New methods for designing and analyzing microarray experiments for the detection of differential expression

Justin C. Recknor
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Bioinformatics Commons, and the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/3049

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
New methods for designing and analyzing microarray experiments for the detection of differential expression

by

Justin C. Recknor

A thesis submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Statistics;
Bioinformatics and Computational Biology

Program of Study Committee:
Dan Nettleton, Co-major Professor
James Reecy, Co-major Professor
Phillip Dixon
Ken Koehler
Chris Tuggle

Iowa State University
Ames, Iowa
2006

Copyright © Justin C. Recknor, 2006. All rights reserved.
# TABLE OF CONTENTS

Abstract ................................................................. v

General Introduction ..................................................... 1

1 Microarrays ................................................................. 1
  1.1 Cell life ............................................................... 1
  1.2 Spotted Microarrays ................................................. 3
  1.3 Oligonucleotide Microarrays ....................................... 5

2 Gene Ontology Consortium .............................................. 6

3 Dissertation Organization ............................................... 8

4 References ................................................................. 10

Design and Analysis of Two-Variety Two-Dye Microarray Experiments
for the Detection of Differentially Expressed Genes ................. 12

Abstract ................................................................. 12

1 Introduction ............................................................... 12

2 Gene-Specific Models for Two-Variety Two-Dye Microarray Experiments 16
  2.1 Example Experiments ............................................... 16
  2.2 A Mixed Linear Model for the DBWU Design .................. 17
  2.3 A Mixed Linear Model for the DBWV Design .................. 18

3 Methods of analysis ................................................... 19
  3.1 Mixed Linear Model Analysis Using SAS ......................... 19
  3.2 Analysis by Differences ........................................... 22
Microarray Experimental Design Selection Accounting for Mixed Model Variability

Abstract ................................................................................. 38

1 Introduction ........................................................................... 38

2 Benefit of Using Mixed Model Analysis ................................. 43

3 Mixed Model design comparisons ......................................... 46

3.1 Equation Comparisons ...................................................... 46

3.2 Comparing Designs on the Basis of Pilot Data ..................... 47

3.3 Benefits in Using Mixed Model Design Comparisons .......... 49

4 Examples ................................................................................ 50

4.1 Mixed Model Design Selection Versus Array Differences ...... 50

4.2 Comparison of Designs Which Cannot Be Considered Using Array Differences ........................................ 52

5 Design Comparisons Accounting for Cost ............................. 55

6 Considerations ....................................................................... 62

7 Conclusion ............................................................................. 64

8 References .............................................................................. 65
Identification of Differentially Expressed Functional Categories in Microarray Studies Using Nonparametric Multivariate Analyses

Abstract

1 Introduction

2 Multi-Response Permutation Procedure (MRPP)
   2.1 MRPP in Microarray Analysis
   2.2 P-values
   2.3 Considerations for Variance Differences Between Genes
   2.4 Benefits

3 Resampling Based P-value Test Method
   3.1 Testing Interaction

4 Regulation

5 Influential Genes

6 Examples
   6.1 RBPT versus MRPP

7 Conclusion

8 References

General Conclusion

Acknowledgments

Appendix
   1 Example Code
   2 Containment Method - Order Dependency
Abstract

This thesis is divided into three sections all pertaining to microarray experimental design and analysis. Microarrays are a tool used in biological research which enables scientists to measure the relative level of expression many genes within an organism at the same time. Microarrays have also opened new research areas in statistics which are currently being investigated concerning different aspects of data normalization, experimental design, and analysis.

The first chapter entails a comparison of two commonly used experimental designs in two-dye microarray experiments. Both designs are applicable only to experiments containing treatments with two levels. One design is shown to be more powerful when constrained by the number of arrays. Also, mixed model analysis is often used for both designs. With small sample sizes, mixed model analysis is shown to give inaccurate results under certain conditions. Due to this problem, an alternative method of analysis is proposed for both experimental designs which eliminates this concern.

Two-dye microarray experiments require special consideration in design since they have multiple random effect in the model. This is because arrays are usually viewed as a random factor that should always be contained in a model for the data. Research has been done on comparing two-dye microarray experimental designs by requiring calculation of array differences. This is shown to inhibit the power of the analysis by removing inter-block information. There are also experimental designs that are viable options which can not be compared using this method. An alternative method of analysis is proposed which allows for multiple random effects in the model. Under certain condi-
tions, this method is shown to choose designs that either would not be chosen, or cannot be considered, when using methods based on array differences.

The third chapter discusses new methods for analyzing microarray experiments by categories. Most commonly, microarray analysis is performed on a gene-by-gene basis with the goal of finding the genes whose expression differ the greatest between varieties of treatments. However, scientists often would like to know what aspect of cell life is affected most by differences in varieties. There could be cases where a group of genes pertaining to the same task are all have a mild change in expression which would not be found using gene-by-gene analysis. Two different resampling based methods are proposed for solving this problem. Both methods are compared and results are visualized on a directed acyclical graph.
General Introduction

1 Microarrays

1.1 Cell life

Microarrays can be used to approximate the protein production of the genes expressed within the cells of an organism. Gene expression determines the actions of a cell and how it interacts with its neighboring cells. Within the nucleus of all cells of both plants and animals, is deoxyribonucleic acid (DNA) which can be viewed as a long linear sequence of bases. This sequence is responsible for hereditary traits passed down to offspring. The DNA is distributed among long linear structures called chromosomes. There can be 1 to more than 50 chromosomes within an organism. The size and the number of chromosomes is the same in each cell of an organism but varies among different organisms. Within DNA, a total of four different bases can be found. They are adenine(A), cytosine(C), guanine(G), and thymine(T). Due to their chemical structure, A will bind to T and C will bind to G. This is important because a sequence of bases will have chemical attraction to its complement. For example if the sequence of one strand is $\cdots$A-G-C-A$\cdots$ than its complement strand is $\cdots$T-C-G-T$\cdots$ and these sequences would naturally bind to each other.

Certain regions of the chromosome are referred to as genes. Within genes, specific base sequence code for specific protein(s). Proteins account for almost all cellular activity while also forming the key elements of a cell, such as the cell walls in plants. The process
of creating proteins from genes is the central dogma of molecular biology and consists of two stages: transcription and translation. Through transcription, mRNA, which is also a sequence of bases, is created. The mRNA is the complement to the gene sequence except that the base thymine, is replaced by uracil (U). An example of transcription is shown in Figure 1.

\[ \begin{align*}
\text{DNA} & \quad \cdots A - C - T - A - G \\
\downarrow & \quad \downarrow \quad \downarrow \quad \downarrow \\
\text{mRNA} & \quad \cdots U - G - U - U - C
\end{align*} \]

Figure 1  Transcription.

The mRNA is then converted to the corresponding protein through a process called translation. A protein is a three dimensional structure consisting of a sequence of amino acids. In mRNA, a sequence of three bases indicates one amino acid. After translation, there is approximately one third the number of amino acids in the protein as there are bases in the mRNA. An example of translation is shown in Figure 2.

\[ \begin{align*}
\text{mRNA} & \quad \cdots A - G - U - U - C - C \cdots \\
\quad & \quad \underset{\text{AA}_1}{A - G - U - U - C - C} \quad \underset{\text{AA}_2}{A - U - C - C} \\
\end{align*} \]

Figure 2  Translation. AA, indicates an amino acid.

The preceding examples are oversimplified. In organisms, such as humans, a chromosome is millions of base pairs long while gene sequences are usually thousands of bases long. More than one protein can be made from a single gene sequence and a large number of mRNA sequences can be translated simultaneously. Different types of cells within the same organism, such as skin cells and bone cells, will express different genes, enabling the cell to perform the correct duties. Cells affected by a stress, such as a disease, will usually change their normal expression patterns. Learning what genes are
expressed in an organism resistant to a disease versus a similar organism that is susceptible can give valuable information toward deriving a cure. This is just one example but there are many other ways in which genome expression information is useful.

Microarray experiments measure the amount of mRNA that is currently present within the cells of the organism that is sampled. Measuring the mRNA does not give a direct measure of the amount of protein being produced because translation is necessary for the protein to be created. It has been shown that mRNA can be transcribed without being translated. However, research has also shown that the amount of mRNA present can give a good estimate of the amount of protein being produced (Lodish et al., 2000). When performing microarray experiments, the mRNA removed from the cells of an experimental unit has to be synthesized to complementary DNA (cDNA). The cDNA is complementary to the mRNA sequence except that where there would be uracil(U) in a complementary mRNA sequence, there is now thymine(T). Converting the mRNA to cDNA makes the sample more chemically stable which creates more consistency in the experimental results.

1.2 Spotted Microarrays

The spotted microarray, or two-channel microarray, is a glass slide containing a large number of small spots. There can be more than 30,000 spots for different genes on a single array. Each spot contains a large number of identical strands that are usually 200-1000 bases long. These strands are made so that they are complementary to the cDNA for a specific gene created from the original mRNA sample. At every spot, a large number of strands which are complementary to a specific gene are applied. The complementary nature of the strands enable the cDNA for that gene to attract to and bind with the strands on the microarray slide. The cDNA from a sample is labeled with a fluorophore, or dye. There are two different colored dyes used in an experiment. One is usually red and the other green. Using different dyes enables the measurement of two
different samples on the same array. Measurements are obtained by exciting the samples with a laser and measuring both the green and red fluorescence at each spot.

As an example, assume there is to be a study analyzing gene expression differences between cells from a person with cancer versus another one without cancer. First, mRNA samples would be taken from each subject. The mRNA would then be converted to cDNA and the red dye would be applied to the sample from the cancer subject while the green dye would be applied to the sample from the subject without cancer. Both dyed samples would then be hybridized to the microarray. Each spot would then be excited by a laser and measurements would be obtained for each gene and each dye. An example of an array is shown in Figure 3.

![Figure 3](image-url)  
Figure 3  A spotted microarray. Bright colors indicate high expression levels, while dark colors indicate low expression levels. The red spots indicate expression predominantly in the subject receiving the red dye while green spots indicate the opposite. Yellow spots indicate roughly equal expression levels.
A problem found when working with spotted microarrays is that the intensity measurement is biased depending upon the dye. A larger measurement can be obtained from one dye versus the other for the same amount of cDNA. Because this is considered a global effect and is introduced because of technical difficulties, normalization is often used prior to the analysis to remove the dye bias. There have been many methods proposed for normalization of the data (Wolfinger et al., 2001), many of which involve some form of loess regression (Smyth and Speed, 2003; Yang et al., 2002). The analysis is then performed on the data resulting from the normalization.

1.3 Oligonucleotide Microarrays

Oligonucleotide microarrays are created using different procedures than those used for spotted arrays. DNA oligonucleotides, which are short sequences of about 25 bases, are synthesized to the surface of a glass slide. The DNA oligonucleotides are created to be complementary to a gene sequence region that is unique. The uniqueness is necessary so that only the sequence of the gene of interest will bind to the oligonucleotides and no other. Probes contain tens of thousands of identical oligonucleotides placed within small square areas on the slide. Each gene will have several different probes and each will be complementary to a different portion of the gene sequence. These probes for one gene are placed on different regions of the slide. The mRNA samples are then tagged with a fluorescent dye and applied to the slides. The quantity is measured by shining a laser upon each probe and the fluorescence of the tagged samples is then calculated. An example of an oligonucleotide microarray is shown in Figure 4.

When using oligonucleotide microarrays, researchers obtain a probe set of observations for each gene as opposed to one measure. There is concern for a possible bias being added to the measurements of each probe according to their location. Methods such as MAS 5.0 (Affymetrix) and RMA (Irizarry et al., 2003) have been proposed to remedy these concerns. One of these normalization methods is chosen for use prior to
the analysis to condense the data and minimize any bias that could be introduced because of technical difficulties. Analysis is then performed on the values received after normalization.

2 Gene Ontology Consortium

The Gene Ontology Consortium was started in 1998 at www.geneontology.org with the goal of aiding in the efficiency of scientific research (Gene Ontology, 2000). A problem biologists can have is finding information of other scientists that pertains to their special research area. Additional problems were generated by inconsistent terminology making information even more difficult to find and share. Gene Ontology (GO) is a collaborative effort to create consistent descriptions of gene products throughout different databases.
GO has developed three separate ontologies or vocabularies which are: biological processes, cellular components and molecular functions. Each ontology is designed to describe gene products in a species-independent manner. The goal of making GO terminology species-independent is because different organisms have large numbers of gene products which accomplish similar tasks. Information can be obtained about the gene product in the organism being studied by its similarity to gene products of other organisms. The biological processes ontology is “a series of events accomplished by one or more ordered assemblies of molecular functions.” (Gene Ontology, 2000) The molecular functions is defined to be “the action characteristic of a gene product.” (Gene Ontology, 2000) The molecular functions and the biological processes do overlap and so there is a rule that the biological processes must have more than one distinct step to help separate them. The cellular component is the part of the cell that the gene product is active within.

Each GO term in each ontology is associated with a unique numerical identifier of the form GO:xxxxxxxx, where the x’s indicate a 7 digit number. Each term also usually has a definition. Every ontology relates the terms, or GO numbers, through a directed acyclic graph. This graph is the same as a phylogenic tree except that a child node can have more than one parent as shown in Figure 2.

The GO Consortium can be used in analyzing microarray experiments where data is usually analyzed on a gene by gene basis. In these experiments, scientists are often looking for networks consisting of many genes to explain biological differences. They can investigate the list of genes whose null hypothesis, of equal expression levels amongst the varieties, are rejected under some criterion and try to associate a network through this list. However, there could be a family of genes, many with relatively small mean differences, that are all having expression changes and could hold valuable information. Since many of the genes within such a family will have small mean differences, most or even all of them may not appear on the list of genes declared to have significant mean
Figure 5  One region of the molecular function hierarchy with the broad-
est point at the bottom and specificity increasing going upward. Each node contains its GO number and has arrows pointing at its ancestor. Nodes colored orange have more than one ancestor.

differences. The family of genes could then go undetected. Alternatively, tests could be performed by the GO categories instead of by the gene. Analysis using this perspective could be used to investigate the entire family simultaneously, giving additional power to finding groupings or families of genes which would not be discovered otherwise.

3 Dissertation Organization

There are three separate sections to this dissertation addressing three issues surrounding microarrays. In Chapter 2, we deal with an experimental design commonly called the reverse dye design and a comparison with an alternative design. Both designs are for experiments where there are only two varieties, or treatment levels. An alternative design is shown to be more powerful, when constrained by the number of arrays, through variance comparison as well as an actual experiment performed twice, once using each design. Also, mixed model analysis is often used in both designs even when the size of the experiment is small. This leaves a large number of random effects relative to the number of observations which is shown to return untrustworthy results. Alternative
analysis methods are proposed for both designs which eliminates this problem.

In Chapter 3, we investigate mixed model analysis is commonly used in microarray data analysis due to the experimental designs. When using spotted two-dye arrays, it could be argued that mixed model analysis should have to be used for two reasons. It has been shown that array effects can be gene specific (Kerr, 2003), thus requiring them to be included in the model for each gene. Arrays, or slides, are usually viewed as random effects because an experiment is commonly performed by purchasing arrays containing sequences specific to the organism of interest. Usually, many other experiments are being performed on other arrays containing the same sequences. This enables the arrays to be viewed as a random sample from a large population indicating that they should be modeled as a random factor. Mixed model analysis is required to correctly handle this. However, there have been many methods proposed in the past to compare fixed effects only models. Various researchers have proposed reducing the data by calculating the differences between the observations from a common array and then analyzing the differences instead of the observed values (Kerr and Churchill, 2001; Yang and Speed, 2002; Glonek and Soloman, 2004). This enables comparison of designs by the fixed effects only methods. However, analyzing the differences is shown to result in a loss of power in certain circumstances. It is shown here that to maximize the power, mixed model analysis should be used. If mixed model analysis is going to be used, then the experimental designs should be compared under these conditions to get correct results. A method of experimental design comparison which compares designs based upon the mixed model requirements is proposed. This method allows more models to be considered, accounts for multiple random effects, and is shown to possibly result in different design selections than the alternative methods.

