2006-11-20

A Comparison of Microarray Analyses: A Mixed Models Approach Versus the Significance Analysis of Microarrays

Nathan Wallace Stephens
Brigham Young University - Provo

A COMPARISON OF GENETIC MICROARRAY ANALYSES: A MIXED MODELS APPROACH VERSUS THE SIGNIFICANCE ANALYSIS OF MICROARRAYS

by

Nathan W. Stephens

A project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

December 2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Nathan W. Stephens

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	David G. Whiting, Chair
Date	G. W. Fellingham
Date	G. Bruce Schaalje

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the project of Nathan W. Stephens in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

David G. Whiting
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and
Mathematical Sciences

ABSTRACT

A COMPARISON OF GENETIC MICROARRAY ANALYSES: A MIXED MODELS APPROACH VERSUS THE SIGNIFICANCE ANALYSIS OF MICROARRAYS

Nathan W. Stephens

Department of Statistics

Master of Science

DNA microarrays are a relatively new technology for assessing the expression levels of thousands of genes simultaneously. Researchers hope to find genes that are differentially expressed by hybridizing cDNA from known treatment sources with various genes spotted on the microarrays. The large number of tests involved in analyzing microarrays has raised new questions in multiple testing. Several approaches for identifying differentially expressed genes have been proposed. This paper considers two: (1) a mixed models approach, and (2) the Significance Analysis of Microarrays.

ACKNOWLEDGMENTS

This document was written in WinEdt and typeset in LaTeX. Portable network graphics were converted to postscript using ImageMagick®. Tables were converted into LaTeX code using XTABLE, a package for the R environment.

I would like to express my gratitude to those who supported me throughout this project: my wife for her enduring patience and her unwavering support; my father for showing me the value of education and for supporting my academic work; my mother for her listening ear and her nurturing heart; Dr. Grimshaw for his guidance and for helping me finish this project; and my committee for their help and their suggestions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genomics, the study of genes and their functions, generates huge quantities of data. A common question in the analysis of genetic data is the identification of differentially expressed genes. Differentially expressed genes are those whose expression levels are associated with a response or covariate of interest. DNA microarrays can be used to simultaneously test thousands of genes. This has raised new interest in multiple testing procedures. The objective in many statistical analyses is to identify many differentially-expressed significant genes while incurring a relatively low proportion of false positives.

One popular approach to analyzing microarray data is the mixed models approach developed by Wolfinger et al. (2001). This method uses linear mixed models to normalize and analyze data. The normalization model attempts to remove experimental effects that are common across all genes. The residuals from the normalization model are then passed to separate gene models. Each gene is modeled with a separate gene model, resulting in as many models as there are genes. All possible pairwise treatment comparisons are tested for significance. Significant genes are those with significant pairwise treatment comparisons. A Bonferroni correction is used to control for the family-wise error rate.

Another popular approach to analyzing microarray data is the The Significance Analysis of Microarrays (SAM) developed by Tusher et al. (2001). The SAM procedure involves computing a relative difference for each gene and then comparing it to its expected value. The expected value is calculated using a resampling technique. Genes with relative differences that fall outside tolerance limits are declared significant. The SAM procedure also estimates the False Discovery Rate (FDR), which is the proportion of null genes among those called significant; the Miss Rate (MR), which is the proportion of significant genes among those called non-significant; and the Tail Strength (TS) which is an overall measure of significance for a set of hypothesis tests.

This project compares the mixed models approach developed by Wolfinger et al. (2001) to the SAM procedure developed by Tusher et al. (2001). Genetic microarray data gathered by the University of Utah are used to make the comparison. Causes for the differences between results are discussed, an attempt is made to correct two of the causes, and inherent and flexible characteristics of both methods are discussed.

Chapter 2

Literature Review

The development and use of DNA microarrays is a relatively new technology for measuring gene expression. A DNA microarray is a large set of cloned DNA molecules spotted onto a solid matrix (such as a microscope slide) for use in probing a biological sample to determine gene expression. Each spot on a microarray contains thousands of samples of a particular isolated area, or genes of single strand DNA.

Microarray gene spots are not printed simultaneously; they are printed in blocks called print-tip groups. Microarrays tend to have a few dozen print-tip groups. All the genes in a single print-tip group are printed simultaneously using the same print-tip. Therefore, print-tip effects tend to be an important source of systematic variability. Correcting for the variability due to print-tips is often an important part of analyzing microarray data.

Microarrays allow researchers to simultaneously test thousands of genes. In order to test the genes on a two-channel microarray experiment, a sample of DNA is taken from a treatment specimen and a sample of DNA is taken from a control specimen. The DNA from the treatment and the control is split into single strand molecules (cDNA) that can combine or hybridize with genes on the microarray. In a two-channel microarray experiment, the cDNA is tagged with florescent dyes for

identification purposes. It is common to use green and red dyes to identify the treatment and control strands.

Both samples from the treatment and the control are then applied to the microarray. The single strand samples from the specimens attempt to bind with the single strand genes on the spotted microarray. The amount of binding between the spots and the samples is a measurement of gene expression. Spots that have heavy binding to the treatment look green under a green laser, whereas spots that have heavy binding to the control look red under a red laser. The measurements collected under the two lasers are called gene intensities.

Scanning software produces foreground images and background images for each dye. The background image represents the baseline value for the gene expression. In nearly all microarray analysis it is common to subtract the background image from the foreground image, creating background corrected gene intensities.

Researchers use gene intensities to quantify gene expression. If R is the measurement of gene expression for the treatment (assume it is labelled with a red dye), and G is the measurement of gene expression for the control (assume it is labelled with a green dye), then a common measurement for the overall gene expression is $M = \log_2(R/G)$. The value M is called the log ratio and it represents the difference between the treatment and the control. Another common measurement for the overall gene expression is $A = \log_2 \sqrt{RG}$. This measurement is the log average and represents the average gene expression. Both M and A-values are common metrics in many two-channel microarray experiments.

2.1 Normalization

Microarray data tend to be very noisy. Some sources of variation include dye effects from the red and green dyes applied to the cDNA strands, print quality effects due to the gene printing on the microarray, scanning error, differing experimental conditions, and spatial effects because the measurements come from a physical slide (Yang et al., 2002). Before testing for differentially expressed genes, microarray data must be normalized. (Note: some researchers such as Huber et al. (2002) prefer the term calibrated.) The purpose of normalization is to adjust for effects that are due to the noise of the experimental setup rather than effects that are due to genetic differences (Smyth and Speed, 2003).

Correcting expression values for the background is typically the first step in normalizing the data. The background corrected expression value is the foreground expression value minus the background expression value. Negative or zero expression values are problematic if logged data or log ratio data are sought. Huber et al. (2002) suggests four methods for dealing with negative expression data: (1) ignore the background estimates, (2) replace the negative estimates by 1 before taking the logarithm, (3) subtract the 5%-quantile, then replace the remaining negative values by 1, and (4) flag them as missing. Smyth and Speed (2003) and Wolfinger et al. (2001) recommend treating negative expression data as missing. Alternatively, expression values can be transformed using a rank transformation or an arsinh transformation (Huber et al., 2002).

Additional normalization often takes place on the background corrected log ratios. The is the approach suggested by Dudoit et al. (2002), Smyth and Speed (2003), and Yang et al. (2002). However, some approaches attempt to normalize the background corrected $\log_2$ intensities. This latter approach is suggested by Huber et al. (2002) and Wolfinger et al. (2001).

Normalization can be applied within a microarray and/or between microarrays. The utility of normalizing between arrays in an experiment has been disputed. Smyth and Speed (2003) suggest scaling the M-values so that each array has the same median absolute deviation (MAD). However, normalizing between arrays can introduce more variability in the data than it resolves. Yang et al. (2002) suggest foregoing between-array normalization in cases where the scale differences are small.

Several techniques have been proposed for normalizing gene intensities within a microarray. Plots of the M-values versus the A-values (MA-plots) are useful in evaluating intensity-dependent biases within arrays. It common for the M-values and the A-values to be strongly correlated. Yang et al. (2002) and Smyth and Speed (2003) suggest fitting lowess curves to MA-plots for each print-tip group. The residuals are called the normalized log ratios and are calculated as $\rightarrow \log_2 R/G - c_i(A)$, where $c_i(A)$ is the lowess fit to the MA-plot for the i^{th} print-tip group. This method has the advantage of correcting both the intensity-dependent bias and spatial effects that are associated with print-tip groups.

There are several different varieties of lowess normalization (Yang et al., 2002; Smyth and Speed, 2003). For example, if there does not appear to be any difference between print-tip groups, then a global lowess normalization is appropriate. A

global lowess normalization takes place at the microarray level instead of the print-tip group level. If there are peculiar spatial effects beyond those associated with print-tip groups, then a two-dimensional lowess normalization is appropriate. The two-dimensional technique uses a lowess surface to fit spatial effects. Smyth and Speed (2003) point out that this method requires more investigation. If the data are especially noisy, then normalizing with respect to the median is appropriate (Smyth and Speed, 2003). Furthermore, composite lowess normalization can be used in conjunction with a known control group. Print-order normalization can be used to correct for substantial print order effects. Finally, spot quality weights can be used to refine the normalization process (Smyth and Speed, 2003).

Huber et al. (2002) take a different approach to normalizing data. They propose a transformation in which the variance of gene intensities becomes independent of the mean. The transformation is of the form $h(x) = \text{arsinh}(a + bx)$, where a and b are tuning parameters. The proposed transformation produces difference statistics (Δh) that may be viewed as a generalization of the log-ratio of two-channel microarray data. The variance of these difference statistics is approximately constant along the whole intensity range. The transformation, which is applied globally, does not adjust for spatial effects.

Tusher et al. (2001) compare the log average intensity for each hybridization to a reference. The reference is calculated by taking the log average of the intensities across all hybridizations. A linear least squares regression is fit to the cube root of both the hybridization and reference log averages. The cube root is used because it

allows negative values to be included. The normalized values are the residuals from the linear least squares fit.

Wolfinger et al. (2001) fit the background corrected $\log_2$ intensities to a mixed model, then use the residuals as normalized values. The model that normalizes the data is called the normalization model and does not include any effects due to the individual genes. The model includes fixed effects due to the treatments, and random effects due to the arrays. The normalization model assumes that the random effects are distributed normally with constant variance.

Finally, normalization should be accompanied by plots of the data. Smyth and Speed (2003) suggest four visualization techniques. First, they use an MA-plot to assess the relationship between dye-bias and intensity. They also recommend adding a trend line to the plot to better assess the relationship. Second, they plot intensities on a row and column grid called an image plot to assess spatial effects in the data. They particularly recommend spatial plots of the background intensities and non-normalized M-values. Third, they use side-by-side box plots of the M-values for each print-tip group in order to visualize print-tip effects. Finally, they use an MA-plot by print-tip to assess the relationship between intensities and print-tip effects.

2.2 Analysis Methods

After microarray data have been normalized, they can be analyzed to identify differential gene expression. Several methods have been proposed. The following section examines a mixed models approach, the Significance Analysis of Microarrays, and a few other techniques.

2.2.1 Mixed Models Approach

Wolfinger et al. (2001) propose a mixed models approach to microarray analysis. This approach involves a normalization model and a gene model. Both models use background corrected $\log_2$ intensities as opposed to log ratios or log averages. There is a separate treatment and reference measurement for each spot on the array. The reference is considered to be another treatment level, allowing the researcher to measure the effect of the reference sample. The structure of the two models depends on the experimental setup and the discretion of the researcher; for example, the researcher can include array effects, print-tip effects, or blocking effects in the models. The normalization model attempts to remove experimental variability not due to gene variability. The residuals from the normalization model are then used in the gene model, which determines the significance of the gene. One gene model is fit for every unique gene.

In the mixed models approach, genes are declared significant if their corresponding gene models contain significant treatment effects. If there are multiple treatment levels, Wolfinger et al. (2001) recommend testing all pairwise comparisons between the treatment levels. The most significant pairwise comparisons identify significant genes. A Bonferroni correction is used to control the family-wise error rate (FWER).

Wolfinger et al. (2001) demonstrate the mixed models approach by analyzing a publicly available two-channel microarray data set. The data include five treatment levels. There are $\binom{5}{2} = 10$ pairwise comparisons for each of the 6,917 genes for a

total of 69,170 tests. A Bonferroni correction is then used to control the FWER at 5%. For some studies the Bonferroni correction can be too conservative, yielding no significant genes. Dudoit et al. (2003) suggest using adjusted p-values instead of a Bonferroni correction.

The genes in the mixed models approach are modeled with independent errors. Assuming heterogeneity among gene variances is reasonable; however, the residuals from the normalization model are correlated by construction. Wolfinger et al. (2001) argue that this concern makes little to no difference in practice. The effects fitted by by the normalization and gene models are orthogonal. Furthermore, the effects in the normalization model are estimated with such great precision that they are practically constants.

The mixed models approach also assumes that the error terms for both the normalization model and the gene models are normally distributed. Wolfinger et al. (2001) recommend performing standard graphical and statistical checks of the residuals from the gene models.

2.2.2 Significance Analysis of Microarrays

Tusher et al. (2001) propose the Significance Analysis of Microarrays (SAM). In the SAM procedure, effect differences are estimated with a statistic called the relative difference. The definition of the relative difference depends on the experimental setup. For example, the relative difference for an experiment with two treatments can be defined as a two-sample t-statistic. For more complicated designs the relative difference can be defined as the researcher sees fit.

The relative differences are calculated for each gene and then ordered. The expected value of the order statistics is calculated by bootstrapping, a technique for assessing statistical accuracy (Good, 2001; Hastie et al., 2001). The order statistics are then compared to their expected values. The order statistics that fall outside the user-defined threshold are called significant. The user defines the threshold based on estimates of the False Discovery Rate (FDR).

One concern for detecting differentially expressed genes is whether or not the sample size is sufficient. Tibshirani (2005) suggests a method for assessing the sample size based off of the FDR and the False Negative Rate (FNR). The user can then get an idea of the power of the current experiment and the number of samples to prepare for future experiments.

The SAM procedure itself does not normalize the data, and Tusher et al. (2001) do not make any specific recommendations about normalization, but demonstrated the SAM procedure on a single-channel Affymetrix® array and calibrated the data using the aforementioned cube root method (page 7).

The SAM procedure requires that the data be organized in a $p \times n$ array where p is the number of genes and n is the number of samples. The SAM procedure will not tolerate missing values. Chu et al. (2005) recommend using a K-nearest neighbor algorithm called KNNimpute, proposed by Troyanskaya et al. (2001), in order to impute missing values.

The SAM procedure attempts to estimate the FDR as defined by Storey (2002) (Chu et al., 2005). SAM does not provide any new procedure for controlling the Type I error rate. The only difference between SAM and other procedures which reject the

null when the tests statistic is more extreme than a constant is the use of asymmetric critical values (Dudoit et al., 2003).

2.2.3 Other Methods

Several approaches using linear models have been proposed to analyze microarray data (see Smyth (2004) for a nice summary). Kerr et al. (2000) propose a general analysis of variance model for the logs of the original fluorescence measurements; however, most approaches involve a separate model for each gene. Chu et al. (2002) propose a mixed model for single-channel experiments. Lönnstedt and Speed (2002), Dudoit et al. (2002), and Smyth (2004) all propose models involving only a single variance term.

Smyth (2004) proposes a technique that builds on the work of Lönnstedt and Speed (2002). It involves hierarchical parametric models using empirical Bayes methods. The method produces moderated t-statistics in which posterior residual standard deviations are used in place of ordinary standard deviation.

Other approaches for analyzing microarrays include discriminant methods, such as nearest neighbor classifiers, linear discriminant analysis, and classification trees. A thorough discussion of these methods can be found in Dudoit et al. (2000). A statistically inferior but common method in assessing gene expression is to consider the fold change. In this method, genes are assessed using a $\log_2$ scale. Genes with high fold change are considered significant without regard to the within-gene variation. Therefore, a gene with high expression and high variation may be considered significant, whereas a gene with low expression but low variation may be ignored.

Tanaka et al. (2000) demonstrate that selecting significant genes based solely on fold change leads to false positives and false negatives.

