incurred at this step are bound to propagate through subsequent data analysis.

2. Hypothesis testing for microarray experiments

Mixed model methods provide a natural framework for analyzing microarray experimental data generated from efficient and flexible experimental designs including those where it is fundamentally necessary to discriminate between biological and technical replication. A widely popular two-step mixed model procedure (step one: normalization model only contains global effects; step two: gene-specific model only contains gene-specific effects) for the analysis of microarray data was first introduced by (Wolfinger et al. 2001) based on separate mixed model ANOVA for each gene. However, the low power for one gene-at-a-time hypothesis testing based on small sample sizes is not well served by this simple approach.

Empirical Bayes estimation is an inference procedure having great practical potential for microarray data by combining information on thousands of gene expression levels, each characterized by limited replicates per treatment group (Efron 2003). The information extracting across genes is summarized as prior distributions which can be then used to improve the estimates for individual genes by shrinking estimates to a common value; hence empirical Bayes is often referred to as shrinkage estimation. Much work has been pursued on shrinkage estimation for simple microarray experimental designs based on a single error structure (Efron et al. 2001; Newton et al. 2001; Broet et al. 2002; Lonnstedt & Speed 2002; Kendziorski et al. 2003; Wright & Simon 2003; Edwards et al. 2005). However, most of these approaches are not readily applicable to experimental designs with hierarchical replication structures. Empirical Bayes extensions of mixed model analysis appears to be a promising strategy as realized in a recent version of LIMMA (Linear Models for Microarray Data) (Smyth et al. 2005) , MAANOVA (Cui

et al. 2005) and another recently published paper (Feng et al. 2006). Nevertheless, there are still unresolved issues with those procedures: strong common correlation assumption in LIMMA, shrinkage on variance components in MAANOVA which may not result in more accurate denominator of F-statistics for hypothesis testing, and arbitrary choice of the posterior degree of freedom in Feng et al., (2006).

3. Specific objectives

We combine the strengths of these two approaches for inference: the mixed effect model and Bayesian methods to improve the precision of identifying differentially expressed genes in microarray study.

The overall aim for this thesis is to improve the efficiency of statistical inference of microarray data, specifically mixed model analysis of efficient experimental designs. The three objectives for this dissertation are as follows:

1) To develop an empirical Bayes extension of mixed model analysis for microarray data by combining information on gene-specific variance components for every random source of variability including blocking, experimental and technical sources.

2) To critically evaluate the empirical Bayes strategy in the popular microarray analysis software LIMMA for managing within-array technical replicates via a comparison with a fully Bayesian method.

3) To study different data features coming from different image analyses software and suggest transformations and models that would be most appropriate for the respective characteristics of data.

4. Dissertation outline

The first major part of the thesis includes a literature review on the biological research underpinnings and required statistical inference relevant for the analysis of cDNA microarrays, including an overview of cDNA microarray technology, experimental design, image analysis methods, data normalization, single gene analysis methods, multiple test adjustment issues and relevant statistical concepts. The second major part of this dissertation consists of three independent papers. Paper I proposes an alternative empirical Bayes strategy for mixed model ANOVA to infer upon differentially expressed genes, and to assess the performance by comparing this strategy with recently proposed methods based on empirical Bayes (Feng et. al., 2006) and gene-specific inference (Wolfinger et. al., 2001). Paper II uses a fully Bayesian model to investigate a strong distributional assumption of LIMMA (Smith et. al., 2005), that being of a constant within-array correlation for technical replicates across all genes. Paper III compares four different image analyses software representing three segmentation methods: adaptive circle, adaptive shape and histogram methods to investigate the variability of data derived from different segmentation methods and its impact on subsequent data analyses. The third major part of this dissertation is the conclusions and areas for future study. The summary of results address the specific objectives stated at the beginning of this dissertation. The implications of this dissertation and proposal for future work are discussed.

BIBLIOGRAPHY

BROET, P. RICHARDSON, S. & RADVANYI, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**(4), 671-683.

CUI, X. G. HWANG, J. T. G. QIU, J. BLADES, N. J. & CHURCHILL, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**(1), 59-75.

EDWARDS, J. PAGE, G. GADBURY, G. HEO, M. KAYO, T. WEINDRUCH, R. & ALLISON, D. (2005). Empirical Bayes estimation of gene-specific effects in micro-array research. *Funct Integr Genomics* **5**(1), 32-39.

EFRON, B. (2003). Robbins, empirical Bayes and microarrays. *Annals of Statistics* **31**(2), 366-378.

EFRON, B. TIBSHIRANI, R. STOREY, J. D. & TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**(456), 1151-1160.

FENG, S. WOLFINGER, R. D. CHU, T. M. GIBSON, G. C. & MCGRAW, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural Biological and Environmental Statistics* **11**(2), 197-209.

GLONEK, G. F. V. & SOLOMON, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**(1), 89-111.

KENDZIORSKI, C. M. NEWTON, M. A. LAN, H. & GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**(24), 3899-3914.

KERR, M. K. & CHURCHILL, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**(2), 123-128.

LONNSTEDT, I. & SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**(1), 31-46.

NEWTON, M. A. KENDZIORSKI, C. M. RICHMOND, C. S. BLATTNER, F. R. & TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37-52.

PENG, X. J. WOOD, C. L. BLALOCK, E. M. CHEN, K. C. LANDFIELD, P. W. & STROMBERG, A. J. (2003). Statistical implications of pooling RNA samples for microarray experiments. *Bmc Bioinformatics* **4**.

SMYTH, G. K. MICHAUD, J. & SCOTT, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**(9), 2067-2075.

WOLFINGER, R. D. GIBSON, G. WOLFINGER, E. D. BENNETT, L. HAMADEH, H. BUSHEL, P. AFSHARI, C. & PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625-637.

WRIGHT, G. W. & SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**(18), 2448-2455.

YANG, Y. H. BUCKLEY, M. J. DUDOIT, S. & SPEED, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**(1), 108-136.

ZAKHARKIN, S. O. KIM, K. MEHTA, T. CHEN, L. BARNES, S. SCHEIRER, K. E. PARRISH, R. S. ALLISON, D. B. & PAGE, G. P. (2005). Sources of variation in Affymetrix microarray experiments. *Bmc Bioinformatics* **6**.

CHAPTER 1. LITERATURE REVIEW

1.1 Introduction

Two color microarrays are widely used as powerful functional genomics tools in the animal sciences. Because of the vast amount of data generated by microarray experiments, the rapid growth of this technology has required necessary scaling developments for statistical inference. Many aspects of statistical methods are driven by this requirement, including experimental design, image analysis and hypothesis testing.

This review discusses widely used methods for segmentation in image analysis software, experimental designs and the statistical analysis of gene expression data from DNA microarrays. Discussion on normalization and multiple testing adjustments for hypothesis testing are also provided.

This review is broken down into two major sections. Section 1, albeit very short, provides a brief general overview of microarray technology. Section 2 contains a general development and discussion of the current issues in the statistical analysis of microarray data and is further subdivided into four parts: Section 2.1, common experimental designs; Section 2.2, image analysis; Section 2.3, data normalization and Section 2.4 statistical inference.

1.2 DNA microarray technology

Microarrays were originally developed to facilitate the measurement of simultaneous expression of thousands of genes (Schena 1995). As a result, various studies have been

conducted to study gene expression profiles under various conditions including the study of complex diseases and developmental processes.

DNA microarrays are typically made of glass slides with orderly arranged spots of DNA fragments. The DNA fragments act as probes, which have various characteristics for different platforms. Before the turn of the millennium, traditional cDNA and short oligonucleotide probe microarrays predominated whereas long oligonuleotide (50-70 mer) platforms have now become more popular (Woo et al. 2004). A standard glass slide is 1×3-inch (Gershon 2002), Since the sizes of the spots are typically less than 200 microns in diameter, microarrays are generally large enough for thousands of spots, thereby allowing an investigator to study the expression of nearly just as many genes within a single slide.

Spotted microarrays are one particular type of microarray where competitive hybridization is used to compare the relative amount of mRNA transcript for each gene from two samples treated under different conditions; e.g. treated vs. control. This process as indicated in Figure 1 starts with reverse transcribing mRNA from each sample into cDNA. Each sample is typically labeled with one of two fluors or dyes. The most common pair of dyes are Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green)). Roughly equal volumes of each sample are generally hybridized together within a single array.

After hybridization, the array or slide is scanned with lasers in order to produce two digital images with Tagged Image File Format (TIFF), one for each dye channel. Different dyes absorb and emit light at different wavelengths. In order to measure the abundance of the two fluorescent dyes for each spot, the scanners are designed to

generate excitation light at different wavelengths and detect different emission wavelengths. The dyes Cy3 and Cy5 have emission in 510-550nm and 630-660nm wavelength ranges, respectively (Yang et al. 2002a). The digital images scanned at the two wavelengths are then used to produce a pseudoimage of the array. If one spot corresponding to a particular gene in the pseudoimage appears to be red, it qualitatively suggests that a higher level of expression for the specific gene at the spot exists within the sample labeled with Cy5 whereas a green spot indicates relatively greater gene expression in the Cy3 labeled sample. Any difference in expression between the two samples (up- or down-regulation) is referred to as differential expression. Qualitatively, non-differentially expressed genes (probes) are expected to yield yellow spots because of the equal expression of Cy3 and Cy5 (Qin et al. 2005).

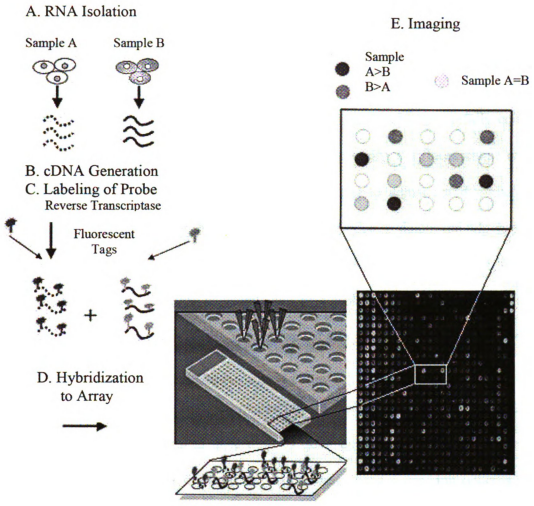


Figure 1.1 Basic process of a comparative hybridization experiment. Figure is taken from www.fao.org/DOCREP/.

1.3 Summary of statistical issues for two color microarray experiments

Consideration of appropriate statistical methods is needed for various stages of any microarray experiment. Firstly, the experimental design should be carefully constructed prior to actually conducting the experiment itself. Secondly, the fluorescence intensities procedure from image analysis should be normalized to adjust for dye-bias and for systematic variation. Thirdly, hypothesis testing is needed to determine which genes are differentially expressed between different conditions (Smyth et al. 2003). Differentially expressed genes identification is usually the first biological concern and also the core goal to be reviewed for this stage. The sections of this review correspond roughly to these various steps.

There are other important analyses conducted for microarray experiments that involve categorizing genes or samples according to gene expression profiles. For example, various clustering methods are used to group similar gene expression patterns across a number of samples (D'Haeseleer 2005). Other computational methods such as gene set enrichment analysis (GSEA) determine if predefined biological classes of genes are differentially expressed in different phenotypes (Shi & Walker 2007), whereas gene expression networks construct paths with links to connect genes having clear dependencies in expression (Khanin et al. 2006). Although these methods are increasingly important, they are beyond the scope of this dissertation and not discussed further.

1.3.1 Experimental Design

Optimizing the design based on the experimental goal is an important part of a successful microarray experiment. There are a number of considerations which should be addressed before conducting an experiment:

1. How many microarray slides need to be used to get balance power against control of false discovery rates (FDR)
2. What should be the relative emphasis of technical to biological replication?
3. What are the most important comparisons?
4. How many experimental factors will be involved?

The answers for these questions can be somewhat addressed by considering the two basic components of experimental design, namely, treatment structure and design structure (Montgomery 1984). Generally speaking, the treatment structure consists of those factors that the experimenter has selected to study; e.g., treatments (patient vs. control), genders (male vs. female). The design structure consists of grouping of the experimental units into homogeneous blocks. Some commonly used design structures are: the completely randomized design (CRD), the randomized complete block design (RCBD), and various deviants of an incomplete block design. Work on microarray experimental designs have been based on applications and extensions of these classical designs. Since a two color microarray experiment is based on hybridizing two samples for pair of comparison on the same slide, it is common to draw a figure with arrows, where each arrow represents one microarray slide and the arrow's tail and head denote the Cy3 (Petersen et al.) and Cy5 (red) labeled samples respectively (see Figure 2).

These tails connect the samples involved in the experiment to provide information about both treatment and design structures.

Kerr and Churchill (2001) first recognized the utility of classical block designs for two color microarray experiments. Figure 2 depicts some simple examples for microarray designs, where caption letters A, B and C refer to treatment assignments for experimental samples and R refers to reference sample. An example of a common reference design is provided in Figure 2a; here, a reference sample is typically defined as a uniform sample that is used for all hybridizations. Since reference designs typically use the log-ratio of the treated over the reference sample as the response variable, the subsequent design can be analyzed as a CRD. A dye-swap design for two treatment comparison is provided in Figure 2b. Here arrays serve as experimental blocks that facilitate within block comparisons somewhat comparable to a RCBD. The connected loop design in Figure 2c is an example of an incomplete block design for comparing three or more treatment groups where only a pair of treatments can be considered for any one hybridization. When considering different levels of replication, the simplified representation in Figure 2 is not specific enough to indicate whether or not the two separate hybridizations involving, for example, Group A, derive from a single biological replicate or from two different biological replicates. This distinction is fundamentally important to determining proper experimental replication, the basis for formal hypothesis testing on treatment mean differences. Figure 3 is a clearer representation of Figure 2 in that the subscripts of each letter A, B and C index the biological replicates; hence the delineation of experimental from technical replication is more clearly represented than in Figure 2.

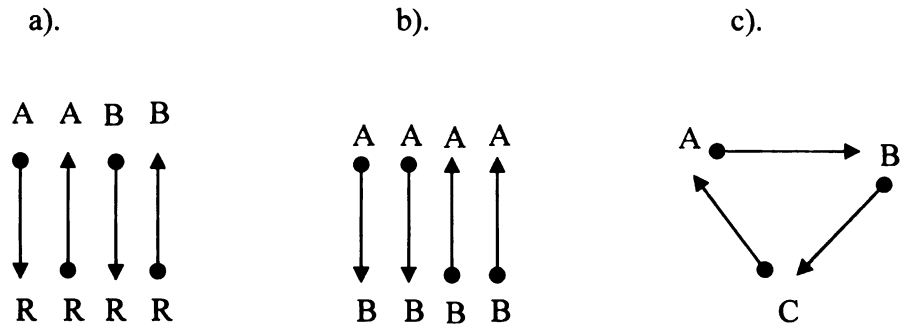


Figure 1.2 The letter A, B and C indicate experimental samples for different groups and R refers to reference sample. Each arrow represents one microarray slide and the arrow's tail and head denote the Cy3 (green) and Cy5 (red) (Petersen et al. 2005): a) common reference design or indirect design. b) direct design. c) connected loop design.

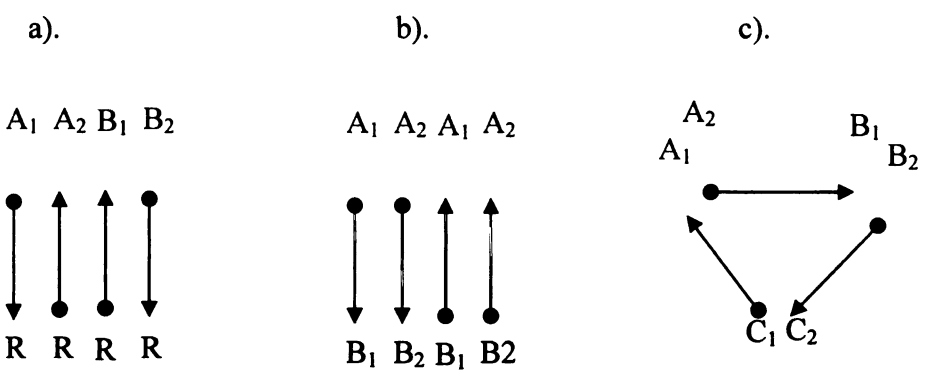


Figure 1.3 The letter A, B and C indicate experimental samples for different groups and R refers to reference sample. The subscripts of each letter represent biological replicates. Each arrow represents one microarray slide and the arrow's tail and head denote the Cy3 (green) and Cy5 (red) (Petersen et al. 2005): This is a more suitable graphical representation of microarray experiments than in Figure 2. a) common reference design or indirect design. b) direct design. c) connected loop design.

The literature on experimental design for microarrays has been extended to consider different sources of variation and hierarchical replication. Hierarchical replication is actually illustrated by Figure 3c) since each sample is hybridized twice such that there is a need to distinguish between the number of samples per treatment (biological replication) and number of hybridizations per sample (technical replication). Kerr and Churchill (2001) first considered A-optimality as a criteria to construct efficient experimental designs; A-optimality pertains to designs where the average squared standard errors of treatment comparisons are minimized. Yang and Speed (2002) emphasized the importance of deciding whether to use direct (within slides) or indirect (between slides) treatment comparisons based on the priority of various research questions. Glonek and Solomon (2004) broadened some simple results established by Yang and Speed (2002) to a conceptual and formal framework by minimizing the standard error of the comparisons of interest as a means to optimize statistical efficiency of factorial and time course designs (Glonek & Solomon 2004). Wolfinger et al. (2001) further extended the approach of Kerr and Churchill (2001) to include random effects in a mixed model ANOVA for microarray data analysis . Tempelman (2005) compared various deviations of reference and loop designs for determinations for statistical precision, power and robustness based on mixed model analysis. Each design deviation was defined by different arrangements of biological replication within the same design layout. Rosa et al. (2005) further reassessed experimental design and analysis of cDNA microarray experiments using mixed effects models

1.3.2 Image Analysis

The process of scanning an array to create a TIFF file is known as image acquisition, whereas the process of converting images to numerical data is referred to as image quantification or processing. The relative abundance of mRNA for any particular gene between the two samples hybridized against each other on a microarray is represented by the relative amount of Cy3 (green) to Cy5 (red) fluorescence at the corresponding spot (Petersen et al. 2005). The data set generated by image quantification provides information about foreground and background intensities and some quality control measures for the red and green channels for each spot on the microarray.

The two major objectives of image analysis are therefore to determine the discrete spot locations and to quantify the spot intensities (Rahnenfuhrer & Bozinov 2004).
The known geometry which places all features presenting the array into a rectangular grid or approximate geometry of the cDNA printing procedure as an input for grid placement enhances the spot-finding procedure (Bozinov & Seidel 2004). This step is typically done automatically using image analysis software along with some user intervention to increase reliability (Smyth et al. 2003). The center of each small square in the grid is an idealized spot center, and the region around each spot center is used to identify the boundaries of the spot in the grid. The pixel values in each grid box are the values used to summarize the expression intensities for each spot (Shaw & Tollett 2001).

The next step is to segment the pixels in each grid box into the precise pixels within spot (the foreground) and those in the background. Various segmentation algorithms have been developed for this kind of spot analysis. These approaches can be broadly classified into methods A. Fixed Circle, B. Adaptive Circle, C. Adaptive Shape and D. the

Histogram methods. A critical assumption invoked for Method A and B is that all spots in the image are circular; in fact, for A, the sizes of all spots are assumed to be the same. The center of each spot and the diameter of the circle can be variable for Method B. Foreground intensity values using Methods A and B are based on the pixels' fluorescence intensities inside the defined circle. Background fluorescence intensities for Method A are simply based on the fluorescence intensities of the pixels outside the fixed circle but inside the grid box. Pixels from the valley spot, which consists of representative pixels from the four corners of the square that encapsulate a given spot, are often used to estimate background fluorescence intensities in Method B. Genepix and MolecularWare are two software programs which implement this algorithm.

Now spots within a microarray image can take shapes other than circular such as ellipses or shapes even more irregular. The adaptive shape segmentation algorithm (Method C) as implemented in Spot uses seeded region growing (SRG) and watershed techniques to deal with different shapes in image segmentation. The assumption of this method requires an initial point, known as the seed. Adjoining pixels are then progressively added to the spot until adjacent spots appear to have distinct pixel value and the running mean of values (Qin et. al., 2005). Therefore, the foreground intensity for a certain spot using Method C is determined by the pixels surrounding this seed as defined by the region growing approach. The background for Method C is estimated by using a non-linear filter called morphological opening. This operation removes all spots, local peaks including artifacts such as dust particles and leaves only the background intensities (Yang et al., 2002a). The histogram-based segmentation uses a target mask which is chosen to be larger than any spot and a histogram is formed from the intensities of the pixels within the mask. A

threshold is computed by using Mann-Whitney test to segment each pixel into foreground or background. Those pixels whose values are greater than this threshold are assigned to foreground region and otherwise as background region. Variations on this method are implemented in Quantarray software, ScanArray Express and Imagene software programs. It should be noted that each of the segmentation techniques work under certain implicit assumptions (Method A and B: shape of each spot as a circle, Method C: the known initial seed, Method D: a suitable mask size) and hence are susceptible to errors when these assumptions are somewhat violated.

It is common to use background corrected fluorescence intensities, i.e. foreground minus background intensities, as microarray data. The motivation of background correction is to obtain a quantification of hybridization not influenced by fluorescence emitted from other chemicals on the glass (Smyth et al. 2003). However, background correction also results in: 1) loss of information associated with low fluorescence intensities due to negative adjustments for spots when foreground intensities are lower than background estimates thereby rendering missing values after log-transformation and 2) greater variability for low intensities where negative adjustments do not occur (Yang et al. 2002a). Therefore, there is some emerging consensus that the background subtraction is not helpful (Allison et al. 2006).

The comparison for different image analysis software programs has been studied mainly based on different segmentation methods (Jenssen et al. 2002; Yang et al. 2002a; Ahmed et al. 2004; Korn et al. 2004; Qin et al. 2005). The precision of the ratio of Cy3 to Cy5 fluorescence intensities based on different segmentation methods associated with different image analysis software has been compared using various aspects such as the

spot to spot variability (Yang et al. 2002a; Ahmed et al. 2004), the correlation coefficient (Jenssen et al. 2002; Ahmed et al. 2004), the repeatability coefficient (Jenssen et al. 2002; Ahmed et al. 2004) and the intra-class correlation coefficient for replicates (Korn et al. 2004). Similar comparisons have also been used for test reproducibility across different microarray platforms (i.e., cDNA, Oligonucletide, and Affymetrix GeneChip) (Woo et al. 2004; Petersen et al. 2005; de Reynies et al. 2006).

1.3.3 Data normalization and transformation

Normalization refers to the process of removing global systematic effects (Cui et al. 2003) after an appropriate data transformation, typically a log transformation to base 2. Data normalization always includes a calibration of the signals from different microarrays to put all signals on a comparable scale (Cui et al. 2003). Data normalization approaches have been proposed based on different assumptions. However, one of the most common assumptions in microarray experiment data is that the majority of the genes are not differentially expressed between comparative conditions (Cheadle et al. 2003) and that the relative number of upregulated and downregulated genes between the two samples on a slide is roughly the same.

There are a large number of data normalization procedures that have been proposed. For example, global ANOVA methods (Kerr et al., 2000; Jin et al., 2001; Wolfinger et al., 2001) have been proposed to adjust for overall effects of array and dye or/and other systematic effects across genes. Wang et al. (2002) further proposed an iterative regression normalization algorithm to unify the tasks of estimating normalization coefficients of regression mapping from gene expression values of reference channel to

other channels and identifying the control gene set to normalize microarray expression data.

The choice of appropriate data transformations may depend upon various features of microarray data. There are several strategies to transform microarray data in order to remove dependence on mean-scale dependencies that are often observed with microarray fluorescence intensities. One transformation involves the shift method, which adjusts the signals of the two channels using an additive constant prior to logarithmic transformation; this constant is typically estimated by minimizing the absolute deviation of each log ratio from the median log ratio of the array (Newton et al. 2001; Kerr et al. 2002). Other examples include curve-fitting strategies, which use local (on intensity axis) regression to estimate a standard curve and then re-center the data (Yang et al. 2002b)). Rocke and Durbin (2003) and Huber et. al. (2002)independently introduced a family of transformations (the generalized-log family (glog)) to further stabilize the variance of microarray data . Ishwaran and Rao (2003) used permutation methods to cluster genes with similar variance and rescaled gene expression within each cluster to stabilize variance. A z-score transformation was introduced by Cheadle et al. (2003) to standardize the data across genes and arrays.

In this section, I will review some commonly used data normalization and transformation methods for cDNA microarray data.

1.3.3.1 LOWESS:

Sometimes the nature of required normalization is different from either removing mean-scale dependencies or global systematic effects. Consider the plot of the log

fluorescence ratio $M = \log_2(Cy5/Cy3)$ against the average log fluorescence intensity $A = (\log_2(Cy5) + \log_2(Cy3))/2$ for each spot between the two samples hybridized together on a cDNA microarray. Under the assumption that most genes are not differentially expressed and/or roughly equal number of genes are upregulated and downregulated between the two samples, this plot should ideally resemble a uniformly spaced horizontal band of points. However, most M vs. A plots are somewhat curvilinear in shape with some evidence of variance heterogeneity (Figure 4). A seemingly effective normalization to remove the dye-intensity bias is to fit a smooth LOWESS curve between M and A such that the corrected M values are expressed as deviations from this curve to reconstitute the LOWESS -corrected Cy3 and Cy5 logarithmic intensities (Figure 5).

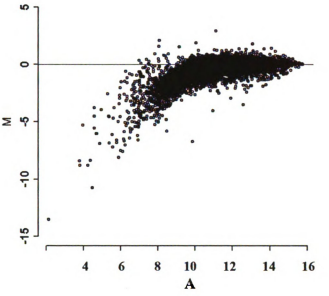


Figure 1.4 M is log-ratio of two expression intensities (Cy3/Cy5) and A is mean log-expression of the two. M vs A plot for one array before LOWESS normalization (data source: (Wade et al. 2005)).

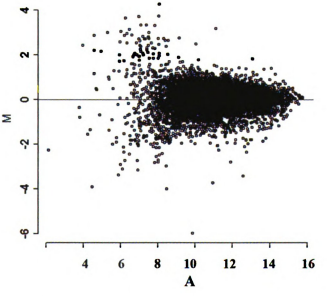


Figure 1.5 M is log-ratio of two expression intensities (Cy3/Cy5) and A is mean log-expression of the two. LOWESS-corrected M versus A plots.

Since differentially expressed genes may appear as outliers on M vs A plots, robust fitting procedures are preferred. LOWESS fitting requires a choice of the "span" which determines which data are local relative to the estimated fit. If the span is too large, the curvature cannot be removed effectively. If the span is too small, the data will be over-fitted. The choice of span is generally subjective; however, usually 20% of the points are chosen for providing a local fit (Yang et al. 2002b). In theory, the largest span that removes the obvious intensity-dependence of the log ratios is ideal, but this may be difficult to assess. Hence, the LOWESS data fitting procedures are straightforward yet a bit perilous as there is risk of overfitting the data and introducing errors larger than those removed (Cui et al. 2003). One possible way to alleviate the problem may be to optimize the span value which minimizes the bias corrected Akaike Information Criteria (AIC) (Hurvich et al. 1998).

1.3.3.2 Variance stabilization transformation

As indicated previously, a logarithmic transformation is typically used to break apart the mean variance relationship inherent with microarray data whereas subsequent LOWESS normalization removes the dependencies of the logarithm of fluorescence intensities on the average fluorescence intensities. It has been recently proposed by Rocke and Durbin (2003) that a particular stochastic model may be responsible for both phenomena. This model decomposes the measurement error into additive and multiplicative errors as follows:

$$y_{ik} = \alpha_i + b_i X_k e^{\eta_{ik}} + \varepsilon_{ik}, \qquad (1)$$

where y_{ik} is the measured raw expression level for channel i and spot k, α_i—mean background intensity of channel i, X_k—the true gene expression level of spot k. Furthermore, $\eta_{ik} \sim N(0,\sigma_{i\eta}^2), \varepsilon_{ik} \sim N(0,\sigma_{i\varepsilon}^2)$.

It can be mathematically derived from Model (1) that a quadratic relationship between variance and intensity of microarray signals. Figure 6 demonstrates that this might be a reasonable assumption for one particular example. Two similar transformations have been developed independently to break apart this dependency. The glog transformation proposed by Rocke and Durbin (2003) has the expression as follows:

$$h_\lambda(y_{ik}) = \ln[\frac{(y_{ik} - \alpha_i + \sqrt{(y_{ik} - \alpha_i)^2 + \lambda_i})}{2}],$$

where $\lambda_i = \sigma_{i\varepsilon}^2 / (e^{\sigma_{i\eta}^2}(e^{\sigma_{i\eta}^2} - 1))$.

A similarly effective arsinh transformation was proposed for microarray data by Huber et. al. (2002):

$$Z_{ik} = \log(b_i Y_{ik} + C_i + \sqrt{(b_i Y_{ik} + C_i)^2 + 1})$$

$b_i = e^{\sigma_{i\eta}^2}(e^{\sigma_{i\eta}^2} - 1)/\sigma_{i\tau}^2$ and $c_i = -\alpha_i(e^{\sigma_{i\eta}^2}(e^{\sigma_{i\eta}^2} - 1))/\sigma_{i\tau}^2$.

Here, the associated parameters can be estimated through a robust variant of maximum likelihood estimation (Cui et. al., 2003), where λ_i is a maximum likelihood estimation.

The two transformations (glog and arsinh) are essentially equivalent, being reparameterizations of each other; nevertheless, the arsinh transformation can be easily implemented using the R package VSN. Both transformations rely upon the assumption of a quadratic relationship between the variance and intensity of the original microarray signals. Nevertheless, the effectiveness of this transformation may further depend on the

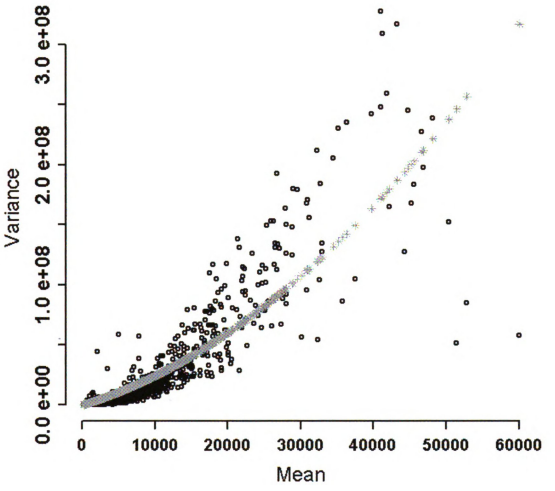


Figure 1.6 Mean intensity vs. variance of intensity plot for each gene (data source:(Wade et al. 2005)). The curve represented by the grey star points is the predictive curve based a quadratic function.

precision of parameter estimation as the curvature correction of this transformation introduces four parameters for each array (Cui et al. 2003).

1.3.4 Identification of differentially expressed genes

Due to the large volume and intrinsic variability of microarray data and the increasing interest to consider many experimental conditions, including different time series, factorial and other more complicated design arrangements, a wide variety of statistical methods have been proposed to infer upon differential gene expression between various treatment groups. Proposed methods range from simple fold change criteria to more general mixed model approaches.

1.3.4.1 Comparison between two treatment groups

Conclusions on differential expression between treatment groups in the earliest microarray experiments were based on simple fold-change criteria not involving any formal statistical inference as noted by Cui & Churchill (2003). For instance, if the average fluorescence intensity ratio between any two conditions exceeded an arbitrarily chosen threshold (i.e., two fold change), one might conclude differential gene expression.

Fold change criteria eventually gave way to simple statistical procedures including non-parametric (e.g. Wilcoxon test) and parametric methods (e.g. Student's *t*-test). Breitling et. al. (2004) proposed an interesting deviation called ranking products from replicate experiments and used permutation based estimation to determine significance levels. A standard *t*-test can be conducted for the log ratios on a gene-by-gene basis. The variances for *t*-test are either estimated from each gene by assuming hetereogeneous

variation across genes (Callow et al. 2000) or a simple pooled variance across all genes (Arfin et al. 2000).

1.3.4.2 Comparison of more than two treatment groups

The fixed effects ANOVA model was first suggested by Kerr et. al., (2000), which was declared to be theoretically reasonable but also realistic for routine implementation to analyze the data from microarray experiments (Kerr et al. 2002). Although it can integrate data normalization and differentially expressed gene identification together (Kerr & Churchill 2001), the fixed effects model can only contain one source of random variation.

1.3.4.3 Mixed model multiple factor

The design structure of microarray experiments may be complicated by multiple factors, each involving multiple levels. The sources of variation could be technical variation or biological variation or both. Models that specify the responses to be functions of fixed systematic effects and random sources of variability are generally referred to as mixed effects models or, simply, mixed models. To fully address the data structure of factorial microarray experiments characterized by multiple random sources of variability, mixed model analysis of variance has been proposed (Wolfinger et. al., 2001) and since then has been widely used and cited in hundreds of papers since it was introduced.

We illustrate the mixed model using, by example, the microarray experimental design used from an earlier experiment available at http://genome-www.stanford.edu/swisnf

(Eisen et al., 1998). The corresponding experiment is based on the comparison of four treatments, each arrayed using three replicates or hybridizations versus a common reference. All treatment samples were labeled with Cy3 against the Cy5 labeled reference sample for all 12 arrays.

Let y_{gij} be the base-2 logarithm of the intensity from gene g (g=1,2,... ,G), treatment i (i=1,2,3,4, R) and array j (j =1,2,...12).

