Investigation and development of statistical method for gene expression data analysis

Nicholas Bradley Larson
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation
Larson, Nicholas Bradley, "Investigation and development of statistical method for gene expression data analysis" (2011). Graduate Theses and Dissertations. 10448.
https://lib.dr.iastate.edu/etd/10448

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 Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Investigation and development of statistical methods for gene expression data analysis

by

Nicholas Bradley Larson

A dissertation 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:
Daniel Nettleton, Co-major Professor
Jack Dekkers, Co-major Professor
Alicia Carriquiry
Peng Liu
Huaiqing Wu

Iowa State University
Ames, Iowa
2011

Copyright © Nicholas Bradley Larson, 2011. All rights reserved.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. v
LIST OF FIGURES ............................................................... viii
ACKNOWLEDGMENTS ............................................................ x
ABSTRACT ................................................................. xi

CHAPTER 1. GENERAL INTRODUCTION ................................. 1
  1.1 Gene Expression ......................................................... 1
    1.1.1 Microarray Technology ............................................. 3
    1.1.2 Next-Generation Sequencing and RNA-Seq ....................... 5
  1.2 Analyzing Expression Data .............................................. 7
    1.2.1 Microarray Data Analysis ......................................... 7
    1.2.2 Analysis of RNA-Seq data ........................................... 8
  1.3 Statistical Challenges of Gene Expression Analysis ............... 9
    1.3.1 Dependence ......................................................... 9
    1.3.2 Small Samples ..................................................... 9
  1.4 Relevant Statistical Methods .......................................... 10
    1.4.1 Multiple Testing and FDR ........................................... 10
    1.4.2 Permutation Testing ............................................... 13
    1.4.3 Bayesian Latent Class Models .................................... 14
  1.5 Thesis Organization ................................................... 16

CHAPTER 2. ESTIMATING THE PROPORTION OF TRUE NULL HYPOTHESISSES WHEN CONDUCTING MANY PERMUTATION TESTS WITH NON-NORMAL LOW-SAMPLE-SIZE DATASETS .............. 27
Abstract ................................................................. 27

2.1 Introduction ....................................................... 27
2.2 Methods ............................................................ 31
  2.2.1 Permutation Testing ........................................ 31
  2.2.2 The Distribution of the Permutation P-value under the Alternative . . 32
  2.2.3 Estimation of the True Null Proportion ..................... 35
  2.2.4 Cross-Validation ........................................... 36
  2.2.5 Simulation Design ....................................... 37
2.3 Simulation Results ............................................... 38
2.4 Case Study ....................................................... 40
2.5 Discussion ...................................................... 43
2.6 Conclusions ..................................................... 47

CHAPTER 3. THE IMPACT OF SIGNIFICANCE THRESHOLD SELECTION ON THE RELATIONSHIP BETWEEN FALSE DISCOVERY RATE ESTIMATES AND THE REALIZED PROPORTION OF FALSE DISCOVERIES ........................................ 51

Abstract ................................................................. 51

3.1 Introduction ....................................................... 51
3.2 Methods ............................................................ 53
  3.2.1 Estimation of False Discovery Rate ......................... 53
  3.2.2 Data Simulation ........................................... 55
  3.2.3 Data Analysis and Threshold Selection ....................... 56
3.3 Simulation Results ............................................... 57
3.4 An Explanation of the Negative Correlation between $\hat{\text{FDR}}$ and $Q$ ................. 60
3.5 A Modeling Approach to Sensitivity Analysis ....................... 62
3.6 The Empirical Null Method .................................... 65
3.7 Discussion ...................................................... 67
3.8 Conclusions ..................................................... 69
3.9 Pearson Correlations from Data Simulations ................... 70
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>METHODOLOGY FOR EVALUATING THE COMPLEMENTATION MODEL OF HETEROsis USING HIERARCHICAL LATENT-CLASS MODELING OF RNA-SEQ DATA</td>
<td>77</td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>78</td>
</tr>
<tr>
<td>4.2</td>
<td>Methods</td>
<td>81</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Latent Class Definitions</td>
<td>81</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Data Model</td>
<td>82</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Bayesian Approach</td>
<td>84</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Model Fitting and Inference</td>
<td>85</td>
</tr>
<tr>
<td>4.3</td>
<td>Application</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation Analysis</td>
<td>89</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>90</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusions</td>
<td>94</td>
</tr>
<tr>
<td>4.A</td>
<td>JAGS Model Code</td>
<td>95</td>
</tr>
<tr>
<td>4.B</td>
<td>Simulation Contingency Tables</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>GENERAL CONCLUSIONS</td>
<td>102</td>
</tr>
<tr>
<td>5.1</td>
<td>Future Research Directions</td>
<td>103</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

1.1 Table of possible outcomes conducting $m$ simultaneous hypothesis tests 11

2.1 Descriptions of each of the six data-based simulations used to evaluate both estimators of $\pi_0$ ...................... 40

2.2 RMSE results from the 6 data simulations involving varying values of $\pi_0$ and distributions $g(\delta)$ ...................... 40

2.3 Estimates of the first five values of $\theta$ for the semiparametric method applied to the mouse data ...................... 43

3.1 Pearson correlation coefficients for the data simulation using $\pi_0 = 0.90$ for each of the five FDR estimators, RND DE selection, and five values of $G$ ...................... 58

3.2 Pearson correlations between $\hat{FDR}_{OR}$ and $Q$ under different values of $\sigma_{\lambda}$ 64

3.3 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.90$ using $\hat{FDR}_{EF}$ for each of the five significance thresholds. 66

3.4 Mean $Q$ and $\hat{FDR}$ for all five estimators, using RND DE selection, $\pi_0 = 0.90$ and $G = 500$ ...................... 68

3.5 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.75$ and RND gene selection ...................... 70

3.6 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.80$ and RND gene selection ...................... 70

3.7 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.85$ and RND gene selection ...................... 71
3.8 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.95$ and RND gene selection ........................................... 71
3.9 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.75$ and GOT gene selection ........................................... 72
3.10 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.80$ and GOT gene selection ........................................... 72
3.11 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.85$ and GOT gene selection ........................................... 73
3.12 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.90$ and GOT gene selection ........................................... 73
3.13 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.95$ and GOT gene selection ........................................... 74

4.1 Definition of all 8 possible state vectors for binary classification of transcriptional activity for the three genotypes ......................... 82
4.2 Prior distribution assignments for hyperparameters of the read count model. Normal distribution is presented using the $N(\mu, \sigma^2)$ parameterization .......................................................... 85
4.3 Posterior means and 95% credible intervals for $\theta^*$ applied to the experimental data setting $\lambda_e = 0.001$ ........................................... 88
4.4 Predicted number of genes exhibiting each state vector based upon $\hat{Z}_i$. 88
4.5 Tabulation of the number of reads associated with observations classified to be transcriptionally inactive fitting the Zea mays data set ............ 89
4.7 Table of posterior means and 95% credible intervals of the model parameters for separate model fits using a grid of $\lambda_e$ values. The columns of results corresponding to model fit using the true value of $\lambda_e$ are indicated in bold ....................................................... 91
4.8 State vector classification results $\hat{Z}_g^*$ from simulation analysis along with true values derived from $\hat{Z}_g$ ................................. 92
4.9  Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.01$. 
92

4.10 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.000001$. 
96

4.11 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.00001$. 
96

4.12 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.0001$. 
97

4.13 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.001$. 
97

4.14 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.1$. 
98
**LIST OF FIGURES**

1.1 Example of a partial DNA sequence (above) with its complementary strand (below). DNA sequences are traditionally read from the 5’ to the 3’ end. ................................................................. 2

2.1 Approximations of the probability mass distributions of $f_{DE}(p \mid \delta)$ obtained via simulation for the case $n_1 = n_2 = 5$ when (a) $\delta = 1$ and (b) $\delta = 2$. Each line represents the probability mass associated with each possible $p$-value $r/R$. ............................................................. 33

2.2 Example linear B-spline basis with $\delta_{max} = 6$ and $K = 7$ ................. 34

2.3 Graph of the three NCP distributions used in simulation, with $N$ (black), $M$ (red) and $F$ (green) shown. Each of the three distributions use $c = 4$, and parameter values $(\alpha, \gamma)$ $N$ being (1,2), $M$ being (2,2) and $F$ being (4,2) ................................................................. 39

2.4 Boxplots of the estimates of $\pi_0$ when the true null proportion is (a) 0.70 and (b) 0.95, with these values indicated by the horizontal dashed line in each plot. The simulation results are indicated by the estimator used (“Semi” or “Non”) along with the distribution on noncentrality parameter $\delta$ ($N=$Near, $M=$Moderate, and $F=$Far) in parentheses . . . 41

2.5 Empirical $p$-value distributions using (a) $t$-tests and (b) permutation testing ................................................................. 42

2.6 Empirical null $p$-value distributions for the null hypotheses for the $t$-test (left) and permutation test (right) procedures ................................. 44
2.7 Semiparametric (semi), compromise (comp), and nonparametric (non)
estimates of $\pi_0$ for simulation data with true $\pi_0 = 0.95$ and $g(\delta) = N$.

3.1 A plot of $\widehat{FDR}_{ST}$ vs. the realized proportion of false discoveries $Q$ for
the simulation $\pi_0 = 0.90$, RND DE selection, and $G = 500$.

3.2 Depiction of $\widehat{FDR}_{ST}$ estimates using $t = p(500)$ across RND and GOT
DE selection methods for the five values of $\pi_0$.

3.3 Histograms of EE $p$-values from microarray data simulation with RND
DE selection and $\pi_0 = 0.90$. For DE gene count $G = 500$, (a) $Q = 0.602$
and $\widehat{FDR}_{OR} = 0.106$, while (b) $Q = 0.104$ with $\widehat{FDR}_{OR} = 0.896$.

3.4 Plots of (a) lowess smoothed EE distributions from one of the simula-
tions (RND, $\pi_0 = 0.90$) and (b) a variety of densities $f_{EE}(x \mid \alpha)$ where
$\alpha \in [-1.5, 1.5]$.

3.5 A plot of $\widehat{FDR}_{EF}$ vs. $Q$ for the simulation $\pi_0 = 0.90$, RND DE selection,
and $G = 500$.

4.1 Basic representation of the genetic concept of complementation.
ACKNOWLEDGMENTS

I would first and foremost like to thank my major advisor, Dr. Dan Nettleton, for all his input and support in making the completion of this degree possible. I would also like to extend my thanks to my co-major advisor, Dr. Jack Dekkers, as well as the remaining members of my program-of-study committee: Dr. Alicia Carriquiry, Dr. Peng Liu, and Dr. Huaiqing Wu. I am also deeply indebted to the professors who provided me with my graduate instruction here at Iowa State University, notably Dr. Karin Dorman, Dr. Heike Hofmann, Dr. Mark Kaiser, Dr. Kenneth Koehler, and Dr. Stephen Vardeman. Finally, I would like to thank my friends and colleagues within the department, who have made these last five years an immensely enjoyable experience, and to my family, for supporting and encouraging me to pursue my goals and attain my doctorate.
ABSTRACT

Gene expression data analysis is a critical component to how today’s researchers comprehend biological function at a molecular level. With the amount of data being generated outstripping the ability to analyze it, it is critically important that the development of statistical methodology keep pace with technological advancement in order to fully take advantage of this wealth of information. In this dissertation, we examine issues that are presented in the context of gene expression analysis and develop new methods to account for these complications via three separate papers, contained in Chapters 2, 3, and 4.

Chapters 2 and 3 are closely related in their relevance to the detection of differential expression and multiple testing procedures for microarray analysis. Specifically, in Chapter 2 we modify an existing semiparametric estimator of the true null proportion of hypothesis tests to make use of permutation testing. We argue that this approach is more appropriate for the typically small sample sizes of microarray experiments, especially since expression data is nonnormal. We show that our modification is more accurate than the original approach using simulated data based upon real microarray expression values, and advocate its use for microarray analysis when small sample sizes are used.

In Chapter 3, we examine the implications of rejecting a fixed number of genes for the detection of differential expression on FDR estimation. We employ a wide variety of estimators which assume a uniformly distributed empirical null $p$-value distribution, and our findings show that there is a strong, negatively correlated relationship with FDR estimates and the true proportion of false discoveries, $Q$, when significance is determined in this fashion. This phenomenon is observed over a wide variety of simulation conditions. We also show that, in conjunction with publication bias, this type of significance threshold selection results in liberally biased estimates of FDR. We contrast these estimators with Efron’s empirical null approach, which produces an FDR estimator which is positively correlated with $Q$. 
Chapter 4 involves the development of a method for simultaneously classifying the transcriptional activity of genes using RNA-Seq data. We specifically consider a crossbreeding experiment involving two inbred lines of maize in order to investigate the complementation model of heterosis. We use a negative binomial distribution to model the read counts, and assume a simple latent class model for transcriptional activity. Application of this model to experimental as well as simulated data provides reasonable classifications and identifies specific genes that appear to be in accordance with the complementation theory of heterosis. We argue for the use of this model in other breeding experiments to further investigate the complementation theory.
CHAPTER 1. GENERAL INTRODUCTION

Since the discovery of the structure of the deoxyribonucleic acid (DNA) molecule by Watson and Crick (1953), the understanding of the molecular mechanisms of genetics has increased exponentially. The benefits have been manifold, with applications ranging from medicine to agriculture. With the advancement of high-throughput technologies, researchers are capable of characterizing entire genomes, as well as simultaneously assess the expression activity of every gene within them. Consequently, the field of biology is becoming more data-driven than ever.