2.3 Multiple Testing

The objective of most microarray experiments is to determine which genes are differentially expressed. In a typical microarray experiment, thousands of genes are simultaneously tested under the null hypothesis that there is no association between the expression levels and the treatments. The biological question of differential expression can therefore be restated as a problem in multiple hypothesis testing. Microarray experiments present new challenges to the application of multiple testing procedures because of the sheer number of tests (Dudoit et al., 2003).

Consider the problem of testing m null hypotheses where m_0 are true and m_1 are not. Table 2.1 summarizes the number of errors that can be committed when testing these hypotheses. The m hypotheses are assumed to be known in advance. The random variables W and R are observable, whereas the random variables U, V, T, and S, are not (Benjamini and Hochberg, 1995).

Number of	**Number not rejected**	**Number rejected**	
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	m_1
	W	R	m

Table 2.1: Outcomes when testing m null hypotheses.

Two types of errors can be committed in a significance test. A Type I error occurs when a gene is called significantly different when it truly is not. A Type II error occurs when a gene is not called significantly different when it truly is. In Table 2.1 the number of Type I errors is represented by V and the number of Type II errors is represented by T. The FWER is the probability of at least one Type I error. In context of Table 2.1, FWER = $\mathrm{P}(V \geq 1)$.

Controlling the number of errors is important when testing many hypotheses. Researchers usually attempt to control the FWER, but there are other less standard Type I error rates such as the Per-Comparison Error Rate and the Per-Family Error Rate (Shaffer, 1995).

There are two levels of control of the Type I error rate. The first, strong control, controls the Type I error rate when any subset of the null hypotheses is true. The second, weak control, controls the Type I error rate only when all the null hypotheses are true. In microarray settings, where it is fairly certain that some null hypotheses are not true, it is important to have strong control (Dudoit et al., 2003).

Several approaches have been proposed to control the Type I error rate. The most common is perhaps the Bonferroni which provides strong control of the FWER. However, the Bonferroni approach can be too conservative in some settings. A less conservative approach is the Šidák procedure (Šidák, 1967), which provides weak control of the FWER under certain circumstances. The Šidák procedure assumes that the p-values are independently distributed uniformly; however, microarray experiments tend to yield p-values that are dependent and have a skewed distribution. The most conservative approach, therefore, is to use the Bonferroni method.

Another useful technique for controlling the Type I error rate is adjusting p-values through step up or step down procedures. These procedures consider the order of the p-values when making adjustments. Westfall and Young (1993) show how to calculate adjusted p-values through resampling. Dudoit et al. (2003) prefer the use of adjusted p-values. One advantage of reporting adjusted p-values is not having to determine the level of the test in advance. Step down procedures provide strong control of the FWER (Westfall and Young, 1993).

Bayesian approaches to the multiple testing problem in microarray analysis also exist. An empirical Bayes approach can be found in Efron and Tibshirani (2002).

2.3.1 The False Discovery Rate

A modern approach to the multiple testing problem is to consider the False Discovery Rate (FDR). The FDR is the expected proportion of Type I errors among the rejected hypotheses (Dudoit et al., 2003). Controlling the FDR is of practical utility when the desire is to produce a large list of significant genes that contain a relatively small number of false positives. In other words, considering the FDR can help researchers identify a large group of significant genes while still providing some control over error rates.

Benjamini and Hochberg (1995) pioneered the work on the False Discovery Rate. Their methodology provides weak control of the Type I error rate. Storey and Tibshirani (2001) later proposed the Positive False Discovery Rate (pFDR). The pFDR provides strong control of the Type I error rate. The difference between the FDR and the pFDR is subtle. The FDR proposed by Benjamini and Hochberg (1995)

is calculated based on the probability that there is at least one significant gene. The pFDR proposed by Storey (2002) assumes that this probability is essentially equal to one. Specifically, the FDR and pFDR are defined as

$$\text{FDR} = E\left(\frac{V}{R}\middle| R > 0\right)\text{P}(R > 0) \tag{2.1}$$

$$\text{pFDR} = E\left(\frac{V}{R}\middle| R > 0\right). \tag{2.2}$$

The FDR and the pFDR are asymptotically equivalent for a fixed rejection region, so for large m the two estimates are basically equivalent. This is due to the fact that for large m $\text{P}(R > 0) \approx 1$ (Storey, 2002). The pFDR is preferred because it has greater power, becomes better as the number of tests increase, provides better estimates for small rejection regions, and provides strong control of the Type I error rate. In this document the term FDR will refer to the pFDR defined by Storey and Tibshirani (2001).

One advantage of the FDR is that it is valid when there are dependencies between p-values (Benjamini and Yekutieli, 2001). The p-values in microarray experiments tend to be dependent due to genetic pathways that interact in order to produce some overall process (Storey and Tibshirani, 2001). The FDR can be estimated by using resampling techniques (Good, 2001) that take these dependencies into account.

Efron (2004) has modified the FDR by using empirical Bayes methods to estimate it in an infinitesimal interval. His method calculates what he calls the local

FDR. The local FDR is closely related to the FDR developed by Benjamini and Hochberg (1995).

2.3.2 The Miss Rate

The *Miss Rate* (MR) complements the FDR. Whereas the FDR measures the proportion of false positives, the MR measures the proportion of false negatives in an interval near the rejection region. In other words, the MR is the proportion of genes that are truly differentially expressed among those declared non-significant for a specified interval (Taylor et al., 2005).

The quantity $E(T/W)$ is what Genovese and Wasserman (2003) call the False Non-Discovery Rate or what Tibshirani (2005) calls the False Negative Rate (FNR). This quantity is typically not of practical importance since most of the expression values in T will tend toward zero. A more useful quantity is the proportion of false negatives in an interval, (c_0, c), near the rejection region.

	$\lvert T_i\rvert$ in (c_0, c)	$\lvert T_i\rvert > c$(Reject)
Null true	U_0	V
Alternative true	T_0	S
	W_0	R

Table 2.2: Outcomes when testing hypotheses in the interval (c_0, c).

Table 2.2 displays various outcomes for testing m genes the interval (c_0, c). The MR is defined as

$$\mathrm{MR}(c_0, c) = E\left(\frac{T_0}{W_0}\right). \tag{2.3}$$

The MR is a cautionary statistic that is to be used when evaluating the False Discovery Rate. It has been shown that when the FDR is very low, the MR can be quite high (Taylor et al., 2005). Therefore, it is advisable that the MR be routinely reported along with the FDR.

2.3.3 Tail Strength

Taylor and Tibshirani (2006) propose an overall measure of significance for a set of hypothesis tests. Their method compares the distribution of p-values to the uniform distribution and calculates a measure called the Tail Strength.

The TS can be thought of as roughly the percentage of significant test statistics due to chance. For example, if $TS = 0.655$, then there are 65.5% more significant test statistics on average than would be expected by chance alone. A uniformly distributed set of p-values would have a TS value close to 0. On the other hand, as p-values bunch up near 0, the TS approaches 1.

In the SAM procedure, the null distribution is simulated through random permutation of the data. Thus, the ordered test statistics can easily be converted to p-values by taking advantage of the simulated null distribution. Taylor and Tibshirani (2006) suggest calculating p-values as the proportion of test statistics in the null distribution that exceed the observed test statistic. Once the test statistics have been converted to p-values the TS can be calculated.

Chapter 3

Materials and Methods

This project compares two popular methods for analyzing genomic data. The first method is a mixed models approach, and the second is the Significance Analysis of Microarrays. These analyses are compared using microarray data from a mouse pregnancy study.

3.1 Mixed Models Approach

The mixed models approach uses mixed models to both normalize and analyze microarray data. A mixed model is a linear model that includes both fixed and random effects. Fixed effects are assumed to be constant. Random effects are assumed to be randomly selected from an infinite population. In their normalization and gene models, Wolfinger et al. (2001) specified fixed treatment effects and random array effects. The treatment effects were assumed to be constant while the array effects were assumed to be randomly selected from all possible array effects.

The mixed models approach is carried out using SAS Proc Mixed® (Littell et al., 2006). The random effects, including the error term, are assumed to come from a normal distribution. The Mixed Procedure estimates the covariance structure

using a restricted maximum likelihood (REML) method. One advantage of the REML method is that it accommodates missing data.

The mixed models approach requires two separate model specifications, one for the normalization model and another for the gene model. The normalization model is applied to the entire data set. It includes effects due to the experimental conditions, not effects due to the actual genes themselves. Because the normalization model is applied to the entire data set, it is computationally advantageous to specify a fairly simple model. Models that include many terms or complex variance structures can take days to run.

The residuals from the normalization model are passed to the gene models, and a separate gene model is run for each gene. The gene model includes effects that are due to the actual genes themselves. The gene models are computationally intensive, because a separate REML optimization problem is solved for each gene. Furthermore, the number of genes that can be analyzed sequentially depends on the amount of available memory. Again, specifying fairly simple models is computationally advantageous.

The mixed models approach identifies significant genes by testing the pairwise treatment comparisons in the gene model. Due to a large number of genes in any given microarray experiment, the number of pairwise comparisons can number in the tens of thousands. Therefore, Wolfinger et al. (2001) suggest using a Bonferroni correction to control the FWER.

3.2 Significance Analysis of Microarrays

The Significance Analysis of Microarrays compares ordered test statistics to their expected values. The expected order statistics are estimated using resampling. The calculations for the SAM procedure are carried out in the R environment using the SAMR package. If desired, the SAM procedure can be manipulated using a Microsoft Excel® interface. Details for the R package and the Excel® interface can be found at www-stat.stanford.edu/∼tibs/SAM.

The SAM procedure requires the data to be in a $p \times n$ array where p is the number of genes and n is the number of samples. Thus, there is one measurement per gene per sample. Furthermore, there can be no missing data. Missing data can be imputed using the IMPUTE package for the R environment.

The SAM procedure itself does not do any normalization. A common normalization procedure is the print-tip lowess normalization proposed by Yang et al. (2002). This procedure is part of the SMA package for the R environment. It has five steps: (1) subtract out the background, (2) calculate the M-values and the A-values, (3) apply lowess curves to the MA-plot for each print-tip, and (4) use the residuals as normalized values. After normalizing each array, the expression values for all arrays should be examined. If there are substantial scaled differences between arrays, then a scale normalization can be considered.

Once the data has been normalized and fit into a $p \times n$ array with no missing values, it can be passed through the SAM procedure. There are seven basics steps in the SAM procedure: (1) compute the observed order statistics for all genes, (2)

estimate the expected order statistics, (3) plot the observed order statistics versus the expected order statistics, (4) choose an arbitrary cutoff (Δ), (5) find the significant genes above and below Δ, (6) estimate the FDR, and (7) calculate the MR.

3.2.1 Compute the observed order statistics for all genes.

In the SAM procedure, effect differences are estimated with a statistic called the relative difference. It has the following form:

$$d_i = \frac{r_i}{s_i + s_0} \ , \tag{3.1}$$

where r_i represents the score, or unstandardized change in gene expression, s_i represents the gene specific scatter, or the standard deviation of repeated gene expression, and i is a gene index. The quantity s_0 is the fudge factor or exchangeability factor and is used in an attempt to make r_i and s_i independent, particularly at low expression levels.

The form of d_i depends on the type of data. For a two-class unpaired experiment the relative difference is basically a two sample t-test statistic. For a multi-class experiment, the relative difference is defined in terms of Fisher's linear discriminant. Fisher's linear discriminant is the linear combination that maximizes the between-class variance with respect to the within-class variance (Hastie et al., 2001). In the SAM procedure the between-class variance (r_i) and the within-class variance (s_i) are

defined as

$$r_i = \sqrt{\{\sum n_k / \prod n_k\} \sum_{k=1}^{K} n_k(\bar{x}_{ik} - \bar{x}_i)^2}, \tag{3.2}$$

$$s_i = \sqrt{\frac{\sum 1/n_k}{\sum (n_k - 1)} \sum_{k=1}^{K} \sum (x_{ij} - \bar{x}_{ik})^2}. \tag{3.3}$$

It is important to note that r_i and s_i for the multiclass design are positive, and therefore d_i can only be positive. This is somewhat unique in that the other functions utilized by SAM can take on both positive and negative values.

The variance of d_i tends to be high for low expression values. The fudge factor s_0 is estimated in order to keep r_i independent of s_i (Tusher et al., 2001). The value s_0 is calculated as the α^{th} percentile of s_i that minimizes the coefficient of variation for the median absolute deviations of d_i. The technique for adding a fudge factor to the denominator is outlined in Storey (2002). Calculation details can be found in Appendix A.2.

It is important to note that (1) there are other ways to make the distribution of d_i independent of the levels of s_i, (2) the SAM software does not warn the user when s_0 has reached the upper bound, and (3) the SAM software does allow the user to override the calculation of s_0. For example, the user can set s_0 to zero if no adjustment to the denominator is desired. Finally, the relative differences, d_i, are ordered to create the observed order statistics $d_{(i)}$.

3.2.2 Estimate the expected order statistics.

Data for a typical SAM experiment can be thought of as a $p \times n$ array where p is the number of genes and n is the number of samples. Note here that typically there are several thousand genes (p), whereas there are typically only a handful of samples (n).

The expected ordered statistics are estimated from random permutations of the columns in the $p \times n$ array. There are five steps in creating the expected order statistics: (1) randomly permute the samples (or columns), (2) calculate the test statistic d_i^* for each gene, (3) order the test statistics statistics $d_{(i)}^*$, (4) repeat B times, and (5) estimate the expected order statistics with $\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$.

In the SAM procedure the samples (or columns) of the data are randomly permuted and the order statistics are recalculated. This process is repeated a sufficient number of times in order to estimate the distribution of the expected value of the order statistics. The random permutations preserve the correlation structure between rows.

3.2.3 Plot observed against expected order statistics.

The observed order statistics ($d_{(i)}$) are plotted against the expected order statistics ($d_{(i)}^*$). If there is no treatment effect, then the plot will fall roughly on a diagonal line. If there is a treatment effect the plot will deviate from the expected line. The genes that fall furthest from the diagonal line indicate significance.

3.2.4 Choose an arbitrary cutoff Δ

Let Δ be the maximum allowable absolute difference between an observed statistic and its expected statistic. For any gene i where $|d_{(i)} - \bar{d}_{(i)}| > \Delta$, gene i is called significant. The user chooses the cutoff with respect to an acceptable FDR. The SAM procedure produces a table showing the relationship between Δ, the FDR, and the number of genes called significant.

3.2.5 Find the significant genes above and below Δ

The genes above and below Δ are declared significant. In the multiclass method all values $d_{(i)}$ are positive, and there is only one cutoff. All values above the cutoff in the multiclass method are declared significant.

3.2.6 Estimate the FDR

The FDR is an estimate of the percentage of false positives. It is calculated by dividing the number of false positives by the number of genes called significant. In other words, SAM estimates the FDR in the following way:

$$\widehat{\text{FDR}} = \frac{\text{Number of false positives}}{\text{Number of significant genes}} \ . \tag{3.4}$$

The number of significant genes is based on the user's choice of Δ. The number of false positives, however, has to be estimated. It is calculated by determining the number of genes in each of the B random permutations with values exceeding Δ (there will be B values). The estimated number of false positives is the median value

from the B permutations. Alternatively, the 90^{th} percentile of the B permutations can be used.

The number of false positives tends to be biased upwards (Storey and Tibshirani, 2001). This is due to the fact that the method used to calculate the number of false positives assumes that there is no treatment effect and no truly significant genes. Typically, however, there is a treatment effect and significant genes are expected to be encountered. Therefore, the number of genes expected to be called significant not only includes the number of false positives but the number of genes that truly are significant.

In order to correct for the upward bias, the proportion of truly null genes (i.e., unaffected genes), π_0, is estimated. If (q_{25}, q_{75}) represent the range of values determined by the first and third quartiles of the expected order statistics, then the number of observed order statistics expected to be found in this interval is one half the total number of genes. The proportion of true null genes, π_0, is estimated as the number of observed order statistics in the (q_{25}, q_{75}) interval divided by half the total number of genes. In mathematical notation, $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\}/(p/2)$. The SAM procedure does not allow for alternative quantiles to be used. If $\hat{\pi}_0$ exceeds one, then $\hat{\pi}_0$ is set to one (Chu et al., 2005). The number of false positives is then multiplied by the estimated proportion of true null genes $\hat{\pi}_0$.