In Chapter 4, we propose an alternative to the usual gene-by-gene analysis performed on microarray experiments. Gene-by-gene analysis leaves researchers with a list of genes to continue further research with after the microarray experiment. An alternative would
be to group the genes based upon their molecular functions, and test this group or category for differential expression between the treatments or varieties. This method would leave scientists with a list of genes to research as well as additional information about molecular functions operating differently between the varieties. There could also be cases where an entire category of genes all change expression between the varieties, but do so at a relatively small magnitude. None of these genes may be found by testing each gene independently. Situations like this would be found when alternatively testing the category. Two different methods for testing gene categories using resampling are proposed. One method deals with the multivariate distribution of the data while the alternative uses comparative measures of the p-values from every gene within a category. Each method has specific situations where one would be preferred over the other. In this dissertation, both methods are used in an example and compared.

4 References

Affymetrix: Statistical Algorithms Description


Design and Analysis of Two-Variety Two-Dye Microarray Experiments for the Detection of Differentially Expressed Genes

Justin C. Recknor and Dan Nettleton

Abstract The reverse dye design has often been used in spotted two-dye microarray experiments when there are only two varieties being studied. An alternative design is shown to be more powerful when the defining factor is the number of arrays through an actual experiment and simulation. The data from either design can be analyzed using mixed model analysis, but we show through simulation that the resulting tests can be too conservative or too liberal when model variance components are estimated to be zero. We show that this problem is eliminated by a simple alternative analysis strategy in which the data from each spot are reduced to a difference prior to the analysis.

1 Introduction

Over the past decade, two-dye microarray technology has become a standard tool in functional genomics research. Researchers use two-dye microarrays to measure gene
expression changes across biological samples of multiple types. By understanding how specific genes change expression across carefully selected sample varieties, researchers gain information about the role of genes in important biological processes.

A single two-dye microarray slide can be used to measure the expression of thousands of genes in each of two samples. The samples compared on a single slide must be dyed with two different fluorescent dyes to obtain separate measures of expression for each sample. Effects on expression associated with sample variety are of scientific interest, but the dyes themselves exhibit effects that can obscure variety effects if experiments are not designed properly (Kerr and Churchill, 2001a). Various normalization strategies have been proposed for removing dye effects prior to statistical analysis (Bolstad et al., 2001; Yang et al., 2002; Smyth and Speed, 2003; Cui et al., 2003). These strategies successfully reduce global effects common to all genes or to groups of genes with similar intensity levels, but gene-specific dye effects will remain after normalization (Kerr and Churchill, 2001b). Thus it is important to account for dye effects in the design and analysis of two-dye microarray experiments.

The simplest and perhaps most common two-dye microarray experiment involves comparing the expression of thousands of genes between two varieties (e.g., treated vs. control, cancerous tissue vs. non-cancerous tissue, mutant vs. wild type, etc.). Dobbin, Shih, and Simon (2003) discussed experimental design issues for such experiments. In particular, they identified the most efficient design for identifying changes in gene expression between two varieties when the number of slides is the limiting factor. The most efficient design is depicted in Figure 1 B. A commonly used alternative design is depicted in Figure 1 A. In this figure, each circle represents a separate experimental unit (in the case of a randomized experiment) or sampling unit (in the case of an observational study). (Henceforth we will use the more generic term unit to refer to either type of unit depending on the type of investigation under consideration.) Each circle is numbered according to its variety. Each arrow in Figure 1 represents a microarray
slide upon which RNA from the two units connected by the arrow will be measured. The direction of the arrow indicates the dye assignment by the following convention: the arrow points from the unit dyed with dye 1 (e.g., green Cyanine 3) to the unit dyed with the dye 2 (e.g., red Cyanine 5). This symbolic notation is motivated by the figures in Kerr and Churchill (2001a). The main difference is that we insist on separate circles for separate units.

![Experimental Designs](image)

Figure 1 Experimental Designs

The adjectives dye-swap, reverse-dye, fluor-reversal, and dye-balanced have been used in the microarray literature to refer to designs like A in Figure 1 (Dobbin, 2002; He et al., 2003; Landgrebe et al., 2006). Both designs in Figure 1 involve direct comparison of varieties on single slides and measurement of each variety an equal number of times with each dye. The feature that distinguishes design A from design B is that in design B each unit is measured only once with only one of the two dyes while in design A each unit is measured twice, once with each dye. Thus design B is dye-balanced within variety (DBWV) while design A is dye-balanced within unit (DBWU). Designs that are DBWU will necessarily be DBWV, but we will treat these descriptors as mutually exclusive by using DBWV to describe the class of designs that are dye-balanced within variety but not dye-balanced within unit.
This paper provides a detailed comparison of the DBWV and DBWU designs described in Figure 1. We focus on analysis of data from such designs, and compare the performance of the two designs using real and simulated data. Our results support the conclusions of Dobbin et al. (2003) in that we find the DBWV design to be superior to the DBWU design when the number of slides is the limiting design factor. There are, however, situations where the number of units is the limiting design factor, and in such cases, the DBWU design is clearly preferable to the DBWV design. Thus it is important to develop appropriate analysis strategies for both designs.

While it is relatively straightforward to specify reasonable mixed linear models for the DBWV and DBWU designs using the approach described by Wolfinger et al. (2001), it is less straightforward to determine the best analysis strategy within the mixed linear model framework. Important issues regarding the determination of denominator degrees of freedom for tests of interest in mixed linear model analysis of microarray data have gone unaddressed in the literature. We compare multiple analysis strategies for the DBWV and DBWU designs and provide recommendations for practitioners that should lead to good performance when testing for differential expression between two varieties using either of the designs in Figure 1.

In Section 2, we describe two datasets generated by a single laboratory using the two competing designs pictured in Figure 1. We present a gene-specific mixed linear model for each design that should capture the important sources of variation in each experiment. Different methods of analysis are then described in Section 3. The power and validity of these methods are examined using simulated data in Section 4. The methods of analysis determined to be most appropriate are then used to compare the effectiveness of designs A and B for the analysis of actual data in Section 5.
2 Gene-Specific Models for Two-Variety Two-Dye Microarray Experiments

2.1 Example Experiments

Two microarray experiments were conducted to identify genes differentially expressed between kernels from normal maize and kernels from maize carrying a mutation to the *Opaque2* gene. Corn from the mutant maize plant is nutritionally more valuable than normal corn because the mutant contains higher levels of the essential amino acids lysine and tryptophan. Unfortunately the mutant corn has other undesirable properties (e.g., softer and more easily damaged kernels and lower yield) that prevent its widespread use as a food source for humans and livestock. The goal of the study was to identify genes differentially expressed in the mutant to better understand the molecular basis of phenotypic differences between the normal and mutant. Such information might eventually be useful for the development of highly nutritious corn without negative characteristics.

Experiments using the DBWU and DBWV designs depicted in Figure 1 were used to compare expression in normal and mutant kernels. For these experiments, each unit (circle in Figure 1) corresponds to a corn plant from which kernels of corn were sampled from a single ear. A total of 18 plants were used (three pairs of plants for the DBWU design and six pairs of plants for the DBWV design). The experiments each used six microarray slides. As is the case in many microarray experiments, the array work was completed over the course of multiple days. In this case, two arrays were processed each day for three days for each design. The dye-reversed slides for a single pair of plants were processed together on a single day when conducting the the DBWU design. Analogously, dye-reversed slides were processed together when conducting the DBWV design so that slides like the first two depicted in Figure 1 would be done on a single day.
2.2 A Mixed Linear Model for the DBWU Design

We propose the following mixed linear model for the normalized log-scale data for single gene from the DBWU design:

\[ y_{dvaub} = \mu + D_d + V_v + A_a + U_u + B_b + \epsilon_{dvaub}, \]  

(1)

where \( y_{dvaub} \) denotes the normalized log-scale expression measure from array \( a \), dye \( d \), variety \( v \), unit \( u \), and block \( b \); \( \mu \) denotes a fixed intercept parameter; \( D_d \) denotes the fixed effect of dye \( d \); \( V_v \) denotes the fixed effect of variety \( v \); \( A_a \) denotes the random effect of array \( a \); \( U_u \) denotes the random effect of unit \( u \); \( B_b \) denotes the random effect of block \( b \); and \( \epsilon_{dvaub} \) denotes the random error associated with the observation from array \( a \), dye \( d \), variety \( v \), unit \( u \), and block \( b \). All the random terms are assumed to be independent and normally distributed with with mean 0 and variance \( \sigma^2_A \), \( \sigma^2_U \), \( \sigma^2_B \), and \( \sigma^2 \) for arrays, units, blocks, and error terms, respectively. To simplify notation, we have omitted a gene-specific subscript that would indicate that all the parameters are unique for each gene.

For the maize DBWU experiment described in the previous subsection, \( d = 1, 2 \) for the 2 dyes; \( v = 1, 2 \) for the normal and mutant genotypes; \( a = 1, \ldots, 6 \) for the 6 arrays; \( u = 1, \ldots, 6 \) for the 6 plants; and \( b = 1, 2, 3 \) for the three processing days used to obtain the microarray data. It is natural to treat the effects of arrays as random because the arrays used in any one experiment are almost always a subset of a larger population of arrays manufactured together for use in many experiments. Similarly the effects of units are treated as random because it is usually reasonable to view the units used in any particular experiment as being like a random sample from a larger population of units that could have been used in the experiment. Furthermore, it is desirable to generalize our conclusions concerning variety effects to the larger population of units. The random block effects in this case correspond to the effects associated with processing days. Alternatively, these block effects could be considered fixed. We discuss
the implications of this modeling choice in Section 3.

If we adopt the convention of ordering the response first by blocks, then by arrays, and then by treatments; the covariance matrix for the DBWU design is block diagonal with blocks of the form

\[
\begin{pmatrix}
\sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_U^2 & \sigma_A^2 + \sigma_B^2 & \sigma_B^2 + \sigma_U^2 & \sigma_B^2 \\
\sigma_A^2 + \sigma_B^2 & \sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_U^2 & \sigma_B^2 & \sigma_B^2 + \sigma_U^2 \\
\sigma_B^2 + \sigma_U^2 & \sigma_B^2 & \sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_U^2 & \sigma_A^2 + \sigma_B^2 \\
\sigma_B^2 & \sigma_B^2 + \sigma_U^2 & \sigma_A^2 + \sigma_B^2 & \sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_U^2
\end{pmatrix}
\]

(2)

Note that this covariance structure is quite complex for dataset with only 12 observations.

### 2.3 A Mixed Linear Model for the DBWV Design

We propose the following mixed linear model for the normalized log-scale data for single gene from the DBWV design:

\[
y_{dvaub} = \mu + D_d + V_v + A_a + B_b + \epsilon_{dvaub},
\]

(3)

where \(\epsilon_{dvaub} = U_u + \epsilon_{dvaub}\) and all other terms are as defined for the DBWU model in (1). Again, a gene-specific subscript on all parameters has been omitted to simplify notation. In contrast to the DBWU design, \(\sigma_U^2\) and \(\sigma^2\) cannot be separately estimated with data from the DBWV design because each unit is measured with only one array. However, the sum \(\sigma_U^2 + \sigma^2\) can be estimated with the DBWV design, and this sum, rather than individual estimates, is sufficient for inference concerning the variety effects of interest.
The covariance structure is again block diagonal with blocks given by

\[
\begin{bmatrix}
\sigma_A^2 + \sigma_B^2 + \sigma_U^2 + \sigma^2 \\
\sigma_A^2 + \sigma_B^2 \\
\sigma_A^2 + \sigma_B^2 + \sigma_U^2 + \sigma^2 \\
\sigma_A^2 + \sigma_B^2 \\
\sigma_A^2 + \sigma_B^2 + \sigma_U^2 + \sigma^2 \\
\sigma_A^2 + \sigma_B^2 \\
\sigma_A^2 + \sigma_B^2 + \sigma_U^2 + \sigma^2 \\
\sigma_A^2 + \sigma_B^2 \\
\sigma_A^2 + \sigma_U^2 + \sigma^2
\end{bmatrix}
\]

(4)

3 Methods of analysis

3.1 Mixed Linear Model Analysis Using SAS

The MIXED procedure in SAS, version 9.1 (SAS Inc., Carey, NC) is perhaps the most widely used code for mixed linear model analysis. The MIXED procedure uses flexible and convenient syntax that permits the fitting of general mixed linear models as described by Littell, et al.; 2006. It is straightforward to specify the mixed linear models for the DBWU and DBWV designs in MIXED syntax and to fit the models to thousands of genes in a sequential manner. Example code is provided in the appendix.

When using the MIXED procedure to fit the DBWU and DBWV models, p-values from F-tests (or, equivalently, t-tests) for variety and dye main effects are automatically generated. The p-values from the tests for variety main effects can be used to identify differentially expressed genes. A common strategy is to reject the null hypothesis of no differential expression for any gene whose p-value falls below a threshold for significance. Usually the threshold for significance is chosen to obtain approximate control of the false discovery rate (FDR). Benjamini and Hochberg (1995) and Storey and Tibshirani (2003) among many others have provided methods for FDR control that can be applied to a collection of p-values. All of these methods require that a p-value from a test with a true null hypothesis be uniformly distributed on the interval [0, 1].

Unfortunately, the assumption that p-values will be uniformly distributed under the
null is in doubt when conducting mixed linear model analysis of DBWU and DBWV experiments like those described in Section 2.1. When sample sizes are small and covariance structures are complex, the restricted (or residual) maximum likelihood (REML) estimators for one or more variance components may fall on the boundary of the parameter space with non-negligible probability. When this happens, the MIXED procedure essentially fits a model that excludes all random factors whose variance components were estimated to be zero. To understand how this can lead to tests whose p-values are not uniformly distributed under the null, it is necessary to consider SAS options for determining the denominator degrees of freedom of the reference distribution used to compute a p-value.

The default SAS method for determining the denominator degrees of freedom is the containment method. With the containment method, the syntax used to specify the model determines the denominator degrees of freedom according to rules described in SAS/STAT (1989). There are some drawbacks to the containment method. First, it is possible that multiple specifications of the same mixed linear model will lead to different inference concerning fixed effects of interest. An example is provided in the appendix. Second, the procedure can result in conservative inference by using substantially fewer denominator degrees of freedom than appropriate when considering the model ultimately fit to the data. For example, consider a gene from our DBWU experiment for which the array, unit, and block variance components are all estimated to be zero when SAS code for the DBWU model (provided in the appendix) is used for analysis. In this case, the MIXED procedure will essentially fit a fixed-effects-only model with three free parameters and 12 observations. An analysis based on this model should yield 9 denominator degrees of freedom for a test of variety main effects. However, the containment method will use only 2 denominator degrees of freedom when determining the p-value for the test. This results in a null p-value distribution that is stochastically larger than the uniform distribution and leads to conservative inference as shown in the simulations of
The other methods for determining the denominator degrees of freedom for models like (1) and (3) are the Satterthwaite method (Satterthwaite, 1946) and the Kenward and Roger method (Kenward and Roger, 1997). For t-tests or F-tests with a single numerator degree of freedom, the Kenward and Roger method is equivalent to Satterthwaite’s method. Thus it is sufficient to consider only Satterthwaite’s method here, given that the test of variety main effects is a single degree of freedom F-test.

Satterthwaite’s method is typically used when the denominator of an F-statistic is a linear combination of mean squares with differing expectations. In such instances, the null distribution of the test statistic will be only approximately F-distributed. The Satterthwaite method uses the observed mean squares involved in the linear combination and their degrees of freedom to determine a denominator degrees of freedom value that can be used to conduct the approximate F-test.