An important aspect of this age of genomics is the comprehension of the transcriptome, the summation of all RNA transcripts within a given cell. While the genome of a given organism represents the catalog of molecular information, how that information is utilized by said organism can be just as, if not more, important. The individual genes within a genome can be expressed via transcription at different rates, and these rates can be up- or down-regulated depending on the stimuli placed upon a cellular system. By being able to quantify gene expression and compare the relative abundance of particular mRNA transcripts, investigators can identify individual genes of interest, allowing them to examine the functionality of the genome.

In this chapter, we introduce the basic biology of gene expression and the importance of its quantification, the recent technologies that have been developed to quantify expression, as well as the statistical concepts, hurdles, and methods pertinent to these tools and the research presented in this dissertation.

1.1 Gene Expression

The “genetic code” of DNA is composed of four chemical building blocks, referred to as nucleotides: adenosine (A), guanine (G), thymine (T), and cytosine (C). These nucleotides in
Figure 1.1 Example of a partial DNA sequence (above) with its complementary strand (below). DNA sequences are traditionally read from the 5’ to the 3’ end.

turn compose polymers, or chains of molecules, which are the essential components of genetic information. The entirety of the genetic information of an organism is contained within its genome, which consists of one or more large molecules of double stranded DNA. The polymeric nature of DNA can also be considered from an analytic standpoint as a finite directional string of characters representing each nucleotide, known as a DNA sequence (Figure 1.1). This sequence can in turn be reduced to multiple subsequences, known as genes, which (most often) contain the coding information for a specific gene product. Gene content may vary greatly from one organism to the next; the total number of annotated genes within a genome ranges from approximately 500 for simple bacteria to over 30,000 for complex animals (Cristianini and Hahn, 2007).

Genetic polymers are often referred to by their lengths, using either the terminology “bp” for base-pairs or “nt” for nucleotides. A large area of computational genomics involves the sequence characterization of the genome, referred to as “sequencing,” with the sequencing of the first simple genomes occurring in the early 1980’s (Sanger et al., 1982) and the human genome being sequenced in 2001 (Venter et al.; International Human Genome Sequencing Consortium). Each individual nucleotide within a sequence is referred to as a base, with most genes consisting of 1-200 kilobases (kb).

While the genome contains nearly all of the genetic information of an organism, how that information is used is defined through the processes of transcription and translation. Individual genes are expressed through transcription, whereby RNA polymerase transcribes complementary ribonucleic acid (RNA) from the DNA template, yielding what is known as a messenger RNA (mRNA) transcript. These single-stranded transcripts exist within the cytoplasm of a
cell, waiting to be converted into proteins through the process of translation.

In eukaryotic organisms, genes themselves are composed of intronic and exonic sequences, also referred to as “noncoding” and “coding” sequences, respectively. For these organisms, once the mRNA transcript is transcribed from a DNA template, it undergoes splicing, whereby introns are removed from the transcript, yielding the final mRNA product. While this phenomenon increases the versatility of a gene (via alternative splicing), it also makes the process of matching transcripts to their respective DNA templates within the genome much more difficult since the transcripts are not verbatim complementary sequences of a gene.

Due to the direct relationship between mRNA and gene products via translation, the abundance of transcripts within a cell is of particular interest in determining the level of activity of individual genes as well as inferring their overall function.

More thorough descriptions of these complex molecular processes as well as other relevant topics in molecular biology can be found in Gibson and Muse (2004).

1.1.1 Microarray Technology

The origin of modern DNA microarray chips, such as the GeneChip® manufactured by Affymetrix (Lockhart et al., 1996) (www.Affymetrix.com), began with the development of the laboratory technique referred to as “Southern blotting” (Maskos and Southern, 1992). Southern blotting involves attaching sampled single-stranded DNA fragments (known as targets) to some type of substrate, which is then washed with a probe DNA molecule to determine the presence of a particular sequence within a sample. This probe molecule is tagged with some type of marker, often a fluorescent dye, for purposes of detection. All types of arrays which involve genetic probes take advantage of the biochemical process of hybridization, whereby complementary DNA polymers bind to each other via hydrogen bonding between complementary nucleotide pairings.

In a similar fashion, microarrays involve attaching known probe molecules to a solid substrate, which is often referred to as a “chip.” The chemical nature of the probes can range over a multitude of possibilities, including oligonucleotides (also referred to as oligos) (Lockhart et al., 1996), antibodies (Rivas et al., 2008), proteins (MacBeath and Schreiber, 2000), and
microRNA’s (Shingara et al., 2005). Oligos are relatively short (≤50 bases) polymers of nucleotides. Oligonucleotide microarrays, such as the aforementioned GeneChip®, are one of the most commonly used chips in the context of gene expression, and will be discussed exclusively throughout the rest of this section.

Affymetrix GeneChip® microarrays are fabricated by attaching oligos 25 bases in length to a high-density glass quartz assay chip (Lipshutz et al., 1999). This is accomplished via in situ synthesis of the oligos using photolithography (Pease et al., 1994). These oligos are grouped in probe sets, which are all related by a specific sequence of interest. For a given probe set, 11-20 pairs of oligos are typically used, each pair consisting of a perfect match (PM) probe and a mismatch (MM) probe. The PM probe is derived directly from an exonic DNA sequence, whereas the MM probe is identical to the PM probe except that the middle (13th) nucleotide is altered for the purpose of quantifying non-specific binding. A given chip can contain thousands of probe sets, and multiple probe sets can represent a single physical gene. For general purposes, however, the terms “probe set” and “gene” are often used interchangeably in reference to the final expression measures obtained in microarray analysis.

The preparation protocol for microarrays involves a multistage process which yields target molecules of complementary RNA (cRNA). Total RNA, the sum of all RNA within a cell, is extracted from a cultured sample of multiple cells. The mRNA transcripts are then isolated by targeting the poly(A) tail, a terminal subsequence found on the 3’ end of all mRNA transcripts. Complementary DNA (cDNA) is generated via reverse transcription of the mRNA, which is rendered into double stranded cDNA. Finally, cRNA is reverse transcribed from the cDNA via in vitro transcription, resulting in the final target molecule product (Lipshutz et al., 1999). This process typically involves a 30-100 fold linear amplification of the original mRNA sample (Wodicka et al., 1997). For the purposes of detection, target cRNA sequences are labelled with a biotin-conjugated nucleotide (Chee et al., 1996).

Once a target sample has been prepared, the appropriate chip is bathed with the target solution and allowed to incubate in order to facilitate hybridization. Visualization of hybridization is attained by subsequent staining of the chip with a fluorescent chemical conjugate, which can then be detected by combined use of laser excitation of the chip and a fluorescent micro-
scope (Müller and Nicolau, 2005). A large presence of a target transcript will result in a higher number of probes being hybridized, and subsequently a higher light emission value under the microscope. Thus, this fluorescence signal serves as an analog metric for quantifying the relative abundance of transcripts in a sample, and the entire image obtained via this process is used for statistical analysis.

Microarrays are used for a variety of comparative research purposes, including the identification of genetic disease profiles (Tung et al., 2001; Golub et al., 1999), tracking the genetic responses to certain pharmaceutical treatments (De Backer et al., 2001), and investigating the developmental stages of gene expression (White et al., 1999; Altmann et al., 2001).

1.1.2 Next-Generation Sequencing and RNA-Seq

In contrast to target-and-probe based approaches used by DNA microarrays, sequencing-based methods of gene expression analysis, notably RNA-Seq (Wang et al., 2009), operate by directly determining the sequences of the mRNA contained within the sample. The protocol for RNA-Seq typically involves obtaining a purified sample of RNA, shearing the transcripts to fragments of a prescribed length (ranging from 20-400 bases (Wang et al., 2009)), generating a cDNA library through reverse transcription of the sample, and sequencing the fragments using some high-throughput platform (Oshlack et al., 2010). Notable options for sequencing include Illumina (Solexa) (Morin et al., 2008), Applied Biosystems SOLiD™ (Cloonan et al., 2008), and Roche 454 pyrosequencing (Emrich et al., 2007). Once a catalog of sequenced fragments, known as reads, has been created, the final stage of RNA-Seq is mapping, whereby the fragments are matched to specific locations on a reference genome. This process ultimately leads to tabulated counts representing the quantity of reads within a given sample mapped to individual genes.

The phrase “next-generation sequencing” (NGS) is in fact a general term for a variety of recently developed high-throughput technology platforms that rapidly sequence DNA in a massively parallel fashion. Traditional DNA sequencing, often referred to as Sanger sequencing, involves the dye-terminator method of determining sequence information (Sanger et al., 1977). While this method is very accurate, it is an expensive and time-consuming process (Schuster,
With growing interest in genomic data, cheaper and faster methods of sequencing have become necessary to keep up with demand, and massively parallel approaches were soon developed (Brenner et al., 2000). In contrast to Sanger sequencing, which can reliably identify 300-1000 bp sequences per reaction, most NGS applications are limited to 400-500 bp sequences per reaction but run many reactions simultaneously. For example, ABI SOLiD™ sequencing involves billions of simultaneous parallel reactions, ranging in length of 50-100 bp, and can yield roughly 20 gigabases of usable data per complete run (Chen et al., 2011). Full protocol descriptions of the leading NGS platforms can be found in Ansorge (2009).

Once a target sample has been fully sequenced, it requires alignment to a reference genome. This is an algorithmic process that involves pairing each individual read with a known exonic sequence within a given genome. Multiple mapping algorithms are available (Smith et al., 2008; Li et al., 2008b,a), each with various options and approaches toward the challenges presented by mass short-read alignment. Once full alignment has been accomplished, the mappings are reduced to counts attributable to each specific gene within the reference genome.

NGS applications of expression data are largely thought to be advantageous over their microarray counterparts. There is little to no background signal, no upper limit to the quantification of expression, and it exhibits very high technical and biological reproducibility (Nagalakshmi et al., 2008). RNA-Seq is also considered to be more versatile than DNA microarrays, since they do not rely upon the preconstruction of an array chip for the purposes of expression measurement. This makes RNA-Seq applications particularly relevant for non-model organisms (Vera et al., 2008). The direct sequencing of the target sample also allows for the discovery of unique isoforms, sequence variation (such as SNPs), and provides single-base resolution of transcribed genomic regions.

That is not to say that RNA-Seq is without its shortcomings. One criticism with this approach is that it relies upon the necessity to fractionate mRNA (or the resultant cDNA) to be compatible with the shorter sequencing length requirements for NGS. The fragmentation process often leads to biases for particular parts of a transcript, and the shorter reads render the mapping of a read equally well to multiple locations on a reference genome (“multireads”) much more probable (Wang et al., 2009).
1.2 Analyzing Expression Data

The analysis of gene expression data can be separated into two distinct stages: preprocessing and statistical inference. In this section we briefly discuss these two components to gene expression analysis for microarray data in the context of a two-treatment comparative study for purposes of exploratory detection of differential expression. We also mention preprocessing methods for RNA-Seq data and analysis options for this data type. For a more thorough discussion of analysis procedures for microarray data beyond the scope of this dissertation, such as cluster analysis, linear model approaches, and two-dye applications, we refer the reader to Speed (2003).

1.2.1 Microarray Data Analysis

Raw microarray probe-level fluorescence values are generally stored in matrix form in an output file referred to as a CEL file. These raw values undergo preprocessing in order to properly translate the fluorescence values into usable gene-wise expression measurements. This process is an exercise in dimensional reduction, distilling all the individual probe set values into singular summary values. Affymetrix provides its own proprietary analysis software, GeneChip® Operating Software (GCOS) (Affymetrix, 2004), which uses the MAS 5.0 algorithm for purposes of preprocessing. Despite this offering, a variety of approaches have been independently developed to analyze GeneChip® data. All these algorithms operate based on four main steps: background fluorescence correction, data normalization for comparable measurements across chip sets (Åstrand, 2003; Li and Wong, 2001), PM-MM correction, and final summary expression value computation.

Current preprocessing algorithms fall into two general categories based upon their treatment of the MM probe data: those which use MM probe data as a correction for non-specific binding (such as MAS 5.0) and those that do not (such as robust multi-array). While MM probes were originally designed to capture data about such false binding, research has indicated that the subtraction of the MM probe fluorescence values from the PM values may be inappropriate for this type of correction by inflating probe-wise variances (Naef et al., 2002; Irizarry et al.,
As such, many current algorithms aside from MAS 5.0 do not make use of MM values at all with respect to non-specific binding. Robust multi-array (RMA) normalization, proposed by Irizarry (2003b) ignores MM probe data in this respect, and is the preprocessing algorithm used in Chapters 2 and 3.

Probe sets are summarised by a single expression value to represent the assigned probe, often a robust average of the normalized values such as the Tukey biweight (MAS 5.0). This reduces the total amount of data to an $m \times k$ matrix, where $m$ is the total number of probes and $k$ is the total number of observed expression profiles.

One of the basic statistical applications of inference for exploratory analysis of microarray data is the detection of gene-wise differential expression (Efron et al., 2001; Storey and Tibshirani, 2003). Typical analysis protocol for the detection of differential expression is the application of individual statistical hypothesis tests for each measured gene, such as a Student’s two-sample $t$-test. These tests are generally summarized by their resultant $p$-values, and multiple testing procedures are applied in the declaration of significance for differential expression.