Storey and Tibshirani (2003) propose an alternative method for calculating $\hat{\pi}_0$ based on p-values alone. The approach assumes that the distribution of p-values under the null hypothesis is uniform. The proportion of null p-values then is estimated from the *level* or *flat* part of the distribution of actual p-values. Specifically, if the

tuning parameter λ is the value where the distribution of p-values levels off, then $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i=1,\ldots,m\}}{m(1-\lambda)}$. The FDR estimated by the SAM procedure is equivalent to setting $\lambda = 0.5$ (Storey and Tibshirani, 2001).

As an example calculation of the FDR, consider the unpaired two-class randomly generated data set in which there are 20 samples of 1000 genes that are permuted 100 times. Suppose the middle half of the permuted order statistics $d^{*b}_{(i)}$ fall between the range of values $(-0.27, 0.27)$. However, only 472 genes with observed order statistics $d_{(i)}$ fall in that same range. Therefore the estimated proportion of true null genes is $\hat{\pi}_0 = 472/500 = 94\%$. In other words, about 94% of the genes are not significant. Furthermore, suppose that when $\Delta = 0.2$ the number of significant genes is 66, and the median number of false positives from the 100 permutations is 15. The False Discovery Rate then is estimated as $\widehat{FDR} = 95\% \times 15/66 = 21\%$. In other words, 21% of the genes found at $\Delta = 0.2$ are estimated to be false positives.

3.2.7 Calculate the Miss Rate

The Miss Rate (MR) is the proportion of genes that are truly differentially expressed among those declared non-significant in an interval near the cutoff determined by Δ. The SAM procedure estimates the MR by the following:

$$\widehat{\mathrm{MR}}(c_0, c) = 1 - \hat{\pi}_0 \frac{\widehat{U}_0}{W_0} \ , \tag{3.5}$$

where $\hat{\pi}_0$ is the estimated proportion of true null genes, W_0 is the total number of genes declared non-significant, and $\widehat{U}_0$ is the estimated number of false negatives.

The interval (c_0, c) defines an interval near the cutoff. The quantity W_0 is observed, whereas the quantity $\widehat{U}_0$ is estimated from the random permutations used in the SAM procedure (Taylor et al., 2005).

It should be noted that the Miss Rate (along with the FDR and local FDR) are unidentifiable unless π_0 is known. This is the reason for finding conservative estimates of π_0 which will in turn lead to conservative estimates of the MR as well as the FDR. The method used in SAM provides conservative estimates of π_0 (Storey and Tibshirani, 2001).

Chapter 4

Mouse Pregnancy Data

Microarray experiments are conducted to identify differential gene expression. Researchers wish to know which genes are involved with which treatments. For example, a gene that is significantly expressed when exposed to one treatment and not to another indicates that the significant gene and treatment combination is involved in some underlying biological pathway. The mouse pregnancy data was designed to find which genes and treatments influence pregnancy.

4.1 Treatments

The mouse pregnancy data came from an experiment performed in 2003 at the University of Utah. There were two types of mice. The wild type mice were genetically unaltered and served as the control. The knock-out mice were missing a pregnancy-related gene that had been genetically removed. The mice were then exposed to either a treatment or a placebo. The treatment was lipopolysaccharide (LPS), a molecule that induces labor. The placebo was a saline solution (SAL). The reference type mice were exposed to neither the treatment nor the placebo. Ribonucleic Acid (RNA) was then sampled from each mouse and applied to a microarray containing known genes.

The RNA combined to the genes on the microarray and the expression value for each gene was recorded.

Five types of mice were used in the experiment, one for each treatment combination. The treatment combinations were: wild type mice receiving lipopolysaccharide (LPS:B6), knock-out mice receiving lipopolysaccharide (LPS:KO), wild type mice receiving saline (SAL:B6), knock-out mice receiving saline (SAL:KO), and reference mice (REF). Messenger RNA (mRNA) was harvested from each type of mouse and reverse transcribed to form cDNA. The cDNA was then tagged using a fluorescent dye. The cDNA from each treatment was tagged with a green dye (Cy3) and hybridized to one of the arrays. The mRNA from separate reference mice was also harvested and reverse transcribed to form cDNA. The cDNA from the reference was tagged with a red dye (Cy5) and applied to all the arrays.

4.2 Microarrays

The mouse pregnancy data tested 21,529 unique genes. The entire complement of genes did not fit onto one array, so three slides (A, B, and C) were used instead of one. The genes were spotted onto the microarrays using a print-tip that can lay down several gene spots at once. Slides A and B were spotted using a 20×20 print-tip. In other words, 400 probes were spotted each time the print-tip touched the array. Slide C was spotted using a 16×21 print-tip. All three slides were printed in a 12×4 array. In other words, the print-tip groups were printed in 12 rows and 4 columns.

In general, a gene was tested on a single slide. However, a significant number of genes were tested on multiple slides. Figure 4.1 shows the gene pattern for slides

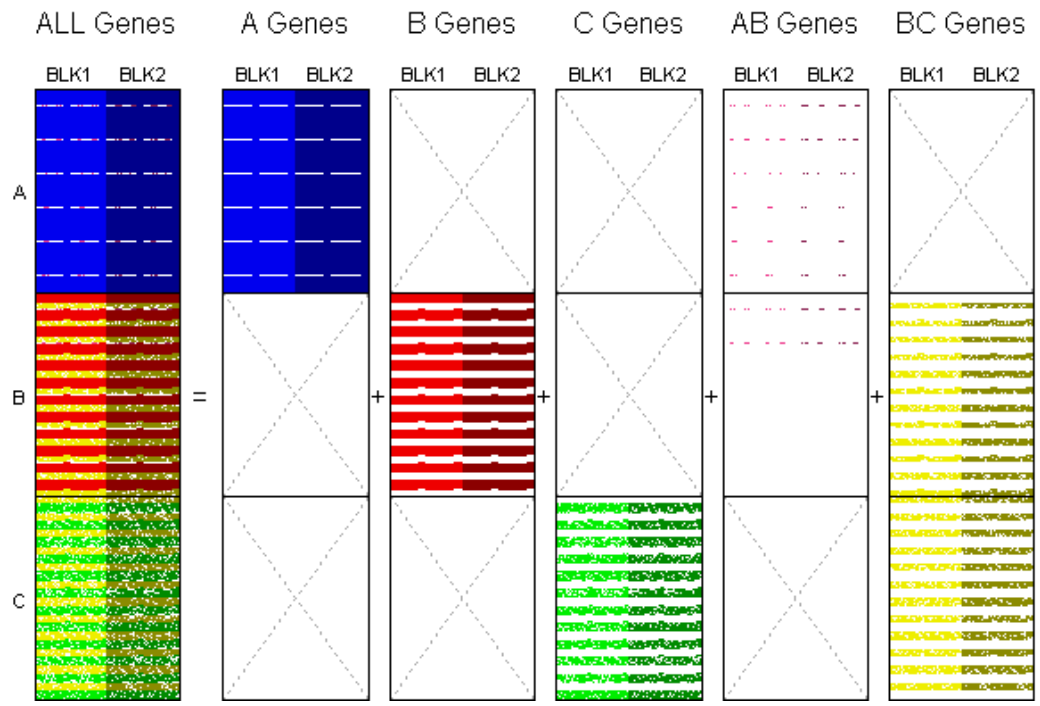

Figure 4.1: Gene pattern for one group of three slides. The gene spots were laid out in two side-by-side blocks, creating at least two replicates of each gene spot. Some gene spots were replicated four times, twice on one array and twice on another. Reference spots (white) also appear on each of the slides.

A, B, and C. The gene pattern on the left half of each slide was duplicated on the right half, forming two blocks on each slide.

Table 4.1 shows the number of genes replicated on each slide. Most gene spots (18,618) were replicated twice on each slide, appearing once in each block. Some spots (2,910) were replicated four times, appearing twice on one slide and twice on another. One spot was replicated 10 times.

Each slide was printed with a significant number of reference spots as well as known genes. These reference spots can be used during the normalization step of

Reps	A Genes	B Genes	C Genes	AB Genes	BC Genes	Total Genes
2	9,214	5,751	3,653	.	.	18,618
4	1	1	.	44	2,864	2,910
10	.	1	.	.	.	1
Total	9,215	5,753	3,653	44	2,864	21,529

Table 4.1: Gene replication. Most genes were replicated twice (once on each block). Some genes were replicated four times, and one gene was replicated 10 times. Slide A contained far more genes than slides B and C.

analysis, but do not correspond to genes of interest. Table 4.2 shows the number of reference spots in each slide. Roughly 10% of the overall spots were reference spots.

Slide	Gene Spots	Reference Spots	Total Spots
A	18,520	680	19,200
B	17,332	1,868	19,200
C	13,034	3,094	16,128
Total	48,886	5,642	54,528

Table 4.2: Reference spots. Each slide contained reference spots that were not associated with genes of interest. The reference spots aid in the normalization process.

4.3 Measurements

Each group of three slides was replicated 16 times, making 48 arrays in all. The arrays were organized into five groups, one group for each treatment. Each treatment (cDNA tagged with green dye) was then hybridized to one group of arrays. The (LPS:B6) treatment was hybridized to arrays $1-12$, the (LPS:KO) treatment was hybridized to arrays $13-24$, the (SAL:B6) treatment was hybridized to arrays $25-36$, the (SAL:KO) treatment was hybridized to arrays $37-45$, and the reference

treatment was hybridized to arrays $46-48$. The reference treatment tagged with red dye was hybridized to all the groups (arrays $1-48$). Thus, each array received one of five treatments (green dye) and the reference treatment (red dye).

The mouse pregnancy data consisted of genes g ($g = 1, \ldots, 21529$), dyes d ($d = 1, 2$), treatments t ($t = 1, \ldots, 5$), arrays a ($a = 1, \ldots, 48$), and gene spots $s(a)$. After the cDNA hybridized with the genes spotted on the microarrays, a foreground and background image was measured for each dye color. The process resulted in four measurements per gene: red foreground (Rf_{gtas}), red background (Rb_{gtas}), green foreground (Gf_{gtas}), and green background (Gb_{gtas}). The background corrected $\log_2$ intensities (y_{gdtas}), the log ratios (M_{gtas}), and the log averages (A_{gtas}) were defined as,

$$
\begin{aligned}
y_{g1tas} &= \log_2(Gf_{gtas} - Gb_{gtas}), && (4.1)\\
y_{g2tas} &= \log_2(Rf_{gtas} - Rb_{gtas}), && (4.2)\\
M_{gtas} &= \log_2 \frac{Rf_{gtas} - Rb_{gtas}}{G_{gtas} - Gb_{gtas}} = y_{g1tas} - y_{g2tas}, && (4.3)\\
A_{gtas} &= \frac{y_{g1tas} + y_{g2tas}}{2}. && (4.4)
\end{aligned}
$$

Notice the background corrected $\log_2$ intensities require an extra subscript, d, for dye. This subscript is use to identify the measurement on arrays $45-48$ which were exposed to the reference treatment in both the red and green dyes. Also notice that the red dye ($d = 2$) will always be associated with the reference treatment ($i = 5$).

The expression values in the mouse pregnancy data were measured using a laser scanner and imaging software. The software measured the quality of each spot

and flagged those spots that did not conform to predefined tolerances. Reasons for flagging low quality spots include background contamination, signal contamination, and shape irregularity. Flags were widespread in the data. Approximately 82% of the genes had at least one flag for one of the expression values. Approximately 2% of the genes had flags for all expression values.

Examining arrays for spatial effects can be very important in a microarray analysis. The mouse pregnancy data showed severe distortions in the background images. Figure 4.2 shows the log ratios for the background measurements only, i.e. $\log_2(Rb_{gtas}/Gb_{gtas})$. There was a bubble-shaped distortion that appeared on all of the B slides. The A slides showed a few smear marks, and the C slides showed a few large spots.

Figure 4.3 shows the M_{gtas} values for the experiment. Most arrays were predominantly green, which indicates that the treatment was expressed more strongly than the reference. There was large variability between arrays.

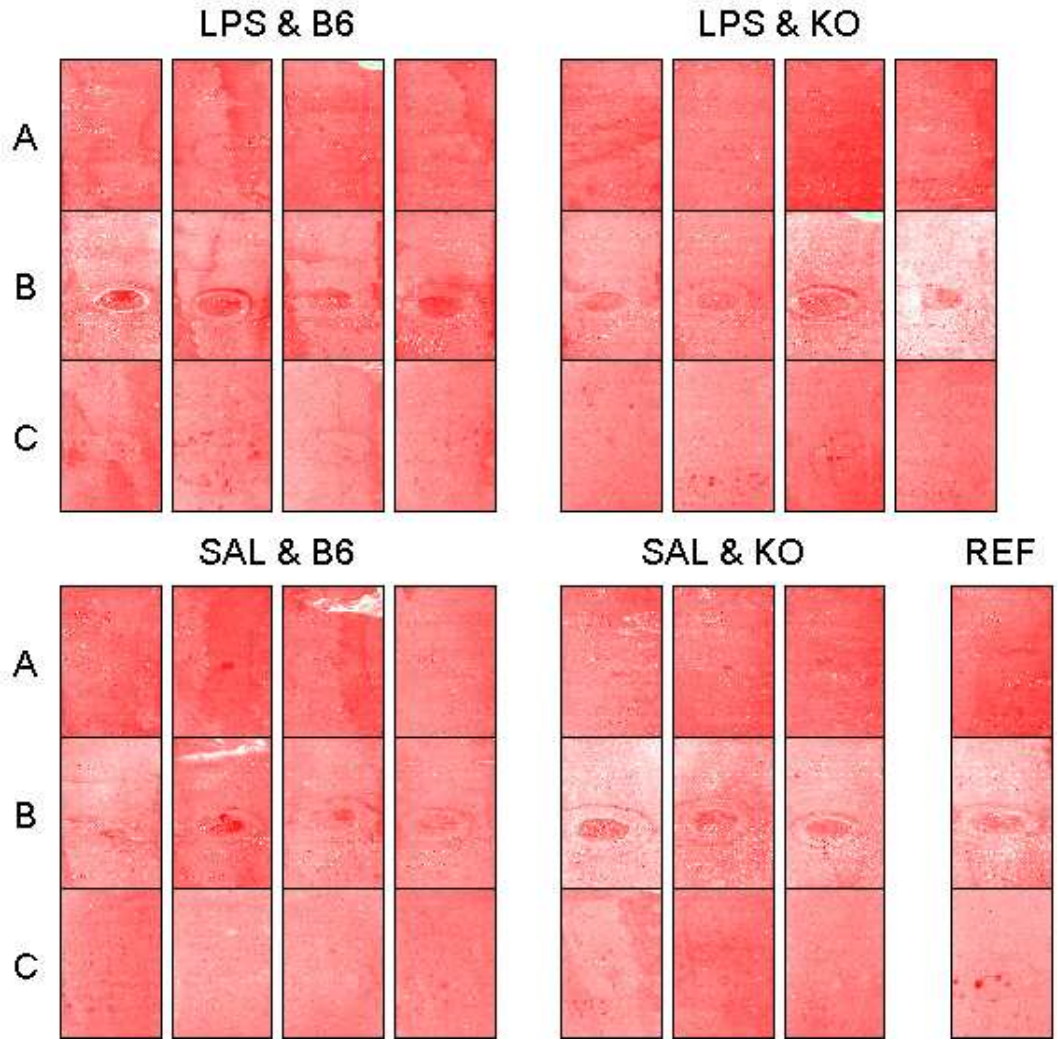

Figure 4.2: Background log ratios. The images show a bubble that runs through the B slides and other distortions in the A and C slides.