If the Satterthwaite method is selected when fitting models (1) and (3) using the MIXED procedure, the tests for variety main effects will be identical to those obtained using the containment method, provided that no variance components are estimated to be zero. On the other hand, if one or more variance components are estimated to be zero, the Satterthwaite denominator degrees of freedom will be redetermined for the model that excludes all random factors with zero variance component estimates. This is in contrast to the containment method where the denominator degrees of freedom depend only on the syntax used to specify the initial model. The redetermined Satterthwaite degrees of freedom are, in many cases, the same as the degrees of freedom that would have been obtained using either method had the random factors with zero variance component estimates been initially excluded from the model. For example, consider again a gene from our DBWU experiment for which the array, unit, and block variance components are all estimated to be zero. When the Satterthwaite method for determining the denominator degrees of freedom is specified for the fit of the initial model (1), the
denominator degrees of freedom will be recalculated to be 9 to match the error degrees of freedom of the fixed-effects-only model that is ultimately fit to the data because of the zero variance component estimates.

When either model (1) or (3) is fit to data for thousands of genes, the Satterthwaite-derived denominator degrees of freedom will vary from test to test depending on which variance components are estimated to be zero. The null distribution of the p-value for the test of variety main effects that results from this testing procedure is unknown and unfortunately intractable. We investigate this distribution via simulation in Section 4.

3.2 Analysis by Differences

As an alternative to the mixed linear model analysis described above, the data can be reduced to differences prior to analysis. Consider first the DBWV design and the differences formed by subtracting the variety 2 normalized log-scale expression measure from the variety 1 normalized log-scale expression measure to obtain one difference for each array. Based on model (3), this will lead to differences of the form

\[
\begin{align*}
\text{diff}_1 &= (V_1 - V_2) + (D_1 - D_2) \cdot (1) + \delta_1 \\
\text{diff}_2 &= (V_1 - V_2) + (D_1 - D_2) \cdot (-1) + \delta_2 \\
&\vdots \\
\text{diff}_{n-1} &= (V_1 - V_2) + (D_1 - D_2) \cdot (1) + \delta_{n-1} \\
\text{diff}_n &= (V_1 - V_2) + (D_1 - D_2) \cdot (-1) + \delta_n,
\end{align*}
\]

where \( n \) is an even integer denoting the number of arrays and \( \delta_1, \ldots, \delta_n \) are independent and identically distributed normal random variables with mean 0 and variance \( 2(\sigma_U^2 + \sigma^2) \).

Note that \( V_1 - V_2 \) and \( D_1 - D_2 \) are the intercept and slope, respectively, of a simple linear regression model with response values \( \text{diff}_1, \ldots, \text{diff}_n \) and explanatory values \( x = 1, -1, \ldots, 1, -1 \). Thus a test of variety effects can easily be obtained via the simple linear regression t-test or F-test for non-zero intercept. Under the null hypothesis of
no variety effects, the test statistic will have a central $t$ or $F$-distribution with $n - 2$ denominator degrees of freedom, and this test will be exact even if one or more of the variance components in model (3) are truly zero. Thus, as long as the assumptions of model (3) hold, the p-value for this variety test will be uniformly distributed on $[0, 1]$.

A similar differencing strategy can be used to analyze data from a DBWU design. However, the array specific differences are not all independent because each experimental unit is measured on two arrays. Thus, the differences from a pair of arrays measuring a pair of experimental units should be averaged prior to analysis to obtain $n/2$ differences that are independent and identically normally distributed with mean $V_1 - V_2$ and variance $2\sigma_U^2 + \sigma^2$. A one-sample $t$-test for non-zero mean with $n/2 - 1$ degrees of freedom can be used to test for variety effects. As was the case for the DBWV design, this test will be exact regardless of the true values of the variance components, and the uniformity of p-values under the null hypothesis is guaranteed whenever model (1) holds.

If no variance components are estimated to be zero with mixed linear model analysis, then the tests for variety effects will be identical for the difference-based analysis and both mixed linear model analysis strategies discussed in the previous subsection. However, when one or more variance components are estimated to be zero, the inferences of interest regarding variety effects may differ drastically between the difference-based approach and mixed linear model analysis. We investigate the relative performances of these methods in the next section.

4 Simulations and Results

4.1 Data simulation

Simulations were ran to compare the discussed methods of analysis for both the DBWU and the DBWV designs and then to use the best method of analysis to compare the powers of the designs. Nine scenarios were simulated as though they consisted of a
10,000 gene experiment performed twice using both designs. Each scenario was created using twelve observations per gene, or six arrays, resulting in 120,000 observations per data set, or experiment. We chose to study six-array experiments because two-treatment experiments of this size are common in practice. Experiments following both the DBWU and the DBWV designs were created using Models 1 and 3 for each simulation setting. To ensure realistic values in creating the data, variance component estimates were obtained from the corn experiment discussed previously. The nine scenarios were created using various combinations of the 25th, 50th, and 75th percentiles of the variance component estimates from the analysis. These combinations can be seen in Table 4.1. Because one of the main interests was in the effect of variance components being estimated to be zero, simulation settings were used which would promote this to occur. For one scenario, the block variance component was actually set to be zero. Simulation setting 2 shown in Table 4.1 was created using variance component settings to minimize zero variance component estimates for the sake of comparison with the other settings in which variance components would be expected to be estimated zero frequently. Each of the 10,000 genes within one scenario, or simulation setting, had the same variance components. To make the DBWU and DBWV values in the same scenario, the exact same values were used for the block, array, variety, dye, and measurement error effects for every observation. This could not be done for the unit effect since the DBWU design uses half the number of units that the DBWV design uses. However, the six unit values used in creating the DBWU data were included in the twelve unit values used in creating the DBWV data. The high level of homogeneity was used to increase the accuracy of the comparison of each design’s power.
<table>
<thead>
<tr>
<th>Random Factor</th>
<th>Dataset Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
</tr>
<tr>
<td>Unit</td>
<td>50</td>
</tr>
<tr>
<td>Array</td>
<td>50</td>
</tr>
<tr>
<td>Block</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 1  Simulation settings in quantiles of variance component estimates from corn experiment.

4.2 Method of Analysis Comparison

4.2.1 Analysis of DBWU data

The first time each simulation setting was analyzed, no variety differences were included. Each simulation setting was analyzed by mixed model analysis using both the containment and the Satterthwaite methods of estimating the denominator degrees of freedom. The model used for this analysis is shown in Equation 1. Analysis was also performed on the differences as described in Section 3.2. Due to the omission of variety effects, the resulting p-values from each analysis would be expected to follow the uniform distribution [0,1]. This had to be true for the analysis of differences because the data was created under the model assumptions. Since there was no difference in variety means included, the null hypothesis was true causing the t statistics to follow the central t distribution resulting in p-values that follow the uniform distribution [0,1]. Mixed model analysis using the containment method for estimating the denominator degrees of freedom resulted in a conservative p-value distribution for every scenario except number 2 where variance components were rarely estimated to be zero. Mixed model analysis with the Satterthwaite estimate also showed conservative results when the unit variance component was very small relative to the measurement error. Figure 2 displays the p-values for each analysis method resulting from simulation setting 6. Each p-value distribution is also shown in Figure 3 as histograms to aid in showing how conservative
the p-values obtained by mixed model analysis are.

Figure 2  P-value distributions attained for scenario 6 using each method of analysis considered for experiments following the DBWU design.

For data from a microarray experiment following the DBWU design, analysis of the differences would be recommended so that the p-value distribution can be assumed to follow the uniform distribution when the null hypothesis is true and the assumptions are met. This enables accurate estimates of the false discovery rate. If using the mixed model analysis, conservative results could occur under certain variance components regardless of which method is used in estimating the denominator degrees of freedom.
4.2.2 Analysis of DBWV data

For the simulation settings following the DBWV design, mixed model analysis was performed using the model shown in Equation 3. Both methods of estimating the denominator degrees of freedom were once again used. Analysis of differences was performed using the method in Section 3.2 for the DBWV design. The p-value distributions for the DBWV design were much different than those obtained for the DBWU design. Here, the p-values obtained through mixed model analysis using either the Satterthwaite or containment methods of calculating the denominator degrees of freedom resulted in p-values...
with a liberal bias, or lower than expected values. This would cause more hypothesis to be rejected than is appropriate, increasing the true false discovery rate. The more often variance components were estimated to be zero, the more liberal the p-value distributions became. Since the simulated data were from a normal distribution and each gene was independent, the test statistic for a variety effect when using analysis of differences did follow the central t-distribution causing the p-values to follow the uniform distribution. The p-value distribution of each analysis method for scenario 9 is shown in Figure 4. Since the denominator degrees of freedom can only be increased when using the Satterthwaite estimate and a variance component is estimated to be zero, this method only increases the liberal bias.

![Figure 4](image)

**Figure 4** P-value distributions attained for scenario 9 using each method of analysis considered for experiments following the DBWV design.
Once again the analysis of differences is recommended in order to obtain dependable results. The p-value distribution when using mixed model analysis can have a liberal bias making p-values and false discovery rates difficult to trust.

4.3 Design Comparison

Both the DBWU and the DBWV experimental designs are for comparing only two varieties. Therefore the only comparison of interest is for an expression difference between the genotype varieties in the corn experiment. The standard error for comparing the differences between the genotypes for the corn experiment when following the DBWU design is \( \sqrt{2\frac{\sigma^2_U}{3} + \frac{\sigma^2_3}{3}} \). \( \sigma^2_U \) represents the variability among the units while \( \sigma^2 \) is the measurement error. The standard error for the differences in genotypes in the corn experiment using the DBWV design is \( \sqrt{\sigma^2_U + \frac{\sigma^2_3}{3}} \). With the DBWU design, there is twice as much unit variability in the standard errors for the differences in the genotypes because this design uses half the number of units to estimate the unit variability.

It has already been shown that the DBWU design outperforms the DBWV design when comparing their variances for variety differences. The amount is dependent upon the relationship between unit variability and measurement error. To measure the benefit, the two designs were compared using real data. The designs were compared by the resulting p-values from analysis of differences to maximize the accuracy of the comparison. Figure 5 displays the additional power achieved when using using the DBWV design over the DBWU design in the corn experiment described previously. Using the DBWU design resulted in 376 genes with p-values less than .01 compared to 1655 genes for the DBWV design. Using a cutoff of 5%, yields 1563 genes with the DBWU design and 3357 with the DBWV design, more than doubling the number of genes detected as differentially expressed.

There are over 5 times the number of genes with p-values less than .01 when using the DBWV versus the DBWU design. There are over twice as many when using a p-value
of .05 as a rejection value.

Figure 5  Design comparison from the transgenic corn experiment.

As an alternative, the variance estimates for the variety differences were also compared for the two designs. This was done by measuring the relative efficiency by calculating $\frac{\hat{\text{Var}}_{DBWU}(V_1 - V_2)}{\hat{\text{Var}}_{DBWV}(V_1 - V_2)}$ for each of the 11,978 genes tested in the corn experiment. Quantiles of the measurements are shown in Table 2. The relative efficiencies were less than one only about 38% of the time indicating that the variance estimate from the DBWU design was only smaller than that of the DBWV design slightly more than one third of the time. The increased accuracy of the variety differences is an additional strength of the DBWV design. The DBWV design also has added denominator degrees of freedom received in the F-test for variety dif-
ferences. Because this will also add to the power of the tests, a relative efficiency that will account for this difference in degrees of freedom was also calculated. The Fisher relative efficiency (Kuehl, 2000) is calculated by multiplying the relative efficiency by 
\((1 + df_{num})(3 + df_{den}) / (3 + df_{num})(1 + df_{den})\). Using the Fisher relative efficiency increased the benefit shown in using the DBWV design. Both of these measurements indicate the additional power received using the DBWV design.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. Eff.</td>
<td>.000038</td>
<td>.1376</td>
<td>.5119</td>
<td>1.643</td>
<td>5.054</td>
<td>14.991</td>
<td>7518.8</td>
</tr>
<tr>
<td>Fisher RE</td>
<td>.000046</td>
<td>.1686</td>
<td>.6094</td>
<td>1.956</td>
<td>6.016</td>
<td>17.846</td>
<td>8950.1</td>
</tr>
</tbody>
</table>

Table 2  Quantiles of the relative efficiencies.

The designs can also be compared by their power to find differences of varying magnitude. Since analyzing the differences has been shown to be the better method of analysis, the distribution of the test statistic is known as long as the distributional assumptions are met. When using the DBWU design with six arrays as in the simulated scenarios, the F-statistic for the test of variety differences will follow the F-distribution with one degree of freedom in the numerator and two degrees of freedom in the denominator and a non-centrality parameter of

\[
\frac{(V_1 - V_2)^2}{\left(\frac{2\sigma_Y^2 + 3\sigma^2_U}{3}\right)}
\]

where \(V_i\) is the mean of variety \(i\), \(\sigma_Y^2\) is the unit variability, and \(\sigma^2\) is the measurement error. When the null hypothesis of equal variety means is true, the non-centrality parameter will equal zero and the F-statistic will follow the central F distribution. When under the same conditions and using the DBWV design, the F-statistic for the test of variety differences will follow the F-distribution with one degree of freedom in the numerator and four degrees of freedom in the denominator and a non-centrality parameter of

\[
\frac{(V_1 - V_2)^2}{\left(\frac{\sigma_Y^2 + \sigma^2}{3}\right)}
\]
Using these known distributions, the power of both designs are compared in Figure 6 using a p-value of .05 as the criterion for rejecting the null hypothesis. The x-axis displays absolute values of variety mean differences and the y-axis shows the power to find a difference of that degree. The design powers are compared using quartiles of the variance components from the corn experiment as values for $\sigma^2_U$ and $\sigma^2$ which are shown on the title of each plot. Often in microarray experiments, researchers wish to find expression differences in terms of fold changes. Because the variance component values are from a natural log scale the differences would have to be exponentiated to determine the fold change. To show the power of the two designs relative to specific fold changes, a blue line was drawn at a two-fold variety mean difference, an orange line at a five-fold mean difference, and a maroon line at a ten-fold mean difference. These figures show a definite increase in power when using the DBWV design over the DBWU design. This benefit is increased as the unit variability increases relative to the measurement error. This would be expected since the DBWV has twice as many units and the DBWU design. Also, under the conditions of the corn experiment, neither design had much power in detecting a two-fold change but the DBWV design shows great power in detecting a five-fold change under any of the variance component values shown.

5 Conclusions

There are two main points to be made: First, the DBWV design is the best to use when the number of slides is the constraint in designing a microarray experiment for two treatments. Regardless of the method of analysis or the values of the random effect parameters used, the DBWV design will lead to more powerful tests than those obtained when using the DBWU design. This is shown using both real and simulated data. However, this does not mean that the DBWU design should not be used. There can be times where the number of units is the constraining factor and repeating the
measures using a DBWU design would add to the accuracy and lower the variability. A benefit of the DBWU design is that a researcher can separate the biological variability, or the variability between the units, from the technical variability coming from the laboratory. The variability between measurements coming from the same unit estimates the technical variability. A combination of technical and biological variability can be estimated by comparing differences between units from the same variety. These can be separated using the estimate of technical variability. This could be valuable information if deciding on different possible laboratories or technical procedures to use.

The second main point is that regardless of which design is used, analyzing the differences should be preferred when the dataset is of limited size. For the DBWV design, analyzing differences can eliminate liberal bias. If using the DBWU design, analysis of differences can eliminate conservative bias. Regardless of the design, analyzing the differences will eliminate additional concerns induced when at least one variance component is estimated to be zero.