### 1.2.2 Analysis of RNA-Seq data

Due to its technological novelty, the statistical methods for RNA-Seq are in their infancy. Preprocessing protocols for RNA-Seq measurements are quite different from microarray data approaches, since the data generating mechanisms are so vastly dissimilar. Raw RNA-Seq data is usually considered to be the alignments of the reads. Common stages of preprocessing procedures include dealing with multireads (Mortazavi et al., 2008; Cloonan et al., 2008), appropriately mapping sequences that span splice junctions (Trapnell et al., 2010), accounting for read location biases (Li et al., 2010), and the optional use of normalization techniques such as the reads per kilobase per million mapped reads (RPKM) transformation (Mortazavi et al., 2008). Other methods eschew normalization algorithms by directly modelling differences in sequence depth across expression profiles through the use of exposure values, particularly models that make use of count-based distributions such as the Poisson or negative binomial (Wang et al., 2010; Anders, 2010).
1.3 Statistical Challenges of Gene Expression Analysis

The research in this dissertation addresses a number of extant challenges that present themselves in the analysis of expression data. Of particular interest are the implications of the typically small sample sizes found in gene expression data analysis and complications due to dependence across genes. In this section we briefly describe both issues and their impact on the analysis of expression data.

1.3.1 Dependence

Many statistical methods which are applied to gene expression data make some independence assumptions across genes. These assumptions often lead to convenient mathematical forms, simplifying the estimation of quantities of interest such as the proportion of equivalently expressed genes (Benjamini and Hochberg, 1995; Storey, 2002; Nettleton et al., 2006). However, genes themselves are intricately related to each other through regulatory networks (Altman and Raychaudhuri, 2001; Wyrick and Young, 2002), yielding correlated expression measurements. This translates to the hypothesis test statistics themselves being correlated, violating the assumption of independence.

This issue has been addressed a variety of ways in research, including presenting robust theoretical properties of methods to gene-wise dependence (Benjamini and Yekutieli, 2001; Storey et al., 2004), modelling the impact of dependence on the distribution of null test statistics (Efron, 2004, 2007), and using kernel methods to account for dependence at the observation level of the data (Leek and Storey, 2008). We examine this issue directly in Chapter 3, exposing the impacts on estimators when biologically-derived covariance structures are preserved in simulation studies.

1.3.2 Small Samples

High-throughput expression profiles are often limited in respect to sample size due to the extreme costs involved in obtaining the measurements (Lee et al., 2000; Qu et al., 2010). This restriction has led to the development of specialized methods which enable the sharing of
information across genes. For example, accurate estimates of gene-wise variance are difficult to attain with small samples, and there are a number of approaches towards the stabilization of these variance estimates (Tusher et al., 2001; Smyth, 2004; Tong and Wang, 2007).

Another concern is that parametric tests for evaluating the difference between expression means may be sensitive to the violation of distributional assumptions, such as asymptotic results derived from the central limit theorem. As such, a variety of nonparametric gene-wise tests for differential expression (Gadbury et al., 2003; Breitling and Herzyk, 2005) have been proposed. We examine this concern in the research conducted in Chapter 2.

### 1.4 Relevant Statistical Methods

In this section, we review the background of some individual statistical concepts that are pertinent to the research discussed in following three chapters, and discuss their relevance to gene expression analysis.

#### 1.4.1 Multiple Testing and FDR

The high-throughput technology of microarrays results in the simultaneous measurement of gene expression across thousands of genes. In the context of comparative analysis across two conditions, this results in a testing multiplicity, where each gene represents an individual hypothesis test. Traditionally, singular hypothesis testing is conducted under the control of the Type I error rate, the probability that the hypothesis is declared erroneously non-null, at some prescribed level $\alpha$. However, when multiple tests are conducted simultaneously, the probability that a Type I error is committed increases with the additional number of tests, and multiple testing procedures are necessary to correct for this occurrence and preserve the control of Type I errors. The classical adaptation of Type I error rate control for multiple testing situations is through procedures designed around the family-wise error rate (FWER), which is defined as the probability of committing at least one Type I error across all tests conducted. These procedures typically involve corrections of $\alpha$, such as the Boneferroni method (Bland and Altman, 1995) and the approach proposed by Holm (Holm, 1979). However, if the number of tests is extremely large, these corrective procedures can severely limit statistical power and make declaration of
significance highly improbable at any reasonable level of $\alpha$ (Benjamini and Hochberg, 1995). For experiments like exploratory microarray analysis, the number of hypotheses is so high that FWER procedures are too conservative for differential expression detection.

Consider a two-treatment microarray experiment involving $m$ genes where each gene is summarized by a $p$-value from a two-sample $t$-test. If we select a significance threshold $t \in (0, 1]$ such that all hypotheses with corresponding $p$-values less than or equal to $t$ are rejected, we obtain results that can be summarized in a manner depicted in Table 1.1. The quantity $V$ represents the total number of Type I errors committed, which we can alternatively refer to as “false discoveries.” $R$ represents the total number of rejected hypotheses, or “discoveries.” Under this notation, we can define the FWER as the probability $P(V \geq 1)$. We can also define $Q = V/R$ as the proportion of discoveries which are false discoveries. Predicated on the work by Spjøtvoll (1972) and Soríc (1989), Benjamini and Hochberg (1995) proposed the “false discovery rate” (FDR), which is defined as:

$$FDR = E\left[\frac{V}{\max\{R,1\}}\right] = E\left[\frac{V}{R} \mid R > 0\right] P(R > 0) \quad (1.1)$$

FDR is an alternative approach to addressing the issue of test multiplicity by allowing one to commit multiple Type I errors during significance classification. In Benjamini and Hochberg, the authors defined an algorithm which they proved would control FDR at a specific rate $\alpha$ under the assumption the null hypothesis $p$-values were independent and identically distributed as Uniform(0,1):

1. Given a set of $p$-values corresponding to $m$ hypothesis tests, $p_1, \ldots, p_m$, let $p_{(1)}, \ldots, p_{(m)}$ represent the ordered $p$-values such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
2. Let \( k = \max \{ p(i) \leq \alpha \frac{i}{m} : i = 1, \ldots, m \} \)

3. If \( k \) exists, then rejection of all hypotheses with \( p \)-values \( \leq p(k) \) controls FDR at rate \( \alpha \)

Commonly known as the Benjamini and Hochberg procedure, this algorithm is known to be quite conservative, since it implicitly assumes that all of the hypotheses are null to maintain strong control of FDR. The algorithm later served as the template for the data-driven adaptive point estimator of FDR, \( \hat{\text{FDR}}(t) \), by Storey (2002), which involved the inclusion of an estimate of the true number of null hypotheses, \( m_0 \). This adaptive procedure is also sometimes written with respect to the proportion of true null hypotheses, \( \pi_0 = \frac{m_0}{m} \). The estimator is defined as:

\[
\hat{\text{FDR}}(t) = \frac{\hat{m}_0 t}{\max \{ R(t), 1 \}}.
\]

The adaptive estimator of Storey has given rise to an entire family of estimators of FDR which differ in their approach to the estimation of \( m_0 \) (Storey and Tibshirani, 2003; Pounds and Cheng, 2006; Nettleton et al., 2006). Estimators of \( m_0 \) (or alternatively \( \pi_0 \)) generally operate on a histogram-based approach by binning the resultant \( p \)-values of a microarray experiment and making assumptions of uniformity in order to approximate the quantity of null hypotheses.

In a similar vein of work, Storey (2002; 2003) also proposed the concept of the “q-value”, an analog to the \( p \)-value which represents the minimum FDR at which the corresponding test is declared significant.

\[
\text{q-value}(p_i) = \min_{t \geq p_i} \hat{\text{FDR}}(t)
\]

These \( q \)-values are a convenient \( p \)-value transformation that makes data-based selection of significance threshold using FDR simple and intuitive.

Alternative approaches to FDR, such as the local false discovery rate, as well as approaches that attempt to directly deal with dependent hypothesis tests (Efron, 2004; Leek and Storey, 2008) have also been developed.
1.4.2 Permutation Testing

Of relevance to Chapter 2, permutation testing involves the process of permuting the observed data to generate a reference distribution of a specific test statistic for the purposes of testing a hypothesis. This statistical test was first mentioned by R.A. Fisher in his seminal work *The Design of Experiments* (8th edition published in 1971), in which he examined Charles Darwin’s plant height data from his analysis of *Zea mays*, testing the difference in means of two groups. Fisher wrote down all possible permutations of the data by hand, dividing them into groups of the original sample sizes and calculating a mean-difference statistic for each permutation. Modern applications of permutation tests follow this approach, enumerating all possible permutations of the observed data and generating the same statistic each time. A hypothesis test $p$-value can then be calculated by dividing the number of permutation statistics as or more extreme than the original statistic by the total number of permutations. These $p$-values are discrete in nature, and their cardinality is limited by the total number of possible permutations of a given set of observations. Since the test assumes that these observations have the exact joint distribution regardless of their random assignment, this approach is valid as long as the observations are exchangeable under the null hypothesis.

Nonparametric testing is a well established approach in the analysis of expression data (Storey and Tibshirani, 2001; Dudoit et al., 2002; Gadbury et al., 2003; Xu and Li, 2003), citing the non-normality of microarray data (Thomas et al., 2001).

For purposes of illustration, let us assume a microarray experiment with two treatment groups, each of sample size $n$. Denote an expression measurement for the $i^{th}$ gene, $j^{th}$ treatment group, and $k^{th}$ replication as $Y_{ijk}$, and define the respective group means for a given gene as $\bar{Y}_{i1}$ and $\bar{Y}_{i2}$. A statistic as simple as $|\bar{Y}_{i2} - \bar{Y}_{i1}|$ can be used to generate the test reference distribution for this type of analysis. There are $\binom{2n}{n}/2 = S$ unique permutations of the data, which we can index using $s$, such that $s = 1, \ldots, S$. Then, if $T_{i,s}$ is the test statistic derived from the $s^{th}$ permutation, and $T_{i,1}$ is the original statistic, then the resultant $p$-value is simply:

$$p_i = \frac{\sum_{s=1}^{S} (I(T_{i,s} \geq T_{i,1}))}{S}$$ (1.4)
where $I(T_{i,s} \geq T_{i,1})$ is the indicator function that yields a value of 1 if $T_{i,s} \geq T_{i,1}$ and 0 otherwise.

Permutation test procedures are advantageous due to their lack of distributional assumptions, which is particularly useful for tests with small sample sizes. However, the enumeration of all possible permutations becomes computationally intensive with increase in sample size, and approximations using Monte Carlo approaches may become necessary. Similarly, if the sample sizes are relatively small, the total number of permutations may be too few, resulting in larger minimal $p$-values that do not afford enough statistical power.

### 1.4.3 Bayesian Latent Class Models

Given that RNA-Seq data represents the direct sequencing of mRNA transcripts, the question of whether or not a gene is actually being transcribed is a topic of interest when the associated number of reads is near 0. Since the transcriptional status of a gene is not a directly observable property, it can be considered a dichotomous latent variable. Latency in this context refers to the fact that is variable can not be directly measured but must rather be inferred in some manner from the available data. The reads themselves then exhibit *incompleteness* with respect to their transcriptional status. We review basic background information on bayesian approaches to latent class (LC) models, specifically data-augmentation approaches discussed by Tanner and Wong (1987) which are applied in Chapter 4.

The frequentist solution to LC model analysis is often the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). An example itself of data-augmentation, the EM algorithm is a two-state iterative process that imputes latent variables via calculating their expectations given a set of model parameters (E-step), and then recalculates point estimates of the remaining model parameters using maximum likelihood conditional on the state of the latent variables (M-step). Dempster et al. provide a multitude of detailed applications of this algorithm in their publication. When a problem is sufficiently complex, however, the use of maximum likelihood may be untenable, and alternative approaches to inference, such as Bayesian approaches, are necessary.

Bayesian statistical inference is predicated on Bayes’ Theorem (Hoff, 2009), which states for given events $A$ and $B$, subject to $p(B) \neq 0$: 
\[ p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \] 

(1.5)

This theorem serves as the foundation for the Bayesian interpretation of probability. Let us rephrase Bayes’ Theorem in the context of a simple parametric statistical model, which we define by \( \theta \). We can then consider \( p(\theta) \) to represent our prior uncertainty regarding the model, often referred to as a prior distribution on \( \theta \). Data, \( y \), is collected, and \( p(y \mid \theta) \), the sampling distribution, represents the likelihood that the observations were an outcome of model \( \theta \). We then use \( y \) to update our knowledge of \( \theta \) by calculating the posterior distribution on \( \theta \), \( p(\theta \mid y) \), as follows:

\[ p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_\theta p(y \mid \theta)p(\theta)d\theta} \] 

(1.6)

The denominator of (1.6) is often referred to as a normalizing constant, since it renders \( p(\theta \mid y) \) as a proper probability. However, in complex statistical problems, this quantity is difficult to calculate since it usually involves difficult integration. In these instances, a popular technique in Bayesian modelling is the use of Markov chain Monte Carlo (MCMC). Bayesian MCMC methods differ from their frequentist counterparts in that they estimate parameters by sampling from their posterior density rather than providing a point estimate (Gelman et al., 2003).