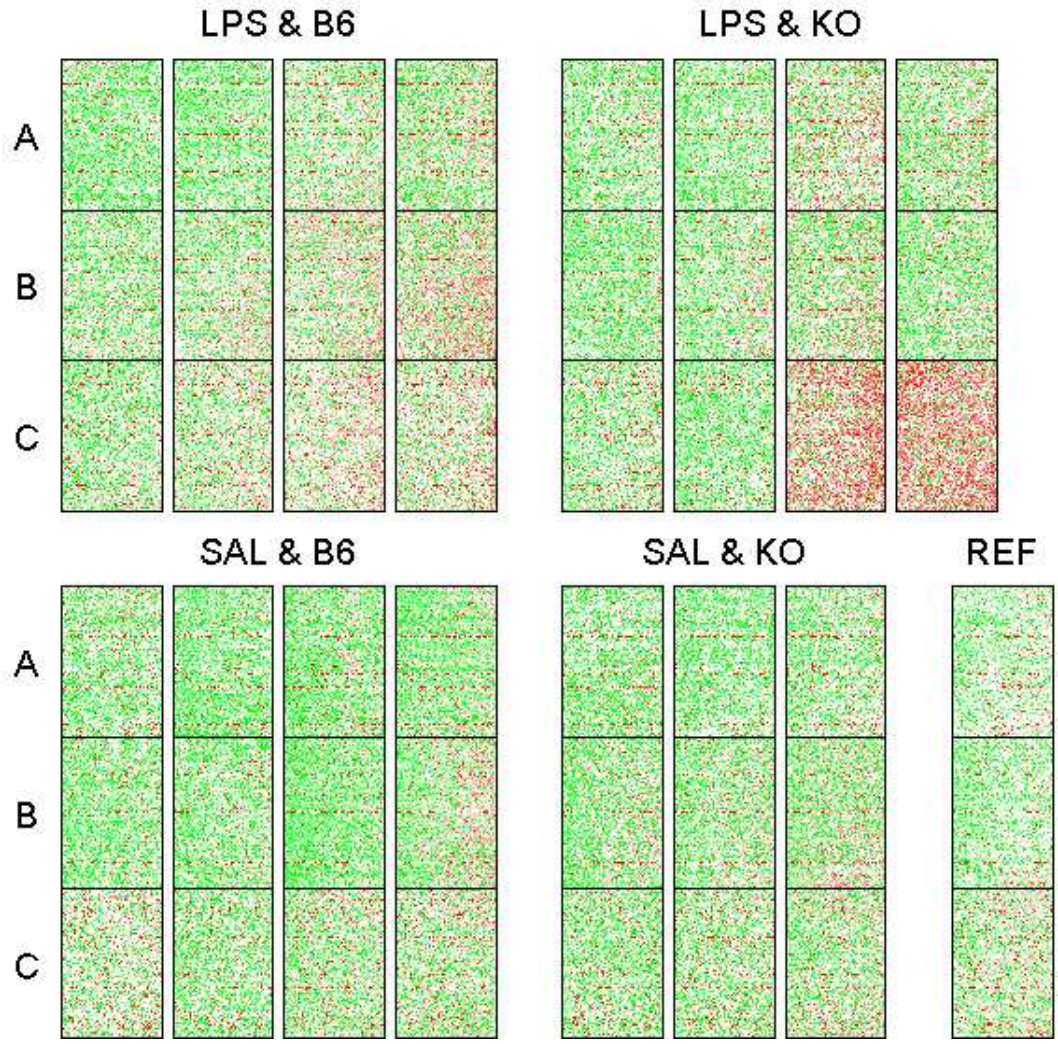

Figure 4.3: Log ratios. Most arrays appear predominantly green, meaning the treatment is expressed more strongly than the reference.

Chapter 5

Mixed Models Analysis

There were four basic steps in the mixed models analysis: (1) organize the data, (2) normalize the data, (3) run the gene models, and (4) identify significant genes. An investigation of treatment effects was also included as part of the mixed models analysis.

5.1 Organize the data

The background corrected $\log_2$ intensities were used in the mixed models analysis. Intensities with negative or zero expression (i.e. the background image was larger than or equal to the foreground image) were not used in the analysis, and flagged data were removed, as suggested by Wolfinger et al. (2001).

The mixed models approach requires the data to be in a tall format (i.e. one measurement per record) as opposed to an array format (i.e. multiple measurements per record). Therefore, each gene spot is associated with two records, one for the treatment (green dye) and one for the reference (red dye).

5.2 Normalize the Data

The purpose of the normalization model is to remove experimental effects that are common across all genes. In other words, normalization attempts to remove

main and interaction effects not associated with individual genes. The mixed model procedure uses one normalization model for the entire set of data. Let y_{gdtas} be the background corrected $\log_2$ intensity for gene g, dye d, treatment t, array a, and spot s. The normalization model was:

$$y_{gdtas} = \mu + T_t + A_a + (TA)_{ta} + \epsilon_{gdtas} \ , \tag{5.1}$$

where μ represents an overall mean value, T is the treatment effect, A is the array effect, and TA is the interaction effect of arrays and treatments. The treatment effect T is a fixed effect. The other effects A, TA, and ϵ are random effects which are distributed normally with mean 0 and variances σ^2_A, σ^2_{TA}, and σ^2_ϵ respectively. The fixed effect T and the random effects A and TA were all highly significant. The normalized values are the residuals from this model, $r_{gdtas} = y_{gdtas} - \hat{y}_{ta}$.

Figure 5.1 shows the normalized intensities by array. Ideally, all the box plots would show similar variances and distributions. However, the reference intensities tend to have a larger interquartile range. The distributions also tend to be negatively skewed.

5.3 Run the Gene Models

After the expression values were normalized, the effects due to each gene were assessed. A separate model was fit to the data associated with each unique gene. Each model had the same form. Let r_{gdtas} be the normalized, background corrected

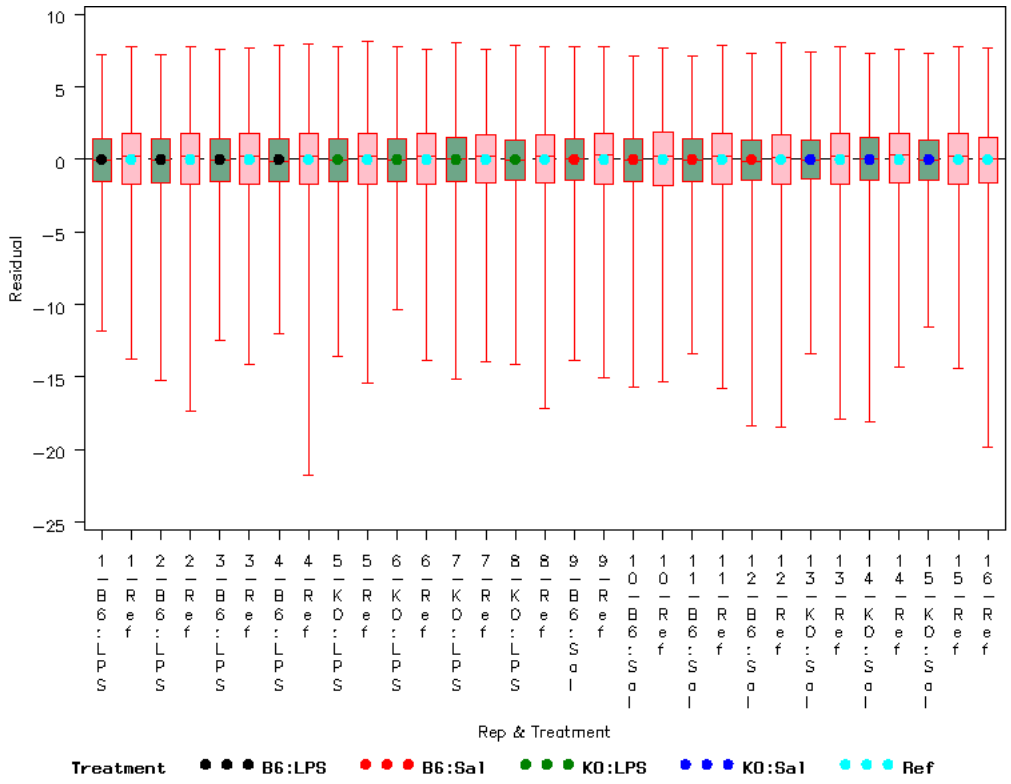

Figure 5.1: Normalized gene intensities by array for the mixed models analysis. The distributions tend to be negatively skewed. The reference distributions tend to have larger interquartile ranges.

$\log_2$ intensities from Model 5.1. The gene model was:

$$r_{gdtas} = G_g + (GT)_{gt} + (GA)_{ga} + (GS)_{gs(a)} + \gamma_{gdtas} \ , \tag{5.2}$$

where G represents the overall mean value for gene g, GT is the fixed effect for treatment t, GA is the random effect for array a, and GS is the random effect for spot. Including G before the effects indicates that all these effects are with respect to gene g. The random effects GA, GS, and γ are distributed normally with mean 0

and variances $\sigma^2_{GA_g}$, $\sigma^2_{GS_g}$, and $\sigma^2_{\gamma_g}$ respectively. The mixed models approach assumes heterogeneous variances among genes.

There were 21,529 unique genes in the mouse pregnancy data. There were 480 genes removed due to negative expression and/or poor spot quality. An additional 338 genes were removed due to insufficient data (i.e. observations with so many negative values and flags that the the parameters could not be estimated). Finally, 908 genes produced F-statistics for the treatment effect that were equal to zero. These genes had zero degrees of freedom for estimating the error. In all, the mixed models approach successfully analyzed 19,803 genes.

Table 5.1 shows the degrees of freedom for a typical gene model. The degrees of freedom for each gene model varied depending on the gene. and some genes were replicated twice as often as other genes. Also, some genes were missing observations due to poor quality or negative expression.

Source	DF
Intercept	1
Treatment	4
Array	15
Spot(Array)	15
Error	29
Total	64

Table 5.1: Degrees of freedom for a typical gene model. Some genes have more degrees of freedom due to extra replication, while others have fewer degrees of freedom due to missing values.

5.4 Identify Significant Genes

All pairwise comparisons for the five treatments were computed on the resulting 19,803 genes. There were $\binom{5}{2} = 10$ comparisons for genes with all five treatments. There were fewer comparisons for genes with missing observations. In all, there were 194,462 pairwise comparisons. A t-test was performed on each comparison. A Bonferroni adjustment was used to calculate the corrected alpha as $\alpha_c = 0.05/194{,}462 = 2.57 \times 10^{-7}$. The cutoff resulted in 9,238 significant pairwise comparisons from 3,542 unique genes. Table 5.2 shows sample output for the mixed analysis. Only the 20 most significant pairwise comparisons are included.

	Gene	Strain1	Strain2	Estimate	StdErr	DF	Tval	−Log10p
1	BG076245	B6:Sal	Ref	−3.55	0.05	55.00	−72.96	55.70
2	BG076245	B6:Sal	KO:LPS	−4.22	0.07	55.00	−61.60	51.71
3	BG076245	B6:LPS	B6:Sal	4.11	0.07	55.00	58.93	50.66
4	BG076245	KO:Sal	Ref	−3.11	0.06	55.00	−54.84	48.97
5	BG076245	KO:LPS	KO:Sal	3.79	0.07	55.00	50.81	47.18
6	BQ550593	B6:Sal	Ref	−3.31	0.08	60.00	−43.95	46.62
7	BG076245	B6:LPS	KO:Sal	3.67	0.08	55.00	48.57	46.13
8	BQ550593	B6:Sal	KO:LPS	−4.03	0.11	60.00	−38.27	43.13
9	BQ554441	B6:Sal	Ref	−1.85	0.04	56.00	−41.24	42.87
10	BQ550593	KO:Sal	Ref	−3.18	0.09	60.00	−36.70	42.08
11	BQ552581	B6:Sal	KO:LPS	−3.52	0.09	55.00	−40.08	41.67
12	BQ550593	B6:LPS	B6:Sal	3.79	0.11	60.00	36.00	41.60
13	BQ552581	B6:Sal	Ref	−2.40	0.06	55.00	−39.76	41.48
14	BQ553339	B6:Sal	Ref	−3.23	0.07	50.00	−43.92	40.88
15	BQ554441	B6:Sal	KO:LPS	−2.35	0.06	56.00	−37.75	40.79
16	BQ550593	KO:LPS	KO:Sal	3.90	0.11	60.00	34.30	40.40
17	BQ553335	B6:Sal	Ref	−2.14	0.06	57.00	−36.28	40.35
18	BQ552581	B6:LPS	B6:Sal	3.18	0.09	55.00	37.44	40.09
19	BQ554441	KO:Sal	Ref	−1.83	0.05	56.00	−36.38	39.93
20	BQ554441	KO:LPS	KO:Sal	2.33	0.07	56.00	35.12	39.11

Table 5.2: Output for the top 20 significant genes in the mixed analysis.

5.5 Investigate Treatment Effects

Table 5.3 shows the number of significant pairwise comparisons for each of the 10 treatment combinations. The highest number of significant differences occurred between the four treatments and the reference. A large number of significant comparisons occurred when comparing mice exposed to LPS versus saline. However, very few significant comparisons occurred when comparing wild type mice to knock-out mice.

	B6:LPS	B6:SAL	KO:LPS	KO:SAL	Total
B6:SAL	668	.	.	.	668
KO:LPS	17	1,023	.	.	1,040
KO:SAL	640	26	843	.	1,509
Ref	945	1,696	1,650	1,730	6,021
Total	2,270	2,745	2,493	1,730	9,238

Table 5.3: Significant pairwise comparisons.

Chapter 6

SAM Analysis

There were five basic steps in the SAM analysis: (1) organize the data, (2) Normalize the data, (3) aggregate, reduce, and impute the data, (4) run the SAM analysis, and (5) identify significant genes. An investigation of treatment effects was also included as part of the SAM analysis.

6.1 Organize the Data

The background corrected $\log_2$ ratios (i.e. M-values) were used in the SAM analysis. Log averages (i.e. A-values) were also used to normalize the data. Gene intensities with negative or zero expression (i.e. the background image was larger than or equal to the foreground image) were not used in the analysis (Smyth and Speed, 2003). Data flagged for poor quality were also left out of the analysis (Wolfinger et al., 2001).

6.2 Normalize the Data

The spots on the mouse pregnancy microarrays were printed in 48 print-tip groups. The data were normalized using print-tip lowess normalization. The A-values were compared to the M-values by print-tip group. Independent lowess curves were

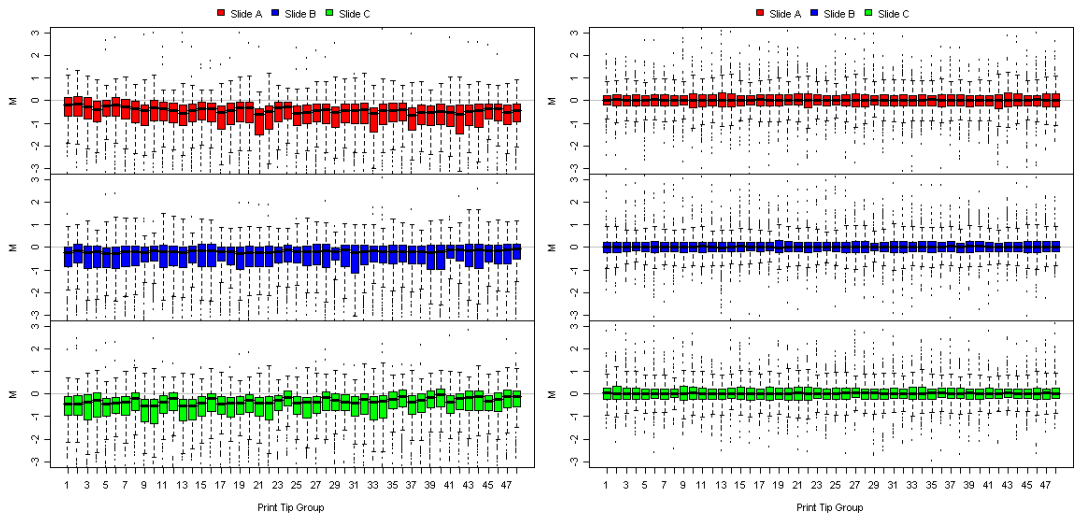

Figure 6.1: Non-normalized print-tip M-plot (Left) and normalized print-tip M-plot (Right) for the first three arrays. The non-normalized plot shows patterns in the expression values for every four print-tip groups. The normalized plot shows expression values that look similar across print-tip groups and between slides.

fit for each print tip group and the residuals were then calculated. The residuals constitute normalized expression values.

Figure 6.1 (Left) shows the non-normalized M-values for the 48 print-tip groups in the first three arrays (i.e. slides A,B, and C in the first sample). There appears to be a pattern in every four print-tip groups, especially for slides A and C. This is probably due to the fact that the microarrays were printed 4 groups wide by 12 groups high. The normalized M-values for the 48 print-tips can be seen in Figure 6.1 (Right). The expression values now appear to be independent of print-tip group. Furthermore, the expression values look similar between slides.