One detail to note is that both the nobound and type3 methods on PROC MIXED were considered as possible solutions. The nobound option allows negative estimates of random effect parameters which prevents any of the parameters from being removed from the model. Instead of a random effect being removed from the model, it is estimated to be zero or less when using the nobound option. However, when using this method, the REML algorithmic procedure failed to converge for some of the simulated genes. Because of this problem, this method was not considered as a solution to the concerns found when using mixed model analysis. Another option was the type3 method. This procedure obtains a method of moments estimator of the variance components by using the Type III expected mean squares and the Type III mean squares values. This method also allows negative estimates of covariance parameters. This method had to be removed from consideration because confounding between the block and array effects created singularity conditions by which the expected mean squares could not be calculated.
6 References


Figure 6  Power charts comparing DBWU design against DBWV design.
Microarray Experimental Design Selection Accounting for Mixed Model Variability

Justin C. Recknor and Dan Nettleton

Abstract

Many microarray experimental designs, when using two-dye spotted arrays, require more than one variance component in the model. In the past, array differences have been used to eliminate this difficulty when possible. However, loss of power is shown to occur when using this data adjustment when compared to using the full data in mixed model analysis. Mixed model experimental design comparison is proposed to allow design selection pertaining to the more powerful mixed model analysis. Optimal designs are shown to rely heavily upon the distribution of the variance components. Due to this factor, prior knowledge of parameter sets are shown to be a valuable asset. Designs are initially compared with a constraint upon the number of arrays possible to use in the experiment. Alternatively, experiments are often constrained by an expenditure limit. A design selection is compared which shows that selecting the optimal design can rely upon the cost ratios of objects required in the experiment as well as the distribution of the variance components.

1 Introduction

Two-color microarray experiments are a valuable resource in biological research today. These experiments enable one to measure the relative level of expression of thou-
sands of different genes in one experimental unit at any specific time. There are different types of microarray experiments that are performed but this paper is only concerned with two-dye spotted microarrays. These microarray experiments begin by taking samples of mRNA from the experimental unit. (The units could also be called observational units when dealing with an observational study, but they will be referred to as experimental units in this paper.) Next, the mRNA is converted to cDNA and a dye is applied to the sample. Both a red and a green dye are usually used. Applying the different colored dyes to samples enables two different samples to be placed on the same array. Samples are referred to within this paper as mRNA which is removed from an experimental unit, has a dye applied, and measured upon a slide. This term is not exchangeable with experimental unit because there can be multiple samples measured from one unit. Slides are produced specifically for the organism that is being studied. Each slide, or array, contains thousands of different spots, each of which has cDNA sequences that are complementary to a proportion of the cDNA sequence of a specific gene. The cDNA from one gene will bind to its complementary sequence at a specific location. Then, lasers excite the samples which enables measurements of the relative quantity of cDNA that is bound to each spot. This measurement is used to estimate the mRNA present within the experimental units that are sampled at that time. Due to the dye which was applied to the samples, two different experimental units can be measured together on a slide. Microarrays enable researchers to learn valuable information about cell life and gene expression under many different conditions. Along with this expanding area of research comes many very large data sets and new problems in how to handle and analyze this data. New statistical techniques are being developed constantly to improve the performance of such experiments.

In the early days of microarray experiments, much of the research being performed was on how to normalize and analyze the data. For the raw data, the effects are assumed to be on a multiplicative scale and a log transformation is used to convert this to an
additive scale for modeling and analysis (Kerr et al., 2000). Variety means are usually compared by ratios, or fold changes, because of the multiplicative scale. Along with the log transformation, normalization of the data is necessary because one dye will get higher measurements than the other when the same amount of cDNA is present. This problem is viewed as technical variability, or a bias introduced through laboratory procedures. However, this is not considered biological variability and data are commonly adjusted prior to being analyzed to remove this hindrance. The dye bias has been shown to be reliant upon the signal intensity which makes it more difficult to handle. A large number of methods have been proposed to remove this, many of which use some form of LOESS regression (Smyth and Speed, 2003; Yang et al., 2002). These processes are meant to remove global dye effects, or non-gene-specific dye biases. Research has shown an interaction between gene and dye to also exist (Kerr, 2003). Since this is gene specific, a dye effect is usually included in the model when doing gene-by-gene analysis to avoid additional variability or biased results.

Kerr and Churchill (2001b) first proposed classical ANOVA as a method of analysis for microarray experiments. This was meant to be performed, on the log transformed data, simultaneously for all genes with global and gene specific effects. An alternative method was proposed which eliminated the normalization step by modeling for both the global and gene specific effects (Wolfinger et al., 2001). This was done by first modeling the entire data set by global effects, or non-gene specific effects such as global dye bias. The residuals were then analyzed by gene to test for the gene specific effects. Experimental design considerations increased as the research into microarrays progressed.

Experimental design is an important part of microarray experiments. With the high cost of arrays, and possibly experimental units, the number of experimental units, samples, and arrays can be very limited in many experiments. Thus, using the best design to maximize the power is imperative to ensure success. Much literature has been published discussing different aspects of microarray experimental design. Early
papers covered different levels of variability and possible dye bias to consider when designing experiments (Kerr and Churchill, 2001a). Other papers proposed designs that were superior to the reference design (Dobbin and Simon, 2002; Townsend, 2003). The reference design received its name because every sample for a variety being studied in an experiment was paired on an array with a reference. The reference itself, is a variety, or treatment, that is not being studied but instead used as a standard. Often the reference would be many different samples such as all the varieties pooled together. The differences between each variety and the reference from the same array were then used in the analysis. This experimental design was frequently used in microarray experiments, especially when there were a large number of treatments being compared. Later articles have been concerned with everything from necessary biological sample sizes (Yang et al., 2003) to comparisons of specific designs (Dobbin et al., 2003).

Two-dye microarray experiments should always have an array effect in the model (Kerr and Churchill, 2001b) and arrays are usually treated as random effects. The arrays are viewed in this manner because an experiment is performed by purchasing arrays containing sequences specific to the organism of interest. Usually, many other experiments are being performed on other identical arrays. This enables the arrays to be viewed as a random sample from a large population. Treating arrays as random forces the analysis of all spotted microarray experiments into having to estimate more than one variance component. Multiple researchers (Glonek and Solomon, 2004; Yang and Speed, 2002; Kerr and Churchill, 2001b) have proposed transforming the microarray data after normalization by calculating the differences between measurements from the same array. For example, all the red dye measurements would have the green dye measurements for the same gene subtracted from them. These differences remove the array effect reducing the number of variance components in the model to one. In order to use this method, the parameters would need to be calculated for each array difference and model the reduced data accordingly. Since the model for the experiment now contains only one
variance component, it can be written in matrix form as $Y = X_D\beta + \epsilon$. $Y$ indicates a vector of responses, or differences, while $X_D$ indicates the experimental design after taking the differences and $\beta$ is a vector of unknown parameters. The vector $\epsilon$ indicates random errors which are assumed to be independently normally distributed with mean 0 and variance $\sigma^2$. Under these assumptions, $E(Y) = X_D\beta$ with a variance of $I_{nxn}\sigma^2$. $X'_D X_D$ will be assumed to be of full rank in this paper, enabling $\beta$ to be estimated using $(X'_D X_D)^{-1} X'_D Y$. If $X_D$ was not full rank then a generalized inverse of $X'_D X_D$ would be used instead. The resulting variance of the estimate of the unknown parameters, $\hat{\beta}$, is $\sigma^2(X'_D X_D)^{-1}$.

Glonek and Solomon (2004) compared many experimental designs by values proportional to the variances of contrasts of interest. Estimable hypothesis tests can be written in the form $H_0 = K'\beta$ and estimated by $K'\hat{\beta}$ which has a variance of $K'(X'_D X_D)^{-1} K \sigma^2$. Since the $\sigma^2$ is the measurement error, it will retain the same value regardless of the experimental design used. This fact allows designs to be compared simply upon the values on the diagonal of the $K'(X'_D X_D)^{-1} K$ matrix by dividing the contrast variance matrix of all designs considered by $\sigma^2$. When computationally practical, Glonek and Solomon would compare all possible designs for a specific number of varieties and arrays. For “all possible designs”, some constraints need to be mentioned. These are all of the designs for which the model, before calculating array differences, contains only two variance components, array and measurement error. This eliminates a large number of designs from consideration, for example, all designs which have experimental units being measured more than once. One design could dominate another one if the corresponding variances for the contrasts were always less than, or equal to, the contrast variances of the other design. However, for at least one of these variance comparisons, the variance of the dominant design would have to be strictly less than that of the other. If a design could not be dominated by any other design, then it was declared admissible. Using this method, only the admissible designs would be considered for any
experiment. Earlier, Kerr and Churchill (2001b) had proposed using the same reduction of the data but using the A-optimality criterion over all pairwise treatment differences in choosing an experimental design. A-optimality requires averaging the variances for all tests of interest and using the design which minimizes this value. Yang and Speed (2002) proposed a similar design selection procedure also using the transformed data. They proposed calculating the variances for some m estimable contrasts \( K_1 \beta \ldots K_m \beta \). Then, for the matrix of estimable contrasts of interest, called \( C \), for which the variance is written as \( V \sigma^2 = C'(X_D X_D)^{-1}C \sigma^2 \), the design to use is selected by minimizing the largest eigenvalue of \( V \).

2 Benefit of Using Mixed Model Analysis

When allowing for random effects, the linear model can be written in matrix form as \( Y = X \beta + Z \gamma + \epsilon \). The \( \beta \) represents the unknown fixed effects parameters while \( \gamma \) designates the random effects. The \( X \) matrix indicates the experimental design for the fixed effects while the \( Z \) matrix, usually a dummy matrix, indicates the design of the random effects. Commonly, it is assumed that \( \epsilon \sim N(0,R) \) and \( \gamma \sim N(0,G) \) with \( \text{Cov}(\epsilon,\gamma)=0 \). Estimable hypothesis tests of fixed effects are still calculated using \( K' \hat{\beta} \), now with a variance of \( K'(X'V^{-1}X)^{-1}K \) where \( V = ZGZ' + R \). The matrix \( X'V^{-1}X \) will be assumed to be of full rank throughout this paper allowing the inverse to be calculated. If this condition is not true, the generalized inverse would be necessary to use instead.

A problem introduced when reducing the data to differences of the data instead of mixed model analysis is the loss of inter-block information. Since each array can measure two varieties, and is treated as a random effect, it can be viewed as a blocking factor. Since the array effect should always be included in the model, all spotted two-dye microarray experiments automatically follow a block design with a block size of
Figure 1  Designs A-C are from Glonek and Solomon (2004). Each arrow indicates an array in which the arrow points from the sample dyed green to the sample dyed red. Each corner indicates a variety level, not an experimental unit, so no experimental unit is sampled more than once. $\alpha$ and $\beta$ indicate variety main effects with $\alpha\beta$ indicating interaction.

two. When there are more than two varieties being studied, the experiment follows an incomplete block design because every treatment cannot be contained within one block simultaneously. Treating arrays as a random effect allows information to be shared about variety differences between blocks, or inter-block. The inter-block information is beneficial when dealing with incomplete block designs. Using the inter-block information can reduce the variance of a test, increasing the power. For example, assume Design C from Glonek and Solomon (2004), which is shown in Figure 1, was used in performing a microarray experiment. In this figure, each circle represents an experimental unit and each arrow represents an array. The direction of the arrow indicates the dye assignment by the following convention: the arrow points from the sample dyed green to the sample dyed red. The expected variances of estimators of the parameters of interest ($\alpha$, $\beta$, and $\alpha\beta$) are shown in Table 1 for both mixed model and array differences methods of analysis.

These equations assume $\sigma_A^2$ symbolizes the array variability while $\sigma^2$ is the measurement error. The expected variance for the test for interaction ($\alpha\beta$) is the same for both methods. However, the expected variances for the main effects ($\alpha$ and $\beta$) differ substantially depending upon the method of analysis used. By an algebraic adjustment to the vari-
<table>
<thead>
<tr>
<th>Parameter Method</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\alpha\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences</td>
<td>$\frac{3\sigma^2}{4}$</td>
<td>$\frac{3\sigma^2}{4}$</td>
<td>$\frac{3\sigma^2}{4}$</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>$\frac{\sigma^2(3\sigma_A^2 + 2\sigma^2)}{4(\sigma^2 + \sigma_A^2)}$</td>
<td>$\frac{\sigma^2(3\sigma_A^2 + 2\sigma^2)}{4(\sigma^2 + \sigma_A^2)}$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

Table 1  Expected variances for parameters of interest for both methods of analysis.

ance equations, a lesser variance estimate is evident for mixed model analysis. This is because

$$\frac{\sigma^2(3\sigma_A^2 + 2\sigma^2)}{4(\sigma^2 + \sigma_A^2)} = \frac{3\sigma^2}{4} \left(\frac{\sigma_A^2 + \frac{2}{3}\sigma^2}{\sigma_A^2 + \sigma^2}\right) < \frac{3\sigma^2}{4}$$

since

$$\frac{\sigma_A^2 + \frac{2}{3}\sigma^2}{\sigma_A^2 + \sigma^2} < 1.$$  

The inequality is strict because $\sigma^2 > 0$.

The extent of a difference the change in variance makes is shown through simulation. Data sets were created using variance component sets from an actual experiment that followed a related design. Mixed model analysis led to a definite increase in power in the tests for main effects. This can be seen in the Receiver Operating Characteristic (ROC) curve for $\alpha$ shown in Figure 2. The data analyzed was simulated by four repetitions of Design C from Figure 1. An effect, which followed a log-normal distribution, was included for $\alpha$, $\beta$, and $\alpha\beta$ in 75% of the genes created. This plot displays the percentage of false conclusions (null hypotheses incorrectly rejected) necessary to find a specific percentage of the true differences (null hypotheses correctly rejected). For example, if it was desired to find 40% of the different population means for $\alpha$ then it would require finding approximately 19% of the means that are not different in this data set when using mixed model analysis. If the differences method of analysis was used instead, then it would require finding approximately 22% of the means that are not actually different in order to find 40% of the different means as desired. The ROC curve for $\beta$ is very similar. These plots show the benefit in using mixed model analysis for the main effects with this experimental design. Both methods of analysis led to essentially the same
results in testing for interaction which is no surprise since both methods have the same expected variances for this parameter. Also, the distribution of the p-values when the null hypothesis was true appeared to follow the uniform distribution for both methods of analysis.

Figure 2  Receiver Operating Characteristic (ROC) curve for $\alpha$ resulting from the simulated data.

3 Mixed Model design comparisons

3.1 Equation Comparisons

If mixed model analysis is going to be used to increase the power of the experiment, the process of design selection should take that into consideration. One method to use
in design selection would be to derive the equations for the variances of the parameters of interest for each design. These equations could then be used to compare the designs of interest. Let $H_0 : K'\beta = 0$ indicate estimable hypotheses of interest pertaining to the fixed effects in the model. Then designs would be compared using the equations from the diagonal of the $K'(X'V^{-1}X)^{-1}K$ matrix. As an example, Designs A and B, shown in Figure 1, are compared by their interaction variances. The equations were calculated and are shown in Equation 1. The array variability is shown by $\sigma_A^2$ while $\sigma^2$ represents the residual variance or measurement error. As can be seen here, although it is possible derive the equations, it can be time consuming and their complexity leaves the design comparison difficult to manage.

$$
\begin{align*}
\text{Design A} &= \frac{(22\sigma_A^2 + 15\sigma^2)(2\sigma_A^2 + \sigma^2)\sigma^2}{30\sigma_A^4 + 41\sigma_A^2\sigma^4 + 14\sigma^2} \\
\text{Design B} &= 16\frac{(\sigma_A^2 + \sigma^2)\sigma^2}{13\sigma_A^2 + 14\sigma^4}
\end{align*}
$$

(1)

3.2 Comparing Designs on the Basis of Pilot Data

To avoid the complexity issues of the equations, designs can also be compared using values of the standard errors of tests of interest without actually calculating the variance equations. At least one set of values of the variance components is necessary to use this method. Following the model described in Section 2, the $X$ and $Z$ matrices need to be created according to the design to study. The $R$ and $G$ matrices will need to be created according to the design while using the predefined values of the variance components. Using these matrices along with estimable contrast values for $K'$, the variance for each test of interest can be retrieved from the diagonals of $K'(X'V^{-1}X)^{-1}K$. The criteria used in comparing the designs would depend upon the goals of the experiment. For example, say an experiment was to have a two factorial treatment structure with modeling Treatments A, B, and their interaction. Assume the interaction was the primary
concern with the effects of Treatment A as the second of interest, and little interest in the effect of Treatment B. If there were two designs to be compared, then a ratio of their interaction variances would be calculated for each set of variance component values. If multiple sets of variance components were used, a graphical representation of the ratios could reveal valuable information about the circumstances where one design would be favored over the other. So to compare two experimental designs, when the test of interest can be written as $H_0 : K'\beta = 0$, the ratio of $Var_{Des1}(K'\hat{\beta})/Var_{Des2}(K'\hat{\beta})$ would be calculated across differing values of variance components. Varying processes could then be used to select a design with these values. One choice could be that if the median ratio $> 1$ then Design 2 would be chosen, otherwise Design 1 would be used. If there were more than two designs being compared, then a rank test could be used. The variances of the test of interest would be ranked by design for every set of variance component values. The design with the minimum mean rank would then be chosen. These are just some options because the method of comparison to use would depend upon the goals of the experiment. If there were multiple effects of equal interest, than a comparison using the A-optimality criterion could be used.