Wolfinger et al. (2001) proposed a two-step mixed model approach. The first stage model is used for normalization and could be described by:

$$y_{gij} = \mu + T_i + A_j + (TA)_{ij} + \varepsilon_{gij}. \quad (3)$$

Here μ corresponds to an overall mean value, T_i is the main effect for treatment i, A_j is the main effect for array j, $(TA)_{ij}$ is the interaction effect of array j and treatment i, and ε_{gij} is stochastic error. This normalization model is fitted for all genes simultaneously (i.e. across all g, i, and j). Let r_{gij} denote the estimated residuals (i.e. estimates of ε_{gij}) from Model (3), determined by subtracting the predicted $\hat{y}_{gij}$, using estimates based on Model (3), from the y_{gij} values. The second stage model which is fitted separately for each gene is then specified as

$$r_{gij} = G_g + (GT)_{gi} + (GA)_{gj} + \gamma_{gij} \quad (4)$$

This model has the similarly written as the normalization model (3), except that now all effects are indexed by g or separately for each gene, $(GT)_{gi}$ is the main effect for treatment i for gene g, which is main interest of the study; $(GA)_{gj}$ is the main effect for array j for gene g, and γ_{gij} is stochastic error for gene g. The array effect GA is crucial to

the model, as it specifies random blocking factor array effect and accounts for the insidious spot-to-spot variability inherent in spotted microarray data. Wolfinger et al. (2001) indicates that the inclusion of this effect allows us to extract appropriate information about the treatment effects and obviates the need to form ratios.

Standard stochastic assumptions are made for the preceding two-stage linear mixed models as with conventional mixed models. In the first stage model (3), random effects A_j, $(TA)_{ij}$, ε_{gij} , are all assumed to be normally distributed with variance components $\sigma_A^2, \sigma_{TA}^2$, and σ_ε^2 whereas in the second stage model, $(GA)_{gj}$, and γ_{gij} , are specified to have gene-specific variance components $\sigma_{GA_g}^2$ and $\sigma_{\gamma_g}^2$, respectively, across all genes $g = 1,2,\ldots,G$.

The method of restricted maximum likelihood (REML) (Searle et al. 1993; Littell et al. 1996) is typically used to estimate variance components for the two-stage mixed models. REML is an alternative to full maximum likelihood estimation. Rather than maximizing the likelihood of the data, REML frees the fixed effects and maximizes the likelihood of the observed residuals over the non-negative space of variance component estimator. REML provide unbiased estimates while ML estimators yield biased estimates of variance components. REML does not always eliminate all the bias since REML can not return negative estimates of variance components and set all negative values to zero (Khattree 1999). This REML method has also been applied in other microarray data analysis studies (Burgueno et al. 2005; Bhowmick et al. 2006; Feng et al. 2006).

1.3.4.4 Shrinkage estimation

Microarray experiments typically include too few replicates to reliably estimate gene-specific differential expression even though these experiments provide information on thousands of genes simultaneously. Empirical Bayes (EB) approaches to inference seem natural for this kind of data feature. EB characterizes the situation when inference on one particular element (e.g. gene) is supplemented by borrowing information from other elements whose effects could be characterized by a distribution. That is, these prior distributions sharpen inference on estimates at the gene level that is superior to gene-specific statistical inference.

The EB strategy of borrowing information across genes has been well developed for simple experimental design structures. For example, SAM (Statistical Analysis of Microarray) by Tusher et al.,(2001) slightly moderates the Student *t*-statistic for any one particular gene by adding a constant to the standard error in the denominator of this statistic. This strategy effectively eliminates more false positives caused by unusually low values of these denominators for individual genes while increasing statistical efficiency in picking up more truly differentially expressed genes. Efron et al. (2001) introduced a simple nonparametric EB model, which is used to guide the efficient reduction of the data to a single summary statistic per gene, and also to make simultaneous inferences concerning which genes were affected by the treatment. Similar nonparametric EB and fully Bayesian approaches were further discussed and compared in Pan et al. (2003) and Do et al. (2005) respectively. Fully Bayesian approaches are somewhat different from more approximate EB procedures in that complete uncertainty in the parameters of the prior distribution of the elements are accounted for; however, they tend to be much more computationally intensive. Lonnstedt and Speed (2002) present an EB log posterior odds

B-statistic for analysis of a simple microarray design comparing two conditions (Lonnstedt & Speed 2002). Smyth (2004) re-parameterized the B statistic into LOD the log odds-ratio *B* and extended the model to accomodate three or more treatments (Smyth 2004). Lonnstedt and Britton (2005) compared fully Bayes and EB approaches for false discovery rates (FDR) and computational tractability and demonstrated that fully Bayesian approaches do not necessarily improve performance in terms of false discovery rates and computer running time to the EB methods for log odd-ratio B based on their data. Lonnstedt et al. (2005) further demonstrated how to convert B statistics to one-way ANOVA F-statistics for detecting differentially expressed genes across several treatment conditions. Ishwaran and Rao (2003) introduced Bayesian ANOVA for microarray (BAM) to make use of a weighted average of generalized ridge regression estimates, which provide benefits of shrinkage and model averaging. In that paper, the ANOVA model was rewritten as a linear regression model. The problem of identifying differentially expressed genes was then transformed into spike and slab variable selection for high-dimensional regression problems(Ishwaran & Rao 2000). The expression "spike and slab" refers to the prior for coefficients in a linear model used in their hierarchical formulation, which was chosen so that each coefficient was mutually independent with a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). Ishwaran and Rao (2003) extended the method for detecting differentially expressed genes between two biological groups to multi-group data. A rescaled spike and slab hierarchical model is developed by grouping variances into clusters with each cluster having a unique value. Other authors have also developed alternative methods that utilize information generated from the whole ensemble of genes

(Newton et al. 2001; Broet et al. 2002; Kendziorski et al. 2003; Wright & Simon 2003; Edwards et al. 2005). All such methods, however, are generally only appropriate for simple designs where shrinkage is only required on a single error structure. It may be difficult to extend their work to more complicated cases, such as experimental designs with technical and biological replicates leading to multiple strata of random variation for each gene. Only mixed effects models can properly delineate between different hierarchical levels of variability.

The unresolved issues related to shrinkage estimation in mixed models was partly realized and considered by Smyth et al. (2005) in an updated version of the popular LIMMA software (Linear Models for Microarray Data) in R, which is a freely available software environment (www.r-project.org) for statistical computing and graphics. LIMMA facilitates a distinction between two particular types of replication, that of within-array technical versus between-array biological replication. A structural mixed model was proposed recently to flexibly model the residual variance that vary across genes and conditions (Jaffrezic et al. 2007). This model was applied to two real data sets and found to perform similarly to LIMMA and better than SAM. Cui et al. (2005) proposed a shrinkage procedure currently implemented in the software MAANOVA, for mixed model analysis of microarrays. They developed an estimator of all variance components based on borrowing information across all genes using the James-Stein-Lindley shrinkage concept to modify F test statistics. The issue of choosing a shrinkage parameter and using the reciprocal of an estimator of the variance instead of an estimator of the reciprocal of the variance was discussed by Tong and Wang (2007). The optimal shrinkage parameters under both Stein and squared loss functions were derived for the

estimator of the reciprocal of the variance. The family of shrinkage variance estimators compared favorably with the shrinkage procedure suggested by Cui et. al. (2005) in terms of both estimation and hypothesis testing for identification of differentially expressed genes. Feng et al. (2006) also derived a promising shrinkage estimation procedure for use with general mixed model analyses of microarrays. The theoretical basis for this procedure was borrowed from Box and Tiao (1973) and further developed by Wolfinger and Kass (2000); this procedure involves use of an independent chains algorithm to estimate the marginal posterior density of the variance components in mixed models. This work extends EB inference for most microarray experimental design layouts, including those considered in Figure 3.

1.3.4.5 Error Control and Multiple Hypothesis Testing

As a typical microarray experiment measures expression levels for thousands of genes simultaneously, a large number of hypotheses are tested within any one experiment. High throughput gene expression microarrays has spurred substantial theoretical work in multiple testing as microarray data has (i) a dimension (i.e. number of genes) generally much larger than the sample size, (ii) the variables (i.e. genes) are often correlated, and (iii) a large proportion of the null hypotheses is generally expected to be true. Traditional approaches to multiple testing were reviewed by Hochberg and Tamhane (1987). More recent developments in the field include resampling methods (Westfall & Young 1999; Pollard & van der Laan 2004); to control the familywise error rate (FWER) and procedures that control the FDR (Benjamini & Hochberg 1995) and positive FDR (pFDR)

(Storey & Tibshirani 2003). Dudoit et al. (2003) recently reviewed and made a comparison of all these procedures.

Given unadjusted p-values for inferences on all gene-specific treatment comparisons, i.e., based on Model (4), multiple test adjustments need to be applied if one wishes to control the FWER for all hypothesis tests. FWER is the probability of committing one Type I error in a series of hypothesis tests. A classical and commonly used procedure for controlling FWER is the Bonferroni test (i.e., Troyanskaya et al. 2002). However, as microarray experiments involve thousands of multiple tests, at least one for each gene on the array, controlling the FWER, particularly with a rather conservative Bonferroni test, is generally considered to be far too insensitive for finding truly differentially expressed genes. Hence, this review will concentrate on a detailed description about pFDR.

Positive False Discovery Rate (FDR) Control

False Discovery Rate (FDR) is a new approach to the multiple comparisons problem. Instead of controlling the chance of *any* false positives (as Bonferroni or random field methods do), FDR controls the expected proportion of false positives within a list of genes declared to be differentially expressed. Estimated FDR's for any particular threshold of statistical significance can be determined from the observed *P*-value distribution for the hypothesis test of interest across genes. The weakness of the classical approach to FDR based on Benjamini and Hochberg (1995) is that the FDR is conservatively assessed by setting $\pi_0 = 1$(the true proportion of all genes that are truly not expressed) without using any information in the data to infer upon π_0. In contrast, pFDR utilizes this information to estimate π_0, thereby yielding a less stringent procedure and

greater sensitivity, while maintaining nominal control of FDR. Suppose that V is the number of true false positive results and R is total number of genes declared to be differentially expressed. Then FDR and $pFDR$ can be defined to be

$$FDR = E(\frac{V}{R} \mid R > 0)\Pr(R > 0) \quad \text{and}$$

$$pFDR = E(\frac{V}{R} \mid R > 0) \text{ (Storey 2002).}$$

The term 'positive' has been added to reflect the fact that we are conditioning on the event that positive findings have occurred; i.e. at least one gene has been declared to be differentially expressed. Therefore, the determination of the probability of differential expression with the probability of 95% or greater should correspond closely to a positive false discovery rate (pFDR) of 5% or less (Storey 2002).

BIBLIOGRAPHY

AHMED, A. A. VIAS, M. IYER, N. G. CALDAS, C. & BRENTON, J. D. (2004). Microarray segmentation methods significantly influence data precision. *Nucleic Acids Research* **32**(5).

ALLISON, D. CUI, X. PAGE, G. & SABRIPOUR, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55-65.

ARFIN, S. M. LONG, A. D. ITO, E. T. TOLLERI, L. RIEHLE, M. M. PAEGLE, E. S. & HATFIELD, G. W. (2000). Global gene expression profiling in Escherichia coli K12 - The effects of integration host factor. *Journal of Biological Chemistry* **275**(38), 29672-29684.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**(1), 289-300.

BHOWMICK, D. DAVISON, A. C. GOLDSTEIN, D. R. & RUFFIEUX, Y. (2006). A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics* **7**(4), 630-641.

BOX, G. E. P. & TIAO, G. C. (1973). *Bayesian inference in statistical analysis*. Wiley, New York.

BOZINOV, D. & SEIDEL, P. (2004). Iterative gridding for automated microarray image analysis. In *Signals, Systems and Computers* pp. 1635 - 1638.

BREITLING, R. ARMENGAUD, P. AMTMANN, A. & HERZYK, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *Febs Letters* **573**(1-3), 83-92.

BROET, P. RICHARDSON, S. & RADVANYI, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**(4), 671-683.

BURGUENO, J. CROSSA, J. GRIMANELLI, D. LEBLANC, O. & AUTRAN, D. (2005). Spatial analysis of cDNA microarray experiments. *Crop Science* **45**(2), 748-757.

CALLOW, M. J. DUDOIT, S. GONG, E. L. SPEED, T. P. & RUBIN, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research* **10**(12), 2022-2029.

CHEADLE, C. VAWTER, M. P. FREED, W. J. & BECKER, K. G. (2003). Analysis of microarray data using Z score transformation. *Journal of Molecular Diagnostics* **5**(2), 73-81.

CUI, X. & CHURCHILL, G. A. (2003). How many mice and how many arrays? Replication in mouse cDNA microarray experiments. In *Methods of Microarray Data Analysis 3* Eds K. F. Johnson & S. M. Lin), pp. 139-154. Norwell, MA: Kluwer Academic Publishers.

CUI, X. G. KERR, M. K. & CHURCHILL, G. A. (2003). Transformations for cDNA Microarray Data *Statistical Applications in Genetics and Molecular Biology* **2**(1).

D'HAESELEER, P. (2005). How does gene expression clustering work? *Nature Biotechnology* **23**(12), 1499-1501.

DE REYNIES, A. GEROMIN, D. CAYUELA, J. M. PETEL, F. DESSEN, P. SIGAUX, F. & RICKMAN, D. S. (2006). Comparison of the latest commercial short and long oligonucleotide microarray technologies. *Bmc Genomics* **7**:51.

DO, K. A. MULLER, P. & TANG, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society Series C-Applied Statistics* **54**, 627-644.

DUDOIT, S. SHAFFER, J. P. & BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**(1), 71-103.

EDWARDS, J. PAGE, G. GADBURY, G. HEO, M. KAYO, T. WEINDRUCH, R. & ALLISON, D. (2005). Empirical Bayes estimation of gene-specific effects in micro-array research. *Funct Integr Genomics* **5**(1), 32-39.

EFRON, B. TIBSHIRANI, R. STOREY, J. D. & TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**(456), 1151-1160.

FENG, S. WOLFINGER, R. D. CHU, T. M. GIBSON, G. C. & MCGRAW, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural Biological and Environmental Statistics* **11**(2), 197-209.

GERSHON, D. (2002). Microarray technology: An array of opportunities. *Nature* **416**, 885-891.

GLONEK, G. F. V. & SOLOMON, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**(1), 89-111.

HOCHBERG, Y. & TAMHANE AC. (1987). *Multiple Comparison Procedures*. Wiley, New York.

HUBER, W. HEYDEBRECK, A. V. SULTMANN, H. POUSTKA, A. & VINGRON, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl 1), S96-104

HURVICH, C. M. SIMONOFF, J. S. & TSAI, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **60**, 271-293.

ISHWARAN, H. & RAO, J. S. (2000). Bayesian nonparametric MCMC for large variable selection probelms - Unpublished manuscript.

ISHWARAN, H. & RAO, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**(462), 438-455.

JAFFREZIC, F. MAROT, G. DEGRELLE, S. HUE, I. & FOULLEY, J. (2007). A structural mixed model for variances in differential gene expression studies. *Genetical Research* **89**(1), 19-25.

JENSSEN, T. K. LANGAAS, M. KUO, W. P. SMITH-SORENSEN, B. MYKLEBOST, O. & HOVIG, E. (2002). Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Research* **30**(14), 3235-3244.

KENDZIORSKI, C. M. NEWTON, M. A. LAN, H. & GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**(24), 3899-3914.

KERR, M. K. AFSHARI, C. A. BENNETT, L. BUSHEL, P. MARTINEZ, J. WALKER, N. J. & CHURCHILL, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**(1), 203-217.

KERR, M. K. & CHURCHILL, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**(2), 123-128.

KHANIN, R. V., V. & E, W. (2006). Reconstructing repressor protein levels from expression of gene targets in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **103**(49), 18592-18596.

KHATTREE, R. (1999). Nonnegative Estimation of Variance Components: A Modification to Henderson's ANOVA Methodology. *The Indian Journal of Statistics* **61**(B), 261-265.

KORN, E. L. HABERMANN, J. K. UPENDER, M. B. RIED, T. & MCSHANE, L. M. (2004). Objective method of comparing DNA microarray image analysis systems. *Biotechniques* **36**(6), 960-967.

LITTELL, R. MILLIKEN, G. STROUP, W. & WOLFINGER, R. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute,Inc.

LONNSTEDT, I. & BRITTON, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics* **6**(2), 279-291.

LONNSTEDT, I. RIMINI, R. & NILSSON, P. (2005). Empirical Bayes microarray ANOVA and grouping cell lines by equal expression levels. *Statistical Applications in Genetics and Molecular Biology* 4 (1), Article 7.

LONNSTEDT, I. & SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**(1), 31-46.

MONTGOMERY, D. (1984). *Design and Analysis of Experiments*. Wiley, New York.

NEWTON, M. A. KENDZIORSKI, C. M. RICHMOND, C. S. BLATTNER, F. R. & TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37-52.

PAN, W. LIN, J. & LE, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics.* **3(3)**(3), 117-124.

PETERSEN, D. CHANDRAMOULI, G. V. R. GEOGHEGAN, J. HILBURN, J. PAARLBERG, J. KIM, C. H. MUNROE, D. GANGI, L. HAN, J. PURI, R. STAUDT, L. WEINSTEIN, J. BARRETT, J. C. GREEN, J. & KAWASAKI, E. S. (2005). Three microarray platforms: an analysis of their concordance in profiling gene expression. *Bmc Genomics* **6:63**.

POLLARD, K. S. & VAN DER LAAN, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* **125**(1-2), 85-100.

QIN, L. RUEDA, L. ALI, A. & NGOM, A. (2005). Spot Detection and Image Segmentation in DNA Microarray Data. *Appl Bioinformatics* **4**(1), 1-11.

RAHNENFUHRER, J. & BOZINOV, D. (2004). Hybrid clustering for microarray image analysis combining intensity and shape features. *Bmc Bioinformatics* **5:47**.

ROCKE, D. M. & DURBIN, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**(8), 966-972.

ROSA, G. J. M. STEIBEL, J. P. & TEMPELMAN, R. J. (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics* **6**(3), 123-131.

SCHENA, M. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467-470.

SEARLE, S. CASELLA, G. & MCCULLOCH, C. (1993). *Variance Components*. New York: Willy.

SHAW, C. & TOLLETT , J. (2001). Image Analysis and Quantitation of cDNA Microarray Images. "http://www.bcm.edu/microarray/imageanalysis.pdf".

SHI, J. & WALKER, M. (2007). Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. *Current Bioinformatics* **2**(2), 133-137.

SMYTH, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1).

SMYTH, G. YANG, H. & SPEED, T. (2003). Statistical Issues in cDNA Microarray Data Analysis. *Methods Mol Biol.* **224**, 111-136.

SMYTH, G. K. MICHAUD, J. & SCOTT, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**(9), 2067-2075.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of The Royal Statistical Society Series B-Statistical Methodology* **64**, 479-498.

STOREY, J. D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9440-9445.

TEMPELMAN, R. J. (2005). Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Veterinary Immunology and Immunopathology* **105**(3-4), 175-186.

TONG, T. & WANG, Y. (2007). Optimal Shrinkage Estimation of Variances With Applications to Microarray Data Analysis. *The Journal of American Statistical Association* **102**(3), 113-222.

TROYANSKAYA, O. GARBER, M. BROWN, P. BOTSTEIN, D. & ALTMAN, R. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data *Bioinformatics* **18**(11), 1454-1461

TUSHER, V. G. TIBSHIRANI, R. & CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**(9), 5116-5121.

WADE, J. PEABODY, C. COUSSENS, P. TEMPELMAN, R. J. CLAYTON, D. F. LIU, L. ARNOLD, A. P. & AGATE, R. (2005). A cDNA microarray from the telencephalon of juvenile male and female zebra finches (vol 138, pg 199, 2004). *Journal of Neuroscience Methods* **142**(2), 327-327.

WANG, Y. LU, J. P. LEE, R. GU, Z. P. & CLARKE, R. (2002). Iterative normalization of cDNA microarray data. *Ieee Transactions on Information Technology In Biomedicine* **6**(1), 29-37.

WESTFALL, P. & YOUNG, S. S. (1999). *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, New York.

WOLFINGER, R. D. & KASS, R. E. (2000). Nonconjugate Bayesian analysis of variance component models. *Biometrics* **56**(3), 768-774.

WOO, Y. AFFOURTIT, J. DAIGLE, S. VIALE, A. JOHNSON, K. NAGGERT, J. & CHURCHILL, G. (2004). A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech* **15(4)**, 276-284.

WRIGHT, G. W. & SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**(18), 2448-2455.

YANG, Y. H. BUCKLEY, M. J. DUDOIT, S. & SPEED, T. P. (2002a). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**(1), 108-136.

YANG, Y. H. DUDOIT, S. LUU, P. LIN, D. M. PENG, V. NGAI, J. & SPEED, T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4).

Chapter 2: A Linear Mixed Model with an Empirical Bayes Adjustment to Detect Differential Gene Expression for Microarray Experiments

Abstract

Analysis of gene expression data using two color microarrays is often complicated by the effects of multiple random experimental sources of variability in addition to the systematic fixed effects of treatments or dyes. Hence microarray data analysis typically necessitates the use of mixed model ANOVA to properly specify correct statistical tests. Some of these random sources of variability are identifiable (e.g. subject on array) and can be modeled explicitly with the remaining sources typically aggregated together as residual effects. Borrowing information on random and residual effects across genes using hierarchical Bayesian techniques then seems desirable given that microarray experiments generally involve inference on thousands of genes, each typically characterized by a limited amount of biological replication across groups or treatments. We propose a hierarchical linear mixed model for each gene that combines gene-specific information with information on ANOVA expected mean squares for random and residual effects across genes. Our procedure leads to a Bayesian ANOVA table that is augmented with posterior sums of squares and posterior degrees of freedom for each random and residual effect.

We compare our method to a recently developed method based on shrinkage of REML estimates of variance components as well as gene-specific mixed model analyses based on REML or ANOVA estimates of VC. Our model was seen to have greatest

power to detect differentially expressed genes while providing correct control of the false discovery rate. We also demonstrate the various methods on two publicly available microarray data sets; in both cases, our proposed method was able to detect more statistically significant genes at the same false discovery rate.

Keywords: Empirical Bayes; cDNA microarray; mixed model; ANOVA component; ANOVA; REML

2.1 Introduction

Microarray technologies have been developed to measure the simultaneous expression of thousands of genes across various experimental conditions. Spotted two-color microarray platforms use competitive hybridization to directly compare the amounts of mRNA transcribed from each gene in the two samples to be compared (Murphy 2002). The cDNA that are reverse transcribed from the mRNA from the two different samples are separately labeled with different fluors (typically, Cy3 versus Cy5) and then hybridized together on a glass slide having separate spots with bounded targets for each complementary expressed sequence tags (EST) or long oligonucleotide sequence of interest. Relative mRNA abundance for each gene is then typically quantified as the average or total fluorescence intensities at the corresponding spots using image scanning and analyses software. In other words, upon appropriate normalization (Yang et al. 2002), the ratio of Cy3:Cy5 fluorescence intensities for each spot is interpreted as the ratio of corresponding mRNA transcript copies in the two samples.

For efficient experimental designs such as the loop design (Tempelman 2005; Vinciotti et al. 2005), several identifiable random sources of variation potentially

influence the fluorescence intensity measurements in such a way that a mixed model ANOVA is most appropriate for statistical analysis for each gene (Wolfinger et al. 2001). However, gene-specific analyses are typically plagued by low power in minimally replicated studies, if replication is properly defined at the biological rather than technical level.

Shrinkage or empirical Bayes estimation has been shown to be statistically efficient and powerful in a number of applications where inference on parameters for a certain class or group of experimental units is based on combining information of that group-specific statistics with information across all other groups (Casella 1985). The idea of modifying estimators of variances for individual genes by borrowing information across all genes was proposed originally for simple designs (Baldi & Long 2001; Lonnstedt & Speed 2002) and later extended for multifactorial studies (Wright & Simon 2003; Smyth 2004). However, these methods fail to borrow information across random effects factors with some special nested design exceptions (Smyth 2004). Since VC for random effects factors generally display varying degrees of heterogeneity across genes in microarray experiments (Cui & Churchill 2003; Chen et al. 2004), this issue is particularly important when random effects factors such as subjects within treatments serve as key experimental error terms for ANOVA F-tests in experimental designs.

There have been at least a couple of recent methods that have considered shrinkage estimation for mixed effects models with applications to microarray experiments. Cui et al. (2005) recently proposed a shrinkage procedure, currently implemented in the software MAANOVA (Wu et al. 2003), for mixed model analysis of microarray data. Their inference procedure is based on the use of permutation testing

which is indeed known to be robust to potential distributional misspecifications (e.g. non-normality) in simple fixed effects models and some special mixed model cases. With regards to the latter, Cui et al. (2005) note that a proper permutation testing strategy should be based on identifying observations that could be deemed exchangeable under the null hypothesis; this is somewhat synonymous to identifying observations that share all of the same random effects. As an example, they indicated that samples hybridized against each on an array are exchangeable since they share the same array effect. However, consider the connected loop design in Figure 2(a) from Tempelman (2005) or the alternating loop design provided in both Figure 1(b) from Kerr and Churchill (2001) and in Figure 3 from Dobbin et al. (2003). In these designs, mRNA from each biological replicate is partitioned into two aliquots, one labeled with Cy3 and the other labeled with Cy5, such that the same biological sample is used in two different hybridizations or arrays. With such designs, it is not readily apparent how to define exchangeable units for permutation testing since there are no two aliquots that jointly share the same random array (block) and biological replicate (experimental error) effects.

Feng et al. (2006) recently introduced a more general and promising procedure for mixed model analysis of microarray data based on shrinkage estimation of REML estimates of VC. However, its statistical properties, e.g. receiver operating characteristics, control of false discovery rates, are not well known relative to methods that use only gene-specific information like classical ANOVA based on ANOVA or REML estimates of VC. In this paper, we also develop a second general shrinkage estimation mixed model procedure which extends the method of Wright and Simon (2003) for each random factor in a mixed model ANOVA and evaluate its statistical properties relative to the

procedure of Feng et al. (2006) and to conventional mixed model analysis based on ANOVA or REML estimation of VC.

2.2 Methods and Materials

2.2.1 Loop Design with Dye Swap

Consider, for example, a published cDNA microarray data set (Liang et al., 2002, GEO Accession: GSE3588) that was used to study mRNA expression of 1751 genes in rat renal medulla associated with the development of salt-sensitive hypertension. The factorial treatment structure was defined by two strains of rats (MCW versus BN13) and two diets (low salt versus high salt) for a total of four treatment groups. A connected loop design (Tempelman, 2005) with dye swaps on the same two samples was used for each of 3 identically constructed loops for a total of 24 arrays as in Figure 1. In other words, the amount of biological replication was 3 rats per treatment. Although the original study by Liang et al. (2002) involved duplicate spots per gene, we study their design further within the context of a single spot per gene for pedagogical considerations.

A mixed model ANOVA can be fitted to the expression data on a gene-by-gene basis. The fixed effects factors are treatment and dye with random effects factors being array, rat within treatment and residual. For the design in Figure 1, the symbolic ANOVA table using the Type III or fully adjusted quadratic forms (Searle et al. 1992) along with expected mean squares is provided in Table 1. Note that arrays are specified as random effects instead of fixed effects to facilitate efficient combined interblock and intrablock analysis for the estimation of treatment contrasts in incomplete block designs (Yates 1940). Also further note that subject with treatment serves as the experimental

unit or error term for treatment. That is, under the global null hypothesis of no treatment differences (the noncentrality parameter $\gamma_{it} = 0$ for treatment with gene i), the ANOVA MS (MS_{it}) for treatment and the MS (MS_{i2}) for subject within treatment share the same EMS. In convention with ANOVA theory, the F-test statistic $F_{i,t} = \frac{MS_{it}}{MS_{i2}}$ is considered to be a random draw from a F distribution with ν_t numerator and ν_2 denominator degrees of freedom (i.e. $F_{it} \sim F_{\nu_t,\nu_2}$). Of course, if the global null hypothesis for treatment effects is false (i.e. $\gamma_{it} \neq 0$), then $F_{i,t} = \frac{MS_{it}}{MS_{i2}}$ is distributed as a noncentral F with noncentrality parameter γ_{it}. The ability (i.e. power) to correctly reject the null hypothesis in this case is increased by larger treatment effects or γ_{it} and larger ν_2. Because the loop design in Figure 1 is an balanced incomplete block design, the VC estimates based on various methods (e.g. REML, Type I or Type III ANOVA) will be necessarily different such that subsequent inference on treatment effects will be necessarily different as well.

Figure 2.1 Connected Loop Design with Dye Swap from Liang et al. (2002) with 24 arrays and 3 rats per each of four treatments defined by a 2 x 2 factorial of strains (SS/Mcw versus SSBN13) and diets (low salt versus high salt). Each oval designates an experimental unit (rat) and each arrow denotes an array with circle end denoting the Cy3 labeled sample and tail denoting the Cy5 labeled sample.

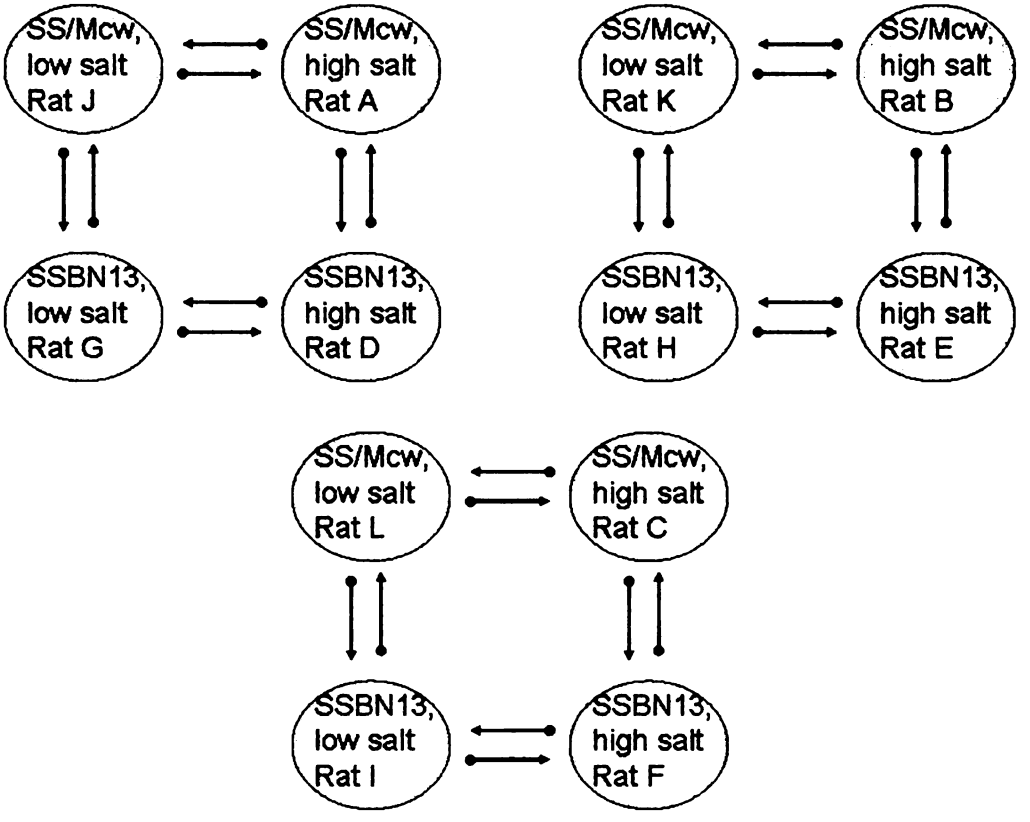


Table 2.1 Classical mixed model ANOVA table based on the fully adjusted (Type III) quadratic forms for the analysis of log fluorescence intensities for any particular gene (i = 1,2....,g) spotted once per array in the connected loop design with dye swap as in Figure 1.

Source	***SS****	***df**†*	***Mean Squares***	***Expected Mean Squares***
Treatment	SS_{it}	ν_t=3	$MS_{it}= SS_{it}/\nu_t$	$0\sigma_{i1}^2+\frac{8}{3}\sigma_{i2}^2+\sigma_{i3}^2+\gamma_{it}$ ‡
Dye	SS_{id}	ν_d=1	$MS_{id}= SS_{id}/\nu_d$	$0\sigma_{i1}^2+0\sigma_{i2}^2+\sigma_{i3}^2+\gamma_{id}$ §
Array	SS_{i1}	ν_1=21	$MS_{i1}= SS_{i1}/\nu_1$	$\phi_{i1}=\frac{12}{7}\sigma_{i1}^2+0\sigma_{i2}^2+\sigma_{i3}^2$
Subject(Treatment)	SS_{i2}	ν_2=6	$MS_{i2}=SS_{i2}/\nu_2$	$\phi_{i2}=0\sigma_{i1}^2+\frac{8}{3}\sigma_{i2}^2+\sigma_{i3}^2$
Residual	SS_{i3}	ν_3=14	$MS_{i3}= SS_{i3}/\nu_3$	$\phi_{i3}=0\sigma_{i1}^2+0\sigma_{i2}^2+\sigma_{i3}^2$

*Sums of squares (first subscript identifies gene; second subscript identifies factor)

†Degrees of freedom (presumed constant from gene to gene with no missing data)

‡ γ_{it} is the noncentrality parameter for treatment for ANOVA of gene i. When there are no treatment mean differences, γ_{it} = 0 such that treatment and subject (treatment) then have the same expected mean square so that $F_{it} = MS_{it}/MS_{i2}$ is a random draw from a F distribution with ν_t numerator and ν_2 denominator degrees of freedom.

§ γ_{id} is the noncentrality parameter for dye such that if there is no dye mean difference, γ_{id}=0 and $F_{id}= MS_{id}/MS_{i3}$ is a random draw from a F distribution with ν_d numerator and ν_3 denominator degrees of freedom

2.2.2 Reference Design with Dye Swap

We also consider the common reference design as used by Pritchard et al. (2001) and further studied by Cui and Churchill (2003). A 5406-clone spotted cDNA microarray was used to quantify transcript levels in the kidney, liver, and testis from each of 6 normal male C57BL6 mice. A common pooled reference sample was created by combining equivalent amounts of mRNA from each of the three organs (treatments) of each mouse and was used for all array hybridizations. Four separate hybridizations were conducted for each organ mRNA sample from each animal against the common reference. For two of these arrays, mRNA samples were labeled with Cy3 dye and paired with the Cy5 labeled reference with dye assignments swapped for the other two arrays (Pritchard et al.,2001). Hence, a total of 72 arrays were utilized in this particular experiment as illustrated by Figure 2. Note that for a common reference design, it suffices to use the logarithm of the ratio of treatment sample over the common reference sample fluorescence intensities at each spot as the response variable for subsequent ANOVA. The ANOVA table based on the analysis of these log ratios is provided in Table 2 for this design. Unlike the loop design considered previously, the expected mean squares are somewhat invariant to the choice of quadratic forms, e.g. Type I, Type III, MIVQUE-0 (Searle et al. 1992), used to derive the SS and their expectations. From Table 2, one should note that subject by treatment interaction MS serves as the experimental error term for the F-test on treatment differences.