In the context of LC modelling, MCMC applications for data-augmentation can follow a similar two-stage algorithmic approach as the EM algorithm. Consider a model with observed data \( y \) and latent data, \( z \), and an assumed data model \( p(y \mid \theta) \). If both \( y \) and \( z \) are observed, then the posterior distribution \( p(\theta \mid y, z) \) is easily defined. However, since \( z \) is missing, the desired posterior distribution \( p(\theta \mid y) \) may prove difficult to obtain. As Tanner and Wong (1987) point out, there is a mutual dependency between distributions \( p(\theta \mid y) \) and \( p(z \mid y) \). To address this issue, they suggest that an analog to the EM algorithm be used, involving an imputation stage (I-step) where samples are drawn from \( p(z \mid \theta, y) \), and a posterior draw stage (P-step) where samples are drawn from \( p(\theta \mid z, y) \).

A good example of an application of this approach is presented by Chung et al. (2006), and additional details on data-augmentation approaches can be found in van Dyk (2001).
1.5 Thesis Organization

In Chapter 2 we modify an existing semi-parametric method of estimating the proportion of equivalently expressed genes to account for the impact of small sample sizes on distributional assumptions. We propose the use of permutation testing as opposed to the traditional t-test, making adjustments to the original algorithm to account for the lack of a distributional basis. We show that this provides a marked improvement over the use of the semi-parametric approach and discuss the appropriateness of its application.

In Chapter 3, we consider the implications of using a standard quantity of genes to determine the significance threshold for differential expression using p-values derived from a microarray experiment with two treatment groups. Of interest in this chapter is the impact of this approach on the estimation of the false discovery rate (FDR). We explore this issue by simulating microarray data using information from actual experiments and evaluate the performance of a multitude of estimators of FDR under a variety of conditions. We uncover a negatively correlated relationship between the estimates and the true values, and offer our thoughts on alternative approaches in light of this phenomenon.

Chapter 4 involves the development of a statistical approach for simultaneously analyzing the transcriptional activity of genes using RNA-Seq data. We examine this next-generation sequencing data in the context of a breeding experiment designed to evaluate the complementation theory of heterosis, and apply a data augmented MCMC approach to fitting a negative binomial model to the data. We apply our model to real experimental data and use simulations to evaluate the adequacy of the model.

Concluding remarks and discussion of the future direction of the research contained within this dissertation can be found in Chapter 5.

Bibliography


to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98:5116–
5121.


Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith,
H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew,
Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C.,
Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts,
R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo,
D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry,
C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V.,
Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z.,
Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielson, A. E., Gan, W., Ge,
W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum,
K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina,
Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H.,
Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter,
D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A.,
Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K.,
C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F.,
Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen,
I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi,


CHAPTER 2. ESTIMATING THE PROPORTION OF TRUE NULL HYPOTHESES WHEN CONDUCTING MANY PERMUTATION TESTS WITH NON-NORMAL LOW-SAMPLE-SIZE DATASETS

Nick Larson and Dan Nettleton

Abstract

When conducting a large number of simultaneous hypothesis tests, the proportion of true null tests, $\pi_0$, is of interest. Recently, a semiparametric method for estimating both $\pi_0$ and the distribution of noncentrality parameters associated with each non-null hypothesis test has been proposed for use with multiple $p$-values from $t$-tests. We adapt this procedure for use with permutation $p$-values and show that our nonparametric adaptation is much more robust to non-normality than the original semiparametric approach. Simulation results and real data analysis illustrate the advantages of our nonparametric procedure.

2.1 Introduction

Microarrays allow researchers to simultaneously assess the expression activity of thousands of genes. One of the more common analysis procedures applied to microarray data is differential expression detection, in which the expression profiles of two treatment groups are compared to test whether genes are equivalently expressed between treatments. This exploratory analysis is usually conducted by examining each individual gene using a statistical test to evaluate a possible difference in expression distributions, with a natural option being the two-sample $t$-test.

One of the principal assumptions of a $t$-test is the normality of the observations; however, it has been shown that microarray expression data are not normally distributed, regardless
of choice of pre-processing methodology (Thomas et al., 2001; Giles and Kipling, 2003; Liu et al., 2003; Hardin and Wilson, 2009). While the $t$-test is typically robust to the violation of the normality assumption when sample sizes are reasonably large (Boneau, 1960; Lumley et al., 2002), the prohibitive cost associated with microarray chips frequently results in small numbers of replications (Lee et al., 2000; Qu et al., 2010). This calls into question the protection against the violation of the normality assumption afforded through the asymptotic results of the central limit theorem. These small sample sizes also make it difficult to verify whether or not the observations are normally distributed. Such parametric approaches may then result in inaccurate $p$-values and be inappropriate for the typically small sample sizes of microarray experiments.

A popular nonparametric approach for testing the difference in expression distributions between two groups is the permutation test. Some of the very first statistical methods for detecting differential expression were in fact designed around permutation testing (Dudoit et al., 2002; Storey and Tibshirani, 2001). Under the null hypothesis that there is no distributional difference in expression levels between the two groups for a given gene, the observations for the given gene are exchangeable. A valid reference distribution for the test statistic can then be directly generated by recalculating the statistic for all permutations of the treatment group assignments to observations.

Using the permutation distribution as the reference distribution yields exact discrete $p$-values, where the number of support points in the permutation $p$-value distribution is determined by the total number of unique test statistic values that arise from permuted data. If we consider continuous observations from a balanced two-sample data set with sample sizes equal to $n$ in each treatment group, then the total number of unique test statistic values in the permutation distribution is $R = \frac{1}{2}(2n\choose n)$ for the test statistic we consider. Under the null hypothesis that both samples are independent draws from the same distribution, the permutation $p$-value will be equally likely to be any of the values in the discrete support set $\{1/R, 2/R, \ldots, R/R\}$.

While conducting an experiment with a large number of simultaneous hypothesis tests, the proportion of true null hypotheses, $\pi_0$, is often a quantity of interest, especially for estimating the false discovery rate (FDR). This is particularly true in exploratory microarray experiments,
where researchers wish to pare down possibly thousands of genes to a few genes of interest, while keeping the expected proportion of false discoveries acceptably low. In this context, we can define the marginal probability mass function of a permutation $p$-value by the mixture distribution

$$f(p) = \pi_0 f_{EE}(p) + (1 - \pi_0) f_{DE}(p), \quad p = \frac{1}{R}, \frac{2}{R}, \ldots, 1 \tag{2.1}$$

where the subscripts EE and DE denote the probability mass functions for equivalently expressed (null) and differentially expressed (alternative) genes. It then follows that $f_{EE}$ is a discrete uniform distribution and $f_{DE}$ is some unknown distribution for permutation $p$-values where the alternative hypothesis is true. Thus, we can simplify (2.1) to

$$f(p) = \pi_0 / R + (1 - \pi_0) f_{DE}(p), \quad p = \frac{1}{R}, \frac{2}{R}, \ldots, 1. \tag{2.2}$$

The quantity $\pi_0$ from (2.2) is not identifiable unless we make some assumptions regarding $f_{DE}$ (Langaas et al., 2005). For example, if we assume that $f_{DE}(1) = 0$, then an unbiased estimator of $\pi_0$ can be obtained as follows. Let $m$ be the number of hypothesis tests, and let $Y_r$ represent the number of permutation $p$-values that are equal to $r/R$ for $r = 1, \ldots, R$. Then, $E[Y_r] = m\pi_0 / R$, and an unbiased estimator for $\pi_0$ is given by

$$\hat{\pi}_0 = \frac{Y_R R}{m}. \tag{2.3}$$

This estimator is a discrete-case analog of the $\pi_0$ estimator proposed by Storey (2002).

Note that if $f_{DE}(1) > 0$, then $E[\hat{\pi}_0] = \pi_0 + (1 - \pi_0) f_{DE}(1) > \pi_0$, so that $\hat{\pi}_0$ is a conservatively biased estimator of $\pi_0$. If there are large differences between expression distributions between treatments for all DE genes, each permutation test will have high power for detecting differential expression, and the probability of a large permutation $p$-values will be small. Thus, $f_{DE}(1)$ will be small, and the bias in $\hat{\pi}_0$ will be negligible. However, in many microarray experiments, there are only subtle differences in expression for many DE genes. In such cases $f_{DE}(1)$ may be substantial, leading to non-negligible bias for $\hat{\pi}_0$. Thus, if we can estimate $f_{DE}(1)$, it may be possible to improve estimation of $\pi_0$. 
More generally, even better estimates of $\pi_0$ can be constructed from $Y_1, \ldots, Y_R$ if we can estimate, for $r = 1, \ldots, R$, the expected contribution to $Y_r$ from DE genes. Note that

$$E[Y_r] = m\pi_0/R + m(1 - \pi_0)f_{DE}(r/R)$$

for all $r = 1, \ldots, R$. Solving for $\pi_0$ yields

$$\pi_0 = \frac{E[Y_r/m] - f_{DE}(r/R)}{1/R - f_{DE}(r/R)}.$$  \hspace{1cm} (2.5)

Replacing $E[Y_r/m]$ with its observed value $y_r/m$ and replacing $f_{DE}(r/R)$ with its estimate would provide estimates of $\pi_0$ for all $r = 1, \ldots, R$, which could then be combined to produce an estimate of $\pi_0$ that makes more use of the data than the simple estimate of $\pi_0$ given in (2.3).

Although we do not explicitly follow such a strategy for estimating $\pi_0$, the key point here is that obtaining an estimator of $f_{DE}(r/R)$ for all $r = 1, \ldots, R$ may enable better estimation of $\pi_0$.

To obtain a framework for estimating $f_{DE}$, we assume that $f_{DE}$ can be adequately approximated by an infinite mixture of probability mass functions. Each probability mass function in the infinite mixture is the probability mass function of a permutation $p$-value for a test of distribution equality between two groups. These probability mass functions are indexed by a noncentrality parameter $\delta$ that summarizes the difference between the distributions corresponding to the two groups. We model the noncentrality parameters as draws from a distribution with unknown density $g$ with support on the interval $[0, \delta_{max}]$. Thus, we have

$$f_{DE}(p) \approx \int_0^{\delta_{max}} f_{DE}(p \mid \delta)g(\delta)d\delta,$$  \hspace{1cm} (2.6)

where $f_{DE}(p \mid \delta)$ denotes the conditional probability mass function of the permutation $p$-value given noncentrality parameter $\delta$.

Our objective in this paper is to use $p$-values $p_1, \ldots, p_m$, obtained from $m$ permutation tests, to simultaneously estimate $\pi_0$ and $g(\delta)$. We develop a method analogous to the semiparametric approach created by Ruppert et al. (2007) for the analysis of multiple $p$-values from parametric $t$-tests. We refer to our approach as a nonparametric estimator – in contrast to the semi-
parametric approach of Ruppert et al. (2007) – because our procedure requires no parametric assumptions for testing or estimation. In Section 2.2, we define the permutation test we apply in our nonparametric estimator of $\pi_0$ and describe our simulation approach to approximating $f_{DE}(p)$. We also design a simulation study which evaluates the performances of the semiparametric and nonparametric approaches simultaneously when the assumption of normality is not met in conjunction with small samples. In Section 2.3, we present the results of our simulations, which show our permutation test approach can be more accurate and exhibit less variance than the semiparametric procedure. In Section 2.4, we provide an application of each approach to a small sample microarray experiment, found in Hand et al. (2009), which measured the difference in gene expression between two groups of mice ($Mus musculus$). The example illustrates some of the complications in using the semiparametric approach, particularly its tendency to severely underestimate $\pi_0$. We further discuss the results of our simulation and our case study application in Section 2.5, outlining the benefits of the nonparametric method, and offer some concluding thoughts in Section 2.6.

2.2 Methods

2.2.1 Permutation Testing

For the purposes of illustration, consider the experimental design conditions of the microarray analysis described in Hand et al. (2009). Suppose we have two sets of observations, each of sample size $n = 5$, respectively denoted $y_{11}, \ldots, y_{15}$ and $y_{21}, \ldots, y_{25}$. Assume that $y_{11}, \ldots, y_{15}$ are independent and identically distributed (iid) $\phi_1$ and $y_{21}, \ldots, y_{25}$ are iid $\phi_2$. To test the null hypothesis $H_0 : \phi_1 = \phi_2$, we first calculate some test statistic, $T$, based upon the original observation groupings. Denote this value $T_1$. For our purposes, we define our test statistic to be

$$T = |\bar{y}_1 - \bar{y}_2|$$

(2.7)

where $\bar{y}_i = \frac{1}{5} \sum_{j=1}^{5} y_{ij}$. This statistic provides good power for detecting location differences between $\phi_1$ and $\phi_2$, which are departures from $H_0$ that are of primary interest when testing for
differential expression. It is straightforward to show that a standard two-sample $t$-statistic will result in the same permutation $p$-value as $T$. $T$ takes a total of $R = \binom{10}{5}/2 = 126$ unique values across all possible divisions of $y_{11}, \ldots, y_{15}, y_{21}, \ldots, y_{25}$ into two groups of 5 observations each. These values, denoted $T_1, T_2, \ldots, T_R$, serve as the reference distribution for the permutation test. To obtain a two-sided $p$-value for our observed test statistic $T_1$, we determine the proportion of permutation statistics generated which are as or more extreme than $T_1$, such that

$$p = \frac{\# \{T_r \geq T_1 : r = 1, \ldots, R\}}{R}. \quad (2.8)$$

Under the null hypothesis, this test leads to a discrete, uniformly distributed $p$-value that is exact, such that $P(p = r/R) = r/R$ for all $r = 1, \ldots, R$.