Estimates of the variance components are required to use this method. They could be obtained in two different manners, selection of which depends on the prior knowledge of the circumstances. If there was little known about the expected variability of the data, then the designs could be compared over a range of parameter values that seemed reasonable. If a similar experiment had been ran previously, then the data set of variance components from that experiment could be used to compare designs under consideration. This method would give a more accurate portrayal of the parameter space by using a more accurate range of values and accounting for possible correlation amongst the variance components. Code was written in R to compare designs using both types of variance component values.

Simply comparing variance estimates simplifies the analysis, over using variance equa-
tions, and has results which are easy to understand. This method was performed in R by creating the corresponding matrices and computing $K'(X'V^{-1}X)^{-1}K$. The variance component values used were obtained in both manners described previously, using a range of reasonable values and by use a variance component estimates from a related experiment.

3.3 Benefits in Using Mixed Model Design Comparisons

The methods discussed previously require transformation of the data with the resulting observations all being independent to be able to compare experimental designs. This criterion does allow consideration of many designs but also eliminates many other designs from the list of possible alternatives. For example, a design commonly used in microarray experiments is the loop design, for which a three treatment example is shown in Design A of Figure 7. In this design, each experimental unit is measured twice by samples on different arrays. Due to the repeated measures, if the differences between measurements on an array were calculated, there would still be a dependence among these values. This can be seen in looking at the first loop of this design letting $D$ be the dye effect, $V$ the variety effect, $A$ the array effect, $U$ the unit (experimental unit) effect, and $\epsilon$ the measurement error. Since array, unit, and $\epsilon$ are all random effects, each would be assumed distributed normal with mean 0, and variances $\sigma^2_A$, $\sigma^2_U$, and $\sigma^2$. An additional assumption of independence between $A$, $U$, and $\epsilon$ would also be made. The model for each observation would be

$$y_{ijkl} = \mu + A_i + D_j + V_k + U_l + \epsilon_{ijkl}$$

with $i$, $k$ and $l$ having values 1-3 and $j$ being either 1 or 2. The array differences would have the form

$$y_{i1kl} - y_{i2k'l'} = D_1 - D_2 + V_k - V_{k'} + U_l - U_{l'} + \epsilon_{i1kl} - \epsilon_{i2k'l'}.$$
such that

$$\text{cov}(y_{1111} - y_{1222}, y_{2122} - y_{2233}) =$$

$$\text{cov}(D_1 - D_2 + V_1 - V_2 + U_1 - U_2 + \epsilon_{1111} - \epsilon_{1222}, D_1 - D_2 + V_2 - V_3 + U_2 - U_3 + \epsilon_{2122} - \epsilon_{2233}) =$$

$$\text{cov}(-U_2, U_2) = -\sigma_U^2 = \text{cov}(y_{1111} - y_{1222}, y_{3133} - y_{3211}) = \text{cov}(y_{3133} - y_{3211}, y_{2122} - y_{2233})$$

This dependence between differences within a loop would eliminate this design from consideration by the previously discussed methods although the loop design is commonly used in microarray experiments. There are many other designs that should be considered in design comparisons that cannot correctly be handled with a single variance component. For many of these designs, mixed model analysis must be used because the data cannot be transformed to eliminate these additional random effects. Comparing designs that allow for multiple variance components permits the consideration of more designs and allows use of the inter-block information in comparing designs.

4 Examples

4.1 Mixed Model Design Selection Versus Array Differences

Selection of experimental designs considering mixed models increases the complexity of the design comparisons. Choosing the “best” experimental design design can now depend upon what region of the parameter space the variance components are within. Although the process may be more difficult, designs can be chosen which fail when using the array differences methods. Both of these conditions are shown in comparing designs A and B shown in Figure 1 that were used by Glonek and Solomon (2004). The designs being compared are for an experiment with a two factorial treatment structure, and 8 arrays to use. The model utilized in analysis is

$$y = \mu + \alpha + \beta + (\alpha \beta) + \epsilon$$
where \( \alpha \) and \( \beta \) are the main effect of each treatment and \((\alpha \beta)\) is their interaction. When using the method proposed by Glonek and Solomon (2004), Design A is dominated by Design B. This means that the variances of Design B, for every parameter of interest, \( \alpha, \beta \) and \((\alpha \beta)\), are less than or equal to those of Design A. Thus, regardless of what parameter is of main concern for a study, Design B would be the design to use. However, when the arrays are treated as random the same cannot be said. The optimal design depends upon the values of the array and measurement error variance components and this is displayed in Figure 3. The red colored regions are where Design A would have a smaller
variance in testing for interaction and would be preferred. The blue and green regions are the opposite, where Design B would be selected. The range of the parameter values used were obtained from an actual experiment to ensure a realistic comparison. When the size of the array parameter is small relative to that of the measurement error, Design A would actually be better to use for estimating the interaction between the treatments. This may not happen as frequently as the converse, but it definitely shows that there can be conditions where Design A would be preferred to Design B.

4.2 Comparison of Designs Which Cannot Be Considered Using Array Differences

Although many designs can be compared using the array differences methods, there are many others which cannot although they can be compared using the mixed models approach. As an example, two designs that require repeated measures were compared by their variances. The two designs are possibilities for use when an experiment has a 2 by 2 treatment structure, 12 arrays, and 12 experimental units. These designs are shown in Figure 4 with arrows showing arrays and circles indicating experimental units as described earlier. Design A is similar to a design examined by Glonek and Solomon (2004). In that case, there were half the number of arrays being used so only one array connected each circle but the dyes were still balanced for each treatment. Under Glonek and Solomon’s conditions, and using their \((X’X)^{-1}\) method of comparison, the design similar to A was shown to be admissible; i.e., no other design could dominate it. Design A was considered due to its similarity to their design but having the arrays repeated with the dyes reversed. Design B is 3 loop designs combined but with the array directions reversing between each loop. Just by looking at the designs, it might appear that Design A would be better for estimating \(\beta\) since there are four arrays directly estimating that parameter while there are only three for Design B although only two experimental units were used for this with Design A while three were used for Design B. Directly estimating
a parameter means that the expected value of the difference between the measurements on an array equals, aside from a dye effect, that parameter. The direction of subtraction, red dye - green dye, can be reversed to prevent expected values of the negative of the parameter. The contrary is true for $\alpha$, with 3 arrays directly estimating this parameter in Design B and 2 in Design A.

Data from a previous experiment was used to approximate the range of each variance component’s values. In that experiment, the experimental unit variance component was estimated to be between .001 and .243. The array variance was estimated to be within .11 and .34, while the measurement error variance was found to typically fall between .09 and .23. Approximately 60000 combinations of parameter values within these ranges were generated. They were obtained by using every value between the upper and lower estimate of each variance component by increments of .005. Each value of each variance component was grouped with every possible combination of the values of the other two variance components. When using this method, Design B always had a lower standard error for estimating $\alpha$ than Design A. Also, the standard error for estimating the interaction was always the same. When estimating $\beta$ Design A had the lowest standard error approximately 97% of the time. In the cases where Design B did
better, the unit effect was large while the measurement error was small relative to other parameter values. The array variability parameter values did not appear to change this relationship. This can be seen in Figure 5 which is a 3 dimensional plot of a subsample of the data.

Figure 5  Simulated data design comparison. Blue pyramids represent the locations where Design B had a lower standard error for estimating beta. Red stars indicate the locations where Design A had a lower standard error.

The data set used to obtain estimates of the variance components yielded 9000 point estimates in three-dimensional space. When these variance component estimates were used, Design A and B always had the same standard errors when estimating the interaction. Design B also always had a lower standard error when estimating $\alpha$. For estimating $\beta$, Design A now had the lower standard error only about 60% of the time. A random subsample of the data is shown in the 3 dimensional plot in Figure 6. When the actual data is used, it can be seen that the size of the unit variability is frequently large relative to that of the measurement error. Consequently, Design B tests as a better
design for estimating $\beta$ more often, but the standard errors are similar the majority of the time. If there was equal interest in estimating both $\alpha$ and $\beta$ and the variance components similar to those estimated from the previous experiment could be expected then Design B would be the logical choice of the two. The benefit in having prior estimates of variance components is apparent in this example. This is not surprising since the better design is dependent upon the correlation of the variance components.

Figure 6  Real data design comparison. Blue pyramids represent the locations where Design B had a lower standard error for estimating beta. Red stars indicate the locations where Design A had a lower standard error.

5  Design Comparisons Accounting for Cost

In the examples until this point, the only constraint upon an experiment is the number of slides, or arrays, that are available to use. However, an experiment could easily be constrained by an expenditure limit. Under these conditions, the optimal design to use could depend upon how much the experimental units cost relative to the cost of the arrays. For example, if a three treatment experiment was being proposed
on a cost constraint basis, three designs that could be considered are shown in Figure 7. Each circle represents an experimental unit and each arrow represents an array, as described previously. Each square represents a possible blocking effect. Blocking effects can be introduced to experiments through a variety of reasons. Blocking effects can be included in designing the experiment, as happens commonly in plant experiments where different fields used in the experiment could be viewed as blocking effects. Blocking effects could also come about during the laboratory work. For example, there are often constraints on the number of slides possible to process in a day’s time. Slides hybridized on the same day would have additional correlation, creating the need for a block effect in the model.

Figure 7 Dye balanced microarray experimental designs considered for a three treatment experiment. Each circle represents an experimental unit and each arrow represents an array with the dye assignment for each array determined by the arrow direction. Squares indicate possible blocks.

Design A is commonly called the loop design and is the only one considered in which an experimental unit is measured more than once. A benefit of this design is allowing for unit variability to be estimated by preventing confounding with measurement error. The unit variability is not estimable with the other two designs since each experimental unit is only measured once which makes distinguishing the unit effect from the measurement
error impossible. The arrow direction is reversed on each replication so that the dyes are balanced for every treatment pair combination whenever there is an even number of blocks. An even number of blocks was not forced to occur when calculating the number of blocks for the loop design. Design B once again uses three experimental units per block, but this design only allows each experimental unit to be measured once. Since there is an odd number of treatments, this forces experimental units from different blocks to be paired together, called across block pairings. A detail to note about Design B is that each pairs of treatments can only be both dye balanced and across block balanced, when there are blocks in increments of twelve. When comparing this design, this was not constrained to occur. Also for Design B, the blocking shown could not be introduced by the lab constraints mentioned previously since an entire array must be made at one time not allowing across block pairings. Design C is simpler in that only one pair of treatments is measured within a block. With this design, it takes at least three blocks for each treatment pair to be measured and six for the dyes to be balanced among the treatment pairs. Since only one array is measured in each block, Design C causes confounding between the block and array effects as well as the unit effect and measurement error. In this example, treatment pairs were forced to be balanced, so blocks were added in groups of three.

Code was created in R, to calculate the number of blocks to use for each design after receiving the experiments total funds along with the cost of arrays and experimental units. The specific constraints mentioned previously were enforced for each design. The number of blocks could differ for the experimental designs since the total cost of the experiment was the only limit.

The three designs were compared at various cost ratios using parameter estimate sets from an actual experiment that was performed using the loop design. The experiment enabled the estimation of all four of the variance components necessary to compare the designs of interest. Since a small adjustment to the total funds available could result
in an additional block for one design but not the other two, the designs were compared multiple times with varying amount of total funds for each array and experimental unit cost. This process was repeated for many different array and experimental unit cost ratios.

Figure 8  Cost comparison of designs A and C. Blue pyramids represent the locations where Design A had a lower standard error for treatment differences. Red stars indicate the locations where Design C had a lower standard error.

The first comparison of the three designs assumed an equal cost between the experimental units and arrays. Saying that a design tested as the best for a cost ratio means that it had the minimum standard error for tests of variety differences the majority of the time. With Design B, the standard errors could differ depending upon which pair of the three varieties were being tested for a difference. This only occurred when the number of inter-block treatment comparisons was not balanced. Although this was not the case for the other two designs, standard errors were compared for each of the three variety pair differences. However, Design B was never chosen to be the best design under
these conditions. So for this discussion, each design was compared using one standard error and the minimum of the three standard errors was used for Design B. Under these constraints, the loop design (A), always tested to be the best design to use regardless of the total funds available. For example, if the total funds available were 50 times either the cost of the experimental units or arrays, then Design A would be favored to use since it has a smaller standard error than Design B approximately 76% of the time. Design A would also have a smaller variance than Design C 83% of the time. The median of the ratio, $s.e_B/s.e_A$ is approximately 1.09 and approximately 1.08 for $s.e_C/s.e_A$. A histogram of the 3730 values of $s.e_C/s.e_A$ is shown in Figure 9. Figure 8 shows that Design C would only be preferred in cases where the unit variability is large relative to the measurement error and slide variability. The same was also true when comparing Design A to Design B. This is not surprising since every experimental unit is measured twice with the loop design, reducing the number of units used in relation to the competing designs. These results are only with respect to the variance component values obtained from the experiment mentioned previously.

![Figure 9 Histogram of the ratio of the treatment difference standard errors for Design C to A.](image)

If the cost of the experimental unit was small relative to the cost of the array,
the optimal design to use would change. Under the same variability conditions as the previous example, but with now having the microarray costing five times as much as the experimental unit, either Design B or Design C would be the best choice. Both Design B and Design C’s standard error would be lower than Design A’s more than 66% of the time. The median of the ratio of $s.e_C/s.e_A$ was approximately .95. The relationship between unit variability and design selection, shown in Figure 8, still is true. The only difference is that with this cost relationship, it does not take as large of a unit effect to choose either Design B or C. For comparing Design C to Design B, Design C’s standard error was less than Design B’s only 50.08% of the time and had a median ratio of .9998. Although both B and C would be selected over A under these conditions, neither show evidence of being better than the other. However, Figure 10 shows that an increase in slide variability relative to measurement error would make choosing Design C more likely. If the opposite conditions were to occur then Design B would most likely be the best selection. An increase in unit variability also increases the likelihood of choosing Design B over Design C slightly. The block effect, which is not shown in any of the figures, was investigated but appears to have very little impact upon design selection.

When accounting for cost, the loop design is the favorable choice a large amount of the time. If the cost design comparison was also performed using a method which requires calculating the differences for each array (Glonek and Solomon, 2004; Yang and Speed, 2002), the loop design could not be considered. Since after calculating the differences, there is still correlation amongst the observations, this design would have to be eliminated. Design B could be investigated if there was not a block effect. Not using mixed model comparisons could eliminate the optimal design from consideration a large amount of the time.