Figure 2.2 Common Reference Design with Dye Swap from Pritchard et al. (2001) with 72 arrays and 3 organs per each of six mice. This figure is an example of arrays performed for one organ (i.e. 24 arrays). Each oval designates an experimental unit (mouse*organ) and each arrow denotes an array with circle end denoting the Cy3 labeled sample and tail denoting the Cy5 labeled sample.

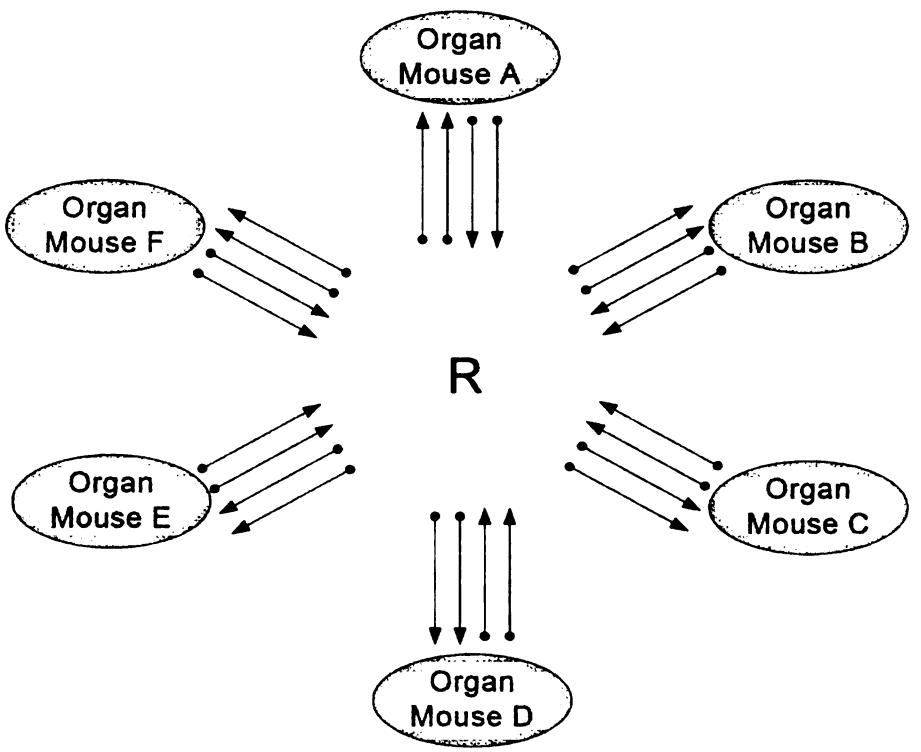


Table 2.2 Classical mixed model ANOVA table for the analysis of log fluorescence intensity ratios (treatment/reference) for any particular gene (i = 1,2....,g) spotted once per array in a dye swapped common reference design.

Source	***SS****	***df**†*	***Mean Squares***	***Expected Mean Squares***
Treatment	SS_{it}	ν_t=2	$MS_{it}= SS_{it}/\nu_t$	$0\sigma_{i1}^2 + 4\sigma_{i2}^2 + \sigma_{i3}^2 + \gamma_{it}$ ‡
Dye	SS_{id}	ν_d=1	$MS_{id}= SS_{id}/\nu_d$	$0\sigma_{i1}^2 + 0\sigma_{i2}^2 + \sigma_{i3}^2 + \gamma_{id}$ §
Subject	SS_{i1}	ν_1=5	$\hat{\phi}_{i1} = MS_{i1}= SS_{i1}/\nu_1$	$\phi_{i1} = 12\sigma_{i1}^2 + 4\sigma_{i2}^2 + \sigma_{i3}^2$
Subject*Treatment	SS_{i2}	ν_2=10	$\hat{\phi}_{i2} = MS_{i2}=SS_{i2}/\nu_2$	$\phi_{i2} = 0\sigma_{i1}^2 + 4\sigma_{i2}^2 + \sigma_{i3}^2$
Residual	SS_{i3}	ν_3=53	$\hat{\phi}_{i3} = MS_{i3}= SS_{i3}/\nu_3$	$\phi_{i3} = 0\sigma_{i1}^2 + 0\sigma_{i2}^2 + \sigma_{i3}^2$

*Sums of squares (first subscript identifies gene; second subscript identifies factor)
†Degrees of freedom (presumed constant from gene to gene with no missing data)
‡ γ_{it} is the noncentrality parameter for treatment for ANOVA of gene i. When there are no treatment mean differences, γ_{it} = 0 such that treatment and subject (treatment) then have the same expected mean square so that $F_{it}= MS_{it}/MS_{i2}$ is a random draw from a F distribution with ν_t numerator and ν_2 denominator degrees of freedom.
§ γ_{id} is the noncentrality parameter for dye such that if there is no dye mean difference, γ_{id}=0 and $F_{id}= MS_{id}/MS_{i3}$ is a random draw from a F distribution with ν_d numerator and ν_3 denominator degrees of freedom

2.2.3 Empirical Bayes ANOVA

Let the total number of random and residual effects factors be denoted by r and the total number of genes under consideration be g. For the moment, we consider the loop design from Figure 1 and Table 1. Note from Table 1 that the random effects factors are numbered from j=1 to r=3 for array (1), subject within treatment (2) and residual (3) according to the subscripts for the ANOVA sums of squares (SS), degrees of freedom (ν), mean squares (MS) and expected mean squares (ϕ). Since VC for these random effects factors appear to be highly heterogeneous across genes (Cui & Churchill 2003; Chen et al. 2004), we assume all of the ANOVA terms in Table 1 to be unique for each gene i except for degree of freedom thereby implying the same experimental design for each gene (i.e. no missing data). In other words, we specify a linear mixed model (Searle et al. 1992) where the data vector for each gene i is written as a linear function of any number of fixed effects (e.g. treatment and dye effects as in Table 1), r -1 sets of *NIID* random effects, each characterized by its own VC σ_{ij}^2, j=1,2,...,r-1, and finally the set of n *NIID* residual effects with variance σ_{ir}^2.

Under a classical mixed model ANOVA, the expected mean square (EMS) components ϕ_{ij} for each set j of random and residual effects based on the ANOVA of each gene i can be written as a linear function of the r VC $\boldsymbol{\sigma}_i = \left[\sigma_{i1}^2 \quad \sigma_{i2}^2 \quad \ldots \quad \sigma_{ir}^2\right]'$ for that gene; i.e.

$$\boldsymbol{\varphi}_i = [\phi_{i1} \quad \phi_{i2} \quad \ldots \quad \phi_{ir}]' = \mathbf{C}\boldsymbol{\sigma}_i \qquad [1]$$

Here **C** is a r x r matrix of known coefficients that depends upon the design. In other words, the elements of **C** are provided under the EMS column within a classical ANOVA table (Cochran & Cox 1957; Hinkelmann & Kempthorne 1994; Giesbrecht & Gumpertz 2004). For example, it can be seen from the loop design of Table 1 that

$$C = \begin{bmatrix} \frac{12}{7} & 0 & 1 \\ 0 & \frac{8}{3} & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad [2]$$

For the common reference design in Figure 2 and Table 2, there are also, coincidentally, $r = 3$ sets of random and residual effects, being subject (1), subject by treatment (2), and the residual effects (3). For that design then, it can be readily seen from Table 2 that:

$$C = \begin{bmatrix} 12 & 4 & 1 \\ 0 & 4 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad [3]$$

Suppose that $\hat{\boldsymbol{\varphi}}$ denotes the vector of ANOVA mean squares (MS) for random effects factors that one would typically find under the MS column of a classical ANOVA table. ANOVA estimates of VC are then derived by equating MS ($\hat{\boldsymbol{\varphi}}$) to EMS; i.e. solving $\hat{\boldsymbol{\varphi}} = C\boldsymbol{\sigma}$ for $\boldsymbol{\sigma}$. In other words, the classic ANOVA estimator of elements of $\boldsymbol{\sigma}$ is $\hat{\boldsymbol{\sigma}} = C^{-1}\hat{\boldsymbol{\varphi}}$.

Our empirical Bayes ANOVA (EB-ANOVA) method is a mixed model extension of the method presented previously by Wright and Simon (2003). From classical mixed model ANOVA theory in Searle (1971), it can be demonstrated that

$$\frac{\nu_j MS_{ij}}{\phi_{ij}} = \frac{\nu_j \hat{\phi}_{ij}}{\phi_{ij}} \sim \chi^2_{\nu_j} \; ; \; j = 1,2,\ldots,r \qquad [4]$$

such that from a Bayesian perspective, the data likelihood for inference on ϕ_j can be written as a scaled chi-squared density

$$\hat{\phi}_{ij} \sim \frac{\phi_{ij}}{df_{ij}} \chi^2_{\nu_j} \; ; \; j = 1,2,\ldots,r \qquad [5]$$

Suppose that the prior distribution for the EMS are specified to be inverted gamma distributed; i.e. $\phi_{ij} \sim IG(\alpha_j, \beta_j)$ with mean $\frac{\beta_j}{\alpha_j - 1}$. Then it can be shown using Wright and Simon (2003) that $\frac{\alpha_j}{\beta_j} \hat{\phi}_{ij} \sim F_{2\alpha_j, \nu_j}$, $i = 1,2,\ldots,g$. This result facilitates the determination of marginal maximum likelihood (MML) estimates of $\hat{\alpha}_j$ and $\hat{\beta}_j$ for α_j and β_j, respectively, by maximizing $\frac{\beta_j}{\alpha_j} F_{2\alpha_j, \nu_j}$ with respect to α_j and β_j. Note that this MML optimization can be conducted independently for each set of random and residual effects $j = 1,2\ldots,r$. Alternatively, a method of moments estimator may be used but we prefer the MML estimator for reasons of statistical efficiency as in Wright and Simon (2003). Using Wright and Simon (2003) further, it can be also demonstrated that $\frac{(\nu_j + 2\alpha_j)\tilde{\phi}_{ij}}{\phi_{ij}} \sim \chi^2_{\nu_j + 2\alpha_j}, j = 1,2,\ldots,r$, where

$$\tilde{\phi}_{ij} = \frac{\nu_j \hat{\phi}_{ij} + 2\alpha_j \left(\frac{\alpha_j}{\beta_j}\right)^{-1}}{\nu_j + 2\alpha_j} \qquad [5]$$

represents a *posterior MS* for random effects factor j for gene i such that the numerator is a *posterior SS* combining gene-specific information $\tilde{\phi}_{ij} = \dfrac{v_j \hat{\phi}_{ij} + 2\alpha_j \left(\dfrac{\alpha_j}{\beta_j}\right)^{-1}}{v_j + 2\alpha_j}$ with the harmonic mean $\left(\dfrac{\alpha_j}{\beta_j}\right)^{-1}$ of ϕ_{ij} for all i with respective weights of v_j and $2\alpha_j$. Note that in [5], α_j and β_j would be replaced by its corresponding MML estimates $\hat{\alpha}_j$ and $\hat{\beta}_j$. Therefore, the less heterogeneity (i.e., larger α_j) in ϕ_{ij} across genes for random effects factor j, the greater the weight (i.e. shrinkage) that the overall harmonic mean has on determining $\tilde{\phi}_{ij}$ and, consequently, the greater the posterior degrees of freedom $v_j + 2\alpha_j$.

The implications for these results for the two designs in Tables 1 and 2 are substantial. Recall previously that subject within treatment serves as the experimental unit for treatment in the loop design (Table 1) whereas subject by treatment serves as the experimental unit for treatments in the reference design (Table 2). Given the borrowing of information on sets of random and residual effects across treatments, the *posterior F-ratio* for treatments can be determined as $F_{i,t} = \dfrac{MS_{it}}{\tilde{\phi}_{i2}}$ in both designs, which can be shown to be a random draw from a $F_{v_t, v_2+2\alpha_2}$ under the global null hypothesis for treatments. If treatment effects do exist (i.e. $\gamma_{it} \neq 0$), then the posterior ANOVA table (as determined by the posterior SS, posterior df and posterior MS for each set of random and

residual effects), should have greater statistical power than gene-specific classical ANOVA (Wolfinger et al. 2001) where information across genes is not borrowed.

2.3 Simulation Study

We studied the loop design more extensively based on a simulation study involving 6000 genes, of which 1100 were specified to be differentially expressed between various treatment groups. Two treatments were specified to have identical means for each one of the 6000 genes. The other two treatments were specified such that one of the treatments was downregulated and the other treatment upregulated with respect to the first two treatments with fold changes being $[1.25^{-1},1]$, $[1.25,1.25^{-1}]$, $[1.5^{-1},1.25]$, $[1.5,1.5^{-1}]$, $[2^{-1},1.5]$, $[2,2^{-1}]$, $[2.5^{-1},2]$, $[2.5,2.5^{-1}]$, $[3^{-1},2.5]$, $[3,3^{-1}]$, and $[1,3^{-1}]$ for each of 11 sets of 100 genes, respectively, relative to the first two treatments.

For each gene i and random effects factor j, VC σ_{ij}^2 were randomly drawn from independent inverted gamma densities $IG\left(\sigma_{ij}^2 \mid \alpha_j^{VC}, \beta_j^{VC}\right)$, i=1,2,...,G as defined by parameters α_j^{VC} and β_j^{VC}, j = 1, 2, 3. Inverted-gamma densities appear to well characterize the distribution of variances for two color microarray data (Wright & Simon 2003). For each VC, we specified either α_j^{VC} = 3 or α_j^{VC} = 12. Since the standard deviation of $IG\left(\sigma_{ij}^2 \mid \alpha_j^{VC}, \beta_j^{VC}\right)$ is $\dfrac{\beta_j^{VC}}{\left(\alpha_j^{VC}-1\right)\sqrt{\alpha_j^{VC}-2}}$ for $\alpha_j^{VC} > 2$, then α_j^{VC} = 3 specifies a situation of highly heterogeneous VC or, synonymously, a high level of

heteroskedasticity across genes for random effects factor j. Values of β_j^{VC} were then specified by equating the VC means to $\frac{\beta_j^{VC}}{\alpha_j^{VC}-1}$; i.e., the mean of $IG\left(\sigma_{ij}^2 \mid \alpha_j^{VC}, \beta_j^{VC}\right)$. It is important to realize that these specifications for α_j^{VC} defining heteroskedasticity of VC for each random effects factor in the simulation study do not directly correspond with the specifications for α_j which define heteroskedasticity for EMS for each random effects factor within our EB-ANOVA model. In other words, the simulation model was designed to not favor any one particular analysis model.

After VC were randomly drawn for each gene, log fluorescence intensities for the connected loop were then simulated from a mixed model conditional on linear combinations of random effects drawn from normal distributions using these VC draws and the specified treatment log fold changes relative to Treatment 1. In all, there were then a total of 2^2 =4 different dataset combinations generated for each design, one for each possible duplet combination of values of $\{\alpha_2^{VC}, \alpha_3^{VC}\}$; i.e. {3,3}, {3,12}, {12,3}, and {12,12}. Here $\alpha_1^{VC} = 3$ always as statistical inference on treatment effects is robust to specifications on VC for blocking factors like array for the loop design and subject for the common reference design. VC estimates for the three methods were compared for their mean absolute deviation (Wolfinger et al. 2001) from the true VC value such that smaller values indicate greater precision of VC estimation. We anticipated that the smaller the MAD of VC estimates, the more likely the corresponding procedure will lead to greater sensitivity and specificity of estimated generalized least squares (EGLS) hypothesis testing on treatment mean differences.

Once VCs were estimated by each method, they were used to provide the corresponding EGLS inferences on treatment effects for each dataset within each of the two designs. Furthermore, the positive control method of GLS (i.e. based on treating the known simulated values of the VC as known) was used for inference on treatment effects. Mixed model EGLS F-test statistics for treatment effects were computed based on the estimated VC for each method. The denominator degrees of freedom for the EGLS tests based on EB-REML were conservatively set to 5 in accordance with recommendations put forth by Feng et al. (2006) since the null distributions of the corresponding F-test statistics are not known. Furthermore, EGLS based on REML, or EGLS-REML, was additionally adjusted using the procedure of Kenward and Roger (1997) which has been noted to lead to substantially better control of Type I error rates for EGLS-REML in recent work (Schaalje et al. 2002; Spilke et al. 2005).

Receiver operating characteristics (ROC) curves were also used to compare the four EGLS methods and the GLS method for the relative frequency of true positives (i.e. differentially expressed) to false positives for every possible gene list as based on all possible thresholds for declaring statistically significant genes. ROC comparisons have been effectively used in other studies for comparing statistical methods and experimental designs in microarray studies (Vinciotti et al. 2005; Feng et al. 2006).

The estimated false discovery rates (EFDR) or q-values for the EGLS F-test on treatments based on the procedures of Storey and Tibsharani (2003) was also computed using the F-test P-values for each of the four EGLS methods and the GLS method. To facilitate a finer comparison of the five methods, the EFDR was estimated based on the true proportion π_0 of non-differentially expressed genes being known; i.e.

$\pi_0 = \frac{4900}{6000} = 0.8167$. For every possible gene list; i.e. all possible thresholds ($0 < q$-value <1) of statistical significance, the true false discovery rate (TFDR) was evaluated for the declared significant genes. Each method was then evaluated for the relative agreement between the EFDR to TFDR across all values of EFDR to assess whether or not the nominal FDR rate was being actually controlled.

A similar but smaller scale simulation study was conducted based on the reference design with dye swap previously mentioned. In that case, we concentrated on one simulated dataset using estimates of α_1 α_2, α_3, β_1, β_2, and β_3 as the corresponding parameters for directly generating the EMS and hence indirectly the VC for the random and residual effects. Analogous to the loop design, we generated log ratios for the reference design based on the linear mixed model implied by Table 2 for 6000 genes, of which 11 sets of 100 genes (i.e. 1100 genes) were differentially expressed. As before, 1100 genes were specified to have fold changes ranging from 3.0^{-1}, 2.5^{-1}, 2.0^{-1}, 1.5^{-1}, 1.25^{-1}, 1, 1.25, 1.5, 2.0, 2.5, and 3.0 with respect to the first two treatments for each of 11 sets of 100 genes each, respectively.

2.4 Data Applications

We also applied each of the EGLS methods to the actual datasets related to the loop design provided by Liang et al. (2003) and for the reference design with dye swap provided by Pritchard et al. (2001). Since there were duplicate spots per transcript in the Liang et al. (2003) array, the Cy3 and Cy5 logarithmic (base 2) intensities were averaged as response variables for each transcript within an array for pedagogical reasons although

certainly a fourth random effects factor (spot within array) could have been additionally modeled without any implications for inference on treatment effects. For both datasets, 5% of the extreme VC estimates from each of the two tails were trimmed for robust empirical Bayes inference using either EB-REML or EB-ANOVA. Each EGLS method was compared for the number of genes declared significant for various FDR thresholds.

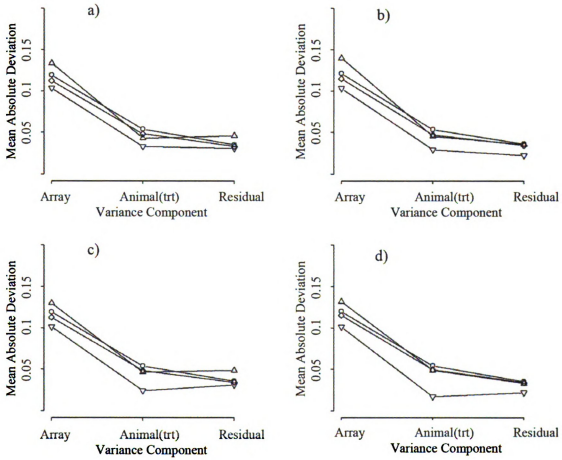
A SAS macro invoking PROC MIXED (version 9.1.3) was used for all analyses.

2.5 Results

2.5.1 Simulation study

The MAD of the VC estimates from their true values are provided for each method (EB-REML, EB-ANOVA, ANOVA and REML) for the 4 simulated datasets as based on 4 possible duplet combinations of α_2 and α_3 are provided in Figure 3. The EB-ANOVA method consistently had the lowest MAD followed closely by REML. The performance and advantage over other methods for EB-ANOVA for MAD was particularly noted for increasing values of α_2^{VC} and α_3^{VC}. This should be anticipated since less heteroskedasticity is specified with larger values of α_2^{VC} and α_3^{VC}, thereby facilitating greater borrowing of information across genes. One possible explanation is that for about 62% of the genes, the estimate of σ_2^2 converged to 0 using REML; this bounded estimate would then cause ripple effects for other VC estimates (Stroup and

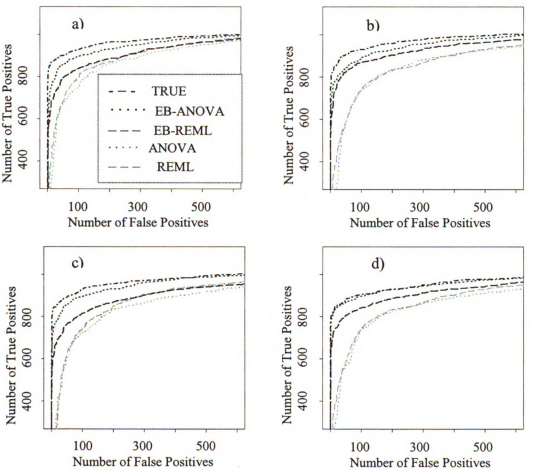
Figure 2.3 Mean absolute deviations for estimates of all variance components (array, animal(trt), and residual) for each of four variance component estimation methods (EB-ANOVA(∇), EB-REML(Δ), ANOVA(o) and REML(◊)): a) $\alpha_2^{VC}=3,\alpha_3^{VC}=3$, b) $\alpha_2^{VC}=3,\alpha_3^{VC}=12$, c) $\alpha_2^{VC}=12,\alpha_3^{VC}=3$ d) $\alpha_2^{VC}=12,\alpha_3^{VC}=12$



Little, 2004). Subsequently, it might not be possible for EB-REML to improve upon such affected estimates appreciably.

The ROC curves for the EGLS *F*-tests for treatments based on four VC estimation methods (EB-ANOVA, EB-REML, ANOVA and REML) and GLS *F*-tests based on the true VC (TRUE) are provided for each of 4 different simulated datasets for the loop design in Figures 4 and 5, respectively. As expected, GLS based on the true VC lead to the best ROC curve (i.e. largest number of true positives for a certain number of false positives within any significant gene list) for all 4 datasets. Among the EGLS methods (i.e. GLS based on use of VC estimates), the EB-ANOVA appeared to clearly outperform all of the other methods. In fact, its performance and advantage over the other methods improved with lower levels of heteroskedasticity as anticipated to the point that its performance was almost indistinguishable from true GLS for $\alpha_2^{VC} = 12, \alpha_3^{VC} = 12$. The EB-REML procedure had the next best ROC performance although it substantially lagged behind EB-ANOVA for most situations except for $\alpha_2^{VC} = 3, \alpha_3^{VC} = 12$. For example, for the situation where $\alpha_2^{VC} = 12, \alpha_3^{VC} = 12$ and a gene list already including 100 false positives, EB-REML would pick up just over 800 true positives whereas EB-ANOVA would pick up roughly 900 true positives.

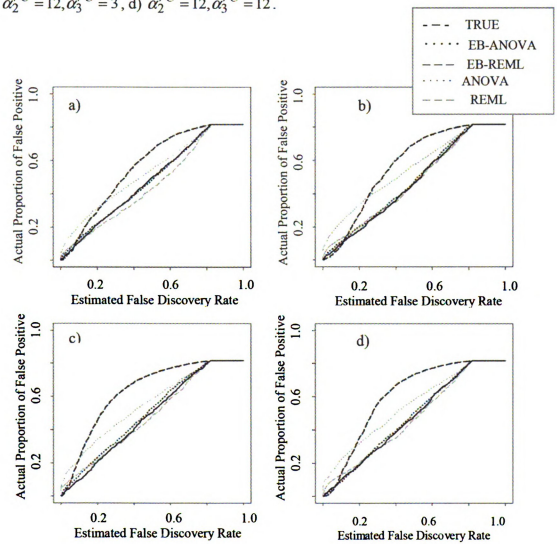
Figure 2.4 Receiver operating characteristic curves for loop design with dye swap (n = 3) using estimated generalized least squares F-tests on treatment effects based on four methods (ANOVA, REML, EB-ANOVA, EB-REML) of variance component estimation and GLS based on known VC (TRUE) for each of 4 simulated datasets defined by 2^2 factorial combination of parameters specifying different levels of heteroskedasticity for random effects subject within treatment (α_2^{VC}), residual (α_3^{VC}) given $\alpha_1^{VC}=3$ for array: a) $\alpha_2^{VC}=3,\alpha_3^{VC}=3$, b) $\alpha_2^{VC}=3,\alpha_3^{VC}=12$, c) $\alpha_2^{VC}=12,\alpha_3^{VC}=3$ d) $\alpha_2^{VC}=12,\alpha_3^{VC}=12$.



The TFDR versus EFDR for four EGLS procedures and the GLS procedure are provided in Figure 5. As anticipated, TFDR ≈ EFDR based on use of GLS for both designs since the test statistics for treatments are exact *F*-tests with infinite degrees of freedom (i.e., equivalent to an exact chi-square test). A general congruence between TFDR and EFDR was also more or less true for EGLS based on REML although it tended to be slightly too conservative (TFDR<EFDR) such that estimated proportion of false positives within a gene list would be understated. Conversely, EGLS based on ANOVA appeared to be substantially liberal (TFDR>EFDR). For example, consider panel c) $\alpha_2^{VC} = 12, \alpha_3^{VC} = 3$ of Figure 4. Using ANOVA, an EFDR of 0.20 actually translates into a TFDR exceeding 0.30; in other words if an investigator decides upon a list of genes based on a EFDR cutoff < 0.20, he or she might expect the true proportion of false positives to be closer to 0.30. Similar conclusions between the effect of REML and ANOVA on Type I error rates in EGLS inference has also been noted for unbalanced data situations by Stroup and Littell (2002).

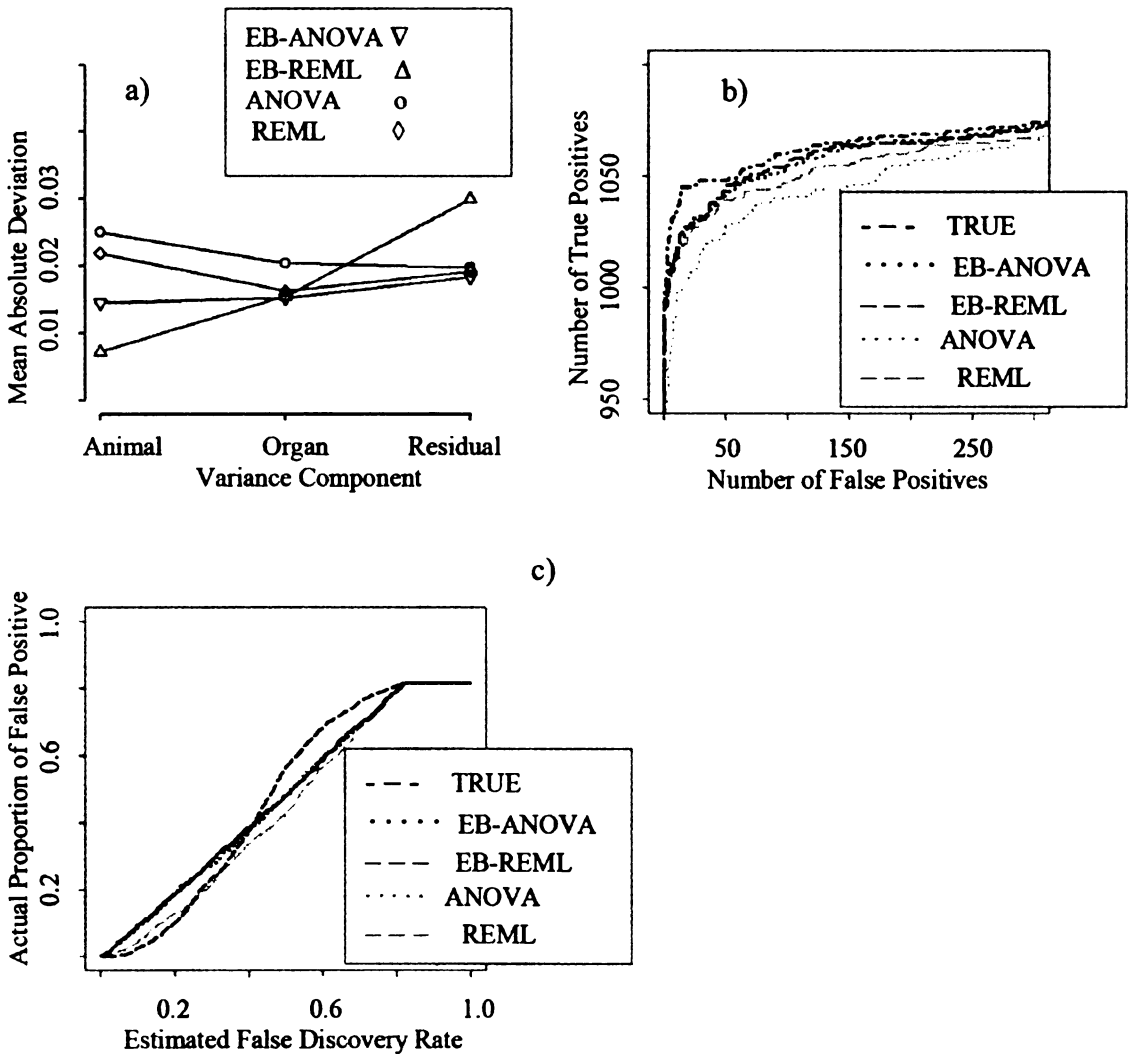
For the two shrinkage procedures, EGLS-REML also tended to be too liberal for 0.10 < EFDR < 0.80 thereby implying that if investigators chose gene lists based on EFDR < 0.10, the EFDR would closely match the actual proportion of genes that are false positives. Conversely, EGLS-ANOVA appeared to lead to effective control of FDR throughout all possible values of EFDR.

Figure 2.5 Actual versus estimated false discovery rate for loop design with dye swap (n = 3) using estimated generalized least squares (GLS) F-tests on treatment effects based on four methods (ANOVA, REML, EB-ANOVA, and EB-REML) of variance component estimation and GLS based on known VC (TRUE) for each of 4 simulated datasets defined by 2^2 factorial combination of parameters specifying level of heteroskedasticity for random effects animal within treatment (α_2^{VC}) and residual (α_3^{VC}) given α_1^{VC} =3 for array a) $\alpha_2^{VC}=3, \alpha_3^{VC}=3$, b) $\alpha_2^{VC}=3, \alpha_3^{VC}=12$, c) $\alpha_2^{VC}=12, \alpha_3^{VC}=3$, d) $\alpha_2^{VC}=12, \alpha_3^{VC}=12$.



Similar results from the simulation study on the reference design with dye swap from Pritchard et al. (2001) using the estimated values of α_1 α_2, α_3, β_1, β_2, and β_3 (provided later) as the true parameters are summarized in Figure 6. Figure 6a) is similar to Figure 3 in that the gene-specific methods (ANOVA and REML) generally had the poorest performance for VC estimation than EB-ANOVA. However, this time, EB-ANOVA did not clearly dominate MAD properties of VC estimation; although EB-REML had the worst performance for estimating σ_3^2 (residual), it had the best performance for estimating σ_1^2 (animal). This result might help explain the ROC performance for the various derivative EGLS methods in Figure 6b). There appeared to be less of a distinction between the two shrinkage methods for each of their ROC curves, with both approaching the ROC curve based on GLS using the true variance component values. Nevertheless, as further indicated by Figure 6c), EB-ANOVA was able to maintain proper control of FDR whereas EB-REML was either too conservative or too liberal.

Figure 2.6 Simulation study results for common reference with dye swap design based on four methods (ANOVA, REML, EB-ANOVA, and EB-REML) of variance component estimation and GLS based on known VC (TRUE) for each of 4 simulated datasets: a) mean absolute deviation of variance component estimates, b) receiver operating characteristic curves and c) actual versus estimated false discovery rates.



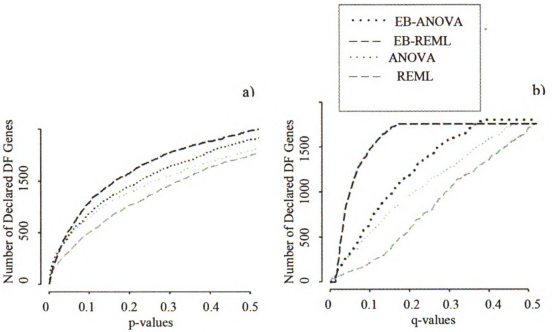
2.5.2 Data analysis

2.5.2.1 Renal data (Loop design with dye swap)

The dataset from Liang et al. (2003) was analyzed using the mixed model ANOVA of Table 1 and using the average Cy3 and Cy5 arcsinh transformed intensities over duplicate spots for each gene within an array as response variables. The arsinh transformation was deemed appropriate for this particular dataset because a quadratic relationship was observed between the variance and intensity of microarray signals (Huber et al. 2002). The MML estimates ± their asymptotic standard errors for α_1 α_2, α_3, β_1, β_2, and β_3 were, respectively, 6.14±0.35, 2.02±0.11, 1.83±0.07, 4.89±0.31, 0.15±0.01, and 0.07 ±0.003.

Figure 7 plots the number of declared significant genes against p-value and q-value cutoffs. It is interesting to note that q-values were less than p-values because of the fact that the estimated proportion of genes that were not differentially expressed was rather low; i.e. π_o=0.17 for EB-REML. EGLS based on REML appeared to detect the least number of genes with EGLS based on ANOVA having a slightly larger number for various P-value and q-value cutoffs; these results are consistent with ROC comparisons previously noted from the simulation study involving this design. Note that EB-REML declared smaller number of genes significant for q-value < 0.02 and more genes for threshold q-value > 0.02 compared to EB-ANOVA. Again, this result may be somewhat consistent with the FDR control issues previously noted from Figure 5 in that EB-REML potentially overestimates FDR for low q-values but underestimates FDR at all q-values > 0.02. EB-ANOVA detects more genes than gene specific methods (ANOVA and REML)

Figure 2.7 Renal data results: Number of declared differentially expressed genes (DF) vs. a) p-values, b) q-values.