The non-normalized M-values and A-values for all 16 replications can be seen in Figure 6.2 (Left). Notice the strong relationship between the A-values and the M-values. Ideally, the expression difference would be independent of the expression

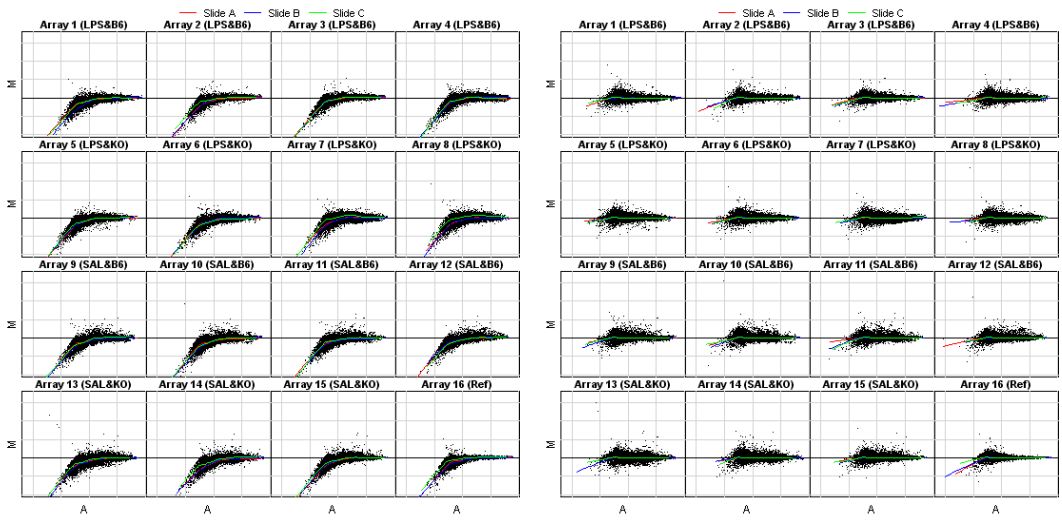

Figure 6.2: Non-normalized (Left) and normalized (Right) MA-plots for all 16 replications. The non-normalized plot shows that small M-values appear to be highly correlated with small A values. The normalized plot shows that the M-values and A-values appear to be much less dependent.

average. However, Figure 6.2 (Left) indicates that M-values tend to be small when the A-values are small. The normalized M-values and A-values can be seen in Figure 6.2 (Right). A slight dependency remains between the log averages and the log ratios. In general, however, the M-values and A-values appear to be less dependent than in Figure 6.2 (Left).

Print-tip normalization is technically a within-array normalization technique. However, it is often sufficient for between-array normalization as well. Figure 6.3 (Left) shows the normalized M-values for all slides. The M-values for each slide center around zero. However, the M-values tend to have different inter-quartile ranges between arrays but not between slides. Each print-tip group for each array in each slide was normalized independently. The pattern between arrays, therefore, suggests that there is an array effect but not a slide effect.

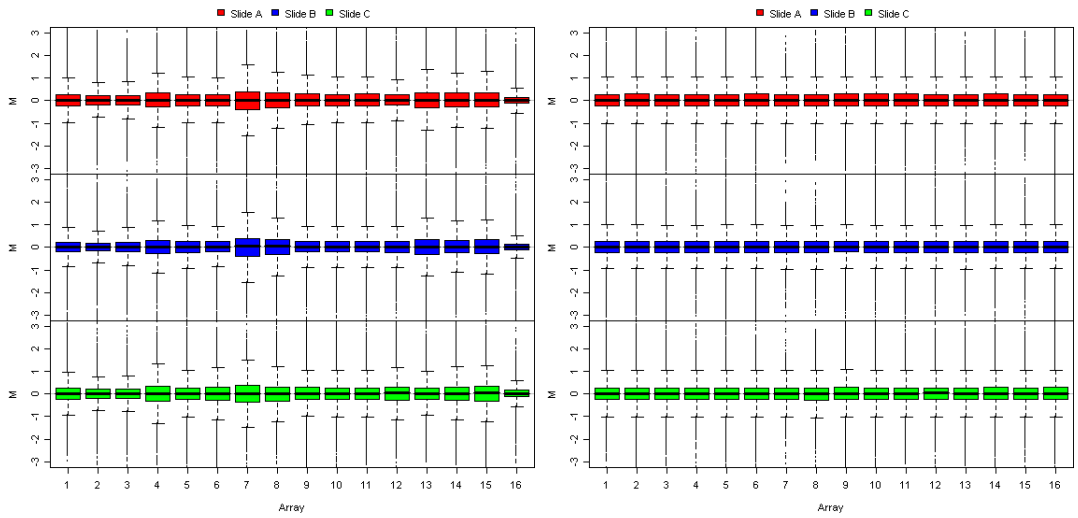

Figure 6.3: Normalized M-plot (Left) and normalized between-array M-plot (Right). The normalized M plot shows large variation between arrays. The normalized between-array plot shows the arrays after the arrays have been scaled to have equal median absolute deviations.

The M-values can be normalized between arrays; however, it has been argued that normalizing between arrays can create more noise than it resolves (Smyth and Speed, 2003), but when there are substantial differences between samples, the samples can be scaled so that they all have the same median absolute deviation. The mouse pregnancy data are very noisy and show slide to slide variability in expression values; therefore, the data were scale normalized to correct for between-array variability. The final normalized M-values can be seen in Figure 6.3 (Right).

6.3 Aggregate, Reduce and Impute the Data

The SAM procedure requires data be in the form of a $p \times n$ matrix where p is the number of genes and n is the number of samples. This structure requires that replicated genes within samples be aggregated into one expression value per

sample. The SAM procedure will not tolerate any missing values; therefore, missing values must be eliminated or imputed. In the SAM analysis 4,179 out of 21,529 genes were thrown out due to a combination of negative gene expression, poor quality, and imputation rules.

The mouse pregnancy data consisted of 16 samples (i.e. 16 replications). Approximately 86% of the genes were replicated twice, while all but one of the other 14% were replicated four times. In order to fit the mouse pregnancy data into a $p \times n$ matrix, it was decided that the replicates would be averaged within each sample, creating one expression value per gene per sample.

The SAM procedure requires that each sample be replicated at least twice. In the mouse pregnancy data, there were four samples for the LPS:B6, LPS:KO, and SAL:B6 treatment groups; there were three samples for the SAL:KO treatment group; and there was one sample for the REF treatment group. The REF treatment group could not be used in the SAM analysis because it was not replicated.

The SAM procedure will not tolerate any missing values in the $p \times n$ data matrix. It is necessary to either throw out genes with any missing observations, or impute values for the missing observations. Missing values occurred in the mouse pregnancy data for two reasons: first, they occurred when either the red or green scan had a background value larger than its foreground value; second, they occurred when intensities were flagged for poor spot quality.

After accounting for flags and background correction, 20% of all observations were missing with 40% of the genes having at least one missing value. It was decided that any gene without at least one observation in each of the four treatment groups

would be discarded from the SAM analysis. This rule eliminated 20% of the total genes.

After eliminating genes without at least one gene intensity per treatment group, 5.6% of the gene intensities were still missing values. A K-nearest neighbor algorithm called KNNimpute proposed by Troyanskaya et al. (2001) was used to impute the remaining missing values. For each gene with missing values the KNNimpute algorithm finds K-similar genes that do not have those same missing values. By default, similarity is defined as Euclidean distance, and K is set to 10. Weights based on the similarity are then calculated for the K-similar genes. A weighted average of the K-similar genes is then used to impute the missing values. The algorithm is robust, with a maximum of 10% decrease in accuracy with 20% of the data missing (Troyanskaya et al., 2001).

6.4 Run the SAM Analysis

The Significance Analysis of Microarrays (SAM) was performed after the mouse pregnancy data had been normalized, aggregated into a $p \times n$ array, and imputed for missing values. The SAM analysis utilized the multiclass method to account for differences in the following four treatments: (1) LPS:B6, (2) LPS:KO, (3) SAL:B6, and (4) SAL:KO. The multiclass method uses Fisher's linear discriminant to essentially compute an F statistic for each gene.

The SAM procedure was used to calculate the expected order statistics. The default, $B = 100$ permutations, was used to estimate the null distribution. The exchangeability factor, s_0, was estimated at 0.171, and the percent of null genes, π_0,

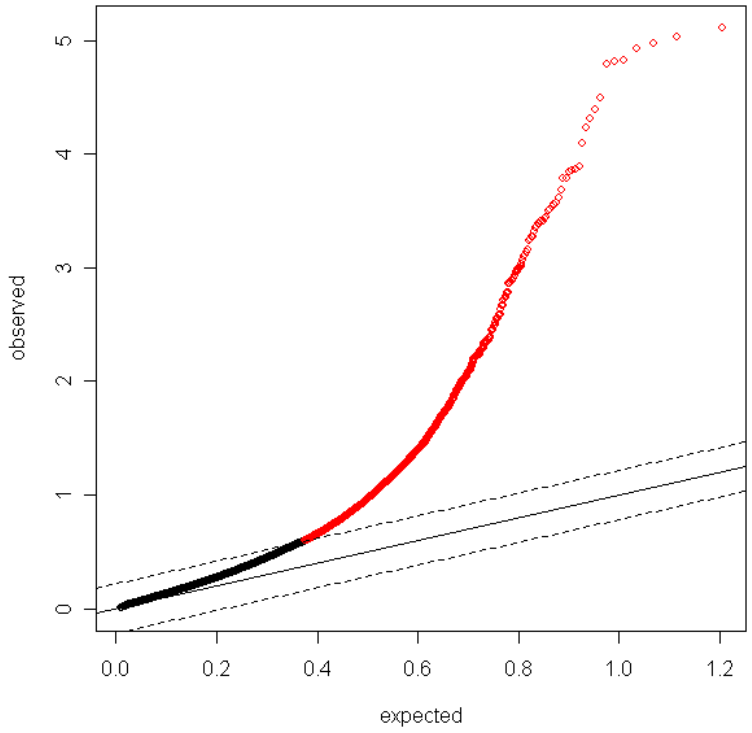

Figure 6.4: SAM plot. The observed order statistics are plotted against the estimated expected order statistics. The dotted line shows $\Delta = 0.55$. Significant genes are those with statistics above the dotted line.

was estimated at 54.8%. Figure 6.4 shows the observed verses the expected statistics. Note that the test statistics are greater than zero due to the multiclass method.

The Tail Strength (TS) for this SAM analysis was 0.56 with a 90% confidence interval of $(0.26, 0.87)$. As the confidence interval suggests, the mouse pregnancy data are extremely noisy.

6.5 Identify Significant Genes

The FDR was calculated for various cutoffs. Table 6.1 shows when $\Delta = 0.0$ that any gene with a relative difference greater than 0.0194 is significant. There were 17,348 such genes. The estimated number of false positives was 9,510. The median

method for estimating the FDR was used. In other words, the estimated number of false positives was the median value of the estimated false positives for each of the B permutations. The FDR for $\Delta = 0$ was 9,510/17,348 = 54.82%.

	delta	# med false pos	# called	median FDR	cuthi
1	0.0000	9510	17348	0.5482	0.0194
2	0.0500	6770	14745	0.4592	0.2037
3	0.1000	2821	10179	0.2772	0.3439
4	0.1500	999	6992	0.1429	0.4555
5	0.2000	351	4863	0.0721	0.5550
6	0.2500	121	3451	0.0350	0.6468
7	0.3000	48	2623	0.0183	0.7279
8	0.4000	8	1673	0.0049	0.8758
9	0.5000	2	1153	0.0014	1.0136
10	0.6000	1	817	0.0007	1.1466

Table 6.1: False Discovery Rates for various cutoffs.

For the SAM analysis, an FDR of 5% was considered satisfactory, so Δ was set to 0.22. The 4,152 genes with test statistics greater than the cutoff (0.595) were considered significant. An FDR of 5% would suggest that 208 out of the 4,152 genes were false positives.

The Miss Rate (MR) estimates the percentage of false negatives near the cutoff. That is, the MR estimates the percentage of genes that are significant, but not declared significant for a given interval. The table of MR percentages is in Table 6.2. The cutoff boundary for the SAM analysis was 0.595. For the genes nearest this boundary, 79.57% are false negatives. The high percentage is due in part to a low choice of FDR. As the FDR increases, the percentage of false positives increases and the percentage of false negatives decreases.

	Cutpoints	Miss Rate(%)
1	0.438 → 0.464	58.01
2	0.464 → 0.493	62.05
3	0.493 → 0.524	68.48
4	0.524 → 0.558	74.42
5	0.558 → 0.595	79.57

Table 6.2: Miss Rates for the the multiclass SAM analysis.

Table 6.3 shows sample output for the SAM analysis. Only the 20 most significant genes are included.

	Row	Gene ID	Rel Diff	Numer	Denom	cont-1	cont-2	cont-3	cont-4
1	13270	BQ552581	5.11	1.70	0.33	−8.56	−8.35	14.18	3.640
2	12465	BQ550593	5.03	1.88	0.38	−8.09	−7.76	10.43	7.220
3	17079	BQ562802	4.98	1.91	0.38	−7.08	−8.22	10.39	6.550
4	14971	BQ557066	4.92	2.33	0.47	−6.95	−6.23	8.81	5.83
5	12722	BQ551218	4.82	3.11	0.64	−5.17	−6.01	7.63	4.74
6	8199	BG076245	4.81	2.06	0.43	−7.06	−6.60	9.32	5.80
7	15236	BQ557743	4.79	2.79	0.58	−5.21	−6.43	7.42	5.61
8	17248	C79033	4.49	1.89	0.42	−7.40	−5.65	7.76	7.04
9	10492	BG084405	4.39	1.74	0.40	−5.92	−7.18	9.11	5.32
10	9578	BG080268	4.31	3.45	0.80	−3.97	−5.40	6.00	4.49
11	3624	BG067921	4.23	2.04	0.48	−5.61	−5.52	7.89	4.32
12	13603	BQ553339	4.09	1.81	0.44	−5.74	−5.60	7.82	4.690
13	6043	BG072303	3.88	1.40	0.36	−6.25	−6.22	8.98	4.66
14	5443	BG071178	3.87	1.33	0.35	−7.48	−5.59	8.65	5.91
15	13454	BQ553020	3.85	2.85	0.74	−4.52	−4.13	5.05	4.800
16	5703	BG071644	3.85	1.99	0.52	−5.64	−4.18	6.39	4.57
17	4254	BG069214	3.79	1.54	0.41	−4.85	−6.38	7.26	5.29
18	17069	BQ562771	3.79	2.06	0.54	−4.29	−5.12	6.43	3.980
19	15908	BQ559603	3.68	1.57	0.43	−5.34	−5.23	6.70	5.170
20	9792	BG081062	3.62	1.12	0.31	−7.60	−6.28	8.90	6.63

Table 6.3: Output for the top 20 significant genes in the SAM analysis

6.6 Investigate Treatment Effects

The multiclass method produces standardized contrasts that are useful for determining which treatment groups contribute to large test statistics. The stan-

dardized contrast (u_{it}) for gene i and treatment t is defined as $u_{it} = (\bar{x}_{it} - \bar{x}_i)/s_i$, where s_i is defined in Equation 3.3. There are four treatments in the SAM analysis, therefore, there are four standardized contrasts per gene.