### 2.2.2 The Distribution of the Permutation $P$-value under the Alternative

If we specify a particular distribution $\phi_1$ and a particular distribution $\phi_2$ such that $\phi_1 \neq \phi_2$, it is straightforward to approximate the distribution of the permutation $p$-value to any desired degree of accuracy by simulation. By repeatedly generating a sample from $\phi_1$ and a sample from $\phi_2$ and computing a permutation $p$-value for the simulated data, we can obtain $\hat{f}^*(p \mid \phi_1, \phi_2)$ as an estimate of $f^*(p \mid \phi_1, \phi_2)$, the probability mass function for the permutation $p$-value under the particular departure from $H_0 : \phi_1 = \phi_2$ specified by the choice of $\phi_1 \neq \phi_2$. If we then view $(\phi_1, \phi_2)$ as random and specify a distribution over $\Omega = \{(\phi_1, \phi_2) : \phi_1 \neq \phi_2\}$, $f_{DE}(p) = E[f^*(p \mid \phi_1, \phi_2)]$.

Obviously, $\Omega$ is an extremely large space, and it is difficult to imagine specifying an appropriate distribution over $\Omega$ in the practical problem of detecting differential expression using microarray data. To make the problem tractable, we assume $f_{DE}(p)$ can be adequately approximated by considering only the subset $\Omega_N$ of $\Omega$ that involves pairs of normal distributions, where the members of each pair differ only in their means. Given that the permutation test statistic we consider is equivalent to a two-sample $t$-statistic, the discrepancy between $\phi_1$ and $\phi_2$ in $\Omega$ can be fully specified by the noncentrality parameter.
Figure 2.1 Approximations of the probability mass distributions of $f_{DE}(p \mid \delta)$ obtained via simulation for the case $n_1 = n_2 = 5$ when (a) $\delta = 1$ and (b) $\delta = 2$. Each line represents the probability mass associated with each possible $p$-value $r/R$.

$$\delta = \frac{|\mu_2 - \mu_1|}{\sigma \sqrt{2/n}},$$

where $\sigma^2$ denotes the variance common to $\phi_1$ and $\phi_2$ and $\mu_i$ and $n_i$ denote the mean and sample size, respectively, for $\phi_i$ ($i = 1, 2$). Thus, we can use $f_{DE}(p \mid \delta) = f^*_i(p \mid \phi_1, \phi_2)$ to denote the probability mass function of the permutation $p$-value for any $(\phi_1, \phi_2) \in \Omega_N$. We can approximate $f_{DE}(p \mid \delta)$ by repeatedly simulating standard normal data, adding $\delta \sqrt{\sigma^2 (1/n_1 + 1/n_2)}$ to the simulated observations from one treatment group, and computing a permutation $p$-value. Figure 2.2 shows approximations of $f_{DE}(p \mid \delta)$ for $\delta = 1$ and 2 and the case $n_1 = n_2 = 5$.

As introduced in (2.6), we assume that $f_{DE}(p)$ can be approximated by $E[f_{DE}(p \mid \delta)]$, where $\delta$ is treated as a random variable with density $g(\delta)$. Rather than attempting to estimate the infinite dimensional parameter $g$, we assume that $g(\delta)$ can be written as a linear spline function $g(\delta; \beta)$, where $\beta$ is a vector of regression coefficients for a set of B-splines basis densities. Define $K$ to be the number of equidistant knots on $[0, \delta_{max}]$, located at values $\kappa_1, \ldots, \kappa_K$, such that the distance between adjacent knots is $d = \delta_{max}/(K-1)$, $\kappa_1 = 0$ and $\kappa_K = \delta_{max}$. The first B-spline
Figure 2.2 Example linear B-spline basis with $\delta_{\text{max}} = 6$ and $K = 7$

basis density, $B_1$, is a linear decreasing spline from $2/d$ to 0 on the interval $[0, \kappa_2] = [\kappa_1, \kappa_2]$. For $k = 2, ..., K - 1$, the basis density $B_k$ increases linearly from 0 to $1/d$ on $[\kappa_{k-1}, \kappa_k]$ and decreases linearly from $1/d$ to 0 on $[\kappa_k, \kappa_{k+1}]$. A depiction of an example set of B-spline basis densities can be found in Figure 2.2. The distribution $g(\delta; \beta)$ can then be written as

$$g(\delta; \beta) = \sum_{k=1}^{K-1} \beta_k B_k(\delta)$$

(2.10)

under the constraints $\beta_k \geq 0$ for all $k$ and $\sum_{k=1}^{K-1} \beta_k = 1$. By replacing $g(\delta)$ with $g(\delta; \beta)$, we rewrite (2.6) as

$$f_{DE}(p) \approx \sum_{k=1}^{K-1} \beta_k \int_{0}^{\delta_{\text{max}}} f_{DE}(p \mid \delta) B_k(\delta) d\delta.$$

(2.11)
2.2.3 Estimation of the True Null Proportion

For \( k = 1, \ldots, K - 1 \), let \( z_{k+1}(r) = \int_{0}^{\delta_{\text{max}}} f_{DE}(r/R | \delta) B_k(\delta) d\delta \). In words, \( z_{k+1}(r) \) represents the probability that a permutation p-value is equal to \( r/R \), where \( r = 1, \ldots, R \), conditional on \( \delta \) from distribution density \( B_k \). We can use simulation to approximate \( z_{k+1}(r) \) for all \( k = 1, \ldots, K - 1 \) and \( r = 1, \ldots, R \). The simulation proceeds exactly as described for approximating \( f_{DE}(p | \delta) \) except that, instead of using a fixed value of \( \delta \), each permutation p-value is simulated using a different \( \delta \) drawn randomly from the distribution with density \( B_k \). We simulated 1,000,000 permutation p-values to approximate each function \( z_{k+1} \) to obtain the results presented in this paper. Exploratory analysis indicated that an appropriate maximum value for probability distribution of \( \delta \), \( g(\delta) \), is \( \delta_{\text{max}} = 6 \). Our simulations determined that for \( \delta > 6 \), \( f_{DE}(p = \frac{1}{R} | \delta) \approx 1 \), which renders p-value distributions corresponding to NCP values of \( \delta > 6 \) indiscernable from those with \( \delta = 6 \). For our purposes, we let \( K = 16 \) to allow for sufficient flexibility of the shape of \( g(\delta; \beta) \).

Now define \( z_1(r) = 1/R \) for \( r = 1, \ldots, R \) and let \( \theta = (\theta_1, \ldots, \theta_K)^T \), where \( \theta_1 = \pi_0 \) and \( \theta_{k+1} = (1 - \pi_0)\beta_k \) for \( k = 1, \ldots, K - 1 \). Then by (2.2), (2.11), and the definitions of \( \theta_k \) and \( z_k \), we have \( f(r/R) \approx \sum_{k=1}^{K} \theta_k z_k(r) \) for \( r = 1, \ldots, R \). A sum-of-squares approach toward estimating \( g \) through \( \theta \) would then be to minimize the quantity

\[
\sum_{r=1}^{R} \left( \frac{y_r}{m} - \sum_{k=0}^{K} z_k(r)\theta_k \right)^2
\]

under the linear constraints on \( \theta \). By fitting such a model, we simultaneously estimate the contributions of \( f_{EE}(p) \) and \( f_{DE}(p) \) to observations \( y_r \) for all \( p \in 1/R, \ldots, 1 \) through our characterization of \( g(\delta; \beta) \) in a manner similar to that described in (2.5).

To impose a roughness penalty in order to avoid “spiky” distribution estimates of \( g(\delta; \beta) \), we penalize deviations of the distribution \( g(\delta; \beta) \) from a linear function using a finite difference approximation to the second derivative of \( g \), defined as

\[
Q(\theta) = (2\theta_2 - 2\theta_3 + \theta_4)^2 + \sum_{k=3}^{K-2} (\theta_k - 2\theta_{k+1} + \theta_{k+2})^2.
\]
Define $z_k$ to be the $R \times 1$ column vector of estimated probabilities $z_k(r)$ for $r = 1, \ldots, R$ and $k = 2, \ldots, K$. Let $z_1 = (1/R)1$. Then, let $Z$ to be the $R \times K$ matrix with $k^{th}$ column $z_k$, and let $y = (y_1/m, \ldots, y_r/m)$. Estimation of $\theta$ is then calculated using a penalized sum of squares regression approach, where $SS(\theta; \tau)$ is defined as

\[ (y - Z\theta)^T(y - Z\theta) + \tau \theta^T(DA)^TDA\theta \]  

(2.14)

where $A = \text{diag}(0, 2, 1, \ldots, 1)$, $\tau$ is the penalization parameter, and $D$ is referred to as a “differencing” matrix, such that $D$ is a $(K - 3) \times K$ matrix with the $i^{th}$ row containing $+1$ in the $i + 1$ and $i + 3$ columns, $-2$ in column $i + 2$, and 0 everywhere else. The minimization of $SS(\theta; \tau)$ is equivalent to minimizing $f^T\theta + 0.5\theta^T H \theta$ where $f^T = -y^T Z$ and $H = Z^T Z + \tau A^T D^T DA$ under the constraints $\theta \geq 0$ and $1^T \theta = 1$. This form of the model can then be solved using a quadratic programming algorithm. Thus, applying this procedure to a vector of observed permutation $p$-values proportions $y$ then simultaneously estimates the density $g$ and $\pi_0$, where

$\hat{g}(\delta) = \sum_{k=2}^{K} \hat{\theta}_k (1 - \hat{\theta}_1) B_k(\delta)$ and $\hat{\pi}_0 = \hat{\theta}_1$.

### 2.2.4 Cross-Validation

Selection of the penalty parameter $\tau$ would typically be conducted under some form of cross-validation (CV), such as in ridge regression (Wahba and Wold, 1975). Leave-one-out cross-validation (LOOCV) requires multiple model fittings by removing each individual observation ($y_r$) and refitting the model such that

\[ \text{LOOCV}(\tau) = \frac{1}{R} \sum_{r=1}^{R} (y_r - \hat{y}_r^{-\tau}(\tau))^2 \]  

(2.15)

where $\hat{y}_r^{-\tau}(\tau)$ is the fitted value for $y_r$ under penalty parameter $\tau$ when observation $y_r$ is removed before fitting the model. Due to computational constraints, however, this approach is not feasible, and we use generalized cross-validation (GCV) (Craven and Wahba, 1979) instead of LOOCV. GCV can be calculated using only one model fit, such that

\[ \text{GCV}(\tau) = \frac{1}{R} \sum_{r=1}^{R} (y_r - \hat{y}_r(\tau))^2 }{(1 - \text{tr}H(\tau)/R)^2} \]  

(2.16)
where \( \text{tr} H(\tau) \) is the trace of the hat matrix of the unconstrained solution of using penalty parameter \( \tau \), such that \( \text{tr} H(\tau) = \) 

One complication with this approach is that the GCV statistic is undefined under the linear constraints imposed by our model, so we use results from the unconstrained form of the objective function. Thus, we approximate GCV(\( \tau \)) such that

\[
GCV(\lambda) \approx \frac{1}{R} \sum_{r=1}^{R} \left( y_r - Z \hat{\theta}(\tau) \right)^2 \left( 1 - \text{tr} H(\tau)/R \right)^2 \tag{2.17}
\]

where \( \hat{\theta}(\tau) \) is the constrained estimate of \( \theta \) using penalty parameter \( \tau \). This statistic is then evaluated over fits using a grid of possible values of \( \tau, \tau \), among which the \( \tau \) value that results in the lowest GCV(\( \tau \)) is used for the final estimation of \( \pi_0 \).

### 2.2.5 Simulation Design

To compare the performances of the semiparametric and nonparametric estimators of \( \pi_0 \), we need to define a method of simulating expression data where the status of DE genes is known. To accomplish this, we obtained an acute myeloid leukemia (AML) dataset from the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) (accession number GSE12417). Raw CEL files were normalized using robust multiarray average (RMA) normalization (Irizarry et al., 2003) with the affy library (Gautier et al., 2004) in R. We used a subset of expression profiles derived from bone marrow samples from 161 patients with muscular diseases (Metzeler et al., 2008). The samples were assessed using the Affymetrix (www.affymetrix.com) U133A GeneChip® platform, measuring the expression levels with 22283 probe sets. Since the expression profiles all correspond to the same treatment group, all genes can be assumed to be EE such that the expression values corresponding to a given gene are derived from the same expression distribution. The large number of expression profiles also provides reliable estimates of the gene-specific expression distributions, including values of the standard deviations.

To simulate microarray data, we generated a matrix of \( m_{\text{sim}} \times 2n \) values, \( S \), from the expression values in the AML dataset, where \( m_{\text{sim}} \leq 22283 \). A random gene, \( j \), is selected without replacement from the AML data to be represented. A total of \( 2n \) expression values are
then randomly selected from the values corresponding to gene $j$ to form the $i^{th}$ row of $S$, for $i = 1, \ldots, m_{\text{sim}}$. This data matrix can then be considered as set of $2n \leq 161$ expression profiles, where two hypothetical treatment groups are each represented by $n$ observations consisting of $m_{\text{sim}}$ simultaneous gene expression measurements.