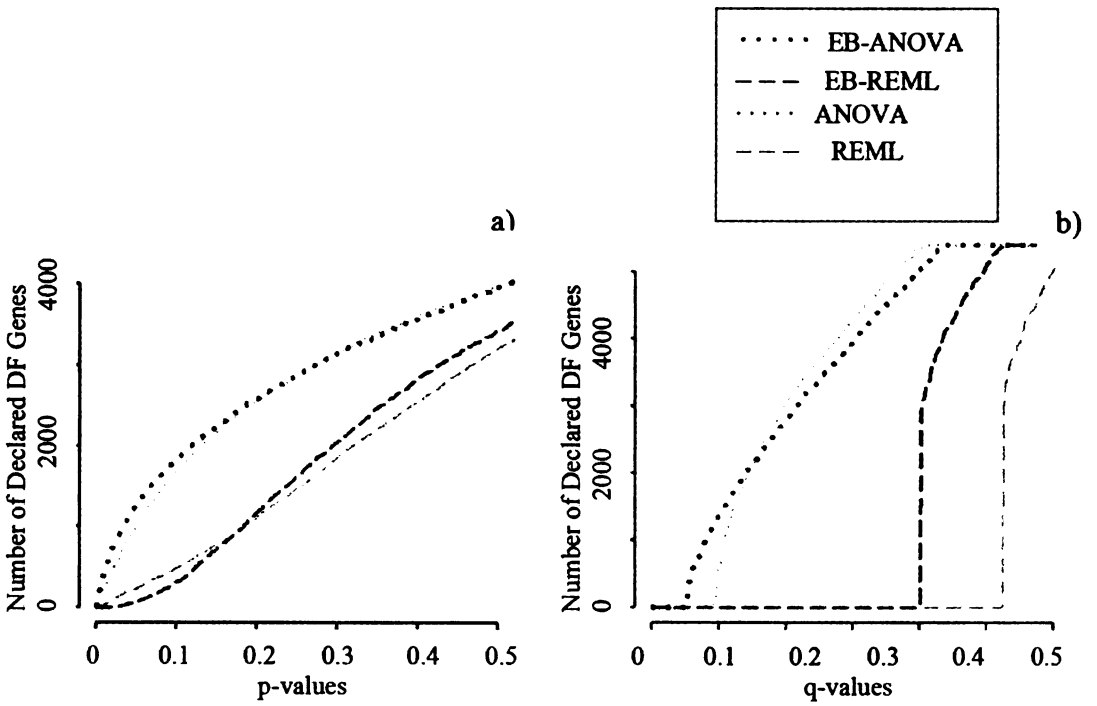
and would be expected to control FDR very well throughout all possible values of EFDR as expected based on our simulation study results.

2.5.2.2 Mouse organ data (Reference design with dye swap)

The data from Pritchard et al. (2001) were analyzed using the mixed model ANOVA of Table 2. Since we did not observe the same quadratic mean-variance relationship for fluorescence intensities like we did for the other dataset, we preprocessed this data using the lowess transformation and scale-adjustment procedure of Yang et al. (2002b). The MML estimates ± their asymptotic standard errors for α_1 α_2, α_3, β_1, β_2, and β_3 were, respectively, 5.29±0.30, 3.75±0.13, 1.63±0.03, 0.86±0.06, 0.58±0.02, and 0.37±0.01.

Figure 8 indicates that EB-REML and REML did not detect any differentially expressed genes if critical q-value was set to be less than 0.4. This attribute may be due to overestimating FDR at EFDR<0.4 for EB-REML and all FDR values for REML as demonstrated by simulation in Figure 6c. EB-ANOVA resulted in the greatest number of differentially expressed genes compared to all other methods for q-value < 0.2. Although Figure 8a) shows that the numbers of differentially expressed genes detected by EB-ANOVA are greater than those detected by ANOVA for p-values <0.5, Figure 8b) appears to have a different pattern where ANOVA seems to detect slightly more genes than EB-ANOVA if the q-value > 0.2. The reason may rely upon the different distribution of p-values and the stochastic effects of non-differentially expressed genes resulting from these two methods.

Figure 2.8 Mouse organ data results: Number of declared differentially expressed genes (DF) vs. a) p-values, b) q-values.



2.6 Discussion

In this paper, we explored alternative mixed model inference methods on two-color microarray data sets under two of the most common designs, namely the loop design and common reference design to demonstrate. In addition to desirable ROC and FDR properties already noted, our proposed shrinkage method EB-ANOVA is relatively easy to implement, requiring only a slight modification of mixed model software such as, for example, SAS PROC MIXED. It can also be expandable for other efficient microarray experimental designs except for repeated measures designs that require specifications of general temporal covariance structures. The underlying assumption of our model that ANOVA components within gene variance are distributed as inverse gamma distributions is appropriate (Box GEP. and Tiao GC.,1973) and appears to be reasonable in actual experiment data. It turns out to be a close approximation in our real data as well.

Our EB-ANOVA method was also demonstrated to detect the greatest number of true positives when controlling false positives for both designs. We believe that one of the reasons for this includes more precise variance component estimates. In gene-specific model, the sample size is usually small such that variance estimates can be imprecise (Wright and Simon, 2003), but the number of genes is very large. Empirical Bayes or shrinkage methods are ideally suited for these situations. Another reason for EB-ANOVA's improved performance is the ability to estimate a posterior degrees of freedom that is best reflective of the distribution of EMS across genes and is generally substantially larger than the classical ANOVA degrees of freedom, the difference coming from the prior distribution for the error term increases the power of hypothesis tests. Finally, we also noted (not reported) slightly more accurate estimates of treatment

mean differences using EB-ANOVA which is partly attributable to the better recovery of the interblock information when variance components are estimated with greater precision.

In a mixed model analysis, negative variance component estimates are possible for ANOVA whereas in REML, otherwise negative estimates are set to zero. These VC estimation situations affect inference on fixed effects in ways that are not generally well-understood (Stroup, 2003). Therefore, we also investigated how ANOVA versus REML as well as EB-ANOVA vs EB-REML influenced EGLS on treatment effects. Our results for the classical REML and ANOVA gene-specific methods agreed with the conclusion in Stroup and Littell (2002). The use of REML tends to deflate statistical power, resulting in a conservative test for EGLS on treatment effects when it is likely to have an excessive number of zero estimates for the VC pertaining to the experimental error term. For microarray data, it is not uncommon to have a larger number of such situations because the variance for technical replicates is often substantially larger than the variance for biological replicates for many genes (Cui and Churchill, 2003).

The null distribution for testing treatment effects may have a direct impact on identification of differentially expressed genes. We extended a formal derivation (Wright and Simon, 2003) to form our null distribution and estimated denominator degrees of freedom for the denominator of F-test from the data. In our simulation study for both designs, we found that the empirical distribution of the F-statistics corresponded well with theoretical F-distributions using $\frac{\alpha_j}{\beta_j}\hat{\phi}_{ij} \sim F_{2\alpha_j,\nu_j}$ in spite of the fact that the simulation model and EB-ANOVA model did not match. However, the EB-REML

procedure (Feng et al., 2006) is based on the use of 5 degrees of freedom because of simulation work that Feng et al. (2006) pursued. This approach may not be reasonable for analysis such as determining differentially expressed genes by pre-set cut-off significance levels, but it may be appropriate to another goal such as gene ranking. Our simulation studies show the poor performance in terms of controlling FDR in both designs.

We have also addressed the issue of multiple testing for the four methods. The positive FDR method (Storey 2002) is commonly used for microarray and other data sets with the large number of comparisons. In our simulation study, we examine the behavior of different methods in terms of controlling FDR. It gives some ideas that the estimated FDR might overestimate or underestimate the true FDR depending on which method is used. In our examples, the EB-ANOVA method has reasonably good FDR control in both designs. Clearly, there is much more to learn, for example, how the distributions of p-values from different methods change with variability among the VC; and those methods differ relative to the efficiency of design and the number of replicates. Therefore, we justify this result with caution to general cases.

BIBLIOGRAPHY

BALDI, P. & LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**(6), 509-519.

CASELLA, G. (1985). An Introduction to Empirical Bayes Data-Analysis. *American Statistician* **39**(2), 83-87.

CHEN, J. DELONGCHAMP, R. TSAI, C.-A. HSUEH, H.-M. SISTARE, F. THOMPSON, K. L. DESAI, V. G. & FUSCOE, J. C. (2004). Analysis of variance components in gene expression data. *Bioinformatics* **20**(9), 1436-1446.

COCHRAN, W. G. & COX, G. M. (1957). *Experimental Designs, Second Edition*. Wiley, New York.

CUI, X. & CHURCHILL, G. A. (2003). How many mice and how many arrays? Replication in mouse cDNA microarray experiments. In *Methods of Microarray Data Analysis 3* Eds K. F. Johnson & S. M. Lin), pp. 139-154. Norwell, MA: Kluwer Academic Publishers.

CUI, X. G. HWANG, J. T. G. QIU, J. BLADES, N. J. & CHURCHILL, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**(1), 59-75.

DOBBIN, K. K. SHIH, J. H. & SIMON, R. (2003). Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute* **95**(18), 1362-1369.

FENG, S. WOLFINGER, R. D. CHU, T. M. GIBSON, G. C. & MCGRAW, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**(2), 197-209.

GIESBRECHT, F. G. & GUMPERTZ, M. L. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Hoboken, Wiley, New York.

HINKELMANN, K. & KEMPTHORNE, O. (1994). *Design and analysis of experiments: Volume I: introduction to experimental design*. Wiley, New York.

HUBER, W. HEYDEBRECK, A. V. SULTMANN, H. POUSTKA, A. & VINGRON, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl 1), S96-104

KENWARD, M. G. & ROGER, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983-997.

KERR, M. K. & CHURCHILL, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**(2), 123-128.

LONNSTEDT, I. & SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**(1), 31-46.

MURPHY, D. (2002). Gene expression studies using microarrays: Principles, problems, and prospects. *Advances in Physiology Education* **26**(4), 256-270.

SCHAALJE, G. MCBRIDE, J. & FELLINGHAM, G. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural Biological and Environmental Statistics* **7**(4), 512-524.

SEARLE, S. R. (1971). *Linear Models*. Wiley, New York.

SEARLE, S. R. CASELLA, G. & MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1), No. 1, Article 3. Available at: http://www.bepress.com/sagmb/vol3/iss1/art3

SPILKE, J. PIEPHO, H. P. & HU, X. Y. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural Biological and Environmental Statistics* **10**(3), 374-389.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of The Royal Statistical Society Series B-Statistical Methodology* **64**, 479-498.

STOREY, J. D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9440-9445.

STROUP, W. W. & LITTELL, R. C. (2002). Impact of variance component estimates on fixed effect inference in unbalanced linear mixed models. In *14th Annual Kansas State University Conference on Applied Statistics in Agriculture* pp. 32-48. Manhattan, KS.

TEMPELMAN, R. J. (2005). Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Veterinary Immunology and Immunopathology* **105**(3-4), 175-186.

VINCIOTTI, V. KHANIN, R. D'ALIMONTE, D. LIU, X. CATTINI, N. HOTCHKISS, G. BUCCA, G. DE JESUS, O. RASAIYAAH, J. SMITH, C. KELLAM, P. & WIT, E. (2005). An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* **21**, 492-501.

WOLFINGER, R. D. GIBSON, G. WOLFINGER, E. D. BENNETT, L. HAMADEH, H. BUSHEL, P. AFSHARI, C. & PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625-637.

WRIGHT, G. W. & SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**(18), 2448-2455.

WU, H. KERR, M. K. CUI, X. & CHURCHILL, G. A. (2003). MAANOVA: A software package for the Analysis of Spotted cDNA microarray experiments. In *The Analysis of Gene Expression Data* Eds G. Parmigiani, E. S. Garrett, R. A. Irizarry & S. L. Zeger), pp. 313-341. New York: Springer-Verlag.

YANG, Y. H. DUDOIT, S. LUU, P. LIN, D. M. PENG, V. NGAI, J. & SPEED, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4).

YATES, F. (1940). The recovery of interblock information in balanced incomplete block designs. *Annals of Eugenics* **10**, 317-325.

Chapter 3: Assessing Shrinkage Procedures for Differential Gene Expression in Microarray Experiments Having Within-Array Replicate Spots

Abstract

Empirical Bayes methods have been promising for moderating test statistics for inference on differential expression in small microarray experiments based on shrinking gene-specific estimates of variance to a common value. Yet many microarray experiments are characterized by both biological and technical replicates, the latter of which may include, for example, several spots per probe on an array, thereby invalidating the use of much available empirical Bayes software. A popular R software package, LIMMA, was recently extended to help addressing this inferential problem for special cases by assuming a constant correlation for the within-array replicates for each gene. We assert that assumption would not be true for most experimental situations and test this in this paper.

Results. A BAYESRATIO model is constructed to generalize the common correlation assumption. We conducted a simulation study based on real data parameter estimates to compare LIMMA software with other alternative methods. Those include the BAYESRATIO model proposed in this paper, the Gene-specific model with ANOVA or REML (Wolfinger et al. 2001), the Empirical Bayes with REML (EB-REML) (Feng et al. 2006) and the Empirical Bayes with ANOVA (EB-ANOVA) (chapter 1). The comparison is based upon various scenarios of heterogeneity of correlation coefficients. The LIMMA package was found to be too liberal in terms of controlling for false discovery rates

(FDR) when the data differs from the common correlation assumption, particularly with increasing numbers of technical replicates per gene. An alternative method, the BAYESRATIO model was shown to have superior performance on ROC curves and FDR control. Moreover, the method EB-ANOVA we proposed in Chapter 1, which is a shrinkage procedure with linear mixed model, has demonstrated robust performance for controlling FDR and has reasonably good ROC curves compared to any other competitive methods. Furthermore, EB-ANOVA is adaptable to any mixed model framework, where the LIMMA package has design limitations.

3.1 Introduction

Gene expression profiling or microarray analysis has enabled the measurement of thousands of genes in a single hybridization experiment. Microarray studies are also stimulating the discovery of new targets for the treatment of disease which is aiding drug development, immunotherapeutics and gene therapy (http://www.microarrayworld.com/). A comparison of gene expression in cells or tissues from different conditions may provide useful information regarding important biological processes and functions. In this type of experimental setup, the main interest is the identification of differentially expressed genes in different conditions (treated vs. untreated samples, diseased vs. normal tissue, mutant vs. wild-type organisms, etc.) (Breitling et al. 2004). The challenge is how to detect those genuine changes from noisy data which still exists, although much attention has been given to the statistical analysis of microarray data (Smyth et al. 2005).

As with all designed experiments, it is necessary to replicate microarray studies in order to infer upon differential gene expression between various treatment groups or conditions (Kerr et al. 2002). There are often two categories of replication in these studies: biological and technical (Churchill 2002). Whereas technical replication is useful to control measurement error, biological replication is vital for valid statistical inference. Mixed model ANOVA is the primary statistical inference tool for experiments with different levels of replication; in other words, mixed model inference is able to disentangle multiple strata of random variation as by different levels of replication or blocking. Typically these analyses are conducted separately for each gene using REML to estimate variance components of each random effect (Wolfinger et al. 2001).

Empirical Bayes (EB) procedures have been popularized for the analysis of microarray studies, since test statistics for differential expression incorporate information across all genes in the study, thereby improving reliability of inference for any one gene (Baldi & Long 2001; Newton et al. 2001; Tseng et al. 2001; Lonnstedt & Speed 2002; Wright & Simon 2003; Smyth 2004). Most such methods have been developed for models assuming a single error strata or residual variance. Hence, these procedures are not readily adaptable to experimental designs characterized by technical and biological replicates thereby leading to multiple strata of random variation for each gene. Cui et al. (2005) proposed a shrinkage procedure, currently implemented in the software MAANOVA (http://www.jax.org/staffh/churchill/labsite/software) for mixed model analysis of microarrays. Their procedure for variance component inference is based on borrowing information across all genes using James-Stein-Lindley shrinkage to modify F test statistics for treatment effects on any one gene. Accurate estimation of variance components using shrinkage estimation should translate into more accurate estimates of ANOVA expected mean squares (EMS) for random effects factors. Since the ANOVA mean squares (MS) for some of these factors constitute the experimental error for treatment effects, greater statistical power should then be afforded for inference on differential expression. This is particularly true, for example, for within-array technical replicates where the array within treatment MS serves as the experimental error term for the analysis of log-ratio data in common reference designs. Feng et al. (2006) recently introduced a promising moderation procedure based on direct shrinkage estimation of ANOVA EMS rather than VC in balanced mixed effects models. Their procedure is based on work by Wolfinger & Kass (2000) who demonstrated that REML could be

specified as a function of independent ANOVA EMS in balanced designs. A major limitation of their procedure, which we denote as EB-REML, is how to appropriately determine the degrees of freedom for test statistics on treatments; they arbitrarily chose 5 degrees of freedom for the example they illustrated in their method (Feng et al. 2006).

In the previous chapter, we proposed an alternative shrinkage estimation procedure for mixed model inference of microarray data. The null distribution for testing differentially expressed genes based on this procedure, that we labeled as EB-ANOVA, was demonstrated to be well defined by F densities having readily estimated denominator degrees of freedom. We subsequently demonstrated superior performance of EB-ANOVA vs. EB-REML for unbiased control of false discovery rates and receiver operating characteristics for inference on treatment effects. Smyth (2005) recently developed a between-gene shrinkage analysis method for a particular type of mixed model design, those where the technical replicates are either within-array (i.e. each gene is spotted more than once on an array) or special cases of between-array replicates where each sample is hybridized more than once. Smyth (2005) invoked the assumption that the correlation of these technical replicates are constant within genes based on a linear model analysis of log-ratios. This modeling strategy is currently available as the *dupcor* option in the popular R software LIMMA. The implications of this assumption are that an investigator can build up degrees of freedom of test by increasing the number of technical replicates while holding constant the number of biological replicates.

In this paper, we develop a fully Bayesian model approach that models variability on the within gene correlation of technical replicates. We compare this approach, which we label the BAYESRATIO model, to the LIMMA *dupcor* procedure in order to evaluate the

robustness of the latter in situations when the correlation is heterogeneous across genes. We also compare both of these methods with those previously considered in Chapter 1, namely EB-ANOVA, EB-REML, and the two non-shrinkage counterparts based on ANOVA and REML estimation of variance components.

The paper is organized as follows. In Section 1, LIMMA (without Empirical Bayesian adjustment) and EB-LIMMA for technically replicated procedures, and the development of BAYESRATIO model, are described. In Section 2, an application is introduced. In Section 3, we simulate data based on this application and varying degrees of heterogeneity of variance ratio and residual variances with average correlation coefficients 0.6 and 0.9 imitating the scenarios of top and bottom halves and side-by-side replicates to explore the robustness of the LIMMA procedure and also to compare its performance with our alternative methods. The results and discussion from simulation study are in Section 4. Real data application is provided in Section 5. Our conclusions are summarized in Section 6.

3.2 Mixed model presentation of Smyth's constant correlation method

We present the approach of Smyth et al. (2005) but in the more general context of a linear mixed model with a single random effects factor. Suppose, as in Smyth et al. (2005) that the data vector y_g, constitutes log-ratios for gene g, $g = 1,2,\ldots,G$, as might be considered appropriate for common reference designs where the ratios are expressed as treatment relative to reference sample for each array. Consider a particular common reference design case where for each of t treatments (not including the common

reference), there are *n* different arrays, experimental or biological replicates, with each biological replicate consisting of *m* technical replicates. Hence the dimension of y_g is *tnm*. We model y_g as a linear function of gene-specific *k* x 1 vector of fixed effects ($\boldsymbol{\beta}_g$) which may include not only treatment effects but other covariates, a *n* x 1 vector of biological or subject effects (u_g), and a *tnm* x 1 vector of residuals (e_g):

$$\mathbf{y_g} = \mathbf{X}\boldsymbol{\beta}_\mathbf{g} + \mathbf{Z}\mathbf{u_g} + \mathbf{e_g} \qquad [1a]$$

with

$$\mathbf{u}_g \sim N(0, \mathrm{I}_n \sigma^2_{u_g}) \qquad [1b]$$

and

$$e_g \sim N(0, \mathrm{I}_{mn} \sigma^2_{e_g}) \qquad [1c]$$

as in Smyth et al. (2005). Here the design matrices **X** and **Z** of dimensions *tnm* x *k* and *tnm* x *n* , respectively, are specified to be the same for all genes. Again, u_g is used to model experimental or biological variability whereas e_g is used to model technical or measurement error variability. Now Smyth et al. (2005) further assumed an effectively known and constant correlation $\rho_g = \dfrac{\sigma^2_{u_g}}{\sigma^2_{u_g} + \sigma^2_{e_g}}$ between technical replicates within experimental replicates across all genes, i.e. $\rho = \rho_1 = \rho_2 = ... \rho_G$. It is well established that when ρ_g is known, or synonymously, when $\lambda_g = \dfrac{\sigma^2_{e_g}}{\sigma^2_{u_g}}$ is known, the best linear unbiased estimator (BLUE) of any linear estimable combination of $\boldsymbol{\beta}_g$, say, $\mathbf{C}'\boldsymbol{\beta}_g$ for **C**

being a known contrast matrix, can be determined as $C'\hat{\boldsymbol{\beta}}_g$ using Henderson's mixed model equations (Henderson 1984):

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+I\lambda_g \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\beta}}_g \\ \hat{u}_g \end{bmatrix}=\begin{bmatrix} X'y_g \\ Z'y_g \end{bmatrix} \qquad [2]$$

As a byproduct of solving [2], $\hat{u}_g$ is the best linear unbiased predictor (BLUP) of u_g which might be of little direct interest in microarray studies. Further note from [2] and from Henderson (1984) that the BLUE of $\boldsymbol{\beta}_g$ does not depend on knowledge of either $\sigma^2_{u_g}$ or $\sigma^2_{e_g}$ when λ_g is known.

Now $\hat{\boldsymbol{\beta}}_g$ can also be shown to be the generalized least squares (GLS) estimator of $\boldsymbol{\beta}_g$ provided by $X'V_g^{-1}X\hat{\boldsymbol{\beta}}_g = X'V_g^{-1}y_g$ where $\text{var}(y_g) = V_g = ZZ'\sigma^2_{u_g} + I\sigma^2_{e_g}$. Writing $V_g = W_g\sigma^2_{e_g}$ with $W_g = \left(ZZ'\lambda_g^{-1} + I\right)$, it can be further confirmed from the GLS estimation equations that $\hat{\boldsymbol{\beta}}_g$ depends only on knowing λ_g in W_g using $X'W_g^{-1}X\hat{\boldsymbol{\beta}}_g = X'W_g^{-1}y_g$

Under the null hypothesis H_o: $C'\boldsymbol{\beta}_g = m$ with **m** typically specified to be a null vector, Henderson (1984) demonstrated that, given known λ_g, the resulting F-test statistic would have denominator degrees of freedom that build up on both increasing *n* and *m*:

$$\left(C'\hat{\boldsymbol{\beta}} - m\right)'\left(C'\left(X'W_g^{-1}X\right)C\right)^{-1}\left(C'\hat{\boldsymbol{\beta}} - m\right)\hat{\sigma}^2_{e_g} \sim F_{rank(K),tnm-rank(X)} \qquad [3]$$

where

$$\hat{\sigma}^2_{e_g} = \frac{\left(\mathrm{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_g\right)' \mathrm{W}_g^{-1} \left(\mathrm{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_g\right)}{tnm - rank(\mathrm{X})} \sim \frac{\sigma^2_{e_g}}{tnm-1} \chi^2_{tnm-1} \qquad [4]$$

Consider a simple design involving $t = 1$ treatment with log-ratios expressed relative to either a common reference or a second treatment in a balanced block design. Again, the design uses n biological or experimental replicates with m technical replicates per each biological replicate. Now suppose that no additional covariates are modeled, i.e. $\mathbf{X} = \mathbf{1}_{nm}$, such that $\boldsymbol{\beta}_g$ is scalar and k=rank($\mathbf{X}$) = 1. Then the contrast scalar $\mathbf{C}' = 1$ with $\mathbf{m} = 0$ in the test for differential mean expression of the treatment of interest relative to the other. Given that special case, the square root of [3] would simplify to Equation (4) of Smyth et al. (2005) noting that the square root of $F_{1,nm-1}$ is a t-test statistic with the same denominator degrees of freedom as provided by Smyth et al. (2005).

Consider again the same experimental design but where now $\boldsymbol{\beta}_g$ includes additional covariates with rank($\mathbf{X}$) = k as in Section 4 of Smyth et al. (2005). As in Smyth et al., (2005), let $\alpha_g = \mathrm{c}'\boldsymbol{\beta}_g$ for $\mathbf{c}$ being a known contrast vector chosen such that it is of interest to test: $\mathrm{H_o}$: $\alpha_g = 0$. Then the square root of the test statistic in Equation [3] is equivalent to the t-test statistic provided near the top of page 2071 in Smyth et al. (2005) with nm-k degrees of freedom. If the design is extended to consider more than $t = 1$ treatments relative to a reference, with again each treatment having n biological replicates with m technical replicates per biological replicate, then the corresponding test statistic would based on tnm-k degrees of freedom as in Equation [3].

Recognizing the utility of further moderating the estimator of $\sigma^2_{e_g}$ in [4], Smyth et al. (2005) proposed an inverse Gamma prior on $\sigma^2_{e_g}$. Using the inverse Gamma specification $\mathrm{IG}(\alpha_e, \beta_e)$ such that the prior expectation $\sigma^2_{e_g}$ is $\mathrm{E}\left(\sigma^2_{e_g}\right) = \frac{\beta_e}{\alpha_e - 1}$. Combining the likelihood contribution from [4] with this inverse prior, the posterior density of $\sigma^2_{e_g}$ is

$$\frac{(tmn - 1 + 2\alpha_e)\tilde{\sigma}^2_{e_g}}{\sigma^2_{e_g}} \sim \chi^2_{tmn-1+2\alpha_e} \qquad [5]$$

where

$$\tilde{\sigma}^2_{e_g} = \frac{(tmn-1)\hat{\sigma}^2_{e_g} + 2\alpha_e\left(\frac{\alpha_e}{\beta_e}\right)^{-1}}{(tmn-1)+2\alpha_e} \qquad [6]$$

is a combination of the data information $\hat{\sigma}^2_{e_g}$ and prior harmonic mean $\left(\frac{\alpha_e}{\beta_e}\right)^{-1}$ weighted by data degrees of freedom $tmn - 1$ and prior degrees of freedom $2\alpha_e$, respectively. Note that the posterior estimate in [6] is similar to that provided as $\tilde{s}^2_g$ in Smyth et al. (2005) except for different parameterization; i.e. their d_o is our $2\alpha_e$ and their s^2_o is our $\left(\frac{\alpha_e}{\beta_e}\right)^{-1}$.

3.3 Accounting for variability in within-array replicate correlation across genes

Recognizing that $\lambda_g = \frac{1-\rho_g}{\rho_g}$, we model variability in ρ_g by modeling variability in $\tau_g = \lambda_g^{-1}$. Again, we start with the linear mixed model in [1]. Let the prior distribution for $\sigma^2_{e_g}$ be $IG(\alpha_e, \beta_e)$ as in the previous section and further let the prior distribution for τ_g be $IG(\alpha_\tau, \beta_\tau)$ such that $\mathrm{E}(\tau_g) = \frac{\beta_\tau}{\alpha_\tau - 1}$. Since we also aim to infer upon α_e, β_e, α_τ, and β_τ, we specify the same proper yet vaguely informative prior $p(\theta) \propto \frac{1}{(1+\theta)^2}$ that seems useful for parameters defined on the positive real line (Albert 1999) for each of these four parameters.

We pursue a Markov Chain Monte Carlo (MCMC) approach (Gelman et al. 1995) to implement fully Bayesian inference with this model. In MCMC, one samples from the joint posterior density of all unknown parameters by simply generating random samples from the full conditional distributions (FCD) for each unknown parameter or groups thereof conditional on simulated values for all remaining parameters and the data $\mathrm{y} = \left[\begin{array}{ccccc} \mathrm{y}_1' & \mathrm{y}_2' & \cdots & \mathrm{y}_{G-1}' & \mathrm{y}_G' \end{array} \right]'$. In our example, these parameters include $\boldsymbol{\beta}_g, \mathbf{u}_g$, τ_g, and $\sigma^2_{e_g}$ for g = 1,2,...,G as well as α_e, β_e, α_τ, and β_τ. In the following specifications of the FCD, we use "*ELSE*" to specify all parameters other than those that we are specifying the FCD for. For example, it can be readily shown using results from Wang et al. (1993) that the joint FCD of $\boldsymbol{\beta}_g, \mathbf{u}_g$ is multivariate normal

$$\boldsymbol{\beta}_g, \mathbf{u}_g \mid \mathrm{y}, ELSE \sim N\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}_g \\ \hat{\mathbf{u}}_g\end{bmatrix}, \begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\tau_g^{-1}\end{bmatrix}^{-1}\right) \quad [7]$$

for gene $g = 1,2,\ldots,G$. Note that scalar FCD sampling of individual elements of $\boldsymbol{\beta}_g$ or $\mathbf{u}_g$ are possible using developments in Wang et al. (1994) if a joint sample from [7] is computationally intractable. The FCD of $\sigma^2_{e_g}$, =1,2,...,G, can also be shown to be an inverted gamma density with parameters $\frac{(tm+1)n}{2}+\alpha_e$ and $\frac{\mathbf{u}_g'\mathbf{u}_g}{2\tau_g}+\frac{\mathbf{e}_g'\mathbf{e}_g}{2}+\beta_e$ for $\mathrm{e}_g = \mathrm{y}_g - \mathrm{X}\boldsymbol{\beta}_g - \mathrm{Z}\mathrm{u}_g$. Furthermore, the FCD of τ_g (g=1,2,...,G) is also inverted gamma with parameters $\frac{n}{2}+\alpha_\tau$ and $\frac{\mathbf{u}_g'\mathbf{u}_g}{2\sigma_{e_g}}+\beta_\tau$.

The FCD of α_τ and β_τ can be written proportionately as follows.

$$p(\alpha_\tau \mid \mathrm{y}, ELSE) \propto \frac{(\beta_\tau)^{\alpha_\tau G}}{(\Gamma(\alpha_\tau))^G}\left(\prod_{g=1}^{G}(\tau_g)^{-(\alpha_\tau+1)}\right)\frac{1}{(1+\alpha_\tau)^2}, \quad [8a]$$

and

$$p(\beta_\tau \mid \mathrm{y}, ELSE) \propto (\beta_\tau)^{\alpha_\tau G}\left(\prod_{g=1}^{G}\exp\left(-\frac{\beta_\tau}{\tau_g}\right)\right)\frac{1}{(1+\beta_\tau)^2}, \quad [8b]$$

with the FCD of α_e and β_e being similarly written as for α_τ and β_τ in [8a] and [8b], respectively. Since the FCD for these four parameters were not recognizable, we embedded a Metropolis-Hastings step (Chib & Greenberg 1995) for sampling from each of these four FCD within the MCMC scheme. Doing so, we found that α_e with β_e and α_τ with β_τ were highly correlated, as might be expected since α_e and β_e jointly

determine the mean residual variance across genes whereas α_τ with β_τ jointly determine the mean variance ratio across genes. This high correlation substantially limits MCMC mixing and hence the number of effectively independent samples that one might obtain for a fixed number of MCMC samples. To improve MCMC mixing, we drew Metropolis Hastings samples from $\mu_\tau = \frac{\beta_\tau}{\alpha_\tau - 1}$ $\mu_e = \frac{\beta_e}{\alpha_e - 1}$, α_e, and α_τ and determined the correlations between the samples from these four parameters to be substantially dampened, thereby improving MCMC mixing.

3.4 Description of experimental design and simulation study

We based our comparison of various methods on the cDNA microarray design utilized by Wade et al. (2004, 2005) who investigated differences in gene expression in telencephalon brain tissue between the two sexes of juvenile (25-day old) zebra finches. A balanced block design with n = 8 birds per sex was used whereby the mRNA sample from one male was hybridized against that for one female for each of 8 different slides. That is, four of the arrays involved a Cy3 labeled male sample hybridized against Cy5 labeled female sample whereas the opposite dye assignments on bird sex were used for the remaining four arrays. One of the 8 slides was eventually discarded because of quality control flags leaving a total of 7 slides.

Now each microarray was spotted with 2399 cDNAs of which two (β-actin and GAPDH) were controls that were printed with each of the 32 print-tips arranged in a 8 metarow x 4 metacolumn grid, thereby leading to a 8 x 4 patch arrangement on the

microarray. The remaining 2397 genes were duplicate spotted on the array with both duplicates being drawn from the same 384 well source plate. Each such pair of wells on the corresponding source plates were located within the same column but two rows apart such that duplicates were printed within the same meta-column but two meta-rows apart on the 8 x 4 microarray grid. Given the potential spatial variability that exists on a slide, we deem this strategy to be sounder for spotting duplicate spots relative to having each duplicate spot located nearly adjacent to each other on the slide.

We decided to model our simulation study on the same 8 slide design as that considered in Wade et al. (2004) but based on G=6000 duplicated spotted genes. Now the expected values of the residual variances and variance ratios were based on those estimated from the data of Wade et al. (2004), being $\mathrm{E}\left(\sigma_{e_g}^2\right) = 0.03$ and $\mathrm{E}\left(\tau_g\right) = 1.95$ respectively, on the logarithmic to base 2 scale. One thousand of these genes were specified to be differentially expressed based on fold changes for 10 equally-sized groups of differentially expressed genes having fold changes 1.25, 1.5, 2, 2.5, 3, 1.25^{-1}, 1.5^{-1}, 2^{-1}, 2.5^{-1}, and 3^{-1} for each of 100 genes between the two treatments. Hence the proportion of genes that were not differentially expressed between the two treatments (sexes) was $\pi_0 = \frac{5000}{6000} = 0.83$.