Accounting for the costs can have a valuable impact upon experimental design selection. However, it is not easy to make a standard rule for how costs should be accounted for. Every experiment could have different costs that could change which design is se-
Figure 10  Cost comparison of Design’s B and C. Blue pyramids represent the locations where Design B had a lower standard error for treatment differences. Red stars indicate the locations where Design C had a lower standard error.

lected. For example, there could be a block cost. An example would be an experiment where offspring where treated and measured making them the experimental unit. There would be correlation amongst the siblings, causing blocking by the parents. Now, Designs A and B would both use 3 offspring per litter while Design C would only use 2. This would create a need for more litters with Design C causing an increase in expenditure. Another possible cost consideration could occur when an experiment is intended to study transgenic organisms. Transgenic organisms will often cost much more than nontransgenic organisms. Taking this into consideration could change the designs to consider and the optimal design to use. Considerations such as these would need to be accounted for in the cost comparisons which make standard methods difficult to present.
6 Considerations

A possible problem with this method is that when comparing designs relative to a test of interest using an F-test, not only can the variability change but also the denominator degrees of freedom. This should be considered as well since one design could have a slightly smaller expected variance but the alternate design could also have smaller denominator degrees of freedom which could cancel or even reverse which model is preferred. One option to account for this is to use Fisher’s relative efficiency (Fisher, 1960) to compare designs instead of the variances alone. Fisher’s relative efficiency differs from the usual ratio of design variances by also accounting for the degrees of freedom. The formula is

\[
\frac{\left(1 + df_{Des1}\right)\left(3 + df_{Des2}\right)}{\left(3 + df_{Des1}\right)\left(1 + df_{Des2}\right)} \times \left(\frac{Var_{Des1}}{Var_{Des2}}\right)
\]

where \(Var_{Desi}\) indicates the variance of the estimator of the parameter of interest for design \(i\). A drawback to this method is that the impact of the difference in denominator degrees of freedom is being estimated.

To exactly account for the difference in denominator degrees of freedom, a method described by Stroup (2002) for power tests could be used instead. This method is described using the PROC MIXED procedure on SAS (SAS Institute, Cary, NC) although the SAS program is not required for this test. This was also discussed for use in microarray experimental designs comparisons recently (Rosa et al., 2005). It is assumed that the model can be written as \(Y = X\beta + Z\gamma + \epsilon\) with \(\epsilon \sim N(0,R)\) and \(\gamma \sim N(0,G)\) with \(\text{Cov}(\epsilon,\gamma)=0\), and testing the null hypothesis of \(K'\beta = 0\), with the assumption that \(K'\beta\) is estimable. Stroup’s method requires knowledge of \(X, K,\) and \(V\) where \(V = ZGZ' + R\). All of these can be obtained from information about the experimental design and estimates of the variance components for use in the design comparison. It is assumed that
the null hypothesis is being tested using

\[ F_{\text{rank}(K),v} = \frac{\left(K' \hat{\beta}\right)' \left[ K' \left(X'\hat{V}^{-1}X\right)^{-1} K \right]^{-1} \left(K' \hat{\beta}\right)}{\text{rank}(K)} \]

which will follow, or approximately follow, the F distribution with \text{rank}(K) numerator degrees of freedom and \( v \) denominator degrees of freedom. When the test statistic does not truly follow the F distribution, the denominator degrees of freedom can be approximated using either the Satterthwaite (1946) or Kenward-Roger (1997) methods, among others. When the null hypothesis is false, this test statistic will follow, or approximately follow, the F distribution with non-centrality parameter \( \lambda \), where

\[ \lambda = (K' \beta)' \left[ K' \left(X'V^{-1}X\right)^{-1} K \right]^{-1} (K' \beta). \]

Experimental designs are then compared based upon their power to pick up a difference of the magnitude desired. This is done by first calculating \( F_{\text{crit}} \) by

\[ P(F_{\text{rank}(K),v,0} > F_{\text{crit}}) = \alpha \]

where \( \alpha \) is the desired Type 1 error rate. The power to find the difference of interest is then calculated by

\[ P(F_{\text{rank}(K),v,\lambda} > F_{\text{crit}}) \]

by using the proper values of \( X, V, v \) and \( \lambda \). The “best” design is then chosen by which has more power for the majority of the parameter value sets considered.

Stroup’s method does allow accurate consideration of both the covariance parameter values and the degrees of freedom. A benefit to this method is that quite often with mixed model designs, the test being performed is not truly an F test but is approximated by the F distribution. To do this, the denominator degrees of freedom are necessary to estimate. Stroup’s method makes this easy by simply indicating in the MIXED procedure the method of estimating the denominator degrees of freedom that is desired to use. One drawback is that the magnitude of the power increase is dependent upon the difference value that is chosen. That is not a problem when comparing standard errors however. When comparing designs by the standard errors, the magnitude is meaningful and may have a significant impact when comparing closely competing designs.
7 Conclusion

The number of microarray experiments being performed keeps growing every day. The information obtained aids in the understanding of organisms, organs, and disease, amongst many other areas of research. Using mixed model analysis for microarray experiments can increase the power obtained relative to analyzing array differences. Lessening the variability of the tests of interest with mixed model analysis may enable one to find important genes which otherwise would not be found.

Proper experimental design selection requires accounting for how the data is to be analyzed. If mixed model analysis is to be used, that should also be accounted for in the design selection process. The variance equation complexity makes simulation a beneficial alternative when comparing mixed model designs. This does make choosing a design more difficult because there may no longer be a specific set of designs that should be used for all experiments with specific treatment structures. Design selection depends upon the distribution of the true values of the variance components. This can make design selection additionally difficult and increases the benefits of acquiring information from similar experiments previous to the choice of experimental design.

Various costs can also make a large impact upon design selection. In cases where there is a specific amount of money dedicated to the microarray phase of an experiment, the optimal design can rely heavily upon the costs of the arrays and experimental units along with the distribution of variance components. Mixed model design comparisons allow all of these factors to be considered in selecting a design. Using all of this information can make a substantial increase in power for an experiment.


8 References


Identification of Differentially Expressed Functional Categories in Microarray Studies Using Nonparametric Multivariate Analyses

Justin C. Recknor, Dan Nettleton, and James Reecy

Abstract

Tests of differential expression across groups of genes, within a functional category, are performed using a method motivated by Barry, Nobel, and Wright (2005). Rather than comparing marginal distributions on a gene-by-gene basis across treatment groups, we use a test statistic that can detect general changes in multivariate distributions across treatment groups. Resampling-based methods and multiple-testing adjustments are used to obtain simultaneous inference for multiple groups of genes. Results are visualized on a directed acyclical graph, and new methods for pinpointing genes of greatest interest are provided.

1 Introduction

Advances in science have enabled researchers to measure a large number of genes’ transcription expression levels simultaneously by use of microarrays. The level of ex-
pression of a substantial proportion of an organism’s entire genome can be measured on a single array or slide. Microarray experiments are designed to find genome expression differences between at least two classes, often to get a better understanding of an impairment in hopes of finding a remedy. This process can empower scientists to obtain valuable information about diseases, animal breeds, and other biological interests.

Experiments can be performed using oligonucleotide microarrays to measure gene expression levels for a biological subject. To perform an experiment, mRNA is extracted from a subject and labeled with a fluorescent dye. The sample is then placed on an array containing a large number of probes. A probe is a short sequence of cDNA that is complementary to a unique region of the mRNA for a specific gene which will make the according mRNA sample bind to it. There are many probes, often 25, placed on an array for one gene resulting in multiple measurements of the same gene on one array. The measurement values for each probe are obtained by applying a laser to them and measuring the fluorescent signal returned. These gene measurements come from multiple locations upon the slide where the probes are located. Procedures such as MAS 5.0 (Affymetrix) and Robust Multi-array Average (RMA) (Irizarry et al., 2003) are often used to condense the data down to one measurement for each gene. The resulting data is then analyzed on a gene-by-gene basis.

Researchers obtain a vast number of measurements for the organism of interest. Since analysis is performed individually for each gene, one experiment will usually consist of 20,000 or more tests. This leaves a very large number of statistics or p-values to deal with. If one was to just declare every gene with a p-value less than a chosen threshold significant, the study could suffer from a high number of false conclusions. To prevent this, the p-values are often adjusted for the multiple comparisons using a method such as the Bonferroni adjustment to control the Type I error rate. This method of p-value adjustment is designed to control the family-wise error rate which means to prevent any of the null hypotheses from being incorrectly rejected. When working with a large
number of tests, as in microarray experiments, controlling the family-wise error rate can result in very conservative results. In order to have a high level of confidence in every null hypothesis rejected, a large number of differentially expressed genes must also go undetected.

To prevent overly conservative results, false discovery rates are used in microarray experiments as an alternative (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). Instead of controlling the probability that all the null hypotheses rejected are truly false, the FDR estimates the expected proportion of the null hypotheses which are rejected that are actually true. For example, if one was to use a 5% FDR and there were 100 significant genes at that level, then the null hypothesis would be expected to be true for approximately 5 of the genes. This method usually allows for a much larger number of null hypotheses to be rejected.

Using either the family-wise error rate or the false discovery rate on a successful experiment, leads to a list of genes that are declared to be differentially expressed for a study. Often scientists are left with a very large number of genes to interpret. Deciding upon the meaning of, and further testing a proportion of thousands of genes can be a very daunting task. On the other hand, after accounting for the multiple comparisons for each gene, microarray experiments can also lead to a very small number of genes testing significant. This list could contain genes with very large fold changes that do not appear to be biologically related. Even worse, genes with very small fold changes which seem to have little biological meaning may test significant with very low p-values due to small variance estimates. This happens often, largely due to a large number of tests with small sample sizes which makes these results difficult to trust. Regardless of any of these situations, only the genes with extreme values of test statistics are chosen for further study.

Often investigators are looking for networks, consisting of many genes, to explain biological differences. There could be a family of genes, many with relatively small
fold changes, that could be important to the study. For example, a group of genes all
working together for the same biological purpose could all experience small fold changes.
All of the genes changing expression could have a large impact upon their common task.
This family of genes would rarely be found using the usual “gene by gene” statistical
procedures and yet could be very important to the study.

Even if a network of genes is discovered in a study, other researchers may have a
difficult time finding the published results due to use of differing terminology. To reduce
the occurrence of this problem, the Gene Ontology (GO) Consortium was established
(Gene Ontology, 2000). GO unites scientists’ vocabularies by making concise annota-
tions of gene functions that are independent of species. GO contains three separate
ontologies pertaining to gene functions. These are the gene’s molecular function, its
biological process, and the cellular compartment that it acts within. A gene’s molecular
function and its biological process can be difficult to separate, but the biological process
is considered more broad and must have more than one step. Each ontology consists of
a hierarchy of terms that start with broad classifications and go to specific terms. This
system does resemble a phylogenetic tree except that one node can have more than one
parent. Genes of more than 87,000 species already have GO annotation and the library
keeps expanding as the research progresses.

Microarray experiments alternatively can be analyzed by a category, such as a GO
classification, instead of by the gene as is usually performed. Analyzing microarray
experiments with the help of GO groupings is not a new area of study. Many papers
have been published discussing different methods for testing GO annotations. Many
of these consist of looking at the quantity of significant genes within each annotation
(Beibarth and Speed, 2004; Berriz et al., 2003; Shah and Fedoroff, 2004). A problem
with these methods is that they rely upon what group of genes are considered to have
“significantly different” expression levels among the treatments. Also, these methods
would not be likely to find gene annotations that consist of many genes having relatively
small fold changes as discussed earlier. Another possible deficiency of these methods are that they often use tests which require independence of the observations to be valid such as Fisher’s Exact Test and hypergeometric probabilities.

Other methods have been proposed that do not have these same weaknesses, one of which is called the Significance Analysis of Functional categories in gene Expression studies (SAFE) (Barry et al., 2005). The SAFE method entails calculating test statistics of interest from the normalized data for each gene. A global statistic is then calculated by measuring the difference between the distribution of statistics within a group (such as a GO annotation) to the distribution of statistics outside of a group using a method such as the Wilcoxon rank sum test statistic or Kolmogorov-Smirnov test statistic. The treatment labels are then permuted by the samples to maintain the correlation between the genes and the global statistics are once again calculated. P-values are computed by measuring the extremity of the actual global statistic versus the permutated values. The groups whose distributions differ the most from the rest of the population are then selected.

An alternative method is called the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). The GSEA method requires first calculating a test statistic for every gene in the study. The statistics are then ordered from least to greatest. Starting with the smallest ordered statistic, a category is given a positive score if it contains that gene, and a negative score if it does not. This is repeated orderly for all the statistics, keeping a running-sum for each category. Every category is then given an enrichment score for its maximum deviation from zero along the process. The enrichment scores are normalized to compensate for different category sizes. P-values are obtained for each category by assuming the enrichment score corresponds to a weighted “Kolmogorov-Smirnov-like” statistic.
2 Multi-Response Permutation Procedure (MRPP)

Multiresponse Permutation Procedures (MRPP) are described in a recent book by Mielke and Berry (2001) which addresses the testing of hypotheses by use of distances between points. MRPP entails measuring the average within treatment distance which gives a measure of the variability within treatments by calculating the average of all pairwise distances between observations within each treatment group. Tests for treatment effects are then performed by comparing the average within treatment distance of the original sample to values obtained through permutation of the treatment labels. P-values are achieved using

$$\sum_{i=0}^{P} I(\bar{D}_i \leq \bar{D}_0) \over P+1$$

where $\bar{D}_i$ is the average within treatment distance of the $i^{th}$ permutation and $P$ represents the number of permutations used. $\bar{D}_0$ is the average within treatment distance received from the original data.

2.1 MRPP in Microarray Analysis

The MRPP method can also be used to test gene categories for evidence of different expression patterns between treatments. Categories can be overlapping, or genes may appear in multiple categories. This is important because multiple GO annotations can be attributed to a single gene which can create high correlations among the categories. It is required that categories be determined a priori, or in a manner unrelated to the data from the experiment to receive valid results. The data from the experiment is assumed to have been normalized previously and laid out in a $G \times N$ matrix. $G$ indicates the number of genes measured. $N$, or number of columns, is equal to the number of samples measured in an experiment. Since independence of the observations is required, the number of samples must also equal the number of experimental or observational units in the experiment. The number of treatments, or varieties, being analyzed is indicated
by T. For the $i^{th}$ treatment, there are $n_i$ samples measured such that $\sum_{i=1}^{T} n_i = N$.

For each category, a subset of the original $G \times N$ matrix of size $G_c \times N$ is obtained where $G_c$ is the number of rows, or genes, within category $c$. Each column consists of measurements from one experimental unit so each column of this matrix can be viewed as one vector in $G_c$ dimensional space. A multidimensional aspect of the MRPP analysis is then used. By associating each column with the appropriate treatment, the average within treatment distance of these vectors is calculated using

$$D = \sum_{i=1}^{T} \frac{n_i}{N} \left( \frac{\sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} d_{ijk}}{\sum_{x=1}^{n_i-1} x} \right)$$

(2)

where $d_{ijk}$ is the distance between replications (vectors) $j$ and $k$ for treatment $i$. The denominator results from there being a total of $\frac{n_i(n_i-1)}{2} = \sum_{x=1}^{n_i-1} x$ column pairs within each treatment group. The distance between samples, or columns, is computed using Euclidean distance, or

$$d_{ijk} = \sqrt{\sum_{l=1}^{G_c} (y_{ijl} - y_{ikl})^2}$$

where $y_{ijl}$ indicates the measurement value for treatment $i$, repetition $j$, and gene $l$. This method of calculating distances is recommended by Mielke and Berry(2001) and is used in all examples shown although other distance measures could be used if desired.