Because each value for a given gene is assumed to be identically distributed, the null hypothesis of distributional equality is true for each row of the data matrix. We then induce differential expression by adding effect sizes to all of the observations associated with the second hypothetical treatment group for selected genes in the same manner described in Section 2.2.3. Let $d_i$ be a simulated NCP assigned to a given hypothesis, $H_i$. Then, for two samples, each of size $n$, the effect size $\epsilon_i = d_i \times \hat{\sigma}_j \times \sqrt{\frac{2}{n}}$ can be added to each observed value in the second treatment group to simulate a difference in population means, where $\hat{\sigma}_j$ is an estimate of standard deviation of the expression distribution corresponding to gene $j$, obtained from the entire data.

For our purposes, the $d_i$ values were sampled from a generalized beta distribution with shape parameters $(\alpha, \gamma)$ on the interval $[0, c]$. We developed three separate distributions for our simulation study, defined in name by the relative distance of the mode of the probability distribution from 0, defined as “Near” (N), “Moderate” (M), and “Far” (F) (Figure 2.3), in order to examine the impact of different effect-size distributions for DE genes.

The set of $p$-values for the semiparametric approach were generated using a two-sample $t$-test under the assumption of equal variance across groups, while the permutation $p$-values were generated from the procedure defined in Section 2.2.1. We also set $K = 16$ for the semiparametric procedure, and set the grid of possible penalty parameters $\tau = (1e-08, 1e-06, 1e-04, 1e-02, 1, 1e02, 1e04, 1e06, 1e08)$. The default value of $N_{\text{bin}} = 2000$ was used for the semiparametric estimator. Fitting the nonparametric method to the simulated data was conducted using the LowRankQP package (Ormerod and Wand, 2009) in R.

**2.3 Simulation Results**

We ran a total of six data simulations, each consisting of 100 replications, with two different true $\pi_0$ values and three total $g(\delta)$ distributions (Table 2.3). For all simulations, we set $n = 5$ and $m_{\text{sim}} = 10000$. We applied both the semiparametric and the nonparametric procedure to
Figure 2.3  Graph of the three NCP distributions used in simulation, with $N$ (black), $M$ (red) and $F$ (green) shown. Each of the three distributions use $c = 4$, and parameter values $(\alpha, \gamma)$ $N$ being (1,2), $M$ being (2,2) and $F$ being (4,2)
<table>
<thead>
<tr>
<th>Case</th>
<th>True $\pi_0$</th>
<th>$g(\delta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>0.70</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>0.70</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 2.1 Descriptions of each of the six data-based simulations used to evaluate both estimators of $\pi_0$

<table>
<thead>
<tr>
<th>$\pi_0$</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
<th>Case 5</th>
<th>Case 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g(\delta)$</td>
<td>Near</td>
<td>Moderate</td>
<td>Far</td>
<td>Near</td>
<td>Moderate</td>
<td>Far</td>
</tr>
<tr>
<td>RMSE(Semi)</td>
<td>0.1380</td>
<td>0.1846</td>
<td>0.1448</td>
<td>0.3631</td>
<td>0.3585</td>
<td>0.3837</td>
</tr>
<tr>
<td>RMSE(Non)</td>
<td>0.0213</td>
<td>0.0153</td>
<td>0.0194</td>
<td>0.0868</td>
<td>0.0465</td>
<td>0.0706</td>
</tr>
</tbody>
</table>

Table 2.2 RMSE results from the 6 data simulations involving varying values of $\pi_0$ and distributions $g(\delta)$

the data, resulting in two sets of 100 estimates of $\pi_0$ for each simulation.

Estimates of $\pi_0$ for each method were generated using $\hat{\theta}_1$. Boxplots of $\hat{\pi}_0$ for each simulation and method are in Figure 2.3 to visually compare the results. To numerically evaluate the efficacy of each estimator, we calculated Monte Carlo estimates of the root mean squared error (RMSE) for each method (Table 2.3).

The boxplots in Figure 2.4 indicate that, under our simulation conditions, the semiparametric approach underestimated the true value of $\pi_0$ on average regardless of effect size distribution and true null proportion. The semiparametric estimates also exhibited much larger variance than the respective nonparametric counterparts.

### 2.4 Case Study

We applied both the semiparametric estimator of $\pi_0$ as well as our nonparametric approach to the mouse microarray experiment previously mentioned in Section 2.2.1. The authors investigated the role of mature miRNA in the liver function. This was done by deriving mice with
Figure 2.4 Boxplots of the estimates of $\pi_0$ when the true null proportion is (a) 0.70 and (b) 0.95, with these values indicated by the horizontal dashed line in each plot. The simulation results are indicated by the estimator used (“Semi” or “Non”) along with the distribution on noncentrality parameter $\delta$ (N=Near, M=Moderate, and F=Far) in parentheses.
nonfunctional Dicer1 enzyme, a protein which is necessary for processing miRNAs. A total of five mice were selected with global loss of Dicer1 function, and compared to a control group of five mice with normal Dicer1 function. RNA isolated from liver tissue were sampled from each observation, and expression was measured using the Affymetrix GeneChip® Mouse Genome 430 2.0 Array (www.Affymetrix.com).

The microarray profile data were downloaded from the GEO database (www.ncbi.nlm.nih.gov/geo) under accession number GSE11899. Raw CEL files were obtained and normalized in the manner described in Section 2.2.5. This yielded expression values for 45101 probe sets for each observation.

We obtained \( t \)-test \( p \)-values for each gene by applying two-sample \( t \)-tests, assuming equal variance, to the set of expression values for each gene to test the difference in means between the two treatments (Figure 2.5(a)). We also obtained permutation \( p \)-values to be used with our nonparametric adaptation of the semiparametric approach (Figure 2.5(b)). We then applied the semiparametric and nonparametric estimators of \( \pi_0 \) to its respective set of hypothesis test \( p \)-values.

Using \( \hat{\theta}_1 \) as an equivalent estimate for \( \pi_0 \) for each method, the semiparametric approach
Table 2.3  Estimates of the first five values of $\theta$ for the semiparametric method applied to the mouse data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>&lt;0.0001</td>
<td>0.2325</td>
<td>0.3817</td>
<td>0.2057</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

resulted in an estimate of $\hat{\pi}_0 < 0.0001$, while the nonparametric method yielded $\hat{\pi}_0 = 0.5159$. While the authors concluded that there was a large amount of differential expression, the fact that the semiparametric method effectively estimates that all genes are DE is likely a vast underestimate of the true null proportion. In Table 2.3, we present the individual estimates for the first five values of $\hat{\theta}$ using the semiparametric approach, which shows a large proportion of hypothesis tests were estimated to be affiliated with very small values of $\delta$.

2.5 Discussion

Our simulation findings and case study suggest that $\pi_0$ may be underestimated if the semiparametric approach is applied to $p$-values from $t$-tests based on non-normal data with small sample sizes. Our simulation results indicate that our adapted approach using permutation $p$-values greatly reduces the RMSE relative to the semiparametric estimator in all simulation cases that were examined. For example, when the true value of $\pi_0$ was set equal to 0.95 and the “Near” $g(\delta)$ distribution was used to generate differential expression effects, we observed a reduction in the RMSE from 0.1380 to 0.0213 using the permutation test approach relative to the semiparametric formulation. Examination of the boxplots shows that this is a result of both reduced bias and reduction in variance of the $\pi_0$ estimator (Figure 2.3).

The case study in Section 2.4 is a prime example of the results we observed in our simulation study. There is a gross discrepancy between the estimates of $\pi_0$ for each method, with the semiparametric estimate being effectively 0. Close examination of Table 2.3 would indicate that a large proportion of the $p$-values are estimated to be associated with the first three linear B-splines basis densities, corresponding to $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$. It would appear that a large proportion of the $p$-values that the nonparametric method predicted to be null may be falsely associated with relatively small values of $\delta$. 
If we examine the distribution of all null $p$-values from our simulation study, we observe a distinctly non-uniform distribution for the $t$-test $p$-values relative to that of the permutation $p$-values (Figure 2.6). The deviation from uniformity is manifold, with underrepresentation of $p$-values occurring near zero, followed by an overrepresentation of $p$-values until approximately $p = 0.50$. The null distribution then continues to decrease nearly monotonically until $p = 1$.

While it is difficult to parse how these complex deviations from uniformity impact the original estimator, it is obvious that they present severe issues and limit its adequacy in estimating $\pi_0$. In particular, the underepresentation of $p$-values near $p = 0$ would serve to negate the presence of non-null $p$-values that would be expected to occur at this location, which would lead one to conclude that the semiparametric estimator would in fact be conservatively biased. However, our results are contrary to this conjecture, and in fact the estimator is actually quite liberally biased relative to the true value of $\pi_0$. We believe this is due largely to the subsequent overabundance of $p$-values between approximately 0.1 and 0.5 and the monotone decreasing pattern from approximately 0.5 to near 1.

Despite the deficit of the very smallest null $p$-values there is a surplus of $p$-values less than or equal to 0.50 and, therefore, fewer null $p$-values than expected in excess of 0.50. Our simulations indicate that approximately 53% of the null $t$-test $p$-values are less than 0.50. Recall
that most current estimators of \( \pi_0 \) use the latter part of the empirical \( p \)-value distribution to estimate \( \pi_0 \) (Schweder and Spjøtvoll, 1982; Turkheimer et al., 2001; Storey, 2002; Storey and Tibshirani, 2003). Thus, the underrepresentation of the largest \( p \)-values relative to the uniform[0,1] distribution leads to negatively biased estimates of \( \pi_0 \) for these types of estimators.

Ruppert et al. (2007) also propose a “compromise” estimator of \( \pi_0 \), such that

\[
\hat{\pi}_0 = \hat{\theta}_1 + \hat{\theta}_2 Z_{N_{\text{bin}},2}.
\] (2.18)

The authors noted that it may be difficult for the estimator to discriminate between hypotheses that are null and those that are non-null with very small associated values of \( \delta \), a possible scenario represented by the case study described in Section 2.4. This compromise estimator then estimates \( \pi_0 \) as the combination of the null and “near null” components of \( \hat{f}_p(1) \), where “near null” indicates \( \delta \) values distributed according to the B-spline basis density \( B_1 \), which concentrates probability on \( \delta \) values near zero. Application of this compromise procedure to the case study yields \( \hat{\pi}_0 = 0.2296 \), a much more reasonable answer in the context of the problem, although still much lower than the nonparametric counterpart.

If we use this compromise procedure in our simulation study, the performance of the semi-parametric estimator does improve relative to when \( \hat{\pi}_0 \) is equivalent to \( \hat{\theta}_1 \). For example, consider the results from the simulation where the true value of \( \pi_0 = 0.95 \) and \( g(\delta) \) is equal to the “Near” distribution. In Figure 2.7, we present boxplots of \( \pi_0 \) estimates for the semiparametric approach, the compromise semiparametric approach, and the nonparametric method.

While we see improvement in the estimation of \( \pi_0 \) when using the compromise approach, our test procedure still performs much better. Comparing estimates of RMSE, the compromise estimator results in a value of 0.0799, an improvement over the 0.1380 observed with the standard semiparametric estimator but still greater than that of the permutation approach (0.0213).

One possible factor for the discrepancy between the two estimators may be the difference in the granularity of the empirical \( p \)-value distributions between the semiparametric and nonparametric approaches. The semiparametric estimator bins the \( t \)-test \( p \)-values into \( N_{\text{bin}} \) subintervals, with the default setting of 2000 bins. In contrast, there are \( R = 126 \) possible
Figure 2.7  Semiparametric (semi), compromise (comp), and nonparametric (non) estimates of $\pi_0$ for simulation data with true $\pi_0 = 0.95$ and $g(\delta) = N$
values of the permutation p-values in our simulation study. It is then possible that with so many bins, there is a greater chance of deviations of the p-value bin counts from \( y_r \), which could serve to destabilize the estimator. In order to determine if there is any benefit of permutation testing over \( t \)-tests in this regard, we ran a number of supplementary simulations by reducing \( N_{bin} \) to 126 for the semiparametric estimator. However, this change had no discernable effect on the accuracy of the semiparametric estimator over setting \( N_{bin} = 2000 \). Interestingly, we found that generating p-value probability mass matrix \( Z \) using the semiparametric approach and applying it to permutation p-values provides very similar results to our simulation based approach of approximating \( Z \), further indicating that it is the inaccuracy of the \( t \)-test p-values that is the issue for the semiparametric estimator.

A concern in the use of permutation tests for small sample sizes is the coarse discreteness of the p-values themselves, or lack of cardinality (Gadbury et al., 2003). This is especially important when \( \hat{\pi}_0 \) is being used to calculate FDR, since the significance threshold \( \alpha \) has a minimum possible value of \( 1/R \). This could result in an untenably high number of rejected hypotheses, yielding unusable results for further analysis. One possible compromise is to use the estimate of \( \pi_0 \) from the permutation test procedure to estimate FDR for \( t \)-test p-values, thereby taking advantage of a more accurate estimate of \( \pi_0 \) but applying it to continuous p-values. Another possibility would be merging the permutation distributions for all test statistics to form one global reference distribution, a popular approach in dealing with the lack of cardinality in gene-wise nonparametric tests (Tusher et al., 2001; Pan, 2003). However, the resulting p-values are not guaranteed to be exact. Pan (2003) shows that combining these permutation distributions can lead to incorrect inference.