	Contrast	Genes	Contrast	Genes	Total
1	-1,-1,1,1	892	1,1,-1,-1	481	1373
2	0,-1,1,0	208	0,1,-1,0	320	528
3	0,-1,1,1	149	0,1,-1,-1	240	389
4	-1,0,0,1	136	1,0,0,-1	150	286
5	0,0,0,-1	191	0,0,0,1	76	267
6	0,-1,0,1	95	0,1,0,-1	142	237
7	-1,0,1,0	105	1,0,-1,0	105	210
8	-1,-1,1,0	95	1,1,-1,0	85	180
9	-1,0,1,1	110	1,0,-1,-1	65	175
10	0,0,-1,0	91	0,0,1,0	73	164
11	0,-1,0,0	57	0,1,0,0	44	101
12	-1,0,0,0	32	1,0,0,0	37	69
13	0,0,0,0	30	0,0,0,0	30	60
14	-1,-1,0,1	16	1,1,0,-1	39	55
15	0,0,-1,1	12	0,0,1,-1	25	37
16	-1,1,0,0	14	1,-1,0,0	11	25
17	-1,1,0,1	7	1,-1,0,-1	1	8
18	0,-1,1,-1	3	0,1,-1,1	4	7
19	-1,1,-1,1	2	1,-1,1,-1	1	3
20	-1,0,1,-1	1	1,0,-1,1	1	2
21	-1,-1,1,-1	1	1,1,-1,1	0	1
22	-1,0,-1,1	1	1,0,1,-1	0	1
23	-1,1,-1,0	1	1,-1,1,0	0	1
24	-1,1,0,-1	0	1,-1,0,1	1	1
25	-1,1,1,0	1	1,-1,-1,0	0	1
26	0,0,-1,-1	0	0,0,1,1	1	1

Table 6.4: Summary of standardized contrasts for the SAM analysis. The standardized contrasts represent treatment means that are exceptionally above (+1) or below (−1) the overall mean. There were 4,152 significant genes.

The standardized contrasts are calculated from the simulated null distribution. The u_{it} that are greater than the 97.5^{th} percentile in the null distribution are flagged with +1, whereas the u_{it} that are less than the 2.5^{th} percentile are flagged with −1.

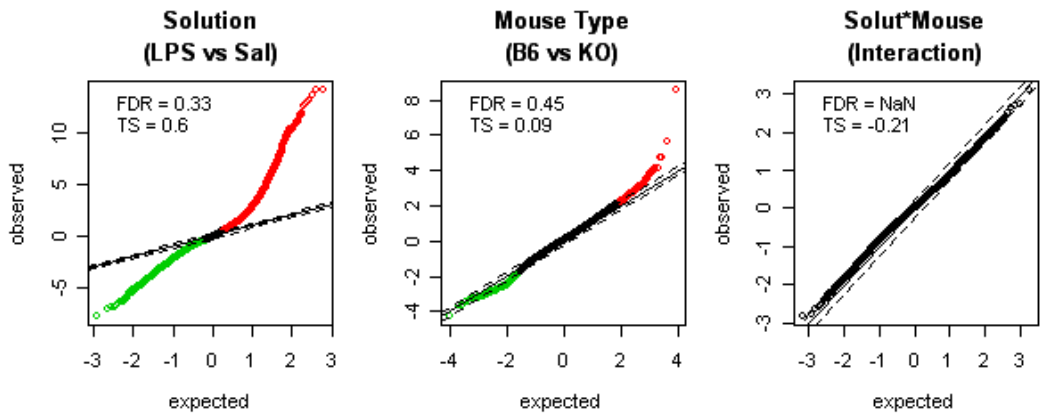

Figure 6.5: SAM plots for treatment effects.

Finally, if a gene did not have an exceptionally high or low u_{it} then that particular treatment group was flagged with 0.

The standardized contrasts for the significant genes in the mouse pregnancy data can be seen in Table 6.4. The largest number of standardized contrasts (1,373) is associated with comparing the first two treatment groups (LPS:B6 and LPS:KO) to the last two treatment groups (SAL:B6 and SAL:KO). This suggests that there is a main effect for the LPS versus the Saline solutions.

The significance of the two main effects and the interaction effects can be examined in separate SAM analyses. Figure 6.5 shows the SAM plots for three pairwise SAM analyses. In these analyses, the two-class unpaired method is used. This method basically employs a two-sample t-test for each gene, and it can be used to compare two effects at a time.

The first follow-up SAM analysis shows the difference between the LPS and Saline solutions. Again, the strong departure from the diagonal line suggests a strong effect difference between LPS and Saline ($\text{TS} = 0.62 \pm 0.33$).

The second follow-up SAM analysis shows the difference between the B6 and Knock-out groups. There is a departure from the expected values, but it is not nearly as strong as the analysis for solution ($\text{TS} = 0.12 \pm 0.38$).

The third follow-up SAM analysis shows the interaction between the four treatment groups. In this analysis there appears to be a slight inversion between observed and expected values. The strong adherence to a diagonal pattern suggests that there is no interaction between the four treatment groups ($\text{TS} = -0.22 \pm 0.31$).

The plots support the conclusion that the solution is driving most of the significance in the analysis, but the mouse type appears to have some influence also. There appears to be no interaction effect.

Chapter 7

Conclusion

The purpose of this project is to compare two popular methods for analyzing genomic data. This paper considered two methods. The first, a mixed models approach, uses a normalization model followed by a gene model. Genes are declared significant when their pairwise comparisons for the treatment effect in the gene model are significant. The second, the Significance Analysis of Microarrays, compares a test statistic to its expected value, which is estimated with resampling. Genes are declared significant when the difference between the observed the expected test statistic exceeds a threshold.

7.1 Comparison of Results

The mouse pregnancy data tested 21,529 genes. Table 7.1 compares the results for the mixed models and SAM analyses. Some genes were not included in the analyses due to poor or insufficient data. The mixed models analysis did not include 1,727 genes, whereas the SAM analysis did not include 4,179 genes. Of the remaining genes, the mixed models analysis produced 3,542 genes with at least one pairwise comparison when the FWER was 5%. The SAM analysis produced 4,152 significant genes when the FDR was 5%. There were 1,827 genes that were significant in both analyses.

	Mixed Models Approach			
SAM Procedure	Not Included	Not Significant‡	Significant†	Total
Not Included	1,722	2,456	1	4,179
Not Significant	4	11,480	1,714	13,198
Significant	1	2,324	1,827	4,152
Total	1,727	16,260	3,542	21,529

Table 7.1: Comparison of significant genes. †At least one significant pairwise comparison. ‡No significant pairwise comparisons.

In general, there was agreement between the two methods. There were 1,827 genes declared significant in both methods, 11,480 genes declared not significant in both methods, and 1,722 genes that were not analyzed in either method, for a total agreement of 15,029/21,529 = 68.8%.

The two methods produced different results. In part this was due to the differences in their methodologies. However, there are several other reasons why the two analyses did not share more results in common: first, the mixed analysis used unbalanced data, whereas the SAM analysis used aggregated and imputed data; second, the mixed analysis assumed normally distributed errors, whereas the the SAM analysis assumed nothing; third, the mixed analysis used background corrected $\log_2$ intensities whereas the SAM analysis used log ratio intensities; fourth, the mixed analysis used all possible pairwise comparisons to test treatment effects, whereas the SAM analysis used an overall test for treatment effects; fifth, the mixed analysis used a mixed model to normalize the data, whereas the SAM analysis used print-tip lowess smoothers; and sixth, the mixed analysis used data from the reference slide, whereas the SAM analysis did not.

The differences between the mixed models approach and the SAM procedure are confounded with the differences in procedure mentioned in the previous paragraph. In order to get a better comparison all extraneous factors should be held constant, allowing the results from the two methods to be more similar; for example, the data could be normalized only once and applied to the two methods, removing the effect of the normalization procedure when comparing the two methods. Unfortunately, some differences cannot be held constant, for example, the SAM procedure cannot handle unbalanced data and the effect of aggregating and imputing the data will be confounded with the differences between the two methods whenever unbalanced data are analyzed.

This project analyzed the mouse pregnancy data with the mixed models approach and the SAM procedure. One important difference between the two analyses was the two different data types. The mixed models analysis used background corrected $\log_2$ intensities, whereas the SAM analysis used log ratio intensities. The mixed models analysis tested whether the treatments, including the reference, were the same, while the SAM analysis tested whether the differences between the treatment and the reference were the same; thus, the nature of the two comparisons were different. If both methods were used to analyze the same type of data, one might expect to see more similar results.

Another important difference between the two analyses was the way in which genes were declared significant. In the mixed models analysis, all pairwise comparisons were tested for significance. There were five treatment levels, creating ten pairwise comparisons associated with each gene. In the SAM procedure only the rel-

ative difference was tested for significance, and there was only one relative difference associated with each gene; unfortunately, there is no method for testing all pairwise differences in the SAM procedure, but the mixed models approach can test the overall effect for each treatment with an F-statistic. If the overall test for treatment in the mixed models analysis is compared to the relative difference in the SAM analysis, one might expect to see more similar results.

In a separate analysis, the mouse pregnancy log ratios were analyzed with the mixed models approach. For each gene the negative $\log_{10}$ p-value from the overall test for treatment was compared to the relative difference from the SAM analysis. Negative $\log_{10}$ p-values were used for convenience only. A Bonferroni adjustment was used to control the FWER in the mixed models analysis. There were 19,466 genes considered in the mixed models analysis, so the cutoff was $-\log_{10}(0.05/19,466) = 5.59$. In the mixed models analysis, the 391 genes with negative $\log_{10}$ p-values greater than the cutoff were declared significant.

For convenience, the cutoff in the SAM analysis was chosen to produce the same number of significant genes as the mixed models analysis. This resulted in a cutoff of 1.46. In the SAM analysis, the 391 genes with a relative difference greater than 1.46 were declared significant.

Figure 7.1 shows the relationship between the negative $\log_{10}$ p-values from the mixed analysis and the relative differences from the SAM analysis. There is an overall agreement between the two methods. The Pearson correlation between the two variables is 0.82. There appears to be a slight curvature to the scatter. The curvature could be due to poor normalization in one of the methods. There appears

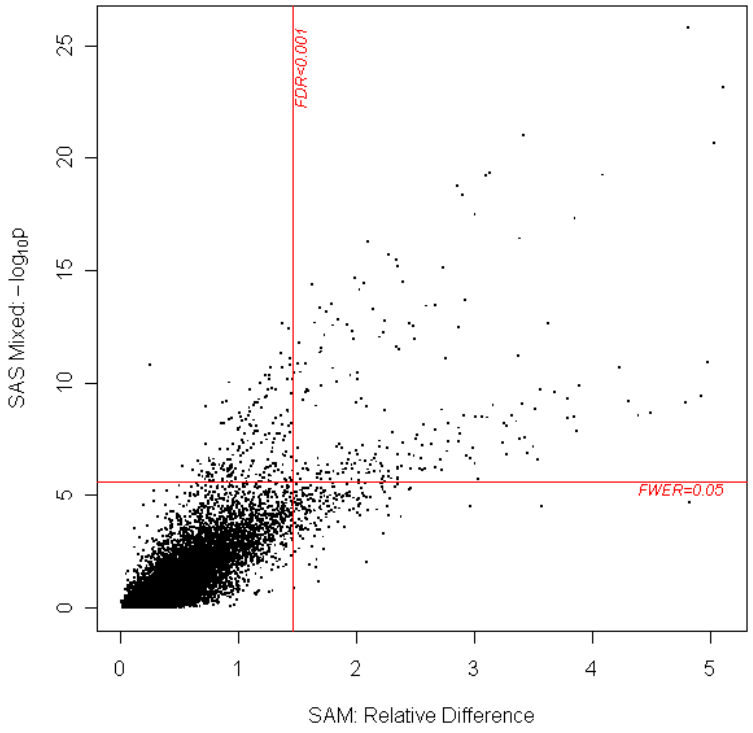

Figure 7.1: Negative $\log_{10}$ p-values from the mixed models analysis versus the relative differences from the SAM analysis. The correlation suggests a general agreement between the two methods.

to be a group of genes that are much more significant in the mixed models analysis and another group that are much more significant in the SAM analysis. The two branches in Figure 7.1 indicate that the two methods are systematically treating two types of genes in different ways. This may be due to the fact that the reference sample (microarrays $45 - 48$) was included in the mixed models analysis but not in the SAM analysis. One branch in Figure 7.1 may be due to genes that are significantly different than the reference sample, and the other branch may be due to genes that are not. Alternatively, the branches could be due to the extra preprocessing in the SAM analysis. Replicated data from the SAM analysis were averaged and missing

data were imputed. Perhaps averaging the replicates removed too much information, or imputing the missing data included the wrong information.

Table 7.2 compares the genes declared significant when the mixed models cutoff is 5.59 and the SAM cutoff is 1.46. There were 17,345 genes that were included in both analyses, 391 of which were declared significant. The percent of genes declared significant in both methods was $205/391 = 52.4\%$. The models had a $(205 + 16{,}768 + 2{,}058)/21{,}529 = 88.4\%$ agreement. The mixed models analysis included 2,121 genes that the SAM analysis did not; however, none of these extra genes were significant. By comparison, the SAM analysis only included 5 genes that were not in the mixed models analysis. These genes were also considered non-significant.

	Mixed Models Approach			
SAM Procedure	Not Included	Not Significant	Significant	Total
Not Included	2,058	2,121	0	4,179
Not Significant	5	16,768	186	16,959
Significant	0	186	205	391
Total	2,063	19,075	391	21,529

Table 7.2: Comparison of significant genes when the mixed models cutoff is 5.59 and the SAM cutoff is 1.46.

The simulated null distribution from the SAM analysis can be used to estimate p-values from the relative differences. A comparison of p-values can be seen in Figure 7.2. The distributions look very similar for both methods. The small p-values in the mixed models method taper off more quickly than the SAM p-values, because the SAM procedure cannot measure extremely small p-values. By definition, the smallest

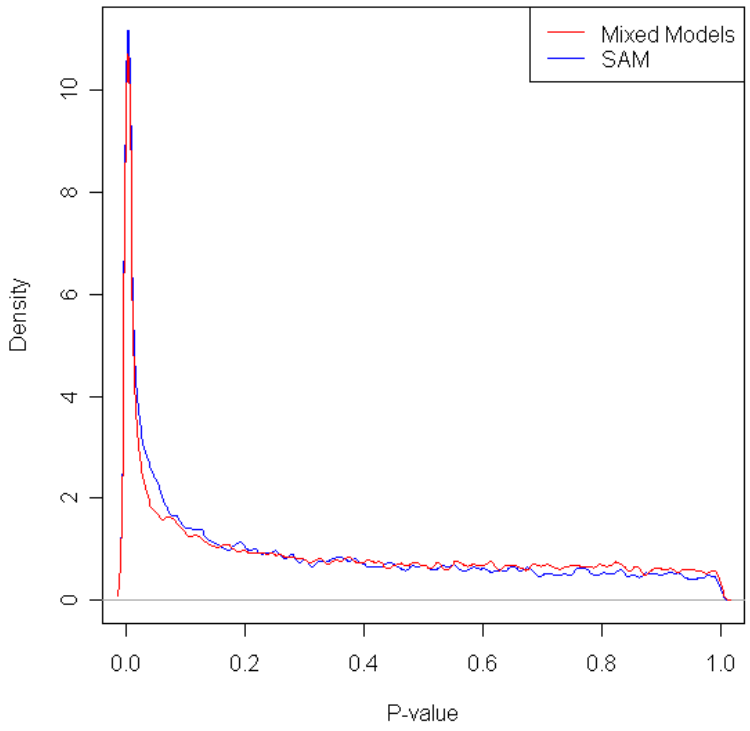

Figure 7.2: Distribution of p-values for the mixed models analysis (red) and the SAM analysis (blue).

p-value in any SAM analysis is $1/n$, where n is the total number of genes. In this analysis, the smallest p-value is $1/17{,}350 = 5.7637 \times 10^{-5}$. By contrast, the smallest p-value in the mixed models analysis is 1.7079×10^{-26}.