We wanted to address several issues with respect to the comparison of LIMMA and the competing alternative BAYESRATIO model that we propose:

1) How does the level of residual heteroskedasticity (α_e) influence the relative performance of the competing methods? We choose to compare $\alpha_e = 3$ (high level of heteroskedasticity) to $\alpha_e = 12$ (mild level of heteroskedasticity),

anticipating that in particular the empirical Bayes based methods should perform better for $\alpha_e = 12$.

2) How does the level of variance ratio heteroskedasticity (α_τ) influence the relative performance of the competing methods? We choose to compare $\alpha_\tau = 3$ (high level of heteroskedasticity) to $\alpha_\tau = 30$ (very low levels of heteroskedasticity), anticipating that in particular the LIMMA-based methods should perform relatively better for $\alpha_\tau = 30$ as homoskedasticity of variance ratios (and hence within-class correlation coefficients) is implied.

3) How does the magnitude of the correlation coefficient influence the comparison of the competing methods? We choose to compare $\mathrm{E}(\rho_g) = 0.6$ to $\mathrm{E}(\rho_g) = 0.9$ with the former being representative of replicated spots being distributed throughout the slide, as based on the described study by Wade et al. (2004), whereas the latter is more representative of replicated spots being spotted by the same print-tip or as also reported as being typical by Smyth et al. (2005). We anticipated that LIMMA was likely to be substantially more liberal compared to other methods with larger $\mathrm{E}(\rho_g)$.

4) How does the number of replicated spots (m=2 versus m=4) per gene influence the comparisons between the competing methods? Again, we anticipated that LIMMA would be likely substantially more liberal compared to other methods with larger m.

Our study was based on generating data from the full complement of $2^4 = 16$ different datasets based on a 2^4 unreplicated factorial for the 4 factors characterized above.

The linear mixed effects model that was used to generate data from this experimental design is based on equation [1] and can be specified in scalar notation as follows:

$$y_{gijk} = \mu_g + \gamma_{gi} + a_{gj} + e_{gijk}\,. \qquad [9]$$

Here y_{gijk} is the normalized logarithmic female vs.male intensity ratio for gene g =1,2,...,6000 at spot k =1,2,...,m within array j=1,2,..,8 and female sample dye assignment of i = 1,2. The fixed effects include the female versus male mean difference μ_g and the effect of dye γ_{gi} i=1,2 whereas array effects a_{gj}, j=1,2....,8 are specified to be random $a_{gj} \sim NIID(0,\sigma^2_{a_g})$ whereas the residuals $e_{gijk} \sim NIID(0,\sigma^2_{e_g})$.

For model [9], it can be demonstrated using classical ANOVA that the denominator mean square MSB_g for the F-test statistic used to test Ho: $\mu_g = 0$ is based on both $\sigma^2_{e_g}$ and $\sigma^2_{a_g}$; i.e.

$$\mathrm{E}\left(MSB_g\right) = \sigma^2_{e_g} + m\sigma^2_{a_g}$$

Hence, the estimates MSB_g for each of the seven methods were compared for their mean absolute deviation $MAD = \dfrac{\sum_{g=1}^{G} |MSB_g - E(MSB_g)|}{G}$ from the true. Methods characterized by smaller MAD should lead to more sensitive and specific hypothesis testing on treatment effects given that the numerator of the associated F test is the same for all methods.

For each method, receiver operating characteristics (ROC) curves were determined by plotting the number of true positives (i.e. truly differentially expressed) genes versus the number of false positive genes. ROC curves provide effective visualizations to compare

methods and experimental designs for the trade-off between false positives and negatives (Vinciotti et al. 2005; Feng et al. 2006). The greater the number of true positives for a fixed number of false positives within a particular gene list, the better the method or design (De Smet et al. 2004).

The estimated false discovery rates (FDR) for the F-test on treatment effects based on the procedure of Storey (2002) were implemented using the fdr.control function in R. In order to facilitate a finer comparison on FDR control between the various methods, the true proportion of non-differentially expressed genes was not estimated but set to be known as $\pi = \frac{5000}{6000} \approx 0.833$ for estimation of FDR for comparison against the true FDR.

3.5 Results and Discussion

The MAD of MSB_g from its expectation are provided for each of the seven methods for the 2^4 simulated datasets in Figures 1 (m=2, ρ = .6), 2 (m = 4, ρ = .6), 3 (m =2, ρ = .9), and 4 (m = 4, ρ = .9). Regardless of the value of m or ρ, EB-REML was consistently the worst performing method for estimating $E(MSB_g)$ except for the situation characterized by the smallest levels of heteroskedasticity for both random effects: $\alpha_\tau = 30$ and $\alpha_e = 12$, in which it slightly outperformed ANOVA and REML. These results suggest that the method as proposed by Feng et al. (2006) has limited merit for microarray data analysis in spite of its shrinkage basis. For all four combinations of α_τ and α_e, the BAYESRATIO method had generally the best MAD performance for estimating $E(MSB_g)$ with relative performance generally improving as α_e and α_τ increased (i.e. decreasing heteroskedasticity) as expected. There were, however, two general exceptions. First, EB-ANOVA outperformed BAYESRATIO whenever $\alpha_\tau = 3$ and $\alpha_e = 12$. Secondly,

whenever $\alpha_\tau = 30$, BAYESRATIO was generally slightly inferior for MAD to E-LIMMA for $m = 2$ but then slightly superior to E-LIMMA for $m = 4$.

With, again the exception of EB-REML, all shrinkage based procedures vastly outperformed the gene-specific methods (LIMMA, ANOVA and REML) for low residual and variance ratio heteroskedasticity ($\alpha_\tau = 30$ and $\alpha_e = 12$). Nevertheless, LIMMA generally outperformed REML and ANOVA and somewhat approached the shrinkage based methods for high and variance ratio heteroskedasticity ($\alpha_\tau = 3$ and $\alpha_e = 3$). REML always slightly outperforms ANOVA but not distinguishable in the plots for estimating MSB_g as was shown in our previous study (Chapter 1).

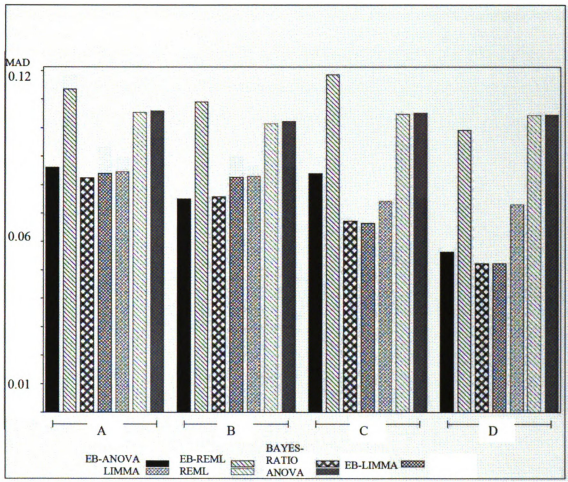


Figure 3.1 MAD (Mean Absolute Deviation) of MSBg from their true values was plotted for seven methods respectively for two replicate spots per gene within slides(mean correlation coefficient =0.6): A). $\alpha_\tau = 3, \alpha_{residuals} = 3$; B). $\alpha_\tau = 3, \alpha_{residuals} = 12$; C). $\alpha_\tau = 30, \alpha_{residuals} = 3$; D). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

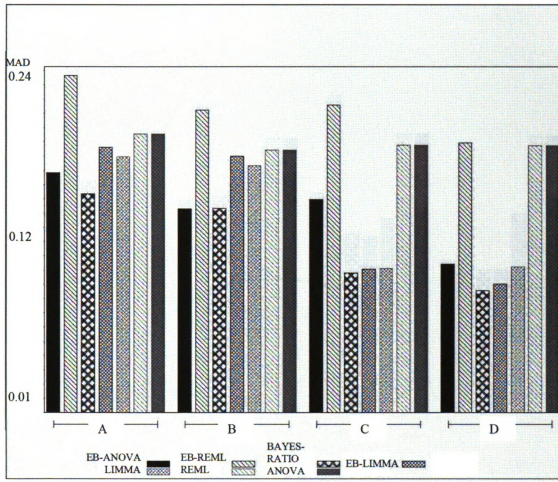


Figure 3.2 MAD (Mean Absolute Deviation) of MSBg from their true values was plotted for seven methods respectively for four replicate spots per gene within slides(mean correlation coefficient =0.6): A). $\alpha_\tau = 3, \alpha_{residuals} = 3$; B). $\alpha_\tau = 3, \alpha_{residuals} = 12$; C). $\alpha_\tau = 30, \alpha_{residuals} = 3$; D). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

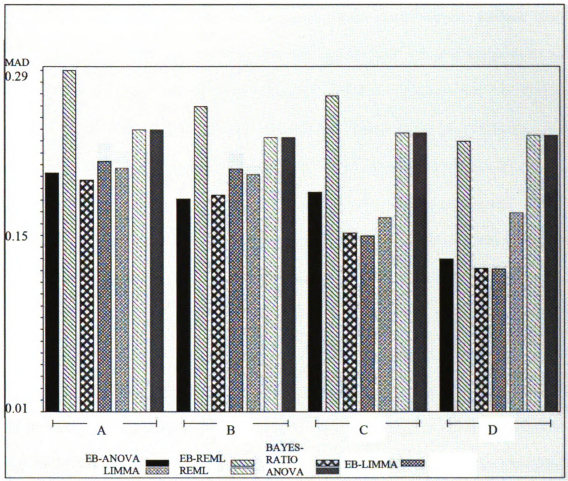


Figure 3.3 MAD (Mean Absolute Deviation) of MSBg from their true values was plotted for seven methods respectively for two replicate spots per gene within slides(mean correlation coefficient =0.9): A). $\alpha_\tau = 3, \alpha_{residuals} = 3$; B). $\alpha_\tau = 3, \alpha_{residuals} = 12$; C). $\alpha_\tau = 30, \alpha_{residuals} = 3$; D). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

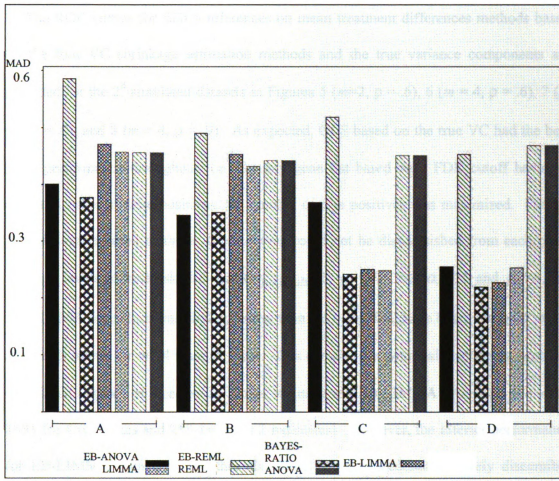


Figure 3.4 MAD (Mean Absolute Deviation) of MSBg from their true values was plotted for seven methods respectively for four replicate spots per gene within slides(mean correlation coefficient =0.9): A). $\alpha_\tau = 3, \alpha_{residuals} = 3$; B). $\alpha_\tau = 3, \alpha_{residuals} = 12$; C). $\alpha_\tau = 30, \alpha_{residuals} = 3$; D). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

The ROC curves for EGLS inferences on mean treatment differences methods based on the four VC shrinkage estimation methods and the true variance components are provided for the 2^4 simulated datasets in Figures 5 (m=2, ρ = .6), 6 (m = 4, ρ = .6), 7 (m =2, ρ = .9), and 8 (m = 4, ρ = .9). As expected, GLS based on the true VC had the best ROC performance throughout; i.e., for any gene list based on a FDR cutoff having a certain number of false positives, the number of true positives was maximized. For the four shrinkage-based methods, these curves could not be distinguished from each other for high levels of heteroskedasticity ($\alpha_{residuals}$=3, α_τ=3). For α_e=12 and α_τ=3, the EB-LIMMA procedure was slightly worse than the other shrinkage based methods m = 2 and much worse for m =4 when ρ = 0.6. This comparison potentially reflects a problem of inflated degrees of freedom for larger m using the EB-LIMMA procedure (i.e. 4*8-1=31 for 4 replicates and 2*8-1=15 for 2 replicates). However, the inferior performance for EB-LIMMA compared to other shrinkage based procedures is barely discernible when ρ = 0.9 (Figures 7 and 8); in other words, regardless of the value of α_τ, for the upper bound of value 1 on ρ creates even less variability in ρ when ρ = 0.9 than when ρ = 0.6 (Figure 5 and 6). For α_τ = 30, the EB-LIMMA and BAYESRATIO methods have similar performance and were somewhat superior to EB-ANOVA and EB-REML, particularly when α_e=3. With the lower level of residual heteroskedasticity (α_e=12), EB-ANOVA had similar ROC performance as EB-LIMMA and BAYESRATIO and was superior to EB-REML. EB- LIMMA was also expected to outperform EB-ANOVA and EB-REML for low levels of variance ratio heteroskedasticity (α_τ = 30).

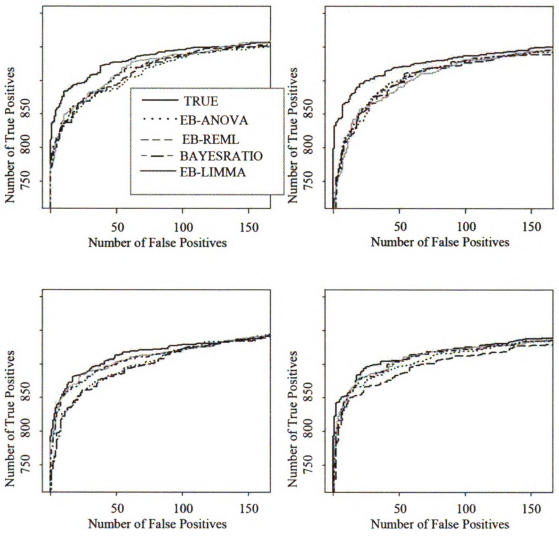


Figure 3.5 Number of true positives vs. Number of false positives obtained by different Empirical Bayes and full Bayes gene selection methods for two replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

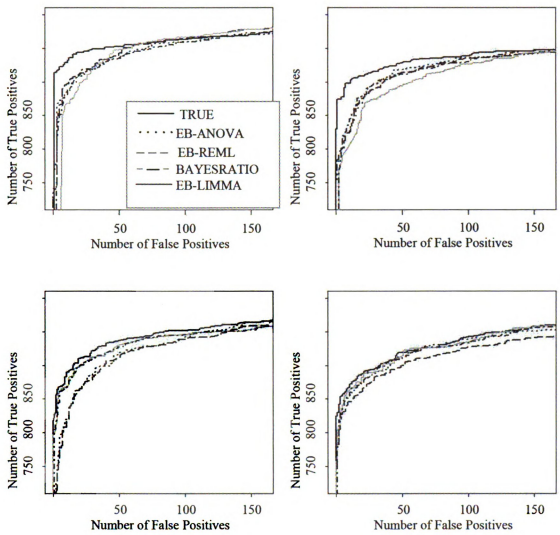


Figure 3.6 Number of true positives vs. Number of false positives obtained by different Empirical Bayes and full Bayes gene selection methods for four replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

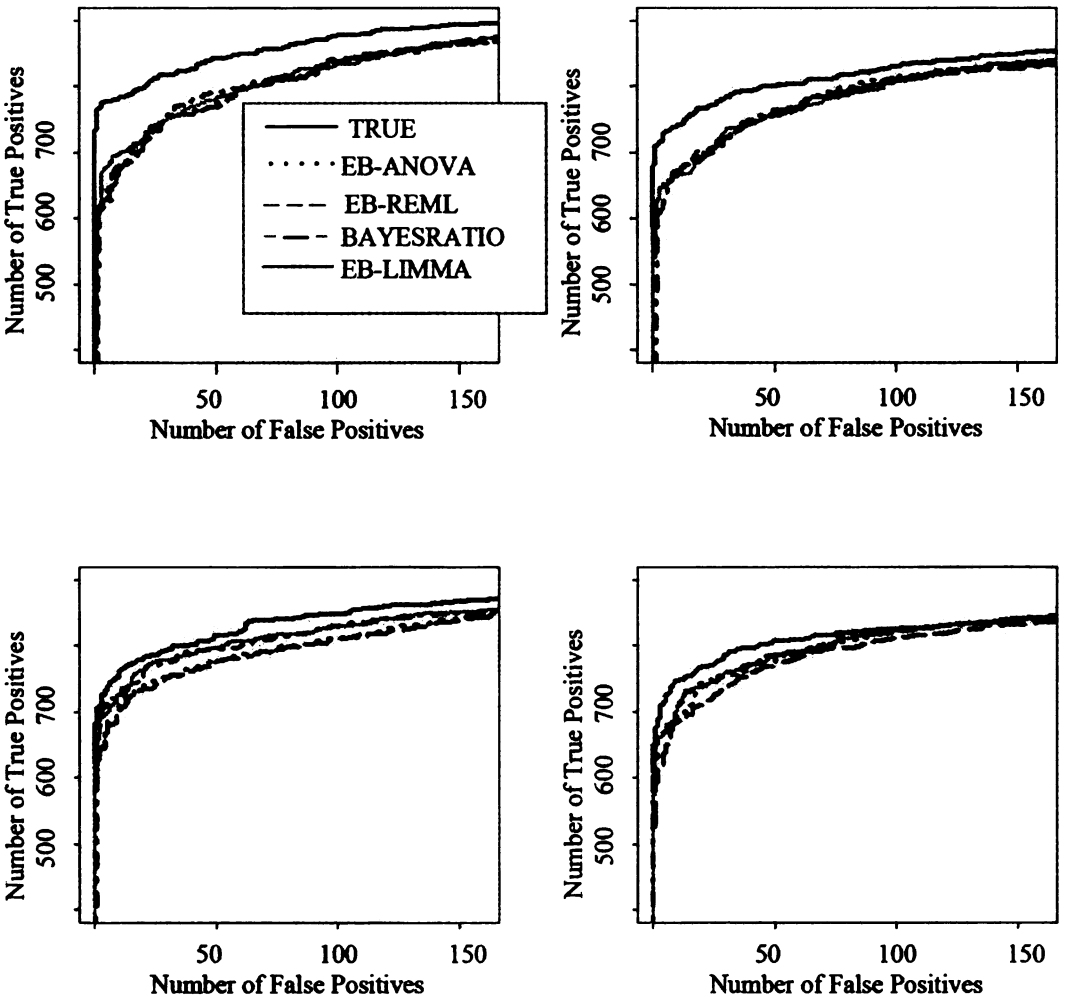


Figure 3.7 Number of true positives vs. Number of false positives obtained by different Empirical Bayes and full Bayes gene selection methods for two replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

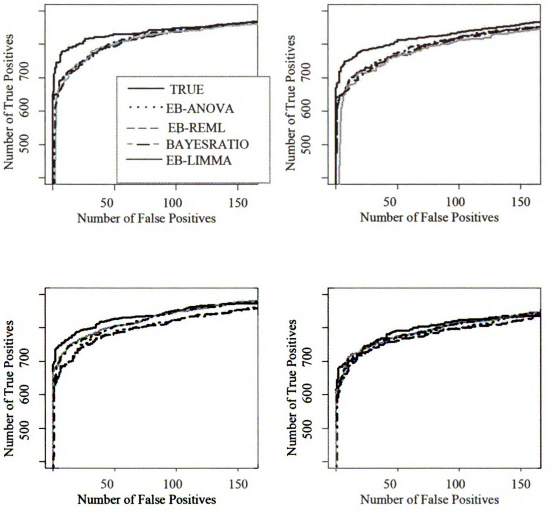


Figure 3.8 Number of true positives vs. Number of false positives obtained by different Empirical Bayes and full Bayes gene selection methods for four replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

The ROC curves for EGLS inferences on mean treatment differences methods based on the three gene-specific estimation methods (ANOVA,REML, and LIMMA) and the true variance components are provided for the 2^4 simulated datasets in Figures 9 (m=2, ρ = .6), 10 (m = 4, ρ = .6), 11 (m =2, ρ = .9), and 12 (m = 4, ρ = .9). In virtually all situations, REML and ANOVA gene specific methods had similar performance in terms of ROC curves. This result was anticipated since precision in estimating the denominators of F-test was shown to be very close for these two methods in MAD comparisons in Figures 1 - 4. These REML method was inferior to LIMMA methods as shown in Smyth et al., (2005).

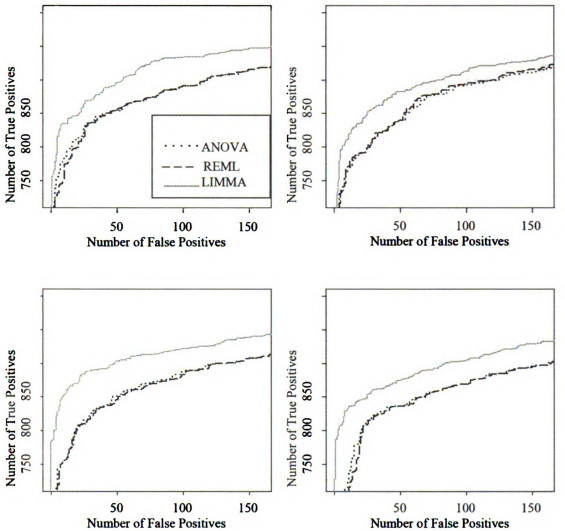


Figure 3.9 Number of true positives vs. Number of false positives obtained by different gene-specific selection methods for two replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

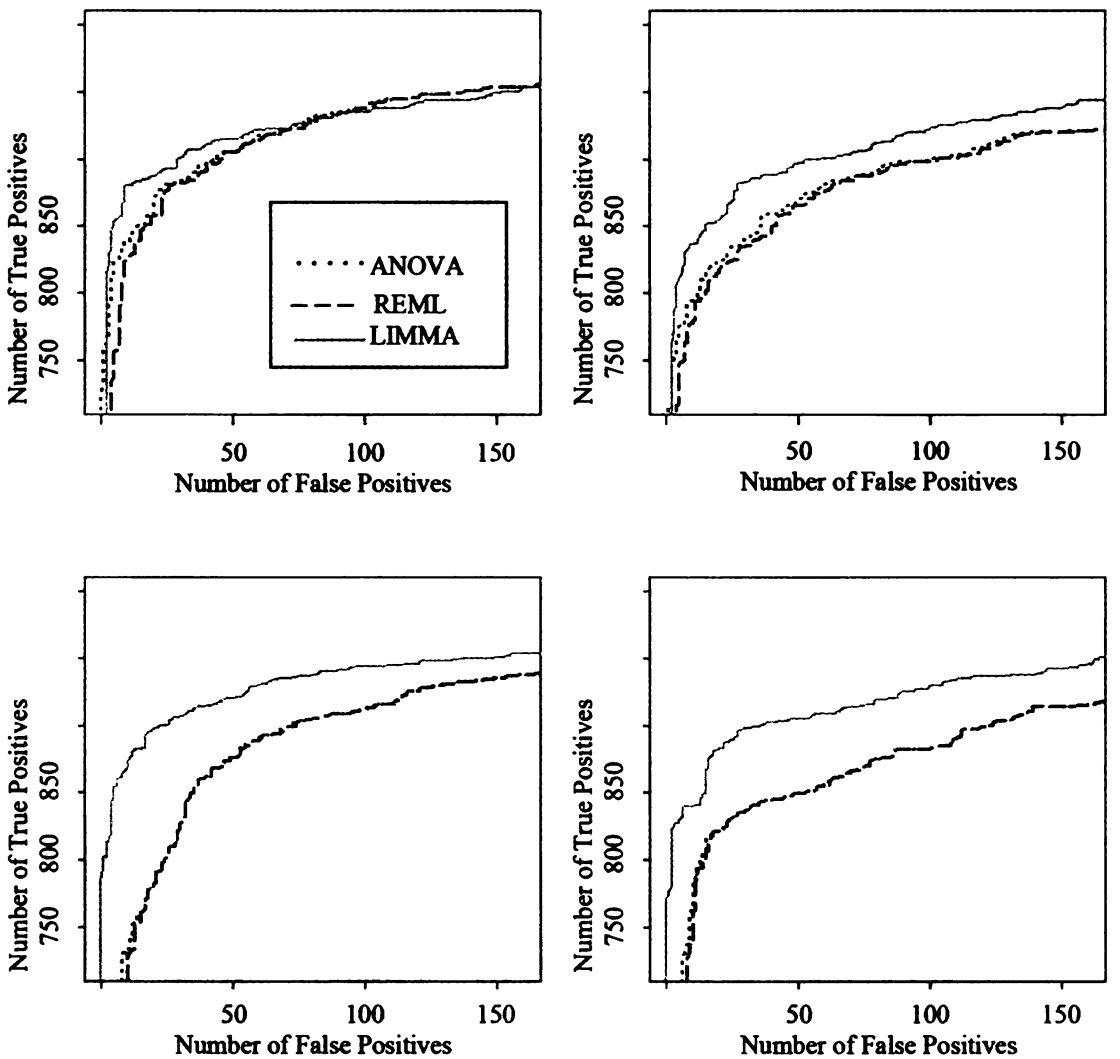


Figure 3.10 Number of true positives vs. Number of false positives obtained by different gene-specific selection methods for four replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

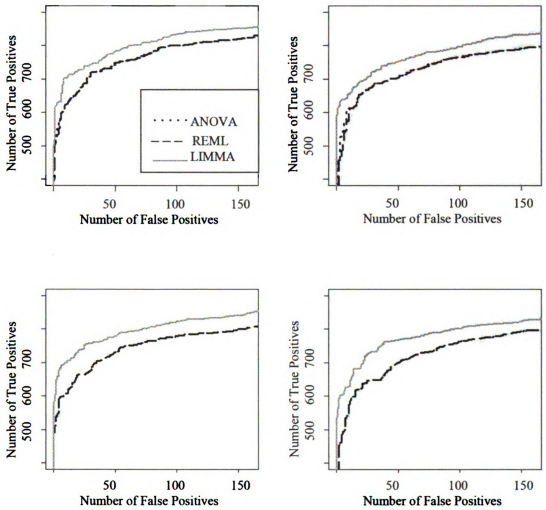


Figure 3.11 Number of true positives vs. Number of false positives obtained by different gene-specific selection methods for two replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

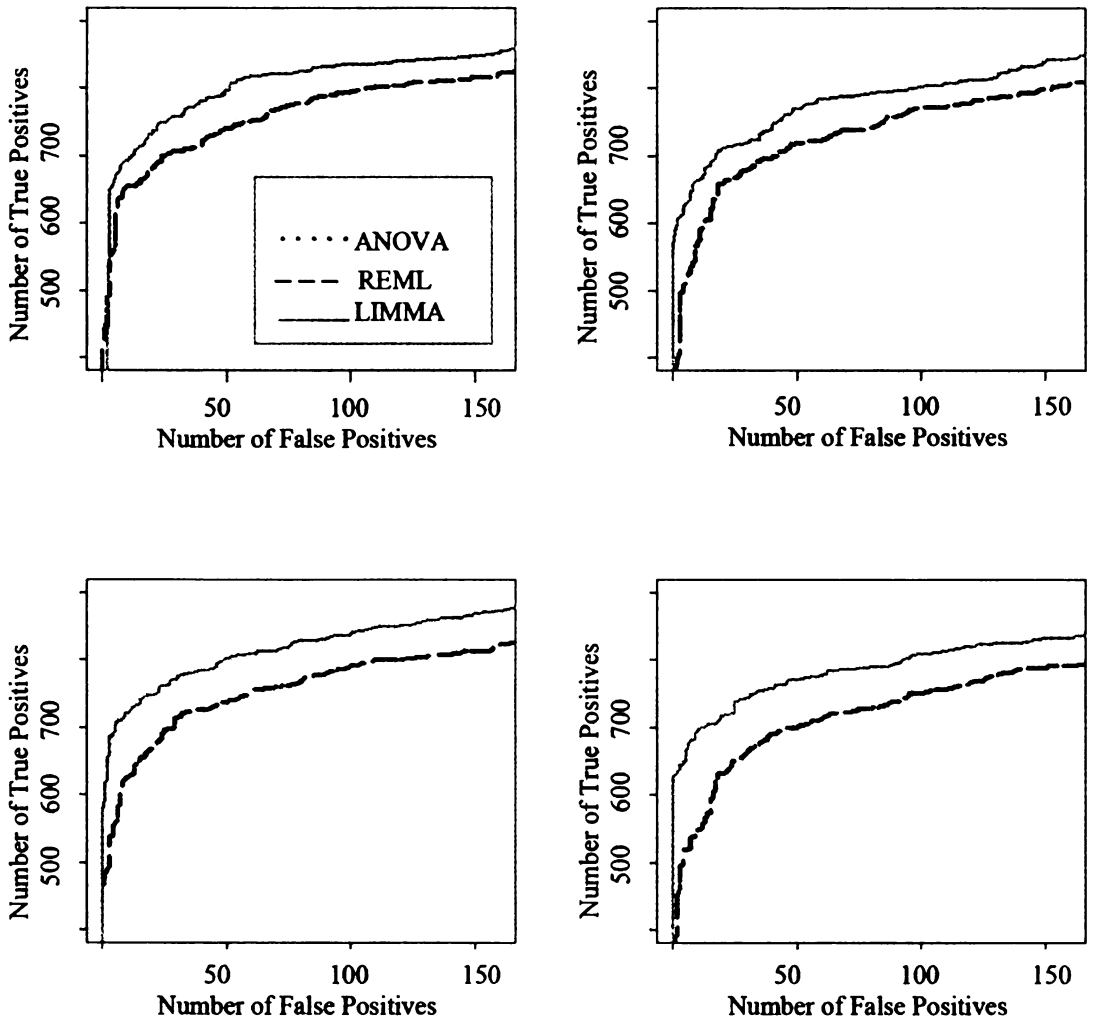


Figure 3.12 Number of true positives vs. Number of false positives obtained by different gene-specific selection methods for four replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

The TFDR (true false discovery rate) versus EFDR (estimated false discovery rate) for GLS and EGLS based on all shrinkage based methods are provided for the 2^4 simulated datasets in Figures 13 (m=2, ρ = .6), 14 (m = 4, ρ = .6), 15 (m =2, ρ = .9), and 16 (m = 4, ρ = .9). As expected, TFDR≈ EFDR for GLS as the test statistics for treatment effects are exact z-tests; a similar performance could be noted for EB-ANOVA. A general agreement between TFDR and EFDR was also observed for BAYESRATIO for all cases; again this might be anticipated since the model generation process matched that of BAYESRATIO. The EB-REML procedure tended to be slightly too conservative (TFDR<EFDR) for low values of EFDR and far too liberal (TFDR>EFDR) for the high values of EFDR in virtually all cases.

Our particular attention was drawn to the EB-LIMMA procedure. It tended to be too liberal particularly when α_τ=3 and when m = 4 (Figures 14,16). Even when the distributional assumptions for EB-LIMMA method were closely matched (α_τ=30), it was still too liberal and again particularly if m = 4 (Figure 14,16). However, even when m = 2, EB-LIMMA was too liberal for α_τ=3 particularly for ρ = 0.9 (Figure 15). In essence then, microarray experiments characterized by high number of within-array replicates (m), high within-array correlation (ρ) between spots, and high levels of variance ratio heteroskedasticity (low α_τ) should generally incur far more false discoveries when using EB-LIMMA.