2.6 Conclusions

In this paper we have proposed and evaluated a permutation test-based modification of the semiparametric estimator of \( \pi_0 \) developed by Ruppert et al. The nonnormality of gene expression data should be a concern to any investigator conducting a microarray experiment with low replication. The results from our simulations suggest that our proposed nonparametric approach using permutation p-values from each hypothesis test results in improved estimation
of the true null proportion $\pi_0$ over the use of traditional $t$-tests.

**Bibliography**


CHAPTER 3. THE IMPACT OF SIGNIFICANCE THRESHOLD SELECTION ON THE RELATIONSHIP BETWEEN FALSE DISCOVERY RATE ESTIMATES AND THE REALIZED PROPORTION OF FALSE DISCOVERIES

Nick Larson and Dan Nettleton

Abstract

The false discovery rate (FDR) is a convenient measure of statistical significance when conducting multiple simultaneous hypothesis tests. This procedure is of particular use in gene expression data analysis, where investigators wish to discover differentially expressed genes. However, if investigators determine their significance threshold a priori by repeatedly selecting a particular number of top differentially expressed genes, the corresponding estimates of FDR may be misleading. We identify a negatively correlated relationship between the estimates of the FDR and the true proportion of Type I errors among all rejected null hypotheses via simulations using experimental microarray data for a variety of estimators and conditions. We argue that this phenomenon is due to the correlated nature of the gene expression values, which can result in an empirical null p-value distribution that is far from uniform. We discuss how this leads to FDR estimates being liberally biased, especially when publication bias is taken into account, and caution against the use of a static gene count significance threshold.

3.1 Introduction

The false discovery rate (FDR), originally proposed by Benjamini and Hochberg (1995), is now a widely used multiple testing error rate. FDR is defined as $E(Q)$, where $Q = V/\max\{1, R\}$ and $V$ is the number of type I errors that occur among a total of $R$ rejected
null hypotheses. Estimates of FDR for various significance thresholds are particularly useful in gene expression analysis when researchers are investigating whether genes are equivalently expressed (EE) or differentially expressed (DE) across conditions of interest. In practice, a variety of strategies can be used to select a threshold for significance. Regardless of the method used, an estimate of FDR is typically reported.

Our interactions with scientists conducting gene expression experiments suggest that threshold selection is often driven implicitly by the number of tests declared significant. Researchers desire to avoid rejecting either too few or too many null hypotheses. If too few are rejected, it may be difficult to understand the general effects of treatments or changing conditions on gene expression. Follow-up analyses, such as clustering of significant genes (Eisen et al., 1998; Getz et al., 2000; McLachlan et al., 2002) or overrepresentation analyses (Draghici et al., 2003) are not likely to be informative if only a few genes are declared DE. On the other hand, too many rejected null hypotheses leads to an untenably long list of genes whose DE status will be difficult to verify. Furthermore, the complexity of a long gene list can cause a lack of focus for follow-up experimentation. Thus, the question of where to draw the line for significance might have more to do with the number of genes on a DE gene list than with the FDR level estimated for that gene list.

This leads to a natural question of how well FDR estimation procedures perform when the FDR threshold for significance is not set a priori but instead is chosen based on the number of genes declared to be significant. Addressing this question is the primary focus of this paper. In particular, we consider the relationship between FDR estimates and the realized proportion of false discoveries across repeated experiments when a fixed number of genes \( G \) is declared significant in each experiment. An example of our main finding is depicted in Figure 3.1 and can be summarized as follows: using many standard methods, false discovery rate estimates are negatively correlated with the proportion of type I errors among all rejected null hypotheses (denoted by \( Q \) henceforth) when the number of significant results is used to select the significance threshold. Thus, the proportion of false positives on a list of genes declared to be significant will tend to be greatest when estimates of FDR are lowest. Although an FDR estimate (\( \hat{FDR} \)) is not intended to be a prediction of \( Q \) in a particular experiment, practitioners
naturally think of an FDR estimate in this way. Thus, this negative correlation can be quite misleading in practice and can be especially problematic when publication bias is taken into consideration.

We argue that the source of the negative correlation between $\hat{FDR}$ and $Q$ is correlation among the gene expression levels. This correlation causes the empirical distribution of the $p$-values associated with EE genes to often deviate far from a uniform distribution in a given experiment, even when the marginal distribution of each $p$-value associated with an EE gene is uniform(0,1). This causes problems with many commonly used estimates of FDR and motivated Efron (2007) to propose estimation of what he described as the “empirical null distribution” for use in estimating local FDR ($lFDR$). We show that Efron’s approach alleviates the problem of negative correlation between $\hat{FDR}$ and $Q$ at the cost of considerably more conservative estimation of FDR when compared with competing methods.

In Section 3.2, we formally introduce the concept of FDR estimation, briefly describe the estimators of FDR that we will consider, and describe the setup of our simulation studies. The results of these studies are presented in Section 3.3. In Section 3.4, we provide evidence that correlation among genes is responsible for the negative correlation between $\hat{FDR}$ and $Q$. We present a simple hierarchical model for $p$-values which illustrates the sensitivity of FDR estimation to the deviation of the empirical null $p$-value distribution from uniformity and its relation to the negative correlation between $\hat{FDR}$ and $Q$ in Section 3.5. In Section 3.6, we illustrate how Efron’s (2007) approach can be used to address the problem of negative correlation between $\hat{FDR}$ and $Q$. We discuss our results in Section 3.7 and offer some concluding remarks in Section 3.8.

3.2 Methods

3.2.1 Estimation of False Discovery Rate

Let $t \in (0,1]$ denote the threshold for significance such that all hypotheses with corresponding $p$-values no larger than $t$ are rejected. Let $m$ be the number of tested null hypotheses, denoted $H_1, \ldots, H_m$, with corresponding $p$-values $p_1, \ldots, p_m$. Suppose $m_0$ of the $m$ null hypothe-
ses are true. Define \( V(t) = \# \{ p_1, \ldots, p_m : p_i \leq t, H_i \text{ true} \} \), \( R(t) = \# \{ p_1, \ldots, p_m : p_i \leq t \} \), and

\[
Q(t) = \frac{V(t)}{\max \{ R(t), 1 \}}.
\]  \tag{3.1}

Then, according to Benjamini and Hochberg (1995), the false discovery rate associated with threshold \( t \) is given by \( \text{FDR}(t) = E[Q(t)] \).

In order to estimate \( \text{FDR}(t) \), Storey and Tibshirani (2003) proposed a modification of the adaptive step-up procedure for controlling FDR described by Benjamini and Hochberg (2000). Let \( \pi_0 = m_0/m \) be the proportion of true null hypotheses among \( H_1, \ldots, H_m \). Then, for some estimate of \( \pi_0 \), denoted \( \hat{\pi}_0 \), the estimate of FDR for \( p \)-value threshold \( t \) is defined as

\[
\hat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m t}{\max \{ R(t), 1 \}}. \tag{3.2}
\]

A quantity closely related to \( \hat{\text{FDR}}(t) \) is the \( q \)-value, originally proposed by Storey (2002) as an FDR analogue of the \( p \)-value. Following equation (11) of Storey et al. (2004), we define

\[
\hat{q}(t) = \min_{t' \geq t} \hat{\text{FDR}}(t'). \tag{3.3}
\]

The \( q \)-value \( \hat{q}(t) \) is then the minimum possible estimate of FDR for all significance thresholds greater than or equal to \( t \). Unlike \( \hat{\text{FDR}}(t) \), \( \hat{q}(t) \) is guaranteed to be a nondecreasing function of the significance threshold \( t \). Thus, \( \hat{q}(t) \) is conceptually preferable to \( \hat{\text{FDR}}(t) \) as an estimator of \( \text{FDR}(t) \) and is more commonly used in practice. For these reasons, we use \( \hat{q}(t) \) as our estimator of FDR throughout the remainder of the paper. Note, however, that our results and conclusions were not sensitive to this choice, as \( \hat{\text{FDR}}(t) \) and \( \hat{q}(t) \) are very similar estimators in practice.

In our analyses, we consider four different estimators of \( \pi_0 \) for use in (3.2): the lowest slope method originally proposed by Schweder and Spjøtvoll (1982) and used in the context of FDR estimation by Benjamini and Hochberg (2000) (hereafter referred to as BH), the spline based procedure developed by Storey and Tibshirani (2003) (referred to as ST), a histogram-based procedure proposed by Mosig et al. (2001) and later simplified by Nettleton et al. (2006) (MN), an approach recommended by Pounds and Cheng (2006) (denoted PC) where \( \pi_0 \) is estimated
by the minimum of 1 and $2\bar{p}$, such that $\bar{p}$ is the mean observed $p$-value, as well as an “oracle” approach which uses the true $\pi_0$ value used in data simulation (OR).

Regardless of which estimator of $\pi_0$ is used, a key assumption that motivates the use of (3.2) and (3.3) to estimate FDR is the assumption that $p$-values from tests with a true null hypothesis are each uniformly distributed on the interval $(0,1)$. It is this assumption that makes the numerator in (3.2) ($\hat{\pi}_0 mt$) a natural predictor of $V(t)$ and, thus, $\hat{FDR}(t)$ a natural estimator of FDR(t). However, even if each $p$-value from a test with a true null hypothesis is marginally uniform $(0,1)$, the empirical distribution of the EE $p$-values may be far from uniform in a given experiment when tests are dependent. As we shall demonstrate, this causes problems for FDR estimators that are based on (3.2) and (3.3). Thus, we also study the method of Efron (2007) that attempts to account for nonuniformity of the empirical EE $p$-value distribution in a given experiment. Details of this method and its performance are presented in Section 3.6 after FDR estimators based on (3.2) and (3.3) have been thoroughly investigated.

### 3.2.2 Data Simulation

To evaluate FDR estimators on simulated datasets with realistic within-gene expression distributions and correlation structures among genes, we made use of previously produced and publicly available microarray data in our simulations. For a given dataset, a subset composed of observations for a single experimental condition was selected as a dataset for the simulations.

For such a “parent” dataset with a total of $N$ expression vectors from $N$ experimental units with $m$ gene probe sets per vector, $2n$ of the $N$ vectors were randomly sampled without replacement to serve as the basis for two hypothetical treatment groups, with $n$ vectors allocated to each group. For the purposes of selecting genes to be designated as DE, we considered two separate strategies: an independent random selection approach, and a gene ontology ([www.geneontology.org](http://www.geneontology.org)), or GO, category approach. The latter strategy was adopted to capture a more biologically realistic correlation structure among DE genes. Under the first procedure, for a selected value of $\pi_0 = m_0/m$, we randomly sampled without replacement $\lfloor(1 - \pi_0) * m\rfloor = K$ genes to be DE. For the GO approach, the GO annotations for each gene were tabulated for all of the probe sets using the **affy** package (Gautier et al., 2004) from Bioconductor.
GO terms were then randomly selected without replacement and all their associated probesets were designated DE. This was conducted in an iterative fashion until the total accumulated DE count was approximately $K$. We only considered GO terms with between 2 and 20 probe sets. This allowed for the selection of genes that were related to at least one other gene, but would avoid GO terms which were too general to be appropriate for simulating genetic co-regulation. We refer to these two DE selection methods respectively as RND and GOT throughout the rest of the paper.

For the $i^{th}$ DE gene, a treatment effect $\tau_i$ was randomly generated from a scaled beta($\alpha, \beta$) distribution with multiplicative scale factor $c\hat{\sigma}_i \sqrt{2/n}$, where $\hat{\sigma}_i^2$ denotes the sample variance of the $N$ observations of the $i^{th}$ DE gene in the parent dataset, and $c$ defines the maximal value of the support of $\tau_i$. This treatment effect $\tau_i$ was added to the $n$ observed expression values allocated to the second hypothetical treatment group. To obtain a range of realistic treatment effects, we set the parameters of the scaled beta distribution to $c = 4$, $\alpha = 2$ and $\beta = 3$.

An acute myeloid leukemia (AML) dataset was obtained from the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) data base (accession number GSE12417) to serve as the parent set. Raw CEL files were obtained and normalized using robust multiarray average (RMA) normalization (Bolstad et al., 2003) with the affy library (Gautier et al., 2004) in R. For this data, mononuclear bone marrow samples were evaluated for 161 patients with muscular diseases (Metzeler et al., 2008). The AML data came from the Affymetrix (www.affymetrix.com) U133A GeneChip® platform, measuring the expression levels with 22283 probe sets.

We generated 1000 dataset replications with sample size $n = 10$, each of the two DE selection methods, and EE gene proportions $\pi_0 = 0.75$, 0.80, 0.85, 0.90, and 0.95 yielding a total of 10 simulations. The $n = 10$ sample size was used to mirror the typically small samples found in microarray experiments (Lee et al., 2000; Qu et al., 2010).

### 3.2.3 Data Analysis and Threshold Selection

After the simulation of a microarray dataset, a vector of $p$-values was obtained by use of two-sample $t$-tests under the assumption of equal variance across treatments (an assumption which is true in our simulation study). Following the computation of $p$-values, we estimated FDR
using the five methods described in Section 3.2.1 (MN,BH,OR,ST,PC). We designate each FDR estimator using a subscript to represent how $\pi_0$ is estimated. For example, $\widehat{\text{FDR}}_{PC}$ uses the PC $\pi_0$ estimation approach. Significance threshold selection at the $p$-value level was determined by the null hypotheses with the $G$ smallest $p$-values. As discussed in the introduction, this is a necessary simplification of the non-algorithmic strategy often used by biological researchers in practice. For $G = 100, 200, 300, 400, \text{and } 500$, we produced an FDR estimate from each of the five estimation methods, along with the corresponding realized proportion of false discoveries, $Q$. The Pearson correlation coefficient was calculated to estimate the correlation between $Q$ and each estimator of FDR for each gene count threshold.