Another way of comparing the results of the two methods is to examine the percent of shared genes. For example, the 10% most significant genes from the mixed analysis and the 10% most significant genes from the SAM analysis share 75.2% of their genes in common. This percentage tends to increase as the percentage of genes considered increases; for example, the 50% most significant genes from both methods share 82.2% in common, and all of the genes from both methods share all

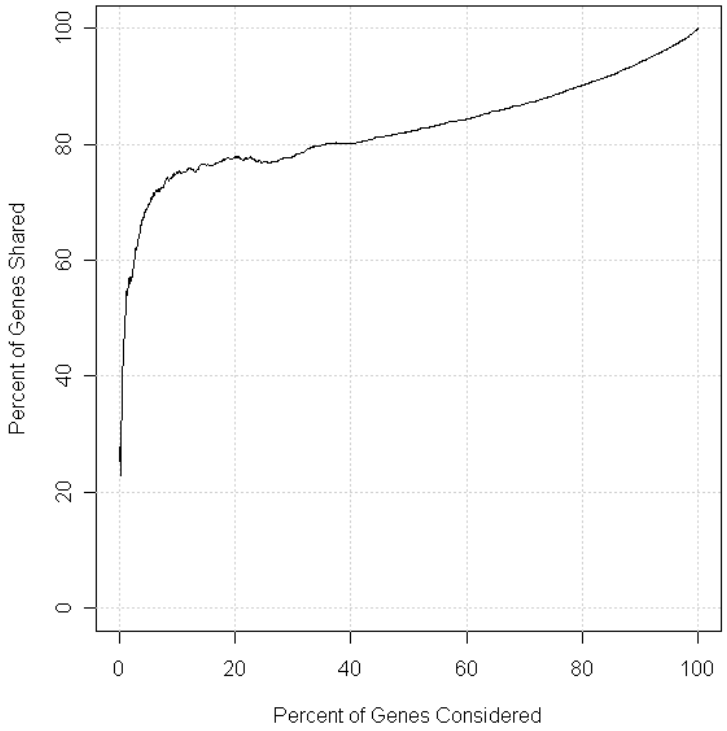

Figure 7.3: Percent of genes shared versus the percent of genes considered.

their genes in common. A plot of the percent of genes shared against the percent of genes considered in both methods is shown in Figure 7.3.

The percent of genes shared is approximately 78% when the percent of genes considered is 20%. This means that there is fairly good agreement between the two methods for the upper 20% of the data. A small dip in the percent of genes shared occurs from the $20\% - 30\%$ range of genes considered. This dip is a result of the mixed models p-values tapering off more rapidly than the SAM p-values at the lower tail of the distributions (see Figure 7.2). This chart shows that the methods tend to have especially strong agreement when at least the top 20% of both gene lists are considered.

The results of analyses were more comparable when the log ratios were analyzed and the overall test for treatment from the mixed analysis was compared to the relative difference from the SAM analysis. The p-values from the mixed models analysis were highly correlated with the relative differences from the SAM analysis. There was a 88.4% agreement between the two methods when the FWER was controlled at $\alpha = 0.05$. The distribution of p-values were very similar. There was a large percentage of genes shared in the top 20% of the most significant genes from both methods.

Although the results from both methods are similar it is still unclear whether the differences are due to the two methods or to some other factors in the two methods, such as the different normalization techniques; also, the reference sample was used in the mixed models analysis but not in the SAM analysis. If the same normalization technique was used for both methods and if the reference sample was eliminated from both methods, then one would expect the results to become more similar. Using the same data type and comparing similar statistics made the results of the mixed models and SAM analyses more similar. It is reasonable to suppose that the more extraneous factors are controlled, the more similar the results will be.

7.2 Comparison of Methods

The mixed models approach and the SAM procedure have fundamentally similar approaches to analyzing microarray data. Table 7.3 lists the critical steps in the two methods. The first step is to organize and prepare the data for analysis. The second step is to normalize the data. Normalization removes effects due to the exper-

imental setup rather than effects due to the genes. The third step is to analyze the data and test for differential gene expression. The fourth step is to finalize the list of significant genes by employing some cutoff rule. The cutoff is based on an error rate such as the FWER or the FDR.

	Mixed Models Approach	**SAM**
1. Organize	Clean the data.	Clean, aggregate, and impute the data.
2. Normalize	Use one mixed normalization model for all genes.	Choose any method.
3. Analyze	Use a separate mixed gene model for each gene.	Calculate the relative difference and estimate its null distribution.
4. Finalize	Identify significant pairwise comparisons.	Compare the observed versus the expected statistics.

Table 7.3: Overview of the mixed models approach and the SAM procedure.

The mixed models approach and the Significance Analysis of Microarrays are two methods for identifying differential gene expression. Determining which method to use depends on the advantages and disadvantages of the two methods. Each has certain characteristics that are inherent and cannot be changed; for example, the mixed models approach uses parametric models to analyze the data, whereas the SAM procedure uses random permutations of the data to simulate the null distribution. Each method also has characteristics that are flexible and can be changed; for example, model specification in the mixed models approach and normalization in the SAM procedure depend entirely on the researcher's discretion. The following sections

will identify inherent and flexible characteristics as well as comment on some of the advantages and disadvantages of both methods.

7.2.1 Inherent Characteristics of the Mixed Models Approach

The mixed models approach has at least four inherent advantages. The first advantage is that the mixed models approach tolerates unbalanced data. Microarray data can be unbalanced for a variety of reasons; for example, measurements can be thrown out due to negative expression or poor quality, or measurements may be replicated unequally. The mixed models approach tolerates unbalanced data. The only requirement for applying the mixed models approach to unbalanced data is that there be enough degrees of freedom to estimate error terms. The capability of analyzing unbalanced data allows the mixed models approach to take advantage of every observation in the data set. It allows the mixed models approach to estimate more genes and more treatment effects. It also allows the precision of variance estimates to be proportional to the amount of data available.

The second inherent advantage of the mixed models approach is the assumption of a heterogeneous error structure. In the mixed models approach, a separate gene model is fit to every gene; thus, every gene has its own set of variance estimates. This is reasonable given the wide variety of gene expression. One advantage of heterogeneous variances is that it produces better specified models. Another advantage is that estimates of the variance terms may be of interest in their own right.

A third inherent advantage of the mixed models approach is flexible model specification. The normalization and gene models can be adapted to a wide variety of

experimental designs. Better model specification can lead to increased precision of the treatment effects. Additionally, non-treatment effects can be tested for significance.

A fourth inherent advantage of the mixed models approach is the ability to estimate pairwise comparisons of the treatment effect. The mixed models approach can not only test for an overall treatment effect, it can also test for pairwise treatment effects. The ability to test pairwise treatment effects allows the mixed models approach to identify exactly which treatments are causing significant gene expression.

The mixed models approach has at least one inherent disadvantage. The normalization and gene models both assume independent, normally distributed errors. Numerous studies show that gene expression is not independent between genes. Genes tend to be correlated due to the genetic pathways they depend on. Furthermore, the error structure of many microarray studies is not normal. Assuming normality may lead to a misspecified model and improper conclusions regarding treatment effects.

7.2.2 Inherent Characteristics of the SAM Procedure

The SAM procedure has at least two inherent advantages. First, the SAM procedure makes no distributional assumptions. The distribution of expected order statistics is estimated with resampling. There are no assumptions to validate, and there is little risk that the expected null distribution is much different than the true null distribution. Because the SAM procedure makes no distributional assumptions, it can be used on a larger group of microarray analyses, including those with irregular distributions.

A second inherent advantage of the SAM procedure is flexible specification of the relative difference. For two-class data the relative difference is essentially a t-statistic. For multiclass data the relative difference is essentially an F-statistic, but the relative difference could be specified in any form; for example, it can be defined to identify genes whose expression correlates with survival time or some other quantitative measure. Allowing the relative difference to be specified in any form allows the SAM procedure to be adapted to a broad range of experiments.

The SAM procedure has at least two inherent disadvantages; first, the SAM procedure does not handle unbalanced data. Instead, it requires the data to be arranged in a $p \times n$ array were p is the number of genes and n is the number of samples, forcing the data to have exactly one gene expression for each sample. This means that preprocessing is often arduous for complex experimental setups. It also means that unbalanced data typically have to be aggregated, removed, or imputed. For example, the mouse pregnancy data had an unequal replication of spots in each each of the 16 samples. Some genes were replicated twice while others were replicated four times, so the replicates in the mouse pregnancy data were averaged across each sample. The mouse pregnancy data also contained a large number of missing values due to negative or poor quality gene expression. Genes without at least one expression value in each treatment group were thrown out. This rule eliminated 20% of the genes. Aggregating the data into a $p \times n$ array also forced the SAM analysis to exclude the reference sample. In order for the reference sample to have been included, it would have needed to be replicated. Missing values in the remaining genes were imputed using a K-nearest neighbor algorithm. The fact that some measurements

represented the average of four replicates while others represented the average of two replicates while still others represented imputed values did not factor into the analysis. The SAM procedure treats all measurement with the same precision. The researcher is forced to assume that the aggregated data is as accurate as the imputed data. Furthermore, the researcher is forced to decide how much missing data can be reliably imputed. Aggregating data removes information, whereas imputing data creates information. The SAM procedure has no way to measure the effect that imputing and aggregating has on an analysis.

A second inherent disadvantage of the SAM procedure is the lack of pairwise comparisons. The SAM procedure produces one statistic per gene. For multiclass data the statistic is essentially an F-value; therefore, the statistic tests whether there is a difference in at least one treatment group. If there is a difference then it is natural to investigate which group or groups cause the difference. The SAM procedure does not test pairwise comparisons. Instead it provides standardized contrasts that indicate which treatments are different; however, standardized contrasts are not statistical tests, they merely aid in identifying which treatment group or groups are different. The lack of pairwise comparisons in the SAM procedure makes identifying treatment effects difficult. It also makes identifying genes with highly significant pairwise comparisons difficult.

7.2.3 Flexible Characteristics of Both Methods

There are a number of flexible characteristics associated with both analysis methods. The first is that both methods can analyze various types of data. The

mixed models analysis used background corrected $\log_2$ gene intensities. The SAM analysis used log ratios. Both methods can be used to estimate either type of data. If the mixed models approach is used to analyze log ratio data, the model specification will need to change and a covariate for the log average, A, should be included. If the SAM procedure is used to analyze background corrected $\log_2$ data, the $p \times n$ data array will appear very differently. There will still be p rows, but n will be twice as large in order to accommodate two measurements per gene spot.

The type of data analyzed depends on the type of hypothesis desired. In two-channel data, each gene spot is exposed to a treatment and a reference, so there are two measurements per gene spot. If the background corrected $\log_2$ data are analyzed, then the reference is considered to be another treatment. If the log ratios are analyzed, then each treatment is compared to its corresponding reference. The treatment effects for the background corrected $\log_2$ data are the effect of each treatment, including the reference treatment, on the response. The treatment effects for the log ratio data are the differences between the treatment and its corresponding reference. The null hypothesis for background corrected $\log_2$ data is that the treatment effects, including the reference, are the same. The null hypothesis for log ratio data is that the differences between the treatment and the reference are the same.

The treatment effects associated with background corrected $\log_2$ data are inherently different than the treatment effects associated with log ratio data. Any statistical test for treatment is also inherently different for the two types of data. One would expect to get different results by analyzing each type of data. The type of data analyzed depends on the type of hypothesis desired.

A second characteristic that is flexible for both methods is specifying the normalization technique. In the mixed models approach, data are normalized with a mixed model, but the data could technically be normalized using any number of techniques; for example, the mouse pregnancy data showed large variability between print-tip groups, so the data could have been normalized with a different mixed model for each print-tip group. In the SAM procedure no normalization technique is specified. The normalization procedure depends entirely on the discretion of the researcher. The mouse pregnancy data were normalized with lowess smoothers; however, the data could have been normalized with a mixed model. Specifying the normalization technique allows researchers to choose the best technique for their data.

A third characteristic that is flexible for both methods is measuring error rates and quantifying genome-wide summary measures. The mixed models approach provides strong control of the family-wise error rate. The SAM procedure, on the other hand, provides weak control of the FDR. However, the FWER or the FDR could be estimated in either method. The mixed models approach could use several techniques for estimating the FDR from p-values, while the SAM procedure can estimate p-values based off of the simulated null distribution. The SAM procedure also produces estimates of a Type II error rate called the Miss Rate and a genome-wide summary measure called the Tail Strength. The mixed models approach does not currently estimate the Type II error rate or genome-wide summary measures, but methods could be developed to estimate them. Measuring error rates and quantifying genome-wide summary measures allow researchers to assess significance and compare studies.

7.3 Comparison of Software

The mixed models approach depends on the SAS® Mixed Procedure. SAS® handles data sets such as the mouse pregnancy data very efficiently. Specifying the model structure in the mixed procedure is also very easy, but specifying complex models in the mixed models approach can prove to be problematic and time consuming. Because the normalization model is applied to the entire data set, it is computationally advantageous to use a simplistic model. Specifying a very complex normalization model can take days to run: it took three days to normalize the mouse pregnancy data when a print-tip effect was included, but only 10 minutes without it. Gene models also benefit from a simplistic design. Since every gene has its own model, a complex specification increases computation time. Convergence criteria are also less likely to be met, and complex models are more likely to include insignificant terms. While the mixed model approach may technically permit any valid specification of the normalization model, a simple model specification provides the quickest results with the least amount of problems. Even with a simplistic model, the mixed models analysis of the mouse pregnancy data experienced problems running on a personal computer, so the mouse pregnancy data were analyzed with an AIX Unix server. The total run time, including reading the data, normalizing it, and running the SAM analysis, was approximately 35 minutes.

The SAM procedure depends on R, an open source environment for statistical computing and graphics. The actual code for the the SAM procedure is part of the SAMR package. This package is available royalty-free to academic users. The

SAMR package can be run through a Microsoft Excel® interface or directly through the R interface. If it is run through R, then the error handler should be deactivated with the following command: `options(error=NULL)`. Neither R nor Excel® handles exceptionally large data sets very well. Data in Excel® is limited to the maximum spreadsheet size. Data in R is limited to the amount of available RAM. The specification of the relative difference cannot be modified in the SAMR package. The user is restricted to one of ten options. The R environment includes a number of tools that are useful for visualizing, normalizing, and analyzing microarray data. The mouse pregnancy data were normalized with a technique used in the Bioconductor software, a suite of packages designed for the analysis of genomic data. The mouse pregnancy data were analyzed with a Microsoft Windows personal computer running an Intel Pentium 4 processor at 3.0GHz. The total run-time, including reading the data, normalizing it, and running the SAM analysis, was approximately 13 minutes.

7.4 Suggestions for Future Research

This project revealed several areas for future research. The normality assumption in the mixed models approach could be investigated. Wolfinger et al. (2001) suggests that the method will serve analysts well in the large majority of cases. However, it is unclear how robust the mixed models approach is.

A more exact comparison between the mixed models approach and the SAM procedure could be investigated. The mouse pregnancy data came from a fairly complex experimental setup. The results of the two analyses were confounded with many factors besides the two methodologies. If a simple data set were chosen and

analyzed, it could provide a more exact comparison. The data set would preferably be balanced with no missing data.

The effect of normalization could be investigated. There have been numerous normalization techniques proposed, but their impact on the mixed models approach and the SAM procedure is unclear. Comparing multiple analyses from different normalization techniques could reveal the effect that normalization has on the results.

The treatment effects could also be investigated. Neither method makes identifying treatment effects easy. Both methods force the user to investigate treatment effects at the gene level. In the mouse pregnancy data, it was found that the labor inducing solution, LPS, caused the vast majority of significant differences. This conclusion was reached only after a superficial examination of the results. Future research could be conducted into quantifying and comparing overall treatment effects on an experiment.

Appendix A

Technical details of the SAM procedure

The data is $x_{ij}, i = 1, 2, \ldots, p$ genes, $j = 1, 2, \ldots, n$ samples, and response data $y_j, j = 1, 2, \ldots, n$ (y_i may be a vector).

A.1 SAM procedure

1. Compute a statistic

$$d_i = \frac{r_i}{s_i + s_0}; i = 1, 2, \ldots p \tag{A.1}$$

   r_i is a score, s_i is a standard deviation, and s_0 is a fudge factor. Details of these quantities are given in A.2 and A.3.

2. Compute order statistics $d_{(1)} \leq d_{(2)} \cdots \leq d_{(p)}$.

3. Take B sets of permutations of the response values y_j. For each permutation b compute statistics $d^{*b}_{(1)} \leq d^{*b}_{(2)} \cdots \leq d^{*b}_{(3)}$.

4. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(i)} = (1/B) \sum_b d^{*b}_{(i)}$ for $i = 1, 2, \ldots, p$.

5. Plot the $d_{(i)}$ values versus the $\bar{d}_{(i)}$.

6. For a fixed threshold Δ, starting at the origin and moving up to the right, find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_1 are called significant positive. Similarly, start at origin, move down to the left and find the first $i = i_2$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_2 are called significant negative.

For each Δ define the upper cut-point $\text{cut}_{up}(\Delta)$ as the smallest d_i among the significant positive genes and similarly define the lower cut-point $\text{cut}_{low}(\Delta)$.