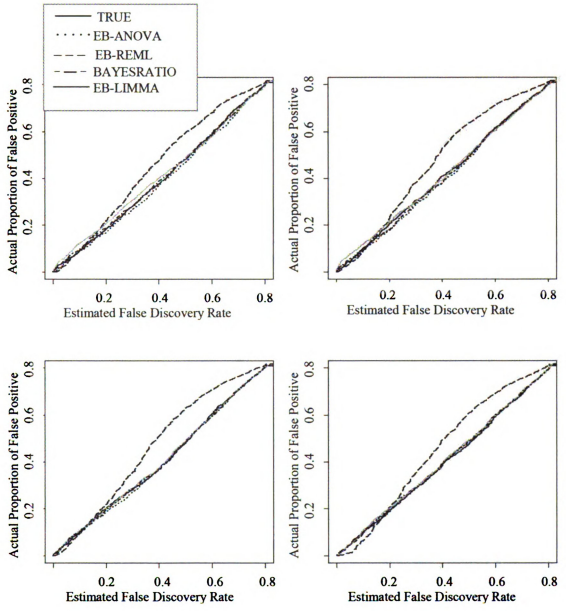


Figure 3.13 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different Empirical Bayes and full Bayes methods for two replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

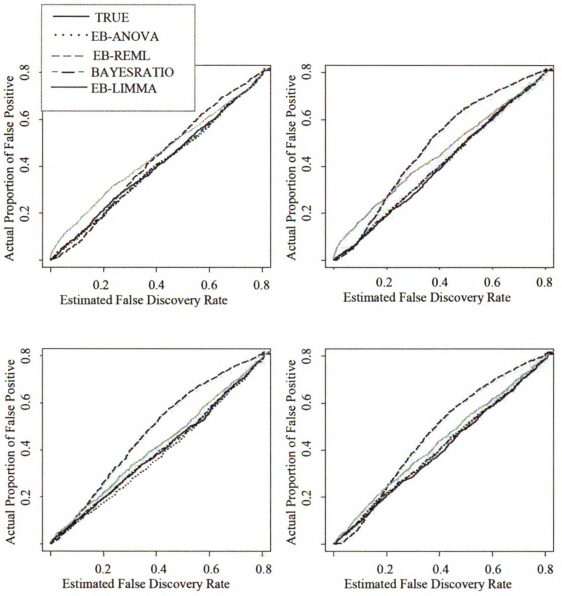


Figure 3.14 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different Empirical Bayes and full Bayes methods for four replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

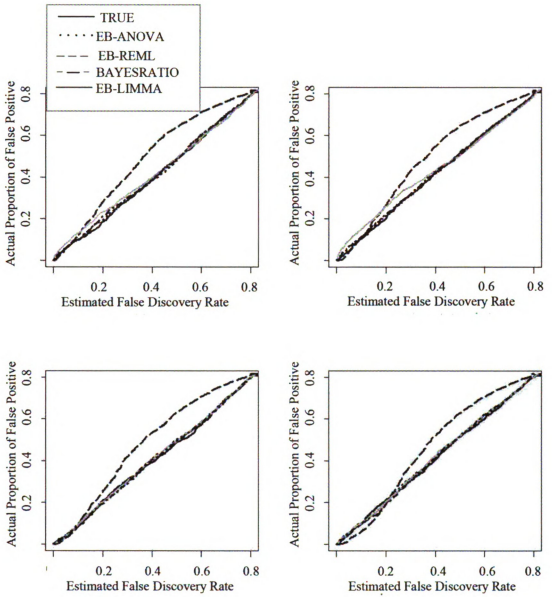


Figure 3.15 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different Empirical Bayes and full Bayes methods for two replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

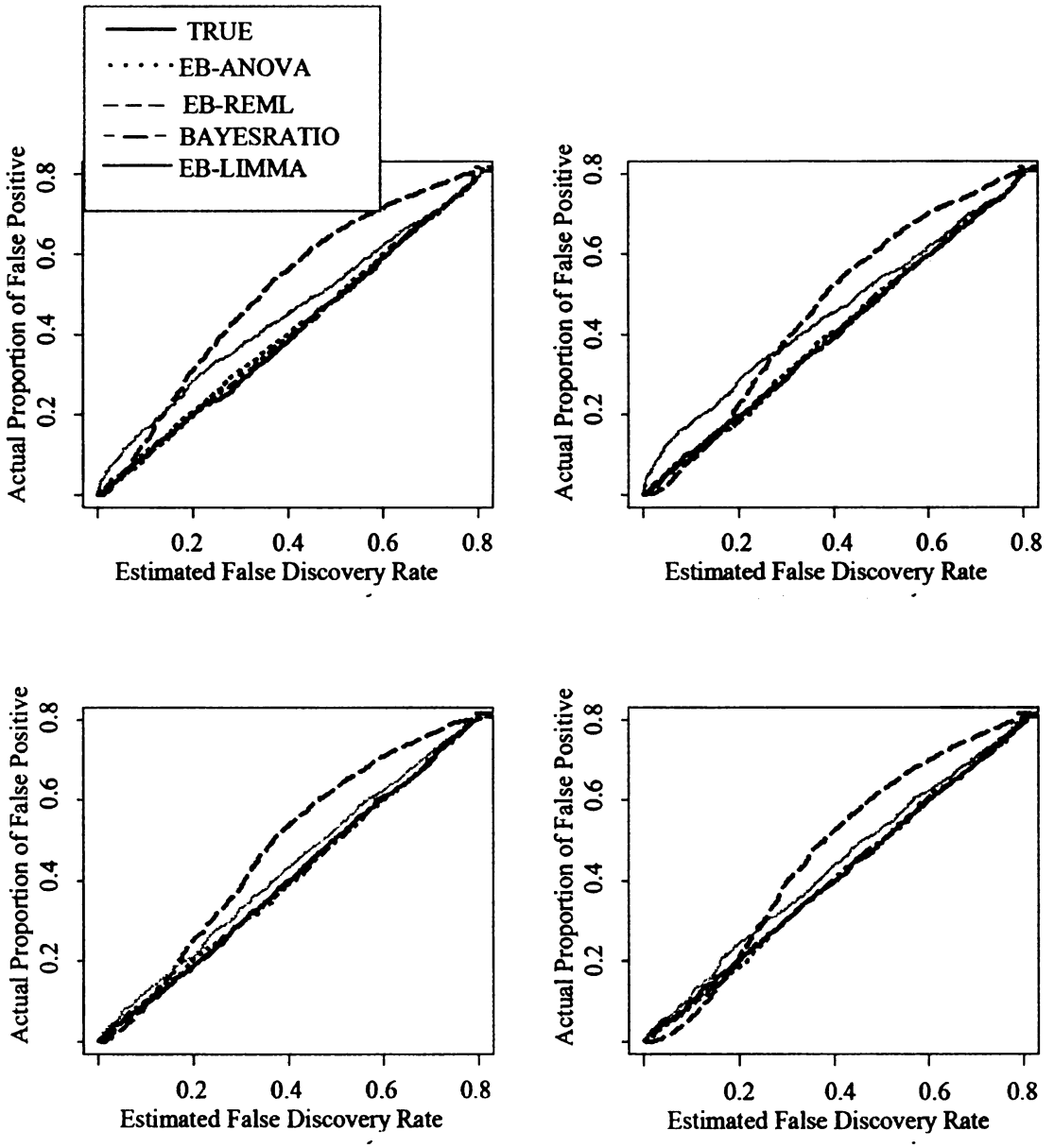


Figure 3.16 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different Empirical Bayes and full Bayes methods for four replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

The TFDR versus EFDR for GLS and EGLS based on all gene-specific methods are provided for the 2^4 simulated datasets in Figures 17 (m=2, ρ = .6), 18 (m = 4, ρ = .6), 19 (m =2, ρ = .9), and 20 (m = 4, ρ = .9). It can be quickly gleaned from these plots that LIMMA is also liberal with FDR assessments, particularly when m = 4 (Figure 18, 20), whereas EGLS based on REML and ANOVA generally demonstrate excellent agreement between TFDR and EFDR. EGLS based on ANOVA had the performance that was expected since the resulting EGLS test statistics are random samples from a F-distribution under the null hypothesis of no differential expression.

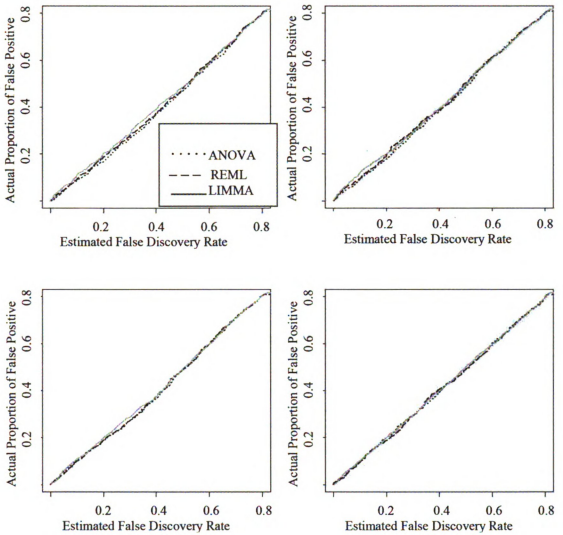


Figure 3.17 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different gene-specific methods for two replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

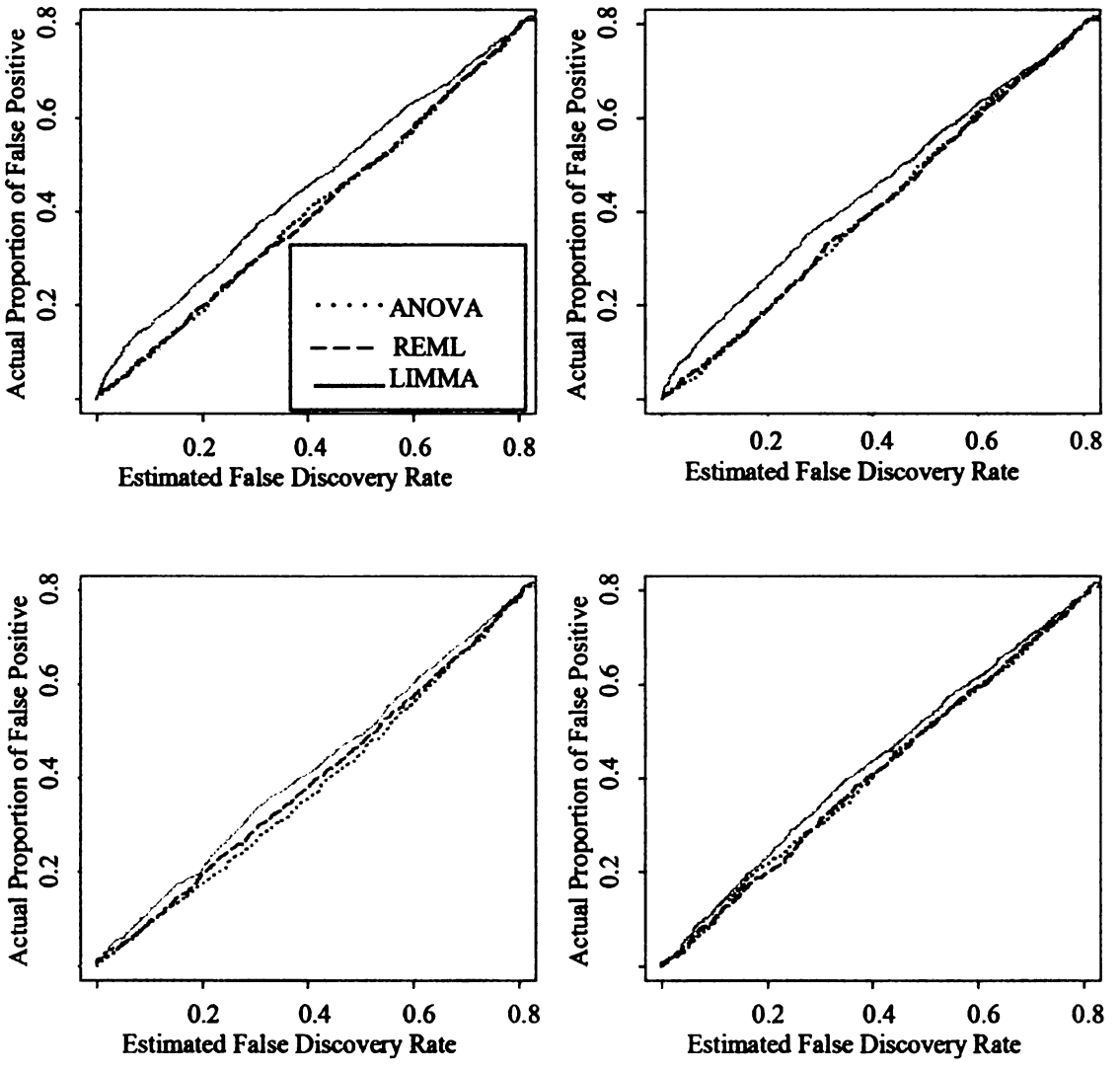


Figure 3.18 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different gene-specific methods for four replicate spots per gene within slides(mean correlation coefficient =0.6): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

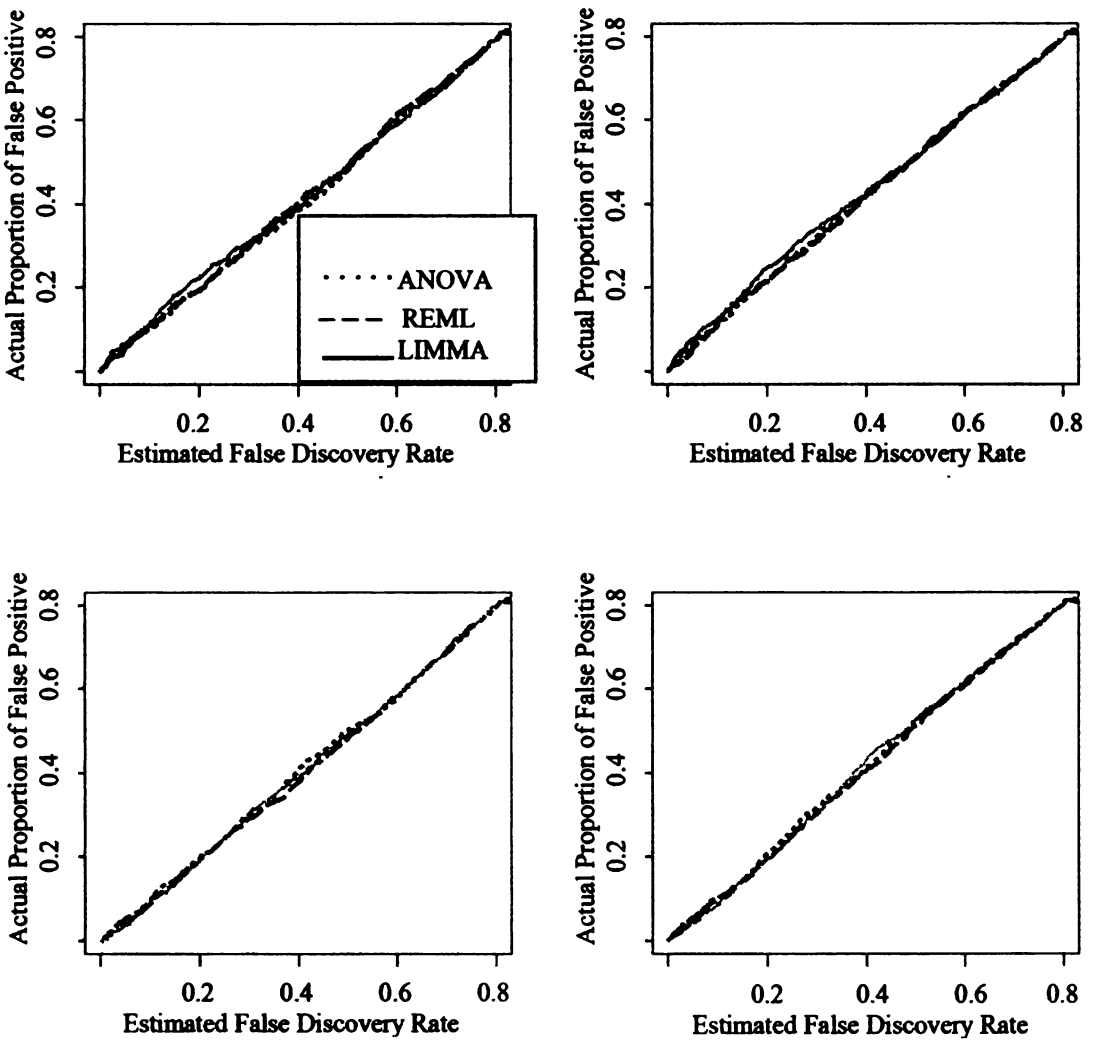


Figure 3.19 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different gene-specific methods for two replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

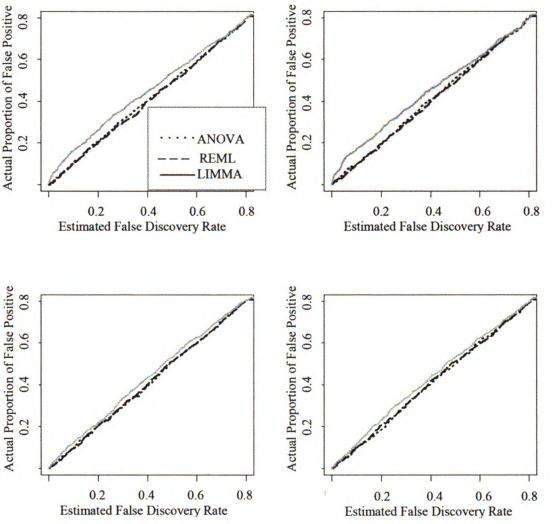


Figure 3.20 Actual Proportion of False Positives vs. Estimated False Discovery Rate for different gene-specific methods for four replicate spots per gene within slides(mean correlation coefficient =0.9): Upper left graph). $\alpha_\tau = 3, \alpha_{residuals} = 3$; Upper right graph). $\alpha_\tau = 3, \alpha_{residuals} = 12$; Lower left graph). $\alpha_\tau = 30, \alpha_{residuals} = 3$; Lower right graph). $\alpha_\tau = 30, \alpha_{residuals} = 12$.

3.6 Data Analysis

We applied all of the seven methods to the data set with within array technical replicates on an existing dataset (Wade et al. 2004; Wade et al. 2005). The dataset was analyzed using the mixed model ANOVA [9] and using the logarithmic female and male lowess normalized intensities as response variables (Yang et al., 2002b).

For the BAYESRATIO model, we estimated the hyperparameters to be $\hat{\alpha}_e = 2.98, \hat{\beta}_e = 0.06$, $\hat{\alpha}_\tau = 2.83$, and $\hat{\beta}_\tau = 3.62$ such that the overall estimated mean residual variance $\widehat{\mathrm{E}(\sigma_e^2)} = \frac{0.06}{2.98-1} = 0.03$ and the overall estimated variance ratio $\widehat{\mathrm{E}(\tau)} = \frac{3.62}{2.83-1} = 1.98$; in other words, a point estimate for the overall estimated within-spot correlation could be determined as $\frac{1.98}{1.98+1} = 0.66$. For EB-LIMMA, we determined an overall correlation estimate of 0.601,estimates of $s_o^2 = 0.051$, thereby defining the estimated harmonic mean of σ_e^2, and $d_0 = 5.74$ which is equivalent to $2\hat{\alpha}_e$. Hence these results are comparable to those attained from BAYESRATIO with the important exception that EB-LIMMA presumes that the estimated correlation between spots within an array to be the same for all genes in the microarray experiment.

For EB-ANOVA, we estimated the hyperparameters to be $\hat{\alpha}_e = 2.35$, $\hat{\beta}_e = 0.042$, $\hat{\alpha}_u = 2.54$, and $\hat{\beta}_u = 0.20$ such that the overall estimated mean residual variance $\widehat{\mathrm{E}(\sigma_e^2)} = \frac{0.042}{2.35-1} =$ 0.031 and the overall estimated expected mean square for array

$\widehat{E(MSB)} = \frac{0.20}{2.54-1} =$ 0.13 being an estimate of $\sigma_e^2 + 2\sigma_u^2$ such that the overall estimate $\widehat{E(\sigma_u^2)} = \frac{0.13-0.031}{2} = 0.05$, thereby translating into an overall correlation estimate of $\frac{0.05}{0.05+0.031} = 0.62$ again leading to inferences on dispersion parameters similar to those for BAYESRATIO. For EB-REML, we estimated the hyperparameters to be $\hat{\alpha}_e = 3.5$, $\hat{\beta}_e = 0.038$, $\hat{\alpha}_u = 2.5$, and $\hat{\beta}_u = 0.10$ such that the overall estimated mean residual variance $\widehat{E(\sigma_e^2)} = \frac{0.038}{3.5-1} =$ 0.015 and the overall estimated expected mean square for array $\widehat{E(MSB)} = \frac{0.10}{2.5-1} =$ 0.068 being an estimate of $\sigma_e^2 + 2\sigma_u^2$ such that the overall estimate $\widehat{E(\sigma_u^2)} = \frac{.068-0.015}{2} = 0.026$, thereby translating into an overall correlation estimate of $\frac{0.026}{0.026+0.015} = 0.63$. $\widehat{E(\sigma_u^2)}$ and $\widehat{E(\sigma_e^2)}$ are 0.046 and 0.031 for ANOVA, 0.047 and 0.030 for REML. Overall correlation would be 0.60 for ANOVA and 0.61 for REML, which correspond the similar dispersion parameter estimates as those for BAYESRATIO.

Plots relating the ANOVA F-test p-value and the corresponding q-values for differential expression versus the number of genes selected for that particular p-value and q-value cutoff are presented for each of the seven methods in Figure 21. It is instructive to note that p-values and q-values did not appear to depend upon shrinkage estimation whatsoever for this dataset. BAYESRATIO tended to detect more genes than EB-LIMMA and the other five methods in p-value plot. Recognizing that most current

experiments would depend upon q-value determinations, the corresponding q-values for differential expression versus the number of genes selected for the particular q-value threshold was shown to have the same pattern as p-value plot but more separate in close-up plot in Figure 21. For example, BAYESRATIO picked up more genes than EB-LIMMA and LIMMA which in turn declared more genes significant than EB-ANOVA. EB-REML, ANOVA and REML were not able to declare any significantly expressed genes for small q-value thresholds.

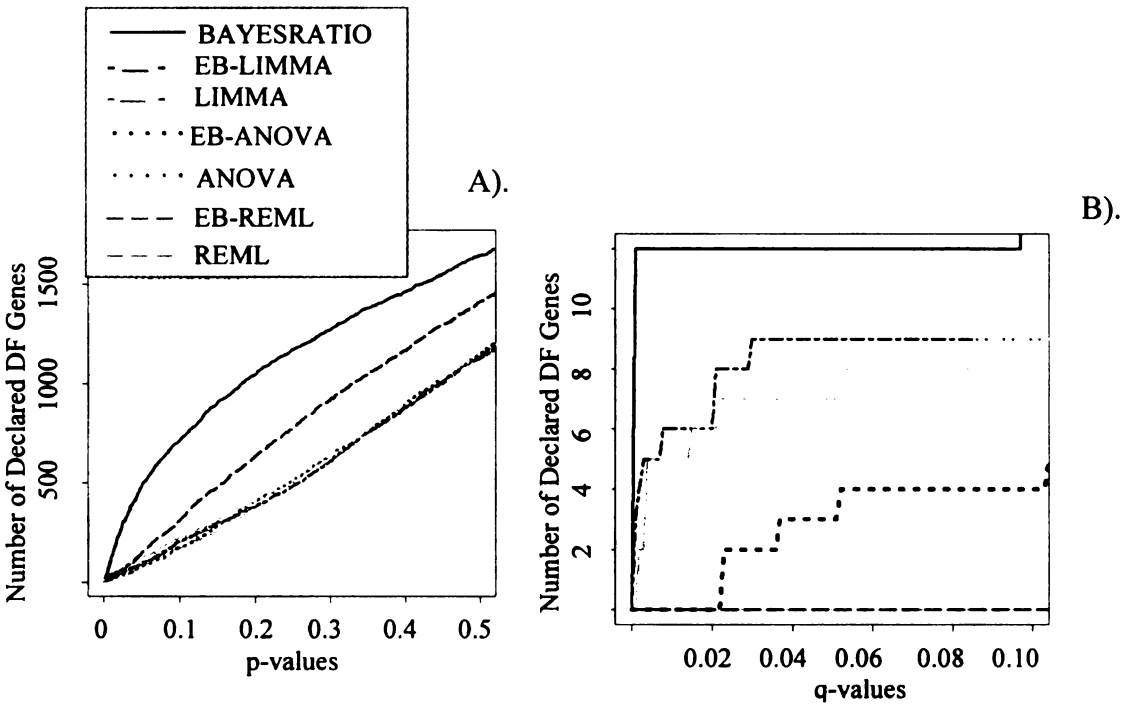


Figure 3.21 Wade data results: A). Number of declared differentially expressed genes (DF) vs. critical p-values, B). Number of declared differentially expressed genes vs. critical q-values.

3.7 Discussion and Concluding Remarks

"Many assumptions that have been made for modeling microarray data have yet to be verified. Hopefully evidence either for or against these assumptions will emerge ..." (Storey et al. 2004). Commonly used microarray data analysis software LIMMA accommodates the analysis of designs that involve technical replicates. The assumption for LIMMA software is that the within array replicate correlation coefficients for each gene are constant across all genes. Using this assumption, the experimental degrees of freedom associated with the test statistics in LIMMA software is more than doubled compared with a regular linear mixed model approach. For example, the experimental degrees of freedom for inference on treatment effects would be specified to be nearly equal for experiments with only two biological replicates per treatment but with eight technical (i.e. spots) replicates per biological replicate compared to another design with sixteen biological replicates per treatment and one technical replicate per biological replicate using the LIMMA procedure. In this paper, we introduce a BAYESRATIO model for generalizing the common correlation assumption in LIMMA. Our motivation was stimulated by the data set we used, which was checked against the common correlation assumption in LIMMA. This case with within-array replicates was found to have high level of heterogeneity for variance ratios, which means that the genewise correlation estimates were too variable across genes to be compatible with a common true correlation. High levels of heteroskedasticity, as shown in our real data, and various levels of heteroskedasticity in the combination of two components variance ratios and residual variances, the magnitude of the correlation coefficient, and the number of technical replicates per gene affected statistical inference. It was shown that the violation

of common correlation assumption results in a poorly controlled false discovery rate and becomes worse with the increase of technical replicates per gene within slides by using the LIMMA procedure. The causes for poorer performance seem to be due to: 1). The denominator degrees of freedom are exaggerated as discussed before. 2). The denominators of F tests for treatment effects are not comparably estimated by seven methods. EB method is compromised as implemented in LIMMA because it forces the variance ratios to equalize across genes.

The BAYESRATIO model we suggested in the paper facilitates describing complicated microarray experimental data with various degrees of heteroskedasticity between genes. It extends the common correlation assumption for the LIMMA procedure to more general cases. We compare the performance of seven methods, which include BAYESRATIO method, EB-LIMMA, EB-ANOVA, EB-REML, LIMMA, ANOVA and REML gene-specific models. To do this we simulate data based on real data structure and scenarios as may occur in real microarray experiments. Some conclusions can be drawn on the basis of results from the representative simulation studies:

1). BAYESRATIO method consistently is superior or at least similar to three other comparative methods: EB-LIMMA, EB-ANOVA and EB-REML. The lowest or slightly second to the lowest MAD for the denominator of F test points to one of the most accurate controls of FDR and the best ROC curves of all the simulation scenarios. The weak assumption of BAYESRATIO method makes the model robust enough to use in any real-world microarray experiment setting with both within-array or between-array technical and biological replicates. The Bayesian approach has several features that make

it advantageous for the analysis of microarray data. These include the incorporation of prior information, flexible exploration of arbitrarily complex hypotheses, easy inclusion of nuisance parameters, and relatively well developed methods for handling missing data (Yang et al. 2004). We are not too worried about its computational complexity because our BAYESRATIO model was implemented in R software in the computer with Pentium (R) 2.4GHz AND 1.00 GB of RAM and took less than 10 hours for 100,000 iterations for 6000 genes in one simulation study.

2). EB-ANOVA we suggest in the first chapter performs reasonably well for any mixed model analyses of microarray experiment design with technical as well as biological replication from the view point of robust-efficiency. The data sets are not simulated on the basis of EMS components (expected mean square components in traditional ANOVA table), which is basic assumption for the model using EB-ANOVA. The results from simulation study support the conclusion of the first chapter, which is that EB-ANOVA performs better than EB-REML, ANOVA and REML gene specific models in terms of precisely estimating variance components and correctly detecting the largest number of differentially expressed genes when controlling for the false discovery rate. For the data sets simulations favoring the BAYESRATIO methods, in which the variance ratio follows inverse gamma distribution, EB-ANOVA performs reasonably well comparing to BAYESRATIO method in terms of ROC curves and controlling FDR for most cases. Nevertheless, BAYESRATIO outperforms EB-ANOVA for the cases with high heterogeneous residual variances ($\alpha_{residuals}$=3) and constant variance ratios (α_{τ}=30) across genes.

The BAYESRATIO model referred to in this paper and the new function duplicateCorrelation in the LIMMA package can't deal with situations which include technical replicates in both within-array and between-array simultaneously. The EB-ANOVA has more flexibility for all kinds of experimental designs because of its shrinkage procedure for EMS components with mixed model approaches. "Essentially, all models are wrong, but some are useful" (Box and Draper, 1987). Since no model can fit all microarray data, the verification of the model assumptions should be the important step to the choice of a good model and then make our models more useful.

BIBLIOGRAPHY

ALBERT, J. (1999). Criticism of a hierarchical Bayes model using Bayes factors. *Statistics in Medicine* **18**, 287-305.

BALDI, P. & LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**(6), 509-519.

BREITLING, R. ARMENGAUD, P. AMTMANN, A. & HERZYK, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *Febs Letters* **573**(1-3), 83-92.

CHIB, S. & GREENBERG, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician* **49**, 327-335.

CHURCHILL, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490-495.

DE SMET, F. MOREAU, Y. ENGELEN, K. TIMMERMAN, D. VERGOTE, I. & DE MOOR, B. (2004). Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer* **91**(6), 1160-1165.

FENG, S. WOLFINGER, R. D. CHU, T. M. GIBSON, G. C. & MCGRAW, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural Biological and Environmental Statistics* **11**(2), 197-209.

GELMAN, A. CARLIN, J. B. STERN, H. S. & RUBIN, D. S. (1995). *Bayesian Data Analysis*. London, UK: Chapman and Hall.

HENDERSON, C. R. (1984). *Applications of linear models in animal breeding*. Guelph, CANADA: University of Guelph.

KERR, M. K. AFSHARI, C. A. BENNETT, L. BUSHEL, P. MARTINEZ, J. WALKER, N. J. & CHURCHILL, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**(1), 203-217.

LONNSTEDT, I. & SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**(1), 31-46.

NEWTON, M. A. KENDZIORSKI, C. M. RICHMOND, C. S. BLATTNER, F. R. & TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37-52.

SMYTH, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1).

SMYTH, G. K. MICHAUD, J. & SCOTT, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**(9), 2067-2075.

STOREY, J. D. TAYLOR, J. E. & SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of The Royal Statistical Society Series B-Statistical Methodology* **66**, 187-205.

TSENG, G. C. OH, M. K. ROHLIN, L. LIAO, J. C. & WONG, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**(12), 2549-2557.

VINCIOTTI, V. KHANIN, R. D'ALIMONTE, D. LIU, X. CATTINI, N. HOTCHKISS, G. BUCCA, G. DE JESUS, O. RASAIYAAH, J. SMITH, C. KELLAM, P. & WIT, E. (2005). An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* **21**, 492-501.

WADE, J. PEABODY, C. COUSSENS, P. TEMPELMAN, R. J. CLAYTON, D. F. LIU, L. ARNOLD, A. P. & AGATE, R. (2004). A cDNA microarray from the telencephalon of juvenile male and female zebra finches. *Journal of Neuroscience Methods* **138**(1-2), 199-206.

WADE, J. PEABODY, C. COUSSENS, P. TEMPELMAN, R. J. CLAYTON, D. F. LIU, L. ARNOLD, A. P. & AGATE, R. (2005). A cDNA microarray from the telencephalon of juvenile male and female zebra finches (vol 138, pg 199, 2004). *Journal of Neuroscience Methods* **142**(2), 327-327.

WANG, C. S. RUTLEDGE, J. J. & GIANOLA, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics, Selection, Evolution* **25**, 41-62.

WANG, C. S. RUTLEDGE, J. J. & GIANOLA, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics, Selection, Evolution* **26**, 91-115.

WOLFINGER, R. D. GIBSON, G. WOLFINGER, E. D. BENNETT, L. HAMADEH, H. BUSHEL, P. AFSHARI, C. & PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625-637.

WRIGHT, G. W. & SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**(18), 2448-2455.

YANG, D. ZAKHARKIN SO., P., GP. BRAND, J. EDWARDS, J. & BARTOLUCCI, A., ALLISON (2004). Applications of Bayesian Statistical Methods in Microarray Data Analysis. *American Journal of PharmacoGenomics* **4(1)**, 53-62.

Chapter 4: Data Comparison for Two-Channel Microarray Image Analysis Methods

Abstract

Image analysis is a key component of microarray experiments. Potentially it has large impact on subsequent data analysis such as identifying differentially expressed genes. Segmentation of the microarray images as foreground and background pixels is an important step influencing data precision. Little consideration has been given to the study of the data features and the statistical modeling to identify differentially expressed genes resulting from three commonly used segmentation methods (adaptive circle, adaptive shape and histogram methods). In this paper, we use four image analysis software programs (Genepix, MolecularWare, Spot and Imagene) representing the three segmentation methods to investigate the variability of data derived from each method. This impacts subsequent data analysis resulting in different numbers of differentially expressed genes. The histogram method (Imagene) gives significantly higher variability across replicate spots compared to other methods. The adaptive shape method (Spot) and the adaptive circle method (Genepix and MolecularWare) share similar data features. Our EB-ANOVA (Chapter 2) is beneficial to the analysis of data generated from all four image software programs combined with different preprocessing methods such as Lowess and Arsinh normalization in identifying differentially expressed genes compared to the gene-specific model.

Keywords: Microarray, Image analysis, Segmentation method, Adaptive circle, Adaptive shape, Histogram, Genepix, MolecularWare, Spot, Imagene, Lowess, Arsinh, EB-ANOVA

4.1 Introduction

Gene expression profiling using microarrays is considered an important tool and powerful technology allowing researchers to study interactions among thousands of genes simultaneously. The cDNA microarray technology is based on an approach where cDNA clone inserts are robotically printed onto a glass slide and subsequently hybridized to two differentially fluorescently labeled probes. The probes are pools of cDNAs which are generated after isolating mRNA from cells or tissues in two states that one wishes to compare (Gilber 2006). Usually, samples from two sources are labeled with different fluorescent dyes (Cy3 and Cy5). The end product of a comparative hybridization microarray experiment is a scanned array image, where the relative intensities between dyes on each spot refer to an indirect measurement of the relative gene expression for further analysis. One of the major challenges of this approach is the image process step. The purpose of this step is to extract information which includes foreground and background intensity estimates and quality measures. The accuracy of image processing has substantial impact on subsequent analyses such as clustering or the identification of differentially expressed genes (Yang et al. 2002a).

The process of image analysis can be categorized into three steps: 1) Gridding: assigning coordinates to individual spots. A precise and automatic microarray gridding method can eliminate the need for human intervention and correct the potential alignment

and rotation problems (Wang 2005); 2) Segmetation: classify pixels either as foreground corresponding to the intensity of interest due to the specific hybridization of the DNA samples or as background; and 3) Quantification: extract intensity information for each spot, which includes calculating foreground fluorescence intensity pairs (Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green)), background intensities and some quality measures (Yang et al. 2002a). Performing the first two steps reliably and accurately results in precise quantification for the subsequent analysis. Most image software provide both manual and automatic gridding procedures which are very diverse and hard to make fair comparisons. Segmentation is supposed to be the most important step for the processing of the microarray images (Istepanian 2003). The methods of summarizing individual pixel data by segmentation could have major effects on the precision of the data (Ahmed et al. 2004).