P permutations of the treatment labels amongst the columns are performed calculating the average within treatment distance ($\bar{D}_i$) for each one. All of the possible permutations should be used but even a modest number of replications per treatment make this too large for realistic use. A random subset of all of the possible permutations can be used as a practical approximation. This entire process is performed for each category using Equation 1 to obtain p-values. The null hypothesis ($H_0$) for this test is equal probability of occurrence for each of the possible R allocations of the N columns.
into the T categories. The total number of possible allocations is

\[ R = \frac{N!}{\prod_{i=1}^{T} n_i!}. \] (3)

### 2.2 P-values

Often, the number of gene categories being investigated can be large. This creates a need for p-value adjustments in order to keep the Type I false discovery rate at an acceptable level. When working with GO annotations, many categories consist of very high proportions of the same genes, some even containing the exact same genes causing high levels of correlation between categories. Research has shown the method of estimating the FDR presented by Benjamini and Hochberg (1995) to give reliable results under certain correlation conditions (Benjamini and Yekutieli, 2001). However, it was also shown that this method gave conservative results under certain highly positive correlation conditions. To eliminate this concern, another method was used as described in the SAFE method paper (Barry et al., 2005). This was an FDR estimation procedure created by Yekutieli and Benjamini (1999) which is specially designed for use with correlated test statistics. This method requires resampling so the average within treatment distance (\( \bar{D} \)) must be retained for every permutation of every category. This creates a \( P^{(1)} \times C \) matrix of average within treatment distances (\( \bar{D}_{pc} \)) where \( P^{(1)} = P + 1 \) with \( P \) being the number of permutations performed. The first row contains the distances obtained from the observed data. This matrix of distances must be converted to a \( P \times C \) matrix of p-values where

\[ p_{p^{(1)}c} = \frac{1}{P^{(1)}} \sum_{x=1}^{P^{(1)}} I(\bar{D}_{xc} \geq \bar{D}_{p^{(1)}c}). \]

Yekutieli-Benjamini’s estimate of the FDR is
\[
\tilde{FDR}_{YB}(p) = \min_{c: p_c \geq p} \left[ \frac{1}{p^{(1)} - 1} \sum_{x=2}^{p^{(1)}} \left( \frac{\hat{V}_x(p_c)}{\hat{V}_x(p_c) + \hat{S}_x(p_c)} \right) \right]
\]

in which

\[
\hat{V}_x(p) = \sum_{i=1}^{C} I(p_{xi} \leq p),
\]

and

\[
\hat{S}(p) = \hat{V}_1(p) - \left[ \left( \frac{1}{p^{(1)} - 1} \right) \sum_{x=2}^{p^{(1)}} \sum_{i=1}^{C} I(p_{xi} \leq p) \right].
\]

Through simulation, this method was shown to increase power while maintaining control over the false discovery rate even in cases of highly correlated observations. The \(\tilde{FDR}_{YB}\) was derived following a resampling based method for controlling the family wise error rate proposed by Westfall and Young (1989). When working with a smaller number of categories, controlling the family wise error rate may be preferred. Westfall and Young’s method

\[
\tilde{FWER}_{WY}(p) = \max_{c: p_c \leq p} \left[ \frac{1}{p^{(1)} - 1} \sum_{x=1}^{p^{(1)}} I \left( \min_{k: p_{kh} \geq p_l} p_{xh} \leq p_c \right) \right]
\]

will prevent the overly conservative results obtained by other methods when adjusting highly positively correlated p-values.
2.3 Considerations for Variance Differences Between Genes

A problem can be introduced by unequal variability between the genes when using the MRPP method to find gene categories with different expression patterns between the treatments. The gene or genes with larger variability can dominate the test results. This is shown in Figure 1. Each plot represents a category of two genes from an experiment with two treatments with ten independent measurements from each. For both plots, the measurements of Gene 2, which are shown on the vertical axis, have 10 times the variability of the observations of Gene 1. In Plot A, there is total separation between the observations of each treatment in terms of Gene 1 whose observations are shown on the horizontal axis colored by treatment. Considering these two genes as an entire category, the MRPP method of analysis was performed using 499 permutations and obtained a p-value of .324. Differences between Treatments 1 and 2 for Gene 1 are obscured because of the larger variability of Gene 2 making the MRPP test fail to find treatment differences for this category. In Plot B, the treatment separation occurs...
for Gene 2 which has the larger variability. There is now total separation between the treatments, so the p-value of the MRPP test is guaranteed to be $1/(P + 1)$ where $P$ is the number of permutations performed showing how more variable genes are allowed to dominate test results.

Two methods were considered to eliminate this problem. A procedure called Euclidean commensuration was proposed by Mielke and Berry (2001) for eliminating result domination by variability. Letting $y_{ij}$ be the expression measurement for gene $i$ and sample, or column $j$, $\phi_i$ is calculated using

$$
\phi_i = \left[ \frac{N-1}{N} \sum_{j=1}^{N} \sum_{k=j+1}^{N} |y_{ij} - y_{ik}|^v \right]^{1/v}.
$$

The observations in each row are standardized using $x_{ij} = y_{ij}/\phi_i$. The resulting data now has the property that

$$
\sum_{j=1}^{N-1} \sum_{k=j+1}^{N} |x_{ij} - x_{ik}|^v = 1
$$

for all $i$ values and any choice of $v$ making every gene have the same total distance using Equation 4 between points. This is called Euclidean commensuration when $v=1$.

A second method was considered which consisted of standardizing the measurements from a common gene by dividing them by their standard deviation. This ensures that every gene will have the same variance. This procedure resulted in the same results as Euclidean commensuration for all analysis ran. Because of this, Euclidean commensuration will just be used throughout the paper.

The data from Figure 1 was transformed using Euclidean commensuration which made the degree of distance separation the same for both genes. These values are shown in Figure 2. Performing the MRPP analysis with 499 permutations on the transformed genes now results in a p-value of 1/500 in both cases, eliminating the dominance of one gene in the outcome allowing separation to be found in both categories.
Figure 2  A plot of expression measurements from two genes. Treatment 1 is shown in blue squares, treatment 2 in red circles. Red and blue marks on the axes indicate the location of the measurements for each gene.

2.4 Benefits

A loss incurred when analyzing the p-value distribution, or the statistics of the genes, as used in either the SAFE or GSEA methods, is the information on the multivariate relationships between the genes in each group. A two dimensional example of this can be seen in Figure 3. Shown are the expression values of two categories of two genes in an experiment with two treatments and five replications. Two genes plotted together indicates that they share a category. When looking at the measurements for Gene 1, which are shown along the x-axis of plot A, there does appear to be some evidence of a difference between the treatments. Performing a t-test results in a p-value of .05383. Gene 2, shown along the y axis, appears to have no difference between the treatments and its t-test yields a p-value of .5531. For a category of size two the p-value distribution may not be convincing although when looking at the two dimensional plot of the values, there does appear to be separation between the treatments. This separation is shown in the results of the MRPP test, which yields a p-value of .0238. Genes 3 and 4, shown
on plot B, both appear to have a difference in expression between the treatments and this is also shown in the resulting p-values for the individual t-tests. The MRPP test also works in finding these differences yielding a p-value of .0079. This is the minimum p-value obtainable since there are only 126 possible permutations.

Figure 3  Plots of two categories consisting of two genes. Treatment 1 is shown in blue squares, treatment 2 in red circles. Red and blue marks on the axes indicate the location of the measurements for each gene. The p-values are from t-tests for treatment differences for that gene.

An additional strong point of the MRPP method is the lack of distributional assumptions which are not necessary when calculating multidimensional distances and acquiring p-values by permutation. This is a valuable asset but it does not aid in separating the MRPP method from the alternatives discussed earlier, which also use permutation tests.

3  Resampling Based P-value Test Method

A limitation of the MRPP method is the experiments to which the procedure can be applied. With the multi-response method of analysis, tests for interaction cannot be obtained. In many studies, the interaction between treatments is the test of interest.
To satisfy these conditions, the Resampling Based P-value Test (RBPT) method of analysis is proposed. This method can also perform tests of treatment differences as performed previously with the MRPP method.

When testing for treatment differences, the data is assumed to be laid out as mentioned previously, in a $G \times N$ matrix where $G$ indicates the number of genes measured and $N$ indicates the number of samples measured. When testing for treatment differences using the RBPT method p-values are obtained separately for every gene in the matrix. These values are often achieved using either a t or an F-test. The treatment labels are then permuted among the samples gathering p-values for every gene and permutation. Resulting, is a $G \times (P + 1)$ matrix of p-values where the first column, numbered 0, contains the p-values from the original data. Permutation is performed by the columns to ensure that the correlation structure between the genes is maintained. When performing all permutations is not feasible, $P$ permutations are selected, allowing equal probability to each of the possible permutations.

When testing for a difference in expression between treatments for a category the null hypothesis is, $H_0 : \mu_{i1} = \mu_{i2} = \cdots = \mu_{iT}$ for $i = 1, \cdots, G_c$. This procedure’s goal is not actually to maximize the power for testing against the null hypothesis of no treatment mean differences for every gene in a category. Instead the goal is to find categories where there is evidence of differential expression for a substantial proportion of the population. The method used in obtaining p-values was created keeping this in mind. First, the $G \times (P + 1)$ matrix of p-values for every gene and permutation is reduced to a $G_c \times (P + 1)$ matrix containing the p-values for every gene in category $c$. The $1 \times G_c$ vectors of p-values obtained from every permutation, and the original data, is condensed to a single value by calculating the median. Other statistics such as the mean, Fisher’s p-value ($\prod_{i=1}^{G_c} p_i$), and Pearson’s p-value ($1 - \prod_{i=1}^{G_c} (1 - p_i)$) were also seriously considered but the median was chosen for it’s robustness to outlying p-values. We did not want to find categories with a very small group of genes with very small
p-values. We would rather find categories with many genes with moderately reduced p-values and using the median enabled this. When $H_0$ is true and the test assumptions are met the resulting p-values will follow the uniform distribution $[0, 1]$. Under these conditions, the median would have an expected value of 50%. When $H_0$ is false the p-values would follow a distribution more concentrated towards zero which would have an expected median less than 50%. If there were only a couple of genes in the category for which the null hypothesis was false, little impact would be implied upon the median statistic for the category. This was a concern since tests receiving small p-values under biologically questionable conditions are not uncommon in microarray experiments as described previously. If many genes in a category had moderately reduced p-values, this would be shown using the median. A graphical representation of the steps of the RBPT test is shown in Figure 4.

Figure 4  RBPT method. Matrices containing the original data values are shown in green while matrices of p-values are shown in blue.
3.1 Testing Interaction

Tests for which the RBPT method can be performed exactly as described previously using permutations are limited. Adjustments are necessary in order to test for interaction. Residual bootstrapping is a reasonable alternative which allows for tests of interaction as shown in Davison and Hinkley (1997). Bootstrapping is similar to permutation except that the data are sampled with replacement.

To test for interaction, an alteration must be made to the RBPT method shown in Figure 4. The only change in this process is in how the p-value matrix is obtained. An important detail when testing for interaction is that no additional assumptions can be made regarding the other parameters, such as the main effects. In order to do this, the main effects must be accounted for in carrying out the bootstrapping. For example, in an experiment following a completely randomized design with a 2 factorial treatment structure, the data would first be modeled on a gene-by-gene basis as:

\[ y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ijk} \]

where \( \alpha \) and \( \beta \) are the treatment main effects and \( (\alpha \beta) \) is the interaction. The \( i \) and \( j \) indicate the treatment levels and \( k \) is the replication number. This model would be fit to the data for every gene keeping the p-values for the test of interaction along with the residuals. Next, each gene is once again analyzed using the model

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \]

in which \( \alpha \) and \( \beta \) remain the treatment main effects and this time keeping the predicted values for every observation. Bootstrapping is performed by randomly sampling from the columns of the residual matrix obtained using Model 5 with replacement \( N \) times. The sampling is done by the column to preserve the correlation amongst the genes. This creates a \( G \times N \) matrix which is added to the matrix of predicted values obtained
using Model 6. A p-value is then obtained for each gene using Model 5 to test the null hypothesis of no interaction. This process is performed \( P \) times preserving the p-values. Through this, the \( G \times (P + 1) \) matrix shown in Figure 4 is obtained and the remaining steps are followed accordingly, to get p-values for each category. Each category is then tested for the null hypothesis of \( H_0 : (\alpha \beta)_1 = (\alpha \beta)_2 = \cdots = (\alpha \beta)_T = 0, \) for \( i = 1, \cdots, G_c. \) Once again, this test is not aimed at maximizing the power in testing against treatment interaction for each gene since the goal is to find categories with substantial evidence of interaction. This method does enable testing strictly for interaction through residual bootstrapping while not making assumptions about the main effects. With some minor restrictions on how the bootstrapping is performed, this method could also be applied to experiments following a randomized complete block design.

4 Regulation

In microarray experiments, interest is not only in selecting which genes have significant changes in expression levels, but also in the direction of the change. When the gene’s production is reduced under the treated conditions, that is called down-regulation. The opposite is called up-regulation. When the MRPP method is used to find gene categories with difference in expression, only the multidimensional distance between points, and not the direction, is used in the analysis. When using the RBPT method, up or down-regulation can be seen in the signs of the t-values. However, these values are not used in the test. Consequently, a second step is necessary to statistically test for regulation direction for either method.

Investigating the type of regulation is only applicable to cases where two treatments are compared. When more than two treatments are being analyzed, only pairs of the treatments can be considered at a time. To test for regulation, after a category has
been declared significant the data are reduced to a $G_c \times N$ matrix containing only the measurements corresponding to the genes within the category of interest. For every gene, t-tests are performed resulting in a $G_c \times 1$ vector of t-values. The number of positive t-values ($t^+$) is then counted. The treatment labels are then permuted $P$ times between the $N$ columns. Permutation by the columns is used to preserve the correlation between the genes as described previously. To test for up-regulation, a p-value is obtained by
\[ \sum_{i=0}^{P} I(t_0^+ \leq t_i^+) / (P + 1), \]
or the percentage of the times the observed $t^+$ value ($t_0^+$) is greater than all of the $t^+$ values obtained through permutation. The p-value for down-regulation is achieved by
\[ \sum_{i=0}^{M} I(t_0^+ \geq t_i^+) / (P + 1). \]
Both tests have a null hypothesis of equal dispersion of up and down regulated genes.

This test is only applied to categories for which the null hypothesis was rejected. This test is just meant to give an idea of categories in which there is some form of regulation. Because of this, no multiple comparison adjustment is made to the resulting p-values. This test also places no bounds on what kind of a disproportion of up or down regulation is necessary to test significant. There have been cases where categories having only 58% of the genes being up-regulated tested significant. This may not be biologically meaningful but the test results do find an imbalance.

5 Influential Genes

After having performed the analysis and decided upon a group of categories to declare as being differently expressed between the treatments, researchers might also wish to identify the genes predominantly responsible for the significant test results to further their research. This may seem as going full circle because the concentration is now back onto individual genes. However, the only genes being investigated are contained within the category of interest. One option would be to simply look at the p-values for each gene in a category. Simply selecting a false discovery rate for these p-values could give a
list to investigate. However, this would ignore the relationships between the genes that
could be responsible for the category’s selection.

An alternative, when dealing with two treatment varieties, would be to use Fisher’s
discriminant function (Fisher, 1938). This finds the linear combination of the response
variables which achieves maximum separation between the treatments in standard de-
viation units. Let $X^{(i)}$ be a $G_c \times n_i$ matrix of observations containing all samples that
received treatment $i$ and all genes in category $c$. Let $X$ be the horizontal concatenation
of $X^{(1)}$ and $X^{(2)}$. The linear combination which achieves maximum separation between
the treatments is found maximizing

$$
\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}
$$

where $y_{1j} = \hat{a}'X_{\bullet j}^{(1)}$ and $y_{2j} = \hat{a}'X_{\bullet j}^{(2)}$ with $j = 1, \cdots, n_1$ and $k = 1, \cdots, n_2$. $X_{\bullet j}^{(i)}$ indicates
the vector of responses for sample $j$ which received treatment $i$. The loadings are
obtained using

$$
\hat{a}' = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_X^{-1}
$$

where $\bar{X}^{(i)}$ is a $1 \times G_c$ vector consisting of the means of $X^{(i)}$ for every gene. $S_X^{-1}$ is the
inverse of the covariance matrix of $X$. The maximum separation is found using $y_{ij} = \hat{a}'X_{\bullet j}^{(i)}$
where $\hat{a}'$ contains the loadings for every gene indicating their value in separating the
varieties. These loadings are then used to indicate which genes play the strongest role
in dividing the varieties.