7. For a grid of Δ values, compute the total number of significant genes (from the previous step), and the median number of falsely-called genes, by computing the median number of values among each of the B sets of $d^{*b}_{(i)}, i = 1, 2, \ldots, p$ that fall above $\text{cut}_{up}(\Delta)$ or below $\text{cut}_{low}(\Delta)$. Repeat this process for the 90th percentile of falsely called genes.

8. Estimate π_0, the proportion of true null (unaffected) genes in the data set, as follows.

   (a) Compute $q25, q75 = 25\%$ and 75% points of the permuted d values (if $p = \#$ genes, $B = \#$ permutations, there are pB such d values).

   (b) Compute $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\}/(0.5p)$ (the d_i are the values for the original data set: there are p such values.)

   (c) Let $\hat{\pi}_0 = \min(\hat{\pi}, 1)$ (i.e., truncate at 1). This estimate of π_0 is analogous to setting $\lambda = 0.5$ in the $\hat{\pi}_0$ proposed in Storey (2002). For multiclass data, the scores are all positive, so we use the 0th and 50th percentiles of the permuted values [NOTE: this was corrected in version 2.0].

9. The median and 90th percentile of the number of falsely-called genes from step 6, are multiplied by $\hat{\pi}_0$.

10. User then picks a Δ and the significant genes are listed.

11. The False Discovery Rate (FDR) is computed as median (or 90th percentile) of the number of falsely called genes divided by the number of genes called significant.

12. **Fold Change**. Suppose $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the average expression levels of a gene i under each of two conditions. These averages refer to raw (not logged) data. If a nonzero fold change t is also specified, then a positive gene must also satisfy $|\bar{x}_{i1}/\bar{x}_{i2}| \geq t$ in order to be called significant and a negative gene must also satisfy $|\bar{x}_{i1}/\bar{x}_{i2}| \leq 1/t$ to be called significant. When a fold change is specified, genes with either $\bar{x}_{i1} \leq 0$ or $\bar{x}_{i2} \leq 0$ (or both) are automatically left off of the significant gene list, as their fold change cannot be unambiguously determined. When such fold changes are reported in output, they are indicated by `NA`.

13. The **local FDR** for a gene is the false discovery rate for genes having a similar relative difference d_i as that gene. It is estimated by taking a symmetric window of 0.5% of the genes on each side of the target gene, and estimating the FDR in that window. If 1.0% times the total number of genes in the data set is less than 50, then the percentage is increased so that the number of genes is 50.

A.2 Computation of s_0

1. Let s^α be the α percentile of the s_i values. Let $d_i^\alpha = \frac{r_i}{s_i + s^\alpha}$.

2. Compute the 100 quantiles of the s_i values, denoted by $q_1 \leq q_2 \leq \ldots \leq q_{100}$.

3. For $\alpha \in (0, 0.05, 0.10 \ldots 1.0)$

   (a) Compute $v_j = \text{mad}(d_i^\alpha \mid s_i \in [q_j, q_{j+1})), j = 1, 2, \ldots n$, where mad is the median absolute deviation from the median, divided by 0.64.

   (b) Compute $cv(\alpha)$ = Coefficient of variation of the v_j values.

4. Choose $\hat{\alpha} = \text{argmin}[\text{cv}(\alpha)]$. Finally compute $\hat{s}_0 = s^{\hat{\alpha}}$. s_0 is henceforth fixed at the value of $\hat{s}_0$.

A.3 Details of r_i and s_i for different response types.

1. **Two class, unpaired data** $y_j = 1$ or 2. Let $C_k = \{j : y_j = k\}$ for $k = 1, 2$. Let $n_k = \#$ in C_k. Let $\bar{x}_{ij} = \sum_{j \in C_1} x_{ij}/n_1$, $\bar{x}_{i2} = \sum_{j \in C_2} x_{ij}/n_2$.

$$r_i = \bar{x}_{i2} - \bar{x}_{i1} \tag{A.2}$$

$$s_i = \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \frac{\sum_{j \in C_1}(x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C_2}(x_{ij} - \bar{x}_{i2})^2}{(n_1 + n_2 - 2)}} \tag{A.3}$$

2. **Multiclass response** $y_i \in \{1, 2, \ldots K\}$. Let $C_k =$ indices of observations in class k, $n_k = \#$ in C_k, $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$, $\bar{x}_i = \sum_j x_{ij}/n$.

$$r_i = \sqrt{\frac{\sum_k n_k}{\prod_k n_k} \sum_k n_k [\bar{x}_{ik} - \bar{x}_i]^2} \tag{A.4}$$

$$s_i = \sqrt{\frac{\sum_k \frac{1}{n_k}}{\sum_k (n_k - 1)} \sum_k \sum_{j \in C_k} [x_{ij} - \bar{x}_{ik}]^2} \tag{A.5}$$

Appendix B

Mixed Models Code

```
/* ------------------ Input Macro ---------------------- */

%macro readdata;

/* Input Gene Names */
data keys;
infile "Key.csv" expandtabs;
input Name Gene;
run;

%do i=1 %to 48;

/* Input Green Channel */
Data green;
infile "Green.txt" expandtabs;
input G Gb GFlag;
run;

/* Input Red Channel */
Data red;
infile "Red.txt" expandtabs;
input R Rb RFlag;
run;

/* Merge Data */
data arrayi;
merge green red keys;
array=&i;
spot=_n_;
if (rflag=0 and gflag=0) then do;
        dye="Green";
        diff=G-Gb;
        if(diff > 0) then logi = log2(diff);
        else logi = .;
        output;
        dye="Red";
        diff=R-Rb;
        if(diff > 0) then logi = log2(diff);
        else logi= .;
        output;
end;
keep array gene name spot dye logi;
run;

proc append base=arrayall data=arrayi;
run;

%end;

%mend;

/* Create Treatment Variable */
```

```
data preg;
set arrayall;
format strain $6.;
set preg;
if dye='Red' then do; strain="Ref"; end;
else do;
    if array<=12 then do; strain="B6:LPS"; end;
    else if array<=24 then do; strain="KO:LPS"; end;
    else if array<=39 then do; strain="B6:Sal"; end;
    else if array<=45 then do; strain="KO:Sal"; end;
    else if array<=48 then do; strain="Ref"; end;
end;
run;

/* ------------------ Normalization Model ---------------------- */

/* Normalize Data */
proc mixed data=preg covtest cl lognote;
class array strain;
model logi = strain / outp=pregnorm;
random array strain*array;
lsmeans strain / diff cl;
run;

/* ------------------ Gene Model ---------------------- */

/* Delete NA Genes */
data pregnorm;
set pregnorm;
if gene='' then delete;
run;

/* Sort For Gene Models */
proc sort data=pregnorm;
by gene array spot;
run;

/* Individual Gene Models */
ods listing close; run;
proc mixed data=pregnorm;
by gene;
class array spot dye strain;
model resid = strain;
random array spot(array);
lsmeans strain / diff;
ods output tests3=overall diffs=pairwise convergencestatus=conver;
run;
ods listing; run;

/* ------------------ Significant Genes ---------------------- */

/* Significant Pairwise Tests */
data pairwisesig;
set pairwise;
log10p = -log10(probf);
if log10p > -log10(0.05/194480);
run;

/* Summary Table of Significant Pairwise Tests */
proc freq data=pairwisesig;
tables _strain*strain;
run;
```

Appendix C

SAM Code

```
###############################################################################
### Input Data
###############################################################################

### Gene Names
gnameA<-read.csv(paste(dir3,'Mouse Slide A Key.csv',sep=''))$GB
gnameB<-read.csv(paste(dir3,'Mouse Slide B Key.csv',sep=''))$GB
gnameC<-read.csv(paste(dir3,'Mouse Slide C Key.csv',sep=''))$GB
glong<-c(as.character(gnameA),as.character(gnameB),as.character(gnameC))
gshrt<-unique(glong)[-1]

### Mouse Grid
mgrid<-list(
    A=list(nspot.r=20,nspot.c=20,ngrid.r=12,ngrid.c=4),
    B=list(nspot.r=20,nspot.c=20,ngrid.r=12,ngrid.c=4),
    C=list(nspot.r=16,nspot.c=21,ngrid.r=12,ngrid.c=4))
msize<-lapply(lapply(mgrid,unlist),prod)

### Setup Empty Mouse Data
matA<-matrix(0,19200,16)
matB<-matrix(0,19200,16)
matC<-matrix(0,16128,16)
mouse<-list(
    A=list(R=matA,G=matA,Rb=matA,Gb=matA,flag=matA),
    B=list(R=matB,G=matB,Rb=matB,Gb=matB,flag=matB),
    C=list(R=matC,G=matC,Rb=matC,Gb=matC,flag=matC))

### Read Mouse Data
for(i in 1:3){
    for(j in 1:16){
        grn<-read.table('grnpath',header=T,sep='\t',skip=42,nrows=msize[[i]])
        red<-read.table('redpath',header=T,sep='\t',skip=42,nrows=msize[[i]])
        mouse[[i]]$R[,j]<-red$Signal.Mean
        mouse[[i]]$G[,j]<-grn$Signal.Mean
        mouse[[i]]$Rb[,j]<-red$Background.Mean
        mouse[[i]]$Gb[,j]<-grn$Background.Mean
        if(!identical(red$Flag,grn$Flag)) stop('Flags not identical')
        mouse[[i]]$flag[,j]<-red$Flag
        }
    }

###############################################################################
### Normalize Data
###############################################################################

### Remove Flags
mdat<-list(A=mouse$A[1:4],B=mouse$B[1:4],C=mouse$C[1:4])
for(i in 1:3) for(j in 1:4) mdat[[i]][[j]][mouse[[i]][[5]]!=0]<-NA

### Normalize Within Arrays
MA.pts<-list()
MA.pts$A<-stat.ma(mdat$A,mgrid$A,norm='p')
```

```
MA.pts$B<-stat.ma(mdat$B,mgrid$B,norm='p')
MA.pts$C<-stat.ma(mdat$C,mgrid$C,norm='p')
for(i in 1:3) for(j in 1:2) colnames(MA.pts[[i]][[j]])<-paste('S',1:16,sep='')

### Normalize Between Arrays
MA.lall<-list(A=alist(A=,M=),B=alist(A=,M=),C=alist(A=,M=))
for(i in 1:3) MA.lall[[i]][['A']]<-stat.norm.exp(MA.pts[[i]][['A']])
for(i in 1:3) MA.lall[[i]][['M']]<-stat.norm.exp(MA.pts[[i]][['M']])
MA.all<-list()
MA.all$A<-stat.norm.exp(rbind(MA.lall$A$A,MA.lall$B$A,MA.lall$C$A))
MA.all$M<-stat.norm.exp(rbind(MA.lall$A$M,MA.lall$B$M,MA.lall$C$M))
for(i in 1:2) rownames(MA.all[[i]])<-glong

### Aggregate
MM.lagg<-apply(MA.all$M,2,function(x) aggregate(x,list(glong),mean,na.rm=T))
for(i in 1:16) print(sum(gshrt!=as.character(MM.lagg[[i]][[1]])))
MM.agg<-matrix(0,21529,16,dimnames=list(gshrt,names(MM.lagg)))
for(i in 1:16) MM.agg[,i]<-MM.lagg[[i]][[2]]
mean(is.na(MM.agg)) # 19.9% of the genes are NA
mean(apply(is.na(MM.agg),1,any)) # 39.9% of the genes have at least one NA

### Reduce
grps<-c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,NA)
grplabs<-c('LPS&B6','LPS&KO','SAL&B6','SAL&KO')
grpcnt<-matrix(0,21529,4,dimnames=list(gshrt,grplabs))
for(i in 1:4) grpcnt[,i]<-rowSums(!is.na(MM.agg[,grps==i]))
grpidx<-apply(grpcnt>=1,1,all)
MM.red<-MM.agg[grpidx,]

### Impute Missing Values
MM<-impute.knn(MM.red,k=15,rowmax=0.75,maxp=nrow(MM.red))

##############################################################################
### SAM Analysis
##############################################################################

### Data Prep
mat1<-MM[,1:15]
tmt1<-c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4)
gid1<-rownames(mat1)
gnm1<-paste('g',gid1,sep='')

### SAM Analysis
dat1<-list(x=mat1,y=tmt1,geneid=gid1,genenames=gnm1,logged2=T)
sam1<-samr(dat1,resp.type='Multiclass',random.seed=2528)

### False Discovery Rate
fdr1<-samr.compute.delta.table(sam1,dels=seq(0,0.8,by=0.1))
del1<-0.22
cut1<-samr.compute.delta.table(sam1,dels=del1)

### Compute Significant Genes
sig1<-samr.compute.siggenes.table(sam1,del1,dat1,fdr1)

### SAM plot
samr.plot(sam1,del1)

### Miss Rate
mis1<-samr.missrate(sam1,del1,cut1)

### Tail Strength
ets1<-samr.tail.strength(sam1)

### Simulated P-values
pva1<-samr.pvalues.from.perms(sam1$tt,sam1$ttstar)
```

Bibliography

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a parctical and powerful apporach to multiple testing," *Journal of the Royal Statistical Society*, Series B, 289–300.

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

Chu, G., Narasimhan, B., Tibshirani, R., and Tusher, V. (2005), *SAM users guide and technical document*, Stanford University Labs.

Chu, T. M., Weir, B., and Wolfinger, R. (2002), "A systematic statistical linear mdeling approach to oligonucleotide array experiments," *Mathematical Biosciences*, 176, 35–51.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2000), "Comparison of discrimination methods for the classifcation of tumors using gene expression data," Tech. Rep. 576, Department of Statistics, UC Berkeley.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple hypothesis testing," *Statistical Science*, 18, 71–103.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002), "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, 12, 111–139.

Efron, B. (2004), "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

Efron, B. and Tibshirani, R. (2002), "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology*, 23, 70–86.

Genovese, C. and Wasserman, L. (2003), "A stochastic process approach to false discovery rates," Tech. rep., Carnegie Mellon University.

Good, P. I. (2001), *Resampling methods*, Birkhauser, 2nd ed.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The elements of statistical learning*, Springer, 1st ed.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002), "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, 18, S96–S104.

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

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006), *SAS for mixed models*, Cary, NC: SAS Institute Inc., 2nd ed.

Lönnstedt, I. and Speed, T. P. (2002), "Replicated microarray data," *Statistica Sinica*, 12, 31–46.

Shaffer, J. P. (1995), "Multiple hypohtesis testing: A review," *Annual Review of Psychology*, 45, 561–584.

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, Article 3.

Smyth, G. K. and Speed, T. (2003), "Normaliaztion of cDNA microaaray data," *Methods*, 31, 265–273.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society*, Series B, 479–498.

Storey, J. D. and Tibshirani, R. (2001), "Estimating false discovery rates under dependence, with application to DNA microarrays," Technical Report 28, Department of Statistics, Stanford University.

— (2003), "Statistical significance for genome wide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Tanaka, T. S., Jaradat, S. A., Lim, M. K., Kargul, G. J., Wang, X., Grahovac, M. J., Pantano, S., Sano, Y., Piao, Y., Nagaraja, R., Doi, H., III, W. H. W., Becker, K. G., and Ko, M. S. H. (2000), "Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray," *Proceedings of the national academy of sciences*, 97, 9127–9132.

Taylor, J. and Tibshirani, R. (2006), "A tail strength measure for assessing the overall univariate significance in a dataset," *Biostatisics*, 7, 167–181.

Taylor, J., Tibshirani, R., and Efron, B. (2005), "The *miss rate* for the analysis of gene expression data," *Biostatistics*, 6, 111–117.

Tibshirani, R. (2005), "A simple method for assessing sample sizes in microarray experiments," Www-stat.stanford.edu/∼tibs/SAM.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing value estimation methods for DNA microarrays," *Bioinformatics*, 17, 520–525.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, 98, 5116–5121.

Šidák, Z. (1967), "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, 62, 626–633.

Westfall, P. H. and Young, S. S. (1993), *Resampling based multiple testing: examples and methods for p-value adjustment*, Wiley series in probability and mathematical statistics, Wiley.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamdeh, H., Bushel, P., Afshari, C., and Paules, R. (2001), "Assessing gene significance from cDNA microarray expression data via mixed models," *Journal of Computational Biology*, 8, 625–637.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, 30, e15:1–e15:11.