The variability arising from image analysis can preclude the yield of meaningful biological information. It is therefore important to understand and reduce the noise introduced from different image processing methods and develop the corresponding data analysis scheme. The segmentation method is supposed to be predominant step among the three sequential steps (gridding, segmentation, quantification) in the processing of the microarray images (Istepanian 2003). Several studies have been published to compare different segmentation methods (Yang et al. 2002a; Ahmed et al. 2004; Korn et al. 2004; Qin et al. 2005). The precision of the ratio measurement from different segmentation methods represented by different image analysis software was compared in terms of spot to spot variability (Yang et al. 2002a; Ahmed et al. 2004), the correlation coefficient (Jenssen et al. 2002; Ahmed et al. 2004), the repeatability coefficient (Jenssen et al. 2002;

Ahmed et al. 2004) and the intra-class correlation coefficient for replicates (Korn et al. 2004). Yang et al. (2002a) discussed the advantages and disadvantages of the different segmentation methods and developed Spot image software based on the adaptive shape segmentation algorithm. That paper also declared that the choice of background correction method has a larger impact on the ratio measurement than the segmentation method. Jessen et al. (2002) introduced the repeatability coefficient as an indicator of internal quality of a single microarray experiment. The reason of using the repeatability coefficient instead of correlation to assess the agreement between two methods is well described in Bland & Altman (1999). Korn et al. (2004) described an objective means of comparing different microarray image analysis systems with a comparison of cDNA microarray data generated by histogram segmentation method in UCSF Spot (University of California, San Francisco (UCSF), San Francisco, CA, USA; http://jainlab.ucsf.edu/) and adaptive circle segmentation method in GenePix (Axon Instruments, Union City, CA, USA). The intra-class correlation is used as one indicator to make comparisons. The results in the paper showed that the adaptive circle segmentation method in Genepix performed slightly better than the histogram segmentation method in UCSF Spot on average. Ahmed et al. (2004) investigated the effect of different segmentation methods on the variability of data and their results in different numbers of the declared differentially expressed genes by comparing the three segmentation methods (adaptive, fixed circle and histogram). The finding that the histogram method gave the lowest variability across replicate spots compared to other methods in Ahmed et al. (2004) is controversial given the results in Korn et al. (2004), which showed that adaptive circle segmentation methods performed better than histogram segmentation methods. However, these studies do not

offer a clear choice of segmentation method in connecting with different statistical modeling to investigate how the image analysis influences the data precision and its ultimate impact on the identification of differentially expressed genes.

We address here a comparison of three segmentation methods (adaptive circle, adaptive shape and histogram method) using a series of slides with duplicate spots. We utilize the methods described in the studies by Yang et al. (2002a), Ahmed et al. (2004), Korn el al. (2004) and Jessen et. al (2002 to compare the data precision for the three segmentation methods. Since the Bayesian model is shown to be more reliable by making use of information generated by the whole set of genes in the study, we also consider the heterogeneous variability across the genes introduced from the competing segmentation methods. The subsequent statistical analyses include: 1) Commonly used normalization methods: Lowess (Yang et al. 2002b) and variance stabilizing transformation: Arsinh (Huber et al. 2002); 2) Gene significance analysis: Gene-specific mixed model (Wolfinger et al. 2001) and EB-ANOVA (Chapter 2). All data sets from the different segmentation methods will be analyzed in all combination of these normalization and significance analysis approaches in order to optimize the statistical modeling scheme for the individual segmentation methods. In addition, the issue of background correction is also taken into account. We use both the background subtracted and the non-background subtracted data for all the analyses. The results of background correction will not be shown in the paper for the purpose of brevity, but the conclusion will be drawn in regard of this issue in the discussion section.

The algorithms used by different segmentation methods have been described in detail in Yang et al. (2002a) and Qin et al. (2005). We will review three representative

categories implemented in four software programs: 1) Histogram-based segmentation (e.g. software Imagene (BioDiscovery, Los Angeles, CA, USA). Histogram segmentation uses a "target" mask (a region larger than any spot) and estimates foreground/background intensity for each spot from the pixel values histogram inside the mask. A threshold using the Mann-Whitney test is computed and applied for assigning pixels for foreground and background estimation; 2) Adaptive circle segmentation (e.g. software: Molecularware (Molecularware Inc, Cambridge , MA, USA), Genepix). The shape of each spot is considered as a circle and the center and diameter of the circle is estimated for each spot. Manual adjustment of the diameter of each spot is allowed in these two software programs; and 3) Adaptive shape segmentation (e.g. software: Spot). Two commonly used methods for adaptive segmentation are Seeded Region Growing (SRG) and Watershed techniques, which are implemented in Spot. One of the advantages of using SRG in microarray image processing is that the location of foreground pixels and background pixels can be estimated (Qin et al. 2005). We will choose this option SRG in Spot to generate the data for further study.

In this paper, we show if the choice of segmentation method results in significantly different data features. These findings have direct, practical implications as the variability in precision between the four methods influence the choice of normalization method and model to get accurate numbers of genes identified as differentially expressed. In order to analyze differences between segmentation methods, (independent of other sources of possible noise), we perform comparisons of data from histogram-based, adaptive circle and adaptive shape segmentation using identical digital image files in Tagged Image File Format (TIFF) . Both the Molecularware and the Genepix software programs were used

for the adaptive circle method because we wanted to check if there is any significant difference for the same segmentation method coming from different image analysis software. Having identified different data features between methods, the benefits of normalization method and Empirical Bayes mixed model ANOVA method are discussed.

4.2 Data

To uncover sexually dimorphic gene expression in the developing zebra finch brain, an experiment was developed to compare gene expression between the sexes using RNA from the telencephalon of males and females on the 25th day, a juvenile stage when song memorization in occurring and morphological differentiation of the song circuit is enhanced. Eight slides were hybridized with females and males using dye-swap design. One array was not used due to a quality issue. Fluorescent dye labeled cDNA probes were hybridized to DNA microarrays containing 2400 cDNAs randomly selected from normalized telencephalic pSport1 library of males and females at posthatching days 10-60. These cDNAs, along with various controls, were located in 32 patches (16 patches in upper section and the duplicated another 16 patches in bottom section) (Wade et al. 2005).

Image processing

The four image software programs are implemented to get data for foreground and background signal intensity summary calculation and quality measures. Gridding refers to the localization of rectangular patches that contain the spots. The gridding template is created by defining the number of metacolumns, metarows, columns and rows within patches, and the information about the spot diameter and distance. A little user interaction

is involved for all four image software to move the patches around to cover all spots with the patches. Segmentation is the process of distinguishing the set of pixels within a probe as foreground or background. GenePix and Molecularware assume that the spots are circular with the centers and the sizes of the circles adjusted automatically or with the additional interaction of the user with individual spots. A little interaction by the user is also involved in using Spot software to adjust individual spots in segmentation step. The data which include mean and median intensities for foreground and background pixels and some quality measures for each spot are extracted automatically after finishing gridding and segmetation process for all four software programs.

4.3 Statistical Methods

All duplicate spots from each array are included in the analysis. The mean foreground pixel intensities and the median background intensities are used for further study (Korn et al. 2004). Data features are assessed by the variability of preprocessed (see below) expression ratio for spots within and between arrays and coefficient of repeatability. Analysis of variance (Ryden et al.) is used to compare the correlation values across these four image analysis software programs. We preprocessed the intensity data using two methods: Lowess normalization combined with scale-adjustment (Yang et al. 2002b) and Arsinh transformation (Huber et al. 2002; Rocke & Durbin 2003) with further normalization model (Wolfinger et al. 2001). The differentially expressed genes are identified by using gene-specific mixed model and Empirical Bayes ANOVA (Chapter 2) with mixed model approach for both preprocessed data sets.

4.3.1 Comparison of foreground mean intensities and local background median intensities across methods

We first visually display foreground mean intensities and local background median intensities from these four image analysis software programs. Scatter-plots of the foreground mean and background median estimates for pairs of methods are produced to see how the estimates from different methods are correlated and if there is any systematically occurring difference across methods. All spot intensities from all arrays are included in each plot.

4.3.2 Intraclass correlation coefficients for genes across arrays

The Intraclass Correlation (ICC) assesses reliability of different measures by comparing the variability of different samples of the same set of genes to the total variation across all biological and technical replicates. To compare the four image analysis software programs, we use one-way ANOVA based on the following model:

$$y_{gkij} = \mu_g + \gamma_{gk} + a_{gi} + e_{gkij}, \qquad (1)$$

where y_{gkij} is preprocessed (Lowess or Arsinh) log ratios of female vs. male intensity for gene g=1,2,...,G; γ_{gk} is the fixed effect of dye, k=Cy3 and Cy5; a_{gi} is the random effect of array, i=1,2,..,n; replicate j=1,2,...,b with the assumption that $a_{gi} \sim N(0, \sigma^2_{u_g})$ and $e_{gkij} \sim N(0, \sigma^2_{e_g})$.

For each gene g, let $\bar{y}_{gi}$ be the sample mean of the replicate observations on array i and $\bar{y}_g$ for the overall sample mean across all arrays. MSB_g and MSE_g for mean square for array and residual respectively,

$$MSB_g = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{b}(\bar{y}_{gi.}-\bar{y}_{g..})^2,$$

$$MSE_g = \frac{1}{n(b-1)}\sum_{i=1}^{n}\sum_{j=1}^{b}(y_{gij}-\bar{y}_{gi.})^2,$$

Standard linear model results show that $\bar{y}_g$, MSB_g and MSE_g are sufficient statistics for μ_g, σ_g and ρ_g (Graybill 1976).

The intraclass correlation for each gene can be written as

$$cor(y_{gij,gij'}) = \rho_g = \frac{\sigma_{u_g}^2}{\sigma_{u_g}^2+\sigma_{e_g}^2} = \frac{MSB_g - MSE_g}{MSB_g + (b-1)MSE_g} \Rightarrow \rho_g = \frac{1}{1+1/\tau_g} \qquad (2)$$

where $\tau_g = \sigma_{u_g}^2 / \sigma_{e_g}^2$ is defined as a variance ratio.

We assume the variance ratios and residual variances have inverse gamma distribution as follows (Chapter 3):

$$\sigma_{e_g}^2 \sim IG(\alpha_e,\beta_e);\ \text{i.e.},\ p\left(\sigma_{e_g}^2 \mid \alpha_e,\beta_e\right) = \frac{\beta_e^{\alpha_e}}{\Gamma(\alpha_e)}\left(\sigma_{e_g}^2\right)^{-(\alpha_e+1)}\exp\left(-\frac{\beta_e}{\sigma_{e_g}^2}\right),$$

$$\tau_g \sim IG(\alpha_u,\beta_u);\ \text{i.e.},\ p\left(\tau_g \mid \alpha_u,\beta_u\right) = \frac{\beta_u^{\alpha_u}}{\Gamma(\alpha_u)}\left(\tau_g\right)^{-(\alpha_u+1)}\exp\left(-\frac{\beta_u}{\tau_g}\right). \qquad (3)$$

From expression (2), we know that variance ratio τ_g is a monotonic increasing transformation of ρ_g, which takes values within the interval (0,1). The procedure

BAYESRATIO we developed in Chapter 3 is used to estimate hyperparameters in expression (3) for the data from four image software programs. Therefore, the heteroskedasticity of the intraclass correlation coefficients across genes can be estimated and compared for these three segmentation methods.

4.3.3 Correlation coefficients within arrays

The correlations between data obtained from within arrays are compared since these data are relatively independent of variations in slide printing or sample preparation (Ahmed et al. 2004). We calculate the Pearson's correlation coefficient (r) for log ratio of duplicate genes within slides as follows:

$$r_i = \frac{\sum_{(j,j')=1}^{G} (y_{ij} - \bar{y}_{ij})(y_{ij'} - \bar{y}_{ij'})}{\sqrt{\sum_{j=1}^{G} (y_{ij} - \bar{y}_{ij})^2 \sum_{j'=1}^{G} (y_{ij'} - \bar{y}_{ij'})^2}} \qquad (4)$$

where array index i=1,2,..,n; gene pairs (j, j')=1,2,…,G and j indicates the genes in top halves of array and j' refers the genes in the bottom halves of array. Furthermore, y_{ij} is log ratios for each spot with $\bar{y}_{ij} = \frac{1}{G}\sum_{j=1}^{G} y_{ij}$ and $\bar{y}_{ij'} = \frac{1}{G}\sum_{j'=1}^{G} y_{ij'}$ where G is the number of genes within slides. There are seven arrays and four different software programs resulting in 28 correlation coefficients. ANOVA model is used for testing the correlation difference between these four methods as follows:

$$r_{si} = \mu + \alpha_s + u_i + e_{si} \qquad (5)$$

where α_s isthe fixed effect of image analysis software, s= Genepix, Imagene, MolecularWare, or Spot; u_i is the random effect of array, i=1,2,...n; r_{si} is the correlation coefficient; and $u_i \sim N(0,\sigma_1^2), e_{si} \sim N(0,\sigma_2^2)$.

4.3.4 Repeatability

Repeatability is relevant to the closeness of agreement between measurements obtained with the same method on identical test material, under the same conditions. The coefficient of repeatability is defined as the range of 95% of the differences between repeated measurements (British Standards Institution 1975). Lower repeatability coefficient represents higher precision. The use of correlation might be misleading because data which seem to be in poor agreement can produce high correlation, for example, a method that consistently overestimates the high expression ratio and underestimates the low expression ratio (Korn et al. 2004). We repeat the analysis as correlation coefficients within arrays using the coefficient of repeatability values to compare among these four image analysis software programs. We calculate repeatability coefficient for duplicate genes within arrays as $2.83\times\hat{\sigma}$, where $\hat{\sigma}$ is as estimated as follows:

$$\hat{\sigma} = \sqrt{\frac{1}{n_s - G}\sum_{i=1}^{G}\sum_{j=1}^{b}(y_{ij} - \hat{\mu}_i)^2}\ , \qquad (6)$$

where n_s is the total number of spots within array; G is the number of genes; b is the number of replicates for gene i; and $\hat{\mu}_i = \frac{1}{b}\sum_{j=1}^{b} y_{ij}$ (Jenssen et al. 2002). The model (5) also applies for testing significant difference among these four competing methods.

4.3.5 Within slide variability

Each slide contains duplicate spots for 2,397 genes. We calculated the within slide standard deviation of the log ratio of intensities for all duplicate spots and for all arrays, which totaled to have 2,397 × 7 = 16,779 estimates of spot-pair deviation for each competing method. A visualized plot is produced based on estimated proportion vs. spot-pair deviation for all four image analysis methods.

4.3.6 Lowess Normalization

Lowess normalization was processed using R as based on the paper by Yang et al. (2002b). A box plot for M-value (log(ratio)) across arrays was produced for each competing image analysis method. Further scale normalization (Yang et al. 2002b) is needed if there is obviously big difference in spreads for different boxes for arrays in the plot.

4.3.7 Arsinh Transformation

It has been demonstrated that there is often dependency between the variance and signal intensity (Rocke & Durbin 2001). When log-transformation is performed, the variance is usually stable at high intensity but vary considerably at low intensity. Arsinh transformations were proposed to stabilize the variance especially at the low intensity end (Huber et al. 2002; Rocke & Durbin 2003). The assumption for Arsinh transformation is that there is a quadratic relationship between mean signal intensities and variances. To visualize the relationship, we plot mean signal intensities vs. variances of genes for the four competing image analysis software programs. Arsinh transformation is processed in R by using the function vsn in bioconductor package if the assumption appears to be not strictly violated. Similar box plots as the above section are provided.

4.3.8 Gene-specific two step mixed model and Empirical Bayes with ANOVA method to identify differentially expressed genes

The significance test used to detect differentially expressed genes is proceeded by the following two approaches:

Two step mixed model by Wolfinger et al. (2001)

1). Step 1: further normalization:

$$Y_{ijlk} = \mu + D_j + T_l + (TA)_{il} + A_i + \varepsilon_{ijlk} \quad (7)$$

Step 2: gene-specific model

$$\varepsilon_{gijlk} = \mu_g + D_{gj} + T_{gl} + (TA)_{gil} + A_{gi} + e_{gijlk} \quad (8)$$

where g=1,2,…,G, i=1,2,..,n; j=Cy3,Cy5; k=1,2,…,b; y_{ijlk} is the preprocessed intensity; D_j is the fixed effect for dye; A_i is the random effect for array and r_{gijk} is the residual from

model (7). The distribution assumption for random and residual terms are specified as follows

$$A_i \sim N(0,\sigma_a^2),(TA)_{il} \sim N(0,\sigma_t^2),\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$$

$$A_{gi} \sim N(0,\sigma_{ga}^2),(TA)_{gil} \sim N(0,\sigma_{gt}^2),e_{gij} \sim N(0,\sigma_{ge}^2).$$

2). Empirical Bayes with ANOVA

Step 1: Use the variance component estimates from model (8) to calculate expected mean squares for Array and Treatment * array for each gene.

Assume:

$$E(MSA_g) \sim IG(\alpha_1,\beta_1) \quad \text{and} \quad E(MS(TA)_g) \sim IG(\alpha_2,\beta_2) \tag{9}$$

$$\frac{df_1 MSA_g}{E(MSA_g)} \sim \chi^2_{df_1} \Rightarrow \alpha_1/\beta_1(MSA_g) \sim F_{df_1,2\alpha_1}$$

ANOVA property

$$\frac{df_2 MS(TA)_g}{E(MS(TA)_g)} \sim \chi^2_{df_2} \Rightarrow \alpha_2/\beta_2(MS(TA)_g) \sim F_{df_2,2\alpha_2}$$

$$\text{then } \tilde{MSA}_g = \frac{df_1 MSA_g + 2\beta_1}{df_1 + 2\alpha_1} \text{ and } \tilde{MS(TA)}_g = \frac{df_2 MS(TA)_g + 2\beta_2}{df_2 + 2\alpha_2} \tag{10}$$

The null t distribution will increase from df_2 to $df_2 + 2\alpha_2$ because the error term for this model is Treatment*Array .

Step 2: Transform the updated mean square estimates of Array and Treatment * array back to variance component estimates for Array and Treatment * array.

Step 3: Rerun model (8) using the results from step 2 and setting variance component parameters as fixed to those results.

After p-values are obtained for Sex*gene effect by using gene-specific model (Wolfinger et al. 2001) and Empirical Bayes with ANOVA approach, q-values for Sex*gene are then calculated by q-value procedure (Storey 2002) in software R to control the false discovery rates. The comparison is based on the numbers of differentially expressed genes identified by these two models applied to the data from these four image analysis methods respectively using the threshold as p-values and their equivalent q-values.

4.4 Results

4.4.1 High correlation for foreground intensities and low correlation for background intensities

Scatter plots of mean foreground and median background estimates for any pair image analysis methods show whether there is systematical agreement for the two methods. Figure 1 indicates that there are high correlations for foreground intensities between any pair of image analysis methods. The red line is the reference line for equal values of x axis and y axis. The foreground intensities from MolecularWare and Imagene have the strongest agreement for high intensity estimates but a relatively large variant with the low intensity estimates among these six pair comparisons. There was consistent agreement between foreground intensities from Genepix and MolecularWare for low and high intensities. The reason is that these two software programs use the same segmentation methods: the adaptive circle method. Spot seems to produce more globally similar data to

the adaptive circle method (Genepix and MolecularWare) than the histogram method (Imagene). Figure 2 provides plots for the median background comparison for the pair image analysis methods. They show very low correlations among these four methods except the pair of MolecularWare vs. Imagene. The correlations for background intensities between data from Spot and other three image software programs are close to zero. This finding further supports the idea that background adjustment may substantially reduce the precision and increase the variability of intensity estimates (Yang et al. 2002a).

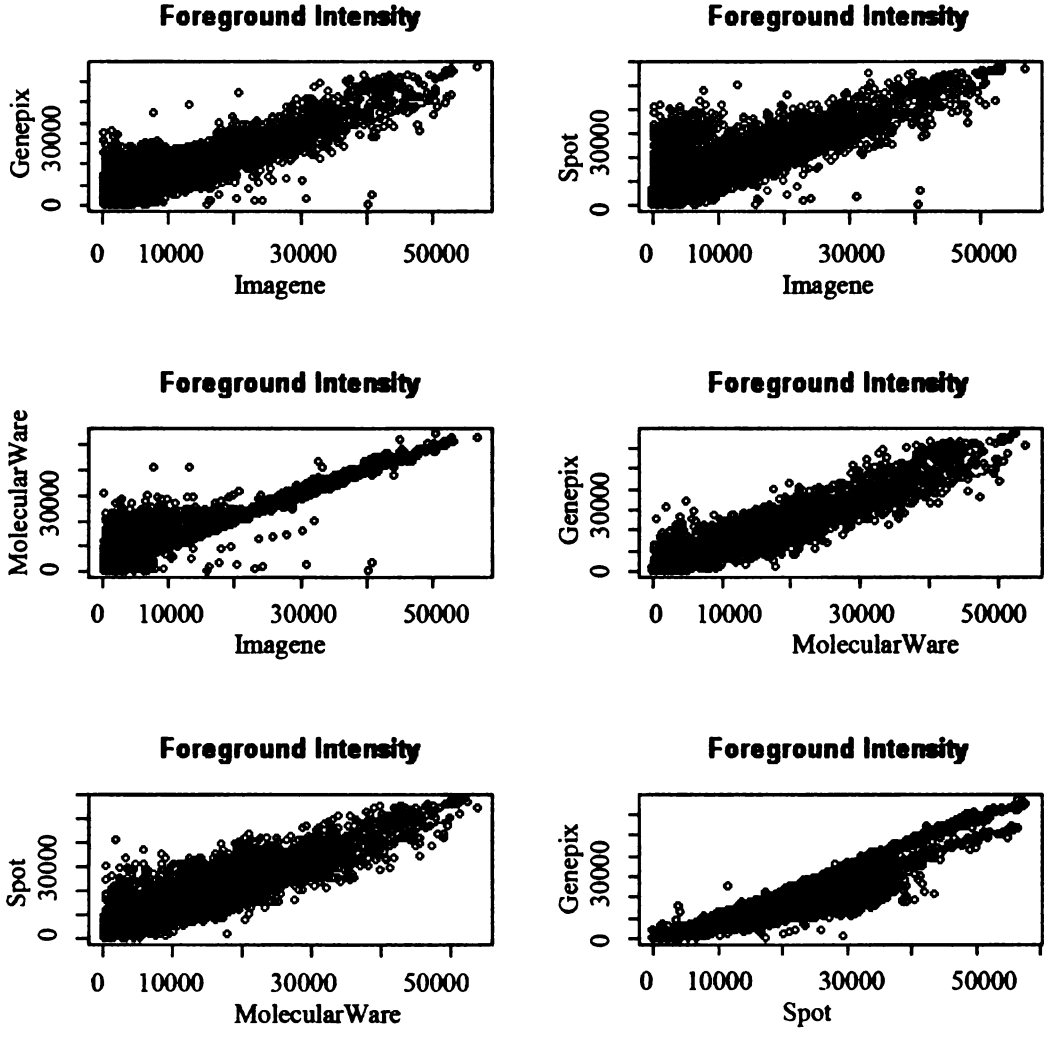


Figure 4.1 Pair wise comparisons for foreground mean intensities.

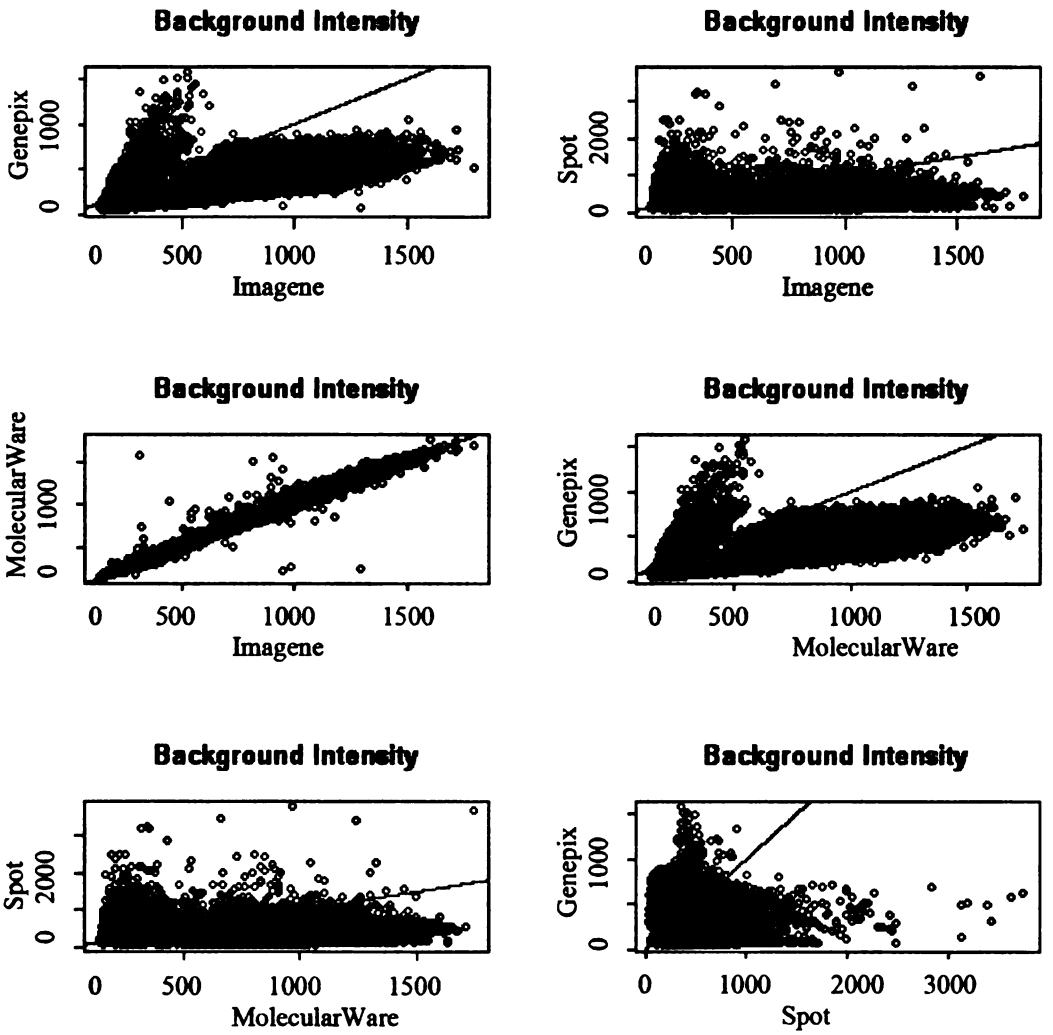


Figure 4.2 Pair wise comparisons for background median intensities.

4.4.2 Segmentation methods influence intra-class (array) correlation

To investigate whether the segmentation methods have important impacts on reliability of measurement, the procedure we developed in Chapter 3 is implemented for the data from these four image analysis methods. Log-ratios of intensities without background adjustment are used as they are more stable and biological meaningful. The hyperparameter estimates for variance ratio, which is equivalent to intraclass correlation coefficient, and residual variance distributions are listed in Table 1 and Table 2 (Arsinh preprocessed data). The shape parameter α in the inverse gamma distribution defines the heterogeneity of the random variable. The first moment of the inverse gamma distribution $\frac{\beta}{\alpha-1}$ is used to represent the expectation of this random variable. Tables 1 and 2 show that the variance ratio distribution for histogram segmentation method by Imagene has the highest $\hat{\alpha}_{\text{var}\,iance-ratio}$ and the second lowest $\hat{\beta}_{\text{var}\,iance-ratio}$ among these four image software programs, which indicates that this method produces the most homogeneous variance ratios with the lowest mean across genes. However, the second lowest $\hat{\alpha}_{residual}$ in Table 1 and the lowest $\hat{\alpha}_{residual}$ in Table 2 and the highest $\hat{\beta}_{residual}$ in both tables for histogram segmentation method result in the highest mean of residual variances. This further explains that the variability for duplicate spots within slides is higher than other methods and the most homogeneous variance ratios imply over-fitting problem across slides. The mean of residual variances for Spot is higher than those from Genepix and MoleculareWare, which use the same segmentation method and have similar hyperparameter estimates. There is no obvious difference for variance ratio

hyperparameter estimates between circle adaptive (Genepix and MoleculareWare) and shape adaptive (Spot) methods.

Table 4.1 Hyperparameter estimates for variance ratio and residual variance distributions in model (3) (Lowess preprocessed data)

Image software	Variance ratio		Residual	
	$\hat{\alpha}_{variance-ratio}$	$\hat{\beta}_{variance-ratio}$	$\hat{\alpha}_{residual}$	$\hat{\beta}_{residual}$
Genepix	2.8300 ± 0.2200	3.6196 ± 0.3771	2.9846 ± 0.1190	0.0572 ± 0.0028
Imagene	5.9247 ± 2.2194	2.3142 ± 0.9678	2.4983 ± 0.0988	0.1536 ± 0.0076
Molecular Ware	2.6150 ± 0.2269	2.0357 ± 0.2423	3.2800 ± 0.140	0.0737 ± 0.0038
Spot	2.6904 ± 0.2238	3.0098 ± 0.3370	2.3228 ± 0.0839	0.0477 ± 0.0022

Table 4.2 Hyperparameter estimates for variance ratio and residual variance distributions in model (3) (Arsinh preprocessed data)

Image software	Variance ratio		Residual	
	$\hat{\alpha}_{variance-ratio}$	$\hat{\beta}_{variance-ratio}$	$\hat{\alpha}_{residual}$	$\hat{\beta}_{residual}$
Genepix	2.6298 ± 0.2057	3.4537 ± 0.3603	2.8819 ± 0.1158	0.0340 ± 0.0017
Imagene	4.2723 ± 2.3051	1.7920 ± 1.1973	2.1980 ± 0.0849	0.0766 ± 0.0038
Molecular Ware	2.2429 ± 0.1674	1.6782 ± 0.1802	2.9539 ± 0.1200	0.0333 ± 0.0017
Spot	2.6442 ± 0.2148	2.6165 ± 0.2913	2.3099 ± 0.0845	0.0298 ± 0.0014

4.4.3 Histogram segmentation method gives lower within-slide correlation

The Pearson's correlation coefficient (r) for preprocessed M ratio values obtained from 2397 pairs of replicate spots for each of seven arrays are calculated for these four image software in Table 3. The histogram segmentation method implemented in Imagene software shows much lower correlation coefficient within arrays than others. We further use the ANOVA model to confirm whether there are significant differences for pair wise comparisons between these four image analysis software in Table 4. After using the Bonferroni adjustment procedure for p-values of pair wise comparisons, we observe that the correlation coefficients from Histogram segmentation method are significantly lower than others. There are no significant differences between the other two segmentation methods: adaptive circle (MolecularWare and Genepix) and adaptive shape (Spot).

Table 4.3 Within-slide correlations between 2397 replicate spots from seven slides by image analysis software and categorized by Lowess and Arsinh preprocessed data

	Lowess preprocessed				Arsinh preprocessed			
array	Genepix	Imagene	Molecular Ware	Spot	Genepix	Imagene	Molecular Ware	Spot
1	0.7751	0.3038	0.6978	0.7581	0.7843	0.3081	0.7329	0.7676
2	0.6994	0.4492	0.7273	0.6401	0.7030	0.4807	0.7479	0.6556
3	0.8485	0.4647	0.7466	0.7219	0.8463	0.4927	0.7716	0.7289
4	0.7522	0.3353	0.6864	0.7721	0.7368	0.3377	0.7007	0.7564
5	0.8104	0.5744	0.6902	0.5981	0.7970	0.4501	0.5672	0.5892
6	0.6443	0.4544	0.6156	0.4503	0.6327	0.4160	0.6153	0.5177
7	0.5641	0.2515	0.4851	0.8082	0.6229	0.3088	0.5555	0.8301

Table 4.4 Significance tests for pair-wise comparisons between four image analysis software (within-slide correlation) and categorized by Lowess and Arsinh preprocessed data

Software	Software	Lowess preprocessed		Arsinh preprocessed	
		Difference	P value	Difference	P value
Genepix	Imagene	0.323	<.0001	0.3327	<.0001
Genepix	Molecular Ware	0.0636	0.2527	0.0617	0.184
Genepix	Spot	0.0493	0.3722	0.0396	0.388
Imagene	Molecular Ware	-0.2594	<.0001	-0.271	<.0001
Imagene	Spot	-0.2736	<.0001	-0.2931	<.0001
MolecularWare	Spot	-0.0143	0.7948	-0.0221	0.6288

4.4.4 Coefficient of repeatability confirms lower precision of the Histogram segmentation method

Although correlation coefficients have a simple interpretation, they may not consistently agree with repeatability as the correlation coefficient is not a measure of sameness (Bland & Altman 1999). We repeated the analysis using the coefficient of repeatability values to compare these four image analysis methods. Table 5 and 6 show the same pattern as Tables 3 and 4, which confirm that the histogram segmentation method has lower precision than the other two segmentation methods. Furthermore, there are no significant differences between the other two segmentation methods.

Table 4.5 Coefficient of repeatability (defined as $2.83 * \hat{\sigma}$) between 2397 replicate spots from seven slides by image analysis software and categorized by Lowess and Arsinh preprocessed data

	Lowess preprocessed				Arsinh preprocessed			
array	Genepix	Imagene	Molecular Ware	Spot	Genepix	Imagene	Molecular Ware	Spot
1	0.1293	0.3431	0.1491	0.1362	0.0934	0.2440	0.0900	0.1011
2	0.1929	0.3236	0.1638	0.2266	0.1435	0.2148	0.1015	0.1684
3	0.1280	0.3074	0.1442	0.1459	0.0942	0.2198	0.0899	0.1063
4	0.1301	0.2776	0.1372	0.2414	0.0998	0.1972	0.0882	0.2162
5	0.2141	0.3067	0.2260	0.2170	0.1939	0.3495	0.2295	0.1792
6	0.1961	0.2754	0.1930	0.2630	0.1618	0.2538	0.1421	0.1867
7	0.2169	0.3660	0.2429	0.1447	0.1507	0.2195	0.1413	0.1017

Table 4.6 Significant tests for pairwise comparisons between four image analysis software (coefficient of repeatability) and categorized by Lowess and Arsinh preprocessed data

Software	Software	Lowess preprocessed		Arsinh preprocessed	
		Difference	P value	Difference	P value
Genepix	Imagene	-0.1418	<.0001	-0.1087	0.0001
Genepix	Molecular Ware	-0.0069	0.7464	0.0078	0.7465
Genepix	Spot	-0.0239	0.2732	-0.0174	0.4741
Imagene	Molecular Ware	0.1348	<.0001	0.1166	<.0001
Imagene	Spot	0.1178	<.0001	0.0913	0.0009
Molecular Ware	Spot	-0.0169	0.4345	-0.0252	0.3024

4.4.5 Histogram segmentation method shows higher proportion of high spot-pair deviation

We summarize the spot-pair deviation of log ratio of intensities as the proportion of pairs falling with equal spaced ranges 0-0.25, 0.25-0.5 and so on. Figure 3 indicates that the Histogram segmentation method implemented in Imagene has higher proportion of large spot-pair deviation than other two methods with three image analysis software programs. There is not much difference among the remaining software programs. We investigate whether scaling might cause the higher spot deviation. Figure 1 does not support this possibility because the other three image software programs tend to have higher foreground intensities for more spots than those from Imagene software. Therefore, the Histogram segmentation method overall introduces more variability than the adaptive circle and adaptive shape segmentation methods.