A problem with using this method is the requirement of calculating the inverse of
the covariance matrix ($S_X^{-1}$). In microarray experiments, the number of genes being
studied is much greater than the number of samples ($G >> N$) and this is also usually
true for the categories of genes being analyzed ($G_c > N$). This causes the covariance
matrix needed to compute Fisher’s discriminant function to be singular which makes
it not possible to invert. A shrinkage approach (Schafer and Strimmer, 2005) has been
proposed as a method to overcome this difficulty. By substituting the shrinkage estimate,
For $S_X$, one is guaranteed that it is positive definite enabling the inverse to be calculated.

To show the information gained using Fisher’s discriminant function, an example was created. A small five gene category was made with the first two genes expression patterns displayed in Figure 5. Genes three through five were randomly simulated following the standard normal distribution. The data was commensurated and analyzed using the MRPP method to yield a p-value of .049. To show which genes were the most active in causing the category’s low p-value, both t-test statistics and the loadings for Fisher’s discriminant function were obtained for each gene. To make them comparable, the values were standardized and can also be seen in Figure 5. A gene’s influence on separating the treatments is indicated by the severity of either its loadings or t-values. The loadings for genes one and two are more extreme than the t-values correctly indicating the two genes which had the larger impact in the category test results. This is just a simple example but the added information obtained from the gene correlations by using Fisher’s discriminant function enables better selection of contributing genes.

6 Examples

Myostatin is a protein which inhibits the rate of muscular cell growth and differentiation. Cattle with mutations in the gene responsible for myostatin production, such as the Belgian Blue and Piedmontese, have increased quantities of muscle mass. Understanding what other proteins are affected by the suppression of myostatin is of interest to science. Transgenic mice that had their myostatin gene knocked out were studied in a recent experiment (Steelman et al., 2006) to study differences in gene expression caused by myostatin production. The experiment compared the expression levels of the transgenic mice to that of normal or wild type mice at three time points. The time points were selected to test the impact of myostatin at both the primary and secondary
Figure 5  Expression values of genes 1 and 2 from a 5 gene category are shown on the left. The red triangles indicate Treatment 1 while the green diamonds indicate Treatment 2. Standardized values of the loadings from Fisher’s discriminant function are shown in blue stars for each gene in the plot on the right. The standardized t-test statistics for each gene are shown in green squares.

stages of muscular tissue formation along with a time of fast muscular growth. This experiment was performed using oligonucleotide arrays on which only one mRNA sample could be measured. A total of five mice of each genotype where measured at each time point. The data used were normalized by MAS 5.0. Additional normalization was performed by subtracting the slide median from every observation on a slide. This made every slide have an equal median value of 0.

One of the interests of the study was to find GO annotations which exhibit expression differences between the transgenic and wild type mice at the time of fast muscular growth. GO annotations have a hierarchical structure that ascends to a single point. Each term can have multiple children but one important difference between a normal hierarchy and the GO consortium is that one term can have more than one parent. For the experiment, the GO classification for each gene was obtained from the site of the company responsible for chip production (www.affymetrix.com). This listing
often gave multiple annotations for each gene. A gene was considered to be within all annotations listed as well as the ancestry of all the annotations listed. Since each ontology is hierarchical to a single GO term, every gene listed within any GO term in an ontology is also contained in the highest ancestral GO term for that ontology. The ancestry of each GO term was obtained by use of the GOstats package in R. The analysis was performed on both the biological process and molecular function ontologies. All GO terms that contained less than forty genes were omitted as they were when testing the SAFE method (Barry et al., 2005).

There were a total of ten mice, five of each genotype, whose expression was measured at the final time point of the experiment. A total of 45,101 different probe sets were measured with 28,110 of them having been assigned to at least one GO molecular function category. A total of 341 different molecular function categories were tested because they met the SAFE method criterion and contained at least forty probe sets. For the biological process ontology, 24,816 probe sets were contained within 739 categories which were also tested for evidence of differential expression between the genotypes.

![Figure 6](image-url)

**Figure 6**  Histogram and box plot of the variance estimates for each probe set within the molecular function ontology. The variance values were transformed using the log, base 10. The actual minimum, median, and maximum values are displayed above the box plot.
The variability within each probe set analyzed was investigated first. This should be done as a prerequisite in order to know whether or not standardization or commensuration is necessary for the MRPP analysis. Figure 6 displays the distribution of the variance estimates of each probe set within the molecular function ontology. An upper quartile of approximately .5 and a lower quartile of approximately .04 of the variances supports the use of Euclidean commensuration of the data which was performed to prevent more variable probe sets from dominating test results.

The data were analyzed using both the RBPT and MRPP methods of analysis. Due to the small number of replications, there were only 126 possible permutations of the data. Thus, every possible permutation was used in analyzing each GO annotation. Using the MRPP method resulted in 223 categories testing significant at a false discovery rate of .006. Because of the possibly high correlation between the categories, the FDR was calculated using the Yekutieli-Benjamini method as discussed previously. These categories all had the minimum p-value of 1/126. With the RBPT test, there were 140 categories with the minimum p-value of 1/126. A comparison of the p-value distribution of both analysis methods, after taking the \( \log_{10} \), is shown in Figure 7. The actual spots have been jittered slightly to show the quantity at each location. Using the minimum p-value as a selection criterion, 135 categories were selected by both tests. There were 5 groups found with the RBPT method but not the MRPP method, while there were 88 categories selected with the MRPP test but not the RBPT test.

A plot of all GO categories tested are colored by selection method and are shown in Figure 8. Each dot in the figure represents one GO annotation. The plot shows the hierarchical structure of the molecular function GO annotation. As the hierarchy branches out, the GO terms become more specific. Arrows point to the ancestor from the descendant. Since there were only two treatments, tests of up and down-regulation were also performed. GO annotations with p-values \( \leq 5\% \) were declared significantly regulated. Up-regulation indicates an increase in gene expression for the wild type
subjects. Terms classified as significant and up-regulated are shown in green while evidence of down-regulation is shown in purple. All up and down-regulated categories tested significant by both methods except for the only up-regulated category which is not connected to another one. This annotation only tested significant by the MRPP method. All terms not testing significant for difference in expression for the treatments are shown by gray ellipses. The results show a great deal of ancestral consistency in that almost all categories that test significant have at least one direct ancestor also testing significant. One possible problem indicated by the hierarchy is that not very many high-level annotations seem to test significant. It was brought to our attention that there has recently been some questions raised about the validity of some of the GO annotations for mice at the higher levels. Both tests were designed assuming a high level of confidence in the GO annotations used. The deficiencies in GO annotations cannot be corrected for but could explain the possible inconsistencies.

By reducing the plots down to a smaller section of the hierarchy as in Figure 9, the specific GO annotations causing the broader terms to test significant can be revealed. Figure 9 reduces the hierarchy to containing only offspring of annotation “GO:0005215”, which is transporter activity. There are six different children that test significant. In order to specify the reason why transporter activity appears to differ between the wild type and transgenic mice, different children can be followed up to a higher specificity. For example, one of the sequences is 05215-05342-46943-15171-15179. These terms stand for transporter activity, organic acid transporter activity, carboxylic acid transporter activity, amino acid transporter activity, and L-amino acid transporter activity. The final term can give scientists a specific group of genes pertaining to one task for further studies and understanding of how this expression change effects muscular cell growth.

Concern could be raised over the percentage of the categories testing significant by both methods in this example. The RBPT method declared over one third of the annotations as having significant expression change between the treatments while the
MRPP method found nearly two-thirds. One reason for this is that with both of these tests, categories testing significant cause all of their ancestors to also test significant increasing the number of categories found to have differential expressions patterns. Also, using a procedure proposed by Nettleton (2006) on the p-values obtained from analysis ran on a gene-by-gene basis for genotype differences, the proportion of the population for which the null hypothesis is false is estimated to be approximately 23%. Taking both of these factors into consideration, a large number of categories testing significant should be expected for a data set such as this one.

6.1 RBPT versus MRPP

The RBPT method is similar to both the SAFE and GSEA methods discussed earlier in that all three require calculating a test statistic for each gene. The main difference is in how the p-values are then calculated. Both the SAFE and GSEA methods compare the distribution of the test statistics from a category, to the distribution of the test statistics from the rest of the population. These procedures find categories whose distribution differs from the rest of the population. In cases where a large number of genes have different rates of expression between the treatments, these methods would still find a small number of categories since testing for population differences. The RBPT method differs in that each category’s statistic is compared directly with statistics obtained using either permutation or bootstrapping. The RBPT method is testing for significant difference in expression between the treatments while SAFE and GSEA are testing for a difference in expression between the treatments that are different than the rest of the population.

An advantage of the MRPP method is the reduction in computing time that is necessary. All the other methods require statistics to be calculated for every gene with every permutation and this process can be computationally intensive. With MRPP method, the Euclidean distance is computed for every column in a category data set.
This requires calculating a total of \( \binom{n_i}{2} \) distances. The Euclidean distances are easily computed by first calculating the differences between two vectors and then squaring those values. The square root of the sum of these values is then calculated taking little computational time. After the distances have been calculated, all the permutations are with respect to the distances between samples which reduces the size of the data set being permuted and computations necessary immensely.

Both the RBPT and the MRPP test appear to get consistent results in that annotations testing significant can be followed up the ancestral tree. The MRPP method does appear to find a larger number of annotations. A possible explanation for this is that the MRPP test can find multidimensional separation between the treatments that involves multiple genes simultaneously. Through simulation it was found that when using commensuration, just a few genes in the population having different expression patterns between the treatments will not make a category test significant. However, if commensuration is not used, large categories can test significant with only one gene actually having treatment expression differences. Since the median is a robust statistic to the influence of a small proportion of the sample, the RBPT method appears to never have this problem through the simulations. A possible weakness the RBPT method is that it shows difficulty in finding a multidimensional separation of the treatments for which the MRPP method detects.

7 Conclusion

Testing for expression changes by category aids in the understanding of what aspects of cell activity are affected by treatment or variety differences. Both methods proposed use resampling to obtain p-values, eliminating any distributional assumptions. Additional information is obtained from the multivariate relationship of samples by use of
the MRPP method. As compared to gene-by-gene analysis methods, the MRPP method reduces computation time while using information normally discarded. One shortcoming is the limitation of experimental designs and tests for which this method can be used. The RBPT method has benefits in being more versatile so it can be an option when looking for effects like treatment interaction.

8 References

Affymetrix: Statistical Algorithms Description


Figure 7  Comparison of p-value distributions for the RBPT and MRPP tests. The p-values have been log_{10} transformed. Spots colored red indicate locations where both p-values were 1/126. Orange locations indicate spots where the MRPP p-values equal 1/126 but RBPT are greater than 1/126 and blue spots indicate the opposite. The orange histogram shows the distribution of the MRPP p-values while the blue histogram indicates the distribution of the RBPT p-values.
Figure 8  Plot of the molecular function hierarchy studied. GO annotations selected by both methods are shown in red. Categories selected by the MRPP method only are orange and blue nodes were selected only by the RBPT method. Green annotations tested significant for up-regulation while down-regulated categories are colored purple.
Figure 9  Smaller hierarchy for GO:0005215
This dissertation researched three areas of experimental design and analysis of microarray experiments. The first paper researched important details about two experimental designs. The second paper proposed a method of design comparison which compensated for mixed model analysis. The third paper proposed methods of analyzing microarray data by groups or categories.

For microarray experiments, it was shown that an experimental design called DBWT is more powerful than the DBWS design when constrained by the number of arrays. This was shown through variance comparison and an actual experiment. We also proposed an alternative method of analysis for each design which prevented a possible bias in the results. Both methods guarantee a uniform distribution of the p-values when the null hypothesis is true and the assumptions hold.

Mixed model analysis was shown to be more powerful than performing ANOVA on the array differences. Showing that mixed model analysis should be used, reinforces selection of designs using the more complicated mixed models. Prior knowledge of the variance component’s distribution is valuable when using mixed model design comparisons since the optimal design depends upon the parameter space of the variance components. Because the number of subjects, relative to the number of arrays used varies for different designs, the number of replications possible for a design can vary by cost limits. This makes the cost of experimental units relative to the cost of arrays have an impact upon design selection. We show that cost ratios along with prior knowledge of the variance components are valuable information when choosing mixed model designs.
that require multiple random effects.

Analysis of microarray experiments can also be performed on groups or categories of genes instead of on a gene by gene basis. High correlation among the groupings because of classification of genes into multiple categories is acceptable. We show that when testing for difference in expression between varieties, additional categories can be found by comparing the multi-dimensional distances between the samples with the MRPP method. Using the distances allows discovery of interaction between the genes to separate the varieties expression patterns. This method finds categories that can’t be found by using tests of each gene within a category. There are experiments which are looking to find factors like interaction between the varieties in which the MRPP method can not be used. For these cases, we propose the RBPT method which uses statistics that are calculated for every gene in a category. P-values are obtained through resampling of the columns, or samples, which preserves the correlation between the genes in the category.

Scientists can also be interested in whether these categories are up or down-regulated and which genes within the category are changing expression the most. Additional procedures were proposed to answer these questions which can be performed regardless of whether the RBPT or MRPP method was used in finding the categories. These discoveries will aid in biological discoveries of the future through design and analysis of microarray experiments.
Acknowledgments

I would like to thank Dr. Nettleton for all his advice and patience along the way. His help and knowledge was essential in enabling the production of my thesis. I would like to thank Dr. Reecy for his advice on the biological side of microarrays as well as allowing the use of his data in the thesis. I would also like to thank Dr. Dixon and Dr. Koehler for their help with both classes and statistical advice throughout my career at Iowa State and Dr. Tuggle for working with me during the analysis phase of the experiments I was involved in.

I would also like to thank my family, especially my parents, for everything they have taught me throughout my life. Last, but definitely not the least, I would like to thank my wife, Jen, for everything that she has helped me with in all phases of my life over the years at Iowa State. Her aid has made everything possible and her emotional support enabled me to stay sane, though that may be debatable, and I owe everything to her!
Appendix

1 Example Code

```plaintext
title "DBWU with Satterthwaite df";
proc mixed;
   by gene;
   class array dye trat unit block;
   model y_dbwu = dye trat / ddfm=satterth;
   random block unit(trat) array;
   run; quit;

title "DBWU with Containment df";
proc mixed;
   by gene;
   class array dye trat unit block;
   model y_dbwu = dye trat ;
   random block unit(trat) array;
   run; quit;

title "DBWU with Satterthwaite df";
proc mixed;
   by gene;
   class array dye trat block;
   model y_dbwu = dye trat / ddfm=satterth;
   random block array;
   run; quit;

title "DBWU with Containment df";
proc mixed;
   by gene;
   class array dye trat block;
   model y_dbwu = dye trat ;
   random block array;
   run; quit;
```
2 Containment Method - Order Dependency

The order of the listing of the random effects can change the denominator degrees of freedom of the test when using the containment method for PROC MIXED in SAS. In Figure 2, the exact same data set, simulated in the form of a DBWU experiment, is analyzed twice using the containment method for estimating the denominator degrees of freedom. The only change in the model is the order of block and unit(trt) in the random effect listing resulting in a change in the denominator degrees of freedom.

```plaintext
proc mixed;
class array dye trt unit block;
model y_dbwu = dye trt;
random block unit(trt) array;
run; quit;
```

**Type 3 Tests of Fixed Effects**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>dye</td>
<td>1</td>
<td>2</td>
<td>9.57</td>
<td>0.0906</td>
</tr>
<tr>
<td>trt</td>
<td>1</td>
<td>2</td>
<td>0.81</td>
<td>0.4638</td>
</tr>
</tbody>
</table>

```plaintext
title "DBWU with Containment df";
proc mixed;
class array dye trt unit block;
model y_dbwu = dye trt;
random unit(trt) block array;
run; quit;
```

**Type 3 Tests of Fixed Effects**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>dye</td>
<td>1</td>
<td>2</td>
<td>9.57</td>
<td>0.0906</td>
</tr>
<tr>
<td>trt</td>
<td>1</td>
<td>4</td>
<td>0.81</td>
<td>0.4198</td>
</tr>
</tbody>
</table>