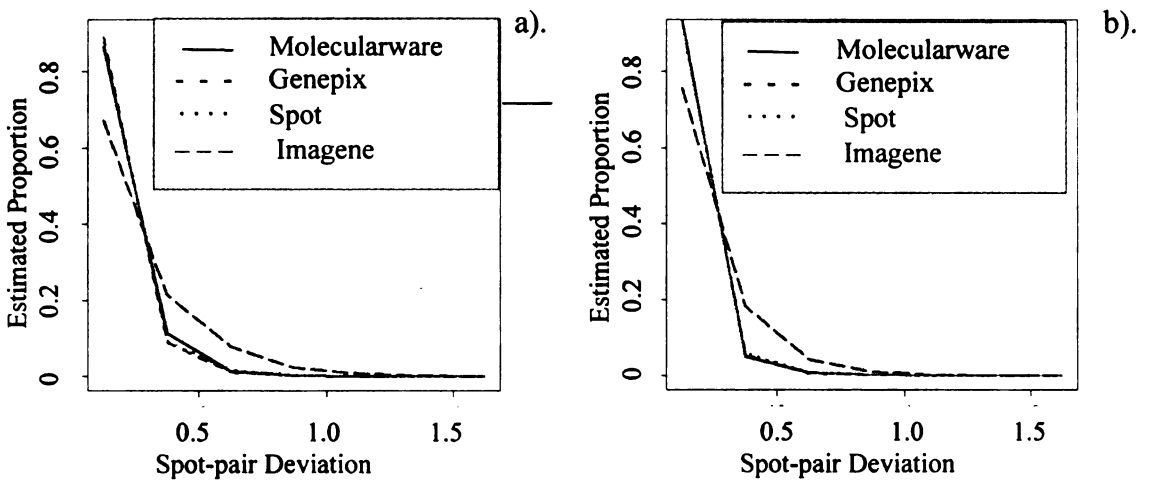


Figure 4.3 The proportion of observations inside fixed width intervals of within gene spot-pair deviation for data sets from the four image analysis software programs: a) Lowess preprocessed data; and b) Arsinh preprocessed data.

4.4.6 Lowess normalization and Arsinh transformation both are applicable for all data sets

Within arrays, the Lowess normalization method is implemented in each array and each image analysis method. Sometime scale normalization is needed to make a series of arrays have the same median absolute deviation if there are substantial scale differences between arrays (Smyth et al. 2003). Figure 4 displays side-by-side box-plots for normalized M-values (log(ratio)) for a series of seven arrays for each image analysis method to visualize if it is necessary to further proceed the scale normalization procedure, which might cause over-fitting problem if done without caution (Yang et al. 2002b). It appears that there is no obvious difference in data distributions across arrays for each image analysis method, so we decide not to go with further scaling procedure for all four competing methods.

Figure 5 shows that the assumption of quadratic relationship between the means and the variances of raw intensities is met by all data sets from the four competing image software programs. Therefore, variance stabilizing transformation Arsinh is conducted for all data sets. The box plots in Figure 6 show the M-values (log(ratio)) across each array after Arsinh transformation.

The box-plots in Figure 4 (after Lowess normalization) and Figure 6 (after Arsinh transformation) show no evidence that one transformation strategy is better than another. Therefore, both preprocessing procedures are used for the data from the four competing image software programs for all comparisons.

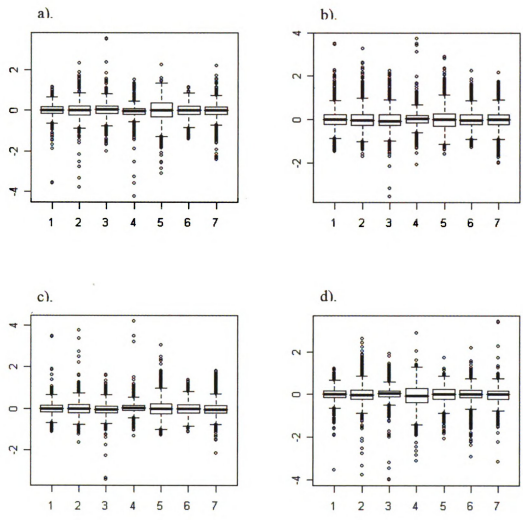


Figure 4.4 Boxplot for M-values across arrays from the four image analysis software programs after Lowess normalization: a) Genepix; b) Imagene; c) MolecularWare; and d) Spot.

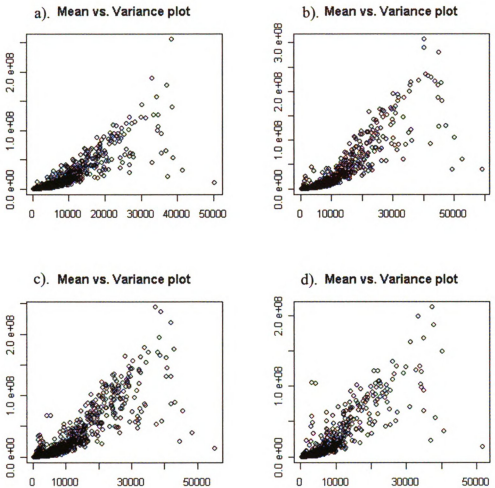


Figure 4.5 Mean intensities vs. variances for all genes from the raw data of the four image analysis software programs: a) Genepix; b) Imagene; c) MolecularWare; and d) Spot.

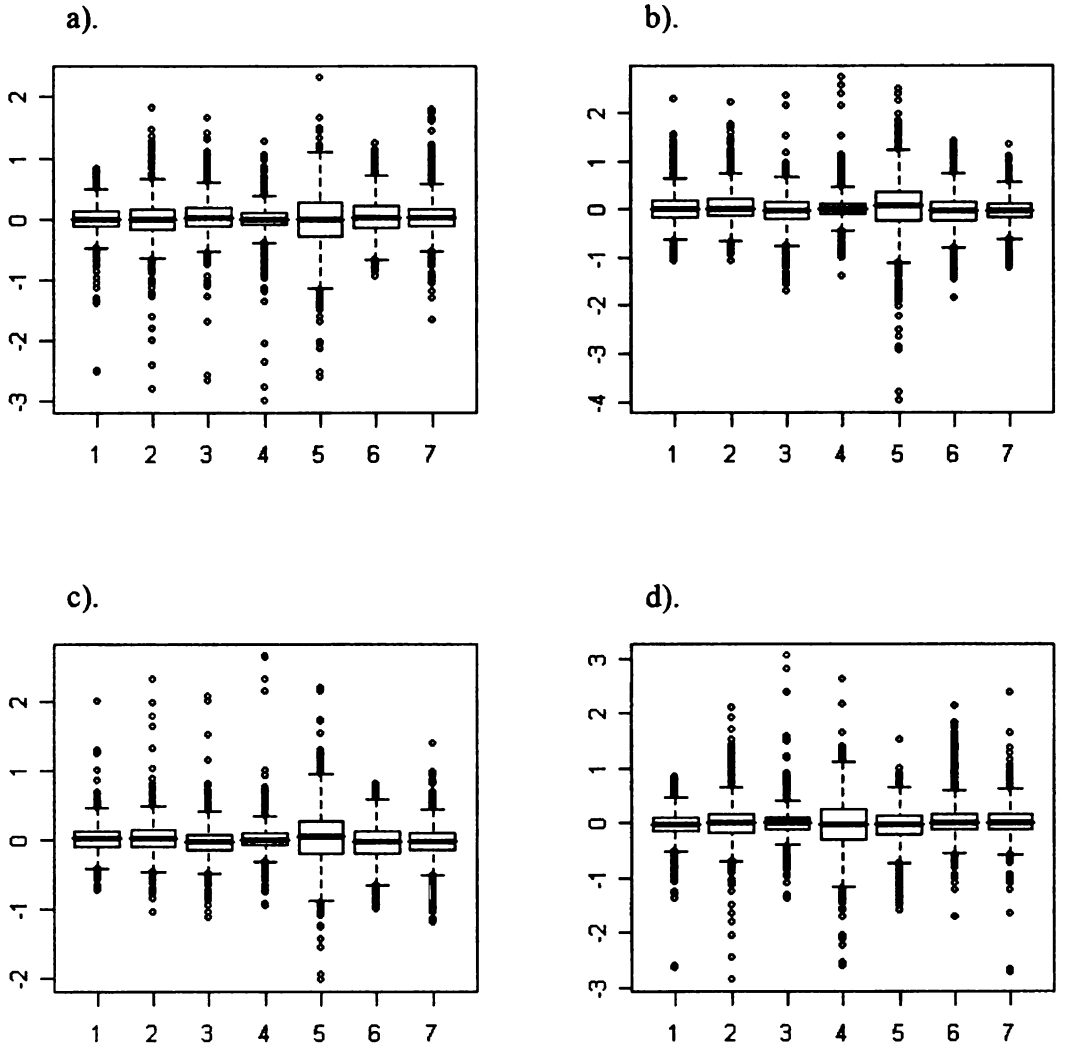


Figure 4.6 Box-plot for M-value across arrays from the four image analysis software programs after Arsinh transformation: a) Genepix; b) Imagene; c) MolecularWare; and d) Spot.

4.4.7 Less numbers of differentially expressed genes are identified by the histogram segmentation method. ANOVA EB has more sensitivity to detect differentially expressed genes across all four image analysis programs

We first use the gene-specific two-stage model introduced by Wolfinger et al. (2001). Stage 1 is designed for further normalization for globally dye and array effects. Stage 2 is used to test differentially expressed genes. The number of genes are 2, 3, 2 and 3 for Genepix, Imagene, MolecularWare and Spot for the data after Lowess normalization, and 1, 2, 1 and 1 for the data after Arsinh transformation respectively at the threshold of p-value<0.001. The differentially expressed gene IDs selected from Arsinh transformed data are subsets of genes from Lowess normalized data for each image analysis software program.

Empirical Bayes with ANOVA method is applied to the stage 2. The marginal maximum likelihood estimates ± their asymptotic standard errors for hyperparameter estimates for each mean square component are listed in Table 7 for Lowess normalized data and Table 8 for Arsinh transformed data. Array*treatment is the experiment unit for this data set. The estimates of mean square of this term array*treatment (MS(array*treatment)) determine the denominator of F test for sex effects. Therefore, the accuracy of this estimation has direct impact on statistical inference. $\hat{\alpha}_{array*trt}$ in Table 7 and Table 8 indicate the degree of heterogeneity across genes for each image software. We use EB ANOVA to combine the prior distribution of mean square of array*treatment with the estimates from gene-specific models. The posterior means are weighted average as expressed in equation (10) and the degree of freedom for F-test is increased by $2*\hat{\alpha}_{array*trt}$. Intuitively, the bigger value of $\hat{\alpha}_{array*trt}$ representing more

homogeneous mean square component has more degree of freedom, which has directly impact on the inference of fixed effect. Table 7 and 8 indicate high heterogeneity exists for the data sets from these four competing image software. The $\hat{\alpha}_{array*trt}$ s are around 3. Imagene seems to have the biggest $\hat{\alpha}_{array*trt}$ with the biggest $\hat{\beta}_{array*trt}$ with Lowess normalized data and the second biggest $\hat{\alpha}_{array*trt}$ with the biggest $\hat{\beta}_{array*trt}$ with Arsinh transformed data. The calculation of the mean based on the first moment $\frac{\beta}{\alpha-1}$ shows that Imagene has larger average MS(array*treatment) across genes than others. MolecularWare shows the slightly smaller average MS(array*treatment) estimates and bigger $\hat{\alpha}_{array*trt}$ than Genepix and Spot for Lowess normalized data and the biggest $\hat{\alpha}_{array*trt}$ for Arsinh transformed data . Overall, Arsinh transformed data have smaller variances than Lowess normalized data, and the data scale for Arsinh transformation is also lower, so the CV (coefficient of variation) is similar between these two methods. It is not surprising that the results from Lowess normalized and Arsinh transformed data are similar across the four competing image analysis software. The number of genes identified as differentially expressed by EB-ANOVA are shown in Table 9. The Figure 7 for Lowess normalized data demonstrates that gene IDs resulting from Genepix and MolecularWare are the same, Spot has one more unshared genes and those from Imagene are the subset of genes from Genepix, MolecularWare and Spot with one extra gene. Actually, the gene lists are shared by the two different preprocessing methods as the smaller numbers are the subsets of the bigger numbers in Table 9 for each image software program respectively.

Table 4.7 Hyperparameter estimates for three MS components in model (9) for Lowess normalization data

	Array		Array*treatment		Residual	
Image software	$\hat{\alpha}_{array}$	$\hat{\beta}_{array}$	$\hat{\alpha}_{array*trt}$	$\hat{\beta}_{array*trt}$	$\hat{\alpha}_{residual}$	$\hat{\beta}_{residual}$
Genepix	2.4598 ± 0.1107	1.2397 ± 0.0665	2.5388 ± 0.1264	0.0984 ± 0.0062	2.6859 ± 0.1000	0.3282 ± 0.0145
Imagene	3.3261 ± 0.1835	2.5186 ± 0.1684	3.4454 ± 0.2075	0.2136 ± 0.0156	2.1938 ± 0.0768	0.3733 ± 0.0159
MoleculareWare	2.7294 ± 0.1378	1.8230 ± 0.1149	3.1382 ± 0.1731	0.1048 ± 0.0071	1.7965 ± 0.0604	0.2614 ± 0.0111
Spot	2.4933 ± 0.1127	1.2710 ± 0.0723	2.4346 ± 0.1179	0.0981 ± 0.0061	2.5144 ± 0.0926	0.3165 ± 0.0140

Table 4.8 Hyperparameter estimates for three MS components in model (9) for Arsinh transformation data

	Array		Array*treatment		Residual	
Image software	$\hat{\alpha}_{array}$	$\hat{\beta}_{array}$	$\hat{\alpha}_{array*trt}$	$\hat{\beta}_{array*trt}$	$\hat{\alpha}_{residual}$	$\hat{\beta}_{residual}$
Genepix	3.1359 ± 0.1573	0.8670 ± 0.0528	2.5271 ± 0.1268	0.0635 ± 0.0041	2.3938 ± 0.0867	0.1724 ± 0.0075
Imagene	3.1711 ± 0.1647	0.9617 ± 0.0608	3.1814 ± 0.1823	0.1219 ± 0.0086	2.1056 ± 0.0732	0.1996 ± 0.0085
MoleculareWare	3.1848 ± 0.1650	0.8345 ± 0.0525	3.3254 ± 0.1874	0.0582 ± 0.0039	1.7773 ± 0.0596	0.1186 ± 0.0050
Spot	2.8975 ± 0.1407	0.8554 ± 0.0511	2.3106 ± 0.1100	0.0600 ± 0.0037	2.2506 ± 0.0812	0.1715 ± 0.0075

Table 4.9 Number of Significant Genes and Estimates of False Discovery Rates (in Parentheses)

	Lowess normalized data				Arsinh transformed data			
Unadjusted p-value	Imagene	Genepix	Molecular Ware	Spot	Imagene	Genepix	Molecular Ware	Spot
0.0001	2 (0.08)	3 (0.06)	3 (0.05)	3 (0.03)	2 (0.03)	3 (0.04)	3 (0.04)	4 (0.03)
0.0005	4 (0.22)	6 (0.18)	7 (0.14)	6 (0.11)	4 (0.29)	6 (0.15)	8 (0.12)	9 (0.13)
0.001	5 (0.4)	8 (0.29)	8 (0.22)	9 (0.23)	4 (0.29)	8 (0.23)	9 (0.21)	10 (0.16)

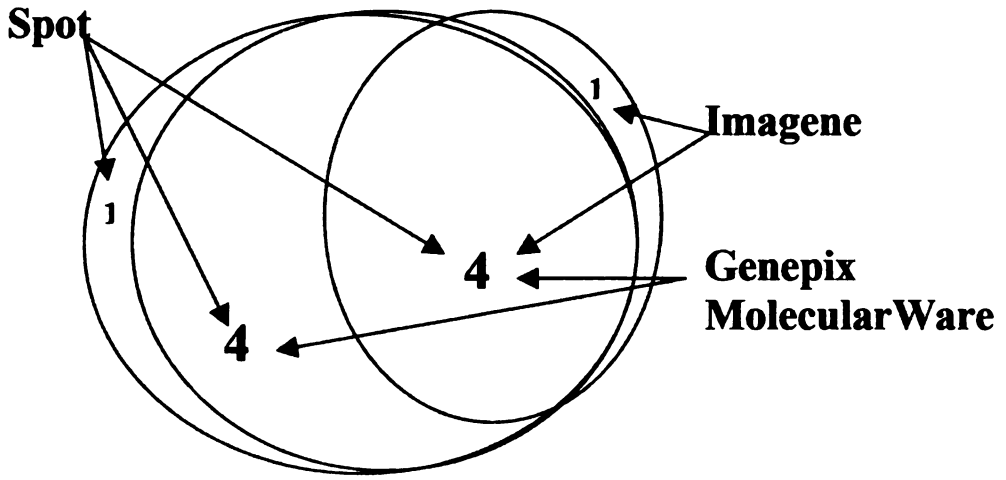


Figure 4.7 Numbers of genes identified by different image software programs at cutoff point of p-value<0.001 for lowess normalized data.

4.5 Discussion

In this paper, we have discussed data features from different image analysis methods with various segmentation approaches (adaptive circle, adaptive shape and histogram). We have compared a number of software (Genepix, Imagene, MolecularWare and Spot) on one experiment with replicated spots on each slide. The comparison indicates that the background intensities vary a lot among these methods but foreground intensities share much more similarity with high correlation. It suggests that the choice of image analysis software would have much smaller impact if we use the data without background corrected intensities for further analysis than the data with background correction. We also used two preprocessed procedures (Lowess and Arsinh) for background corrected intensities to fit BAYSRATIO model and EB-ANOVA model. The results show that there are more heterogeneity for variance ratios, residual variances for background corrected intensities than without background correction and it is also true for mean square components. These findings indicate that shrinkage methods would provide less benefit for background corrected intensities than without background correction. Thus, our advice is to utilize the foreground intensities without background correction as input for identifying differentially expressed genes. This idea is also suggested in (Cui et al. 2003).

Our comparison of different methods for data feature suggests that:

1) The Histogram method yields significantly lower within-slide correlation with higher spot-spot variability than other segmentation methods. The Histogram method defines the ratio of foreground and background as the mean intensities between

predefined percentile values, usually 5%-20% for background, 80%-95% for foreground (Qin et al. 2005), which is expected to have lower variability if the background correction data is compared. Nevertheless, histogram methods have been found to suffer from the difficulty in choosing a suitable mask size, so the foreground intensities resulting from this method might not be as accurate as other recently introduced segmentation methods (Yang et al. 2002a).

2) The lowest mean of common intra-class (slides) correlation coefficients with the most homogeneity further confirms that there might be overfitting problem in the histogram method.

3) The same segmentation method in different image analysis software gives a similar data feature. We have compared correlation, repeatability and spot-deviation for the two software MolecularWare and Genepix which share the same segmentation method (adaptive circle). Results show that there is no significant difference between them and that they share more similar features than other comparing pair software (as expected).

4) The data features from the adaptive circle and adaptive shape methods are not significantly different from each other based on our findings. The reason may lie in that most spots are supposed to be circular for high quality microarray slides.

The EB ANOVA with mixed model approach has more sensitivity to detect differentially expressed genes than gene-specific mixed model for the data from all image analysis software programs. The two different data preprocessing methods identify a similar list of genes for all image analysis software respectively. Lowess normalization

and Arsinh transformation has different algorithms to process data. Lowess is aimed to correct curvatures in MA plots (log ratio vs. mean of log intensity) while Arsinh transformation is intended to stabilize the variances across genes. All comparisons based on these two normalization methods give similar results for the four competing image software programs. Furthermore, Tables 7 and 8 show that the resulting data from these two data preprocessing methods have similar degree of heterogeneity across genes for all mean square components. Therefore, it is not surprising that the two normalization methods do not make much difference in detecting significant genes, and the shrinkage procedure does increase the power by borrowing information across genes for the data from all image software programs. The histogram segmentation method does not benefit as much as the other two segmentation methods, the adaptive circle and adaptive shape methods from shrinkage procedure, because of its more variable data features.

Due to the complexity of images caused by the noise sources inherent in the DNA microarray process, it is challenging to develop tailor-made image processing methodologies (Istepanian 2003). In any case, this paper gives researchers some idea on the choice of image analysis software to match the features of microarray images and subsequent data analysis strategy.

BIBLIOGRAPHY

AHMED, A. A. VIAS, M. IYER, N. G. CALDAS, C. & BRENTON, J. D. (2004). Microarray segmentation methods significantly influence data precision. *Nucleic Acids Research* **32**(5).

BLAND, J. M. & ALTMAN, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods In Medical Research* **8**(2), 135-160.

BRITISH STANDARDS INSTITUTION (1975). Precision of test methods 1: Guide for the determination and reproducibility for a standard test method (BS 597, Part 1). (Ed B. S. Institution): London: BSI

CUI, X. G. KERR, M. K. & CHURCHILL, G. A. (2003). Transformations for cDNA Microarray Data *Statistical Applications in Genetics and Molecular Biology* **2**(1).

GILBER, G. (2006). *Developmental Biology* (8th edition). Sunderland, MA: Sinauer Associates.

GRAYBILL, F. A. (1976). *Theory and Application of the Linear Model.* North Scituate, MA: Duxberry Press.

HUBER, W. HEYDEBRECK, A. V. SULTMANN, H. POUSTKA, A. & VINGRON, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl 1), S96-104

ISTEPANIAN, R. S. (2003). Microarray Image Processing: Current Status and Future Directions. *IEEE Transactions on Nanobioscience* **2**(4).

JENSSEN, T. K. LANGAAS, M. KUO, W. P. SMITH-SORENSEN, B. MYKLEBOST, O. & HOVIG, E. (2002). Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Research* **30**(14), 3235-3244.

KORN, E. L. HABERMANN, J. K. UPENDER, M. B. RIED, T. & MCSHANE, L. M. (2004). Objective method of comparing DNA microarray image analysis systems. *Biotechniques* **36**(6), 960-967.

QIN, L. RUEDA, L. ALI, A. & NGOM, A. (2005). Spot Detection and Image Segmentation in DNA Microarray Data. *Appl Bioinformatics* **4**(1), 1-11.

ROCKE, D. M. & DURBIN, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology* **8**(6), 557-569.

ROCKE, D. M. & DURBIN, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**(8), 966-972.

RYDEN, P. ANDERSSON, H. LANDFORS, M. NASLUND, L. HARTMANOVA, B. NOPPA, L. & SJOSTEDT, A. (2006). Evaluation of microarray data normalization procedures using spike-in experiments. *Bmc Bioinformatics* **7:300**.

SMYTH, G. YANG, H. & SPEED, T. (2003). Statistical Issues in cDNA Microarray Data Analysis. *Methods Mol Biol.* **224**, 111-136.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of The Royal Statistical Society Series B-Statistical Methodology* **64**, 479-498.

WADE, J. PEABODY, C. COUSSENS, P. TEMPELMAN, R. J. CLAYTON, D. F. LIU, L. ARNOLD, A. P. & AGATE, R. (2005). A cDNA microarray from the telencephalon of juvenile male and female zebra finches (vol 138, pg 199, 2004). *Journal of Neuroscience Methods* **142**(2), 327-327.

WANG, Y., SHIH, FY., AND MA, MQ., (2005). Precise Gridding of Microarray Images by Detecting and Correcting Rotations in Subarrays. New Jersey Institute of Technology.

WOLFINGER, R. D. GIBSON, G. WOLFINGER, E. D. BENNETT, L. HAMADEH, H. BUSHEL, P. AFSHARI, C. & PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625-637.

YANG, Y. H. BUCKLEY, M. J. DUDOIT, S. & SPEED, T. P. (2002a). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational And Graphical Statistics* **11**(1), 108-136.

YANG, Y. H. DUDOIT, S. LUU, P. LIN, D. M. PENG, V. NGAI, J. & SPEED, T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4).

Chapter 5: Discussion, Conclusions and Future Work

5.1 Discussion and Conclusions

The main objective of this dissertation is to propose new methods to improve the efficiency of statistical inference for differentially expressed genes using microarray experiments, specifically those based on efficient experimental designs that account for multiple random sources of variation. As elucidated thus far in this dissertation, statistical methods are typically involved in several distinct stages of a microarray experiment such as experimental design, image analysis and class comparison analysis and/or class discovery. Microarray experiments are generally characterized by a very limited number of replicates due to high costs; nevertheless, a large number of genes are typically studied. This data feature of microarrays is often referred to as "large p, small n" issue (West 2003). Here, "p" represents the number of variables, i.e. genes, and "n" is the sample size for each gene. The development of statistical methods for class comparison analysis ranges from non-parametric methods and t-tests for simple designs comparing two conditions, ANOVA for comparing three or more conditions, to mixed effects models for analyzing more efficient and elaborate designs. Empirical and Bayesian methods are natural approaches for microarray data analysis as they facilitate borrowing of information across large p to improve the low power for individual gene tests due to small n. Therefore, Bayes extensions to mixed model analyses could facilitate efficient analyses of microarray experiments characterized by multiple error strata; e.g. biological versus technical replication.

In Chapter 2, an empirical Bayes mixed model with shrinkage on ANOVA components (EB-ANOVA) for estimating variance components was developed and compared with other methods such as an alternative empirical Bayes (EB-REML) mixed model (Feng et al. 2006) and classical mixed model methods (Wolfinger et al. 2001). Two procedures for shrinking variance components across genes were investigated, each based on shrinking expected mean squares for random effects. EB-ANOVA was based on ANOVA inference on variance components whereas EB-REML was based on REML (Feng et al. 2006). The distinction between the two methods is somewhat inconsequential in balanced designs except when ANOVA estimates are negative in which case REML constrains those estimates to zero as with the shrinkage procedure proposed by Feng et al. (2006). Stroup and Littell (2004) determined that this phenomenon might be partly responsible for the differences between the two methods in how they influence Type I error rates on fixed effect inference in unbalanced linear mixed models. They concluded that the standard ANOVA F tests based on ratio of mean squares yield acceptable control of Type I error whereas REML did not for an unbalanced design or models with correlated errors.

This dissertation further addresses the choice between REML and ANOVA when extended to shrinking estimates across many responses (i.e. genes). Data sets representing two popular microarray experimental designs, the loop design and common reference design, were used to compare our proposed method with other alternative methods. Various degrees of heteroskedascity of all random effects were simulated to represent a wide range of scenarios that might be plausible for microarray data. The performance of the competing methods was evaluated by mean absolute deviation for variance

component estimates, and ROC curves and FDR control for inference on differential gene expression. The results from simulation study indicated that the proposed EB-ANOVA method provided better performance in detecting true positives while adequately controlling for false positives in both designs. The study also demonstrated that EB-ANOVA produced more precise variance component estimates and subsequently more accurate treatment effect estimates in loop designs, likely due to the increased efficiency of combining interblock and intrablock information. In addition, it was deemed possible to formally derive the correct ANOVA F-test denominator degrees of freedom for hypothesis testing using EB-ANOVA thereby combining its sensitivity with better control of Type I error and false discovery rates compared to EB-REML.

In Chapter 3, a fully Bayesian method named BAYESRATIO was presented to critically evaluate the empirical Bayes strategy in the popular microarray analysis software LIMMA for managing within-array technical replicates. Microarray experimental designs are often characterized by technical and biological replicates. It is essential to correctly specify the experimental units for statistical analysis, particularly to control Type I error. For those situations where each gene is spotted two or more times on an array, the LIMMA software invokes a very strong shrinkage assumption, that being a constant ratio of within to between slide variability across all genes. In essence then, the treatment of technical replicates is somewhat based on simple moderation methods designed for a single error strata.

Motivated by an application which was determined to violate this common intraclass correlation assumption, BAYESRATIO was proposed to directly model heterogeneity in this intraclass correlation and the residual variance across genes. The performance of

LIMMA procedure and BAYESRATIO were directly compared to each other, our EB-ANOVA model and other conventional mixed model estimation methods. The simulated data sets were based on differing levels of heterogeneity for variance ratios and residual variances across genes. Simulations also differed in the number of technical replicates (spots per gene on an array) and also the magnitude of the correlation coefficients representing the typical top-and-bottom versus side-by-side replicated spots for genes on an array. LIMMA was illustrated to have poor performance in controlling false discovery rates, worsening with larger numbers of spots or technical replicates per gene within slides. Conversely, BAYESRATIO had overall equal or superior performance to any other methods in terms of precisely estimating variance components, balance between false positive and false negative rates and FDR control. Moreover, the weaker assumption of BAYESRATIO regarding the distribution of the intraclass correlation coefficient across genes makes it applicable to microarray data with any level of heteroskedascity and for different design layouts; for example, for those designs where there may be either within slides or between slides technical replicates or both. EB-ANOVA in Chapter 2 was also shown to be robust and effective even for the simulations from Chapter 3.

In Chapter 4, different data features coming from different image analyses software were studied, and transformations and models that would be most appropriate for the respective characteristics of data were suggested. Image processing may have a substantial impact on subsequent analysis such as the identification of differentially expressed genes (Yang et al. 2002). In this chapter, we show that the choice of segmentation methods results in significant data features such as variability in precision which may influence the preferable choice of normalization method and statistical model

for formal inference. TIFF files from one microarray experiment were employed as the basis for comparing four image software programs Genepix, Imagene, Molecularware and Spot. Since background intensities varied substantially among these image software programs whereas foreground intensities were highly correlated among them, analyzing data that is not background corrected has smaller impact caused by different choice of image analysis software than with background correction. Based on TIFF files that were used to make comparisons intensities, the histogram segementation method of Imagene software produced significantly greater variability for duplicated spots and more homogeneous correlation coefficients across genes indicating a potential over-fitting problem. The shape adaptive algorithm by Spot image software and circle adaptive methods by Genepix and MolecularWare software programs were found to share similar data features in our data. The proposed EB-ANOVA model was demonstrated to have greater sensitivity for identifying differentially expressed genes compared with the conventional mixed model (Wolfinger et al. 2001) for data generated from all four image analysis software programs.

5.2 Future Work

In this dissertation, I focused on statistical methodology development and applications for microarray data analysis. An empirical Bayes extension of mixed model analysis of microarrays was proposed in Chapter 2 and a Bayesian method for modeling heterogeneity of intraclass correlation coefficients of arrays characterized by multiple spots per gene presented in Chapter 3 was shown to be more suitable and flexible compared to the increasingly popular LIMMA software. Different segmentation methods

based on different image software programs were compared and subsequent statistical analysis method was suggested in Chapter 4.

All the models considered in this thesis are based on the general mixed model framework. The error term and all random effects are assumed to be normally distributed in conventional mixed models and we did not deviate from that assumption in our Bayesian or shrinkage based extensions. As microarray experiments are characterized by multiple complex steps, it is not uncommon for some data influenced by artifacts; i.e. there are outliers. Hence, one potentially important extension of this work is to formally accommodate outliers. A Student t error specification for the random error terms would be robust method to outlier fluorescence intensities. A hierarchical Bayesian model with t distributed errors has been proposed up by Gottado et al. (2003). Dror et al. (2003) also introduced a model that includes novel features such as heavy-tailed additive noise and a gene-specific bias term. An unresolved issue in these Bayesian-based approaches is that they are very computationally intensive and time consuming. Hence it is necessary to consider alternative and efficient algorithms to fit Student t-error model in a reasonable amount of time for large data sets like those generated from microarray experiments. A ECME (expectation-conditional maximization either) (Liu & Rubin 1995) may be one such efficient algorithm.

The methodology we discuss in this study is developed primarily for two color microarray systems. In short oligonucleotide microarrays (i.e. Affymetrix arrays), the probes, several of which form a single gene, are designed to match parts of the sequence of known or predicted mRNAs. There are several methods proposed for normalizing data at probe level and the expression of each gene can be based on summarizing the values of

distinct probe pairs. After data have been properly normalized, the essential difference between two color microarray data and single-channel microarray is that two sample expression profiles are paired in two color arrays whereas only one sample is hybridized in single-channel microarrays (Fan & Ren 2006). Subsequent statistical analysis is then similar to common reference experiments (Smyth 2005). Therefore, the proposed methods can easily extend to the analysis of short oligonucleotide arrays. It would be useful to explore how beneficial the proposed methods are for improving statistical inference when applied to data generated from short oligonucleotide microarray experiments.

BIBLIOGRAPHY

DROR, R. O. MURNICK, J. G. RINALDI, N. J. MARINESCU, V. D. RIFKIN, R. M. & YOUNG, R. A. (2003). Bayesian estimation of transcript levels using a general model of array measurement noise. *Journal of Computational Biology* **10**(3-4), 433-452.

FAN, J. & REN, Y. (2006). Statistical Analysis of DNAMicroarray Data in Cancer Research. *Clin Cancer Res* **12(15)**

FENG, S. WOLFINGER, R. D. CHU, T. M. GIBSON, G. C. & MCGRAW, L. A. (2006). Empirical Bayes analysis of variance component models for microarray data. *Journal of Agricultural Biological and Environmental Statistics* **11**(2), 197-209.

GOTTARDO, R. PANNUCCI, J. A. KUSKE, C. R. & BRETTIN, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**(4), 597-620.

LIU, C. H. & RUBIN, D. B. (1995). Ml-Estimation of the T-Distribution Using Em and Its Extensions, Ecm and Ecme. *Statistica Sinica* **5**(1), 19-39.

SMYTH, G. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* pp. 397-420. New York: Springer.

STROUP, W. & LITTELL, R. (2004). Impact of variance component estimates on fixed effect inference in unbalanced linear mixed models. In *UF IFAS Statistics*.

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* pp. 733-742. Oxford University Press.

WOLFINGER, R. D. GIBSON, G. WOLFINGER, E. D. BENNETT, L. HAMADEH, H. BUSHEL, P. AFSHARI, C. & PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625-637.

YANG, Y. H. BUCKLEY, M. J. DUDOIT, S. & SPEED, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**(1), 108-136.