### 3.3 Simulation Results

Table 3.1 is an example of the Pearson correlation coefficients calculated between $Q$ and $\widehat{\text{FDR}}$, corresponding to the data simulation using the RND selection method and $\pi_0 = 0.90$ (tabulated values from the other simulations can be found in the supplementary materials). Based upon permutation testing, all these correlation estimates are highly significantly different from 0. Figure 3.1 is typical in appearance of the pattern we observed between $Q$ and $\widehat{\text{FDR}}$ across all methods and conditions. While the Pearson correlation coefficient is not a measure that fully captures all aspects of this type of relationship, it serves as a useful single-number summary.

The results in Table 3.1 illustrate the typical relationship between the relative magnitude of the negative correlation and $G$, the number of genes declared DE; higher values of $G$ were associated with more extreme negative correlations. We also observed that the relationship between $Q$ and $\widehat{\text{FDR}}$ depended on the value of $\pi_0$ used in the simulation. As the true null proportion increased, the negative correlation between $Q$ and $\widehat{\text{FDR}}$ also tended to increase in magnitude. Figure 3.2 illustrates this relationship by plotting the ST estimate for $G = 500$ for each DE gene selection method across all values of $\pi_0$ used in our simulations.

With respect to the five different FDR estimation methods used in our analysis, Table 3.1 is indicative of the stark similarity of their respective results under a given simulation scenario and threshold selection strategy. The OR estimator often resulted in estimates least correlated
Figure 3.1  A plot of $\hat{FDR}_S T$ vs. the realized proportion of false discoveries $Q$ for the simulation $\pi_0 = 0.90$, RND DE selection, and $G = 500$.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.1073</td>
<td>-0.1577</td>
<td>-0.2251</td>
<td>-0.2782</td>
<td>-0.3324</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0795</td>
<td>-0.1277</td>
<td>-0.1914</td>
<td>-0.2455</td>
<td>-0.3015</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0786</td>
<td>-0.1272</td>
<td>-0.1912</td>
<td>-0.2457</td>
<td>-0.3021</td>
</tr>
<tr>
<td>ST</td>
<td>-0.1106</td>
<td>-0.1655</td>
<td>-0.2351</td>
<td>-0.2895</td>
<td>-0.3450</td>
</tr>
<tr>
<td>PC</td>
<td>-0.2631</td>
<td>-0.3239</td>
<td>-0.3710</td>
<td>-0.4004</td>
<td>-0.4264</td>
</tr>
</tbody>
</table>

Table 3.1  Pearson correlation coefficients for the data simulation using $\pi_0 = 0.90$ for each of the five FDR estimators, RND DE selection, and five values of $G$. 
with \( Q \) while the PC approach often yielded the most extreme negative correlations. The GOT approach and the RND approach led to similar results, with the GOT method resulting in slightly less extreme correlations. For example, the OR estimate for \( \pi_0 = 0.90 \) and \( G = 500 \) for the GOT approach was -0.2297 for the GOT approach and -0.3021 for RND.

To investigate this phenomenon under differing effect size distributions, we also conducted smaller scale simulations that altered the parameters of the beta distribution used to generate the treatment effects. Selecting the same example simulation described above (RND approach, \( \pi_0 = 0.90 \)), we reran the simulation in the same manner as previously described except that the \( \alpha \) shape parameter was altered to make treatment effect sizes smaller (\( \alpha = 1 \)) and larger (\( \alpha = 3 \)). Skewing the treatment effect distribution towards smaller values resulted in less significant DE \( p \)-values and an increase in magnitude of the negative estimated correlation coefficients. For example, the correlation for \( \text{FDR}_{OR} \) and \( Q \) for \( \pi_0 = 0.90 \), \( G = 500 \) and RND was -0.4391, compared to -0.3021 in the original simulation. On the other hand, by skewing the treatment effect distribution toward greater values by setting \( \alpha = 3 \), we observed a correlation

---

Figure 3.2  Depiction of \( \text{FDR}_{ST} \) estimates using \( t = p(500) \) across RND and GOT DE selection methods for the five values of \( \pi_0 \).
coefficient for the same conditions to be -0.1356.

There is also indication that this phenomenon is microarray-platform dependent. Each microarray chip measures a specific set of genes dictated by the probe set design of the chip. Since we argue in the next section that the negative correlation between $\hat{FDR}$ and $Q$ is due to correlation in expression levels among genes, it follows that the extent of the negative correlation between $\hat{FDR}$ directly depends upon which genes are measured by the probe sets in a given platform. We examined results from smaller simulations using expression profiles from U133 Plus 2.0, U95A, U95B, U95C, and U95Av2 platforms. Running the simulations under $\pi_0 = 0.90$, RND, and $G = 500$ led to variable results, with correlations between $\hat{FDR}_{OR}$ estimates and $Q$ ranging from -0.7532 for the U95B platform to 0.0398 for U95C.

Although there was variation in the results between methods and across simulation scenarios, the clear message from our simulation work is that $Q$ tends to be negatively correlated with FDR estimates based on (3.2) and (3.3). This negative correlation was present in the overwhelming majority of the simulations we studied. Although we investigated only a few estimators of $\pi_0$ among many that could be used in (3.2) and (3.3), our results for the OR method imply that this negative correlation will persist regardless of how well $\pi_0$ is estimated. Thus, it is not necessary to investigate any of the many other estimators of $\pi_0$ that have been proposed in the literature. Even an FDR estimator based on the original method of Benjamini and Hochberg (1995), which uses 1 to estimate $\pi_0$ regardless of the data, will suffer this negative correlation problem because its estimates are nearly a constant multiple ($1/\pi_0$) of $\hat{FDR}_{OR}$.

### 3.4 An Explanation of the Negative Correlation between $\hat{FDR}$ and $Q$

The simulations from Section 3.3 indicate that the application of the FDR estimators based on (3.2) and (3.3) to microarray data results in a negatively correlated relationship between $\hat{FDR}$ and $Q$ when a fixed number of genes is rejected. Since this relationship is also present in the results for the OR estimator, where the known value of $\pi_0$ is applied in calculating the estimate, any biasedness imposed by inaccurate estimates of $\pi_0$ in the remaining four estimators must only serve to compound an existing issue. To address the potential causes of this phenomenon, we then solely consider the OR estimator.
Figure 3.3  Histograms of EE p-values from microarray data simulation with RND DE selection and \( \pi_0 = 0.90 \). For DE gene count \( G = 500 \), (a) \( Q = 0.602 \) and \( \hat{\text{FDR}}_{OR} = 0.106 \), while (b) \( Q = 0.104 \) with \( \hat{\text{FDR}}_{OR} = 0.896 \).

Let us examine a pair of extreme simulation results where there is a large discrepancy between \( \hat{\text{FDR}}_{OR} \) and \( Q \). Figures 3.3(a) and 3.3(b) depict histograms of the empirical EE p-value distributions for cases where the quantity \( (Q - \hat{\text{FDR}}) \) is highly positive or negative, respectively. Initial examination of these plots show that the empirical null p-value distribution is far from uniform, indicated by respective severe under- and overrepresentation of small p-values relative to the uniform distribution. Note that these histograms do not imply that the marginal distribution of any EE p-value is non-uniform. In fact, the histogram of p-values generated across simulations tends to be very close to uniform for each EE gene. Rather, the departure from uniformity is due to the correlation in the multivariate distribution of the EE p-values. Figure 3.3 shows the empirical summaries of just two draws from the multivariate distribution of the vector of EE p-values. It is clear that marginal uniformity provides no guarantee of uniformity for the empirical distribution of the components of one multivariate observation.

When rejecting a fixed number of genes \( G \) using the OR estimator, (3.2) simplifies to

\[
\hat{\text{FDR}}_{OR} = \frac{\pi_0 m p(G)}{G},
\]  

(3.4)
where \( p(G) \) denotes the \( G^{th} \) smallest \( p \)-value. When combined with DE genes, \( p(G) \) will tend to be larger for EE \( p \)-values like those in Figure 3.3(a), compared to EE \( p \)-values like those in Figure 3.3(b). The opposite is true for \( Q \), which will tend to be smaller when EE \( p \)-values are like those in Figure 3.3(a) compared to 3.3(b), because in Figure 3.3(a), the vast majority of rejected hypotheses will correspond to DE genes due to the underrepresentation of small EE \( p \)-values. Thus, the estimates of FDR for the oracle procedure will tend to be greater when \( Q \) is smaller and smaller when \( Q \) is greater, which leads to the negative correlation we have documented.

It can then be inferred that this negative correlation between \( Q \) and \( \hat{FDR} \) is the result of correlation among measures of gene expression values. To further examine this argument, we ran additional simulations which disrupted the biological correlation structure naturally occurring in the expression profiles. To accomplish this, we generated simulated null data in the same manner described in Section 3.2.2, and then for each gene, randomly permuted the expression values prior to adding the treatment effect to the second set of observations. When we applied the estimators of FDR to these data, we observed empirical null \( p \)-value distributions that were consistently close to the uniform. This disrupted the assumed cause of the negative correlation between \( Q \) and \( \hat{FDR} \), and eliminated the negative correlation coefficient values for data simulations of various sample size and dataset selections.

### 3.5 A Modeling Approach to Sensitivity Analysis

In Section 3.4 we have shown that the negatively correlated relationship between \( \hat{FDR} \) and \( Q \) is due to the deviation of the empirical EE \( p \)-value distribution from uniformity. Particularly, we see that there tends to be either an under- or overrepresentation of small EE \( p \)-values relative to the uniform distribution. We explored a simplified model of the EE and DE \( p \)-value distributions in order to examine the sensitivity of the correlation phenomenon to the magnitude of this deviation from the assumption uniformity.

Our observations of the empirical DE and EE \( p \)-value distributions in our simulation studies led us to believe that while the EE distribution varies greatly in shape (Figure 3.4(a)), the DE distribution remains relatively static. Consequently, we decided upon a hierarchical approach
to modelling the \( p \)-values, where each simulation replication would result in the EE and DE \( p \)-values being distributed uniquely. Taking into account the empirical EE \( p \)-value distributions from our simulations, we devised a novel, single parameter probability distribution to approximate the observed distributional shapes. Let the conditional probability density function for the EE \( p \)-values be given as the following:

\[
f_{EE}(x \mid \lambda_i) = \lambda_i x^2 - 2\lambda_i x + 2\lambda_i/3 + 1
\]

where \( \lambda_i \) is a random shape parameter unique to the \( i \)th simulation replication. This function yields a proper probability density for \( \lambda_i \in [-3/2, 3] \). We derived the above formula based upon three criteria: the derivative of the density at 1 is always equal to 0, the shape of the distribution is parabolic and it is relatively simple in form. Despite its plain format, the distribution has some appealing properties: the density is uniform(0,1) when \( \lambda_i = 0 \), the distribution approximates the empirical EE distributions remarkably well (Figure 3.4(b)) and for any distribution on \( \lambda_i \) symmetric about 0, the marginal distribution of \( x \) is uniform(0,1).

The last property is particularly useful because we have observed that, averaging over all simulations, the EE distribution is actually uniform. Using this hierarchical approach, we emulate the effects of gene-wise correlation by assigning a particular distributional shape to the EE \( p \)-value distribution for each replication of the simulation study.

Similarly, we can adopt a hierarchical approach for the DE \( p \)-values which is must more restrictive, capturing the small amount of variation observed in our data simulation results. Our findings indicate that the shape of the DE \( p \)-value distribution is approximately beta(1,8). We can similarly place distributions on the \( \alpha \) and \( \beta \) parameters such that the DE \( p \)-values for the \( i \)th simulation iteration are distributed as beta(\( \alpha_i, \beta_i \)), for some distributions on \( \beta_i \) and \( \alpha \) fixed at 1.

We conducted simulations similar to those described in Section 3.3 by defining a set number of hypothetical genes, \( m \), and a true value of \( \pi_0 \), which we set at 10,000 and 0.90, respectively. We place a truncated normal distribution on \( \lambda_i \) over the interval [-1,1] with mean \( \mu_\lambda = 0 \), using various values for the normal standard deviation parameter, \( \sigma_\lambda \). As we believe that the EE distributions deviating from the uniform are what induce the negative correlation, by placing
Figure 3.4 Plots of (a) lowess smoothed EE distributions from one of the simulations (RND, π₀ = 0.90) and (b) a variety of densities \( f_{EE}(x \mid \alpha) \) where \( \alpha \in [-1.5, 1.5] \)

<table>
<thead>
<tr>
<th>( \sigma_\lambda )</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.8305</td>
</tr>
<tr>
<td>0.50</td>
<td>-0.7629</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.4999</td>
</tr>
<tr>
<td>0.10</td>
<td>-0.1389</td>
</tr>
<tr>
<td>0</td>
<td>0.00008</td>
</tr>
</tbody>
</table>

Table 3.2 Pearson correlations between \( \hat{\text{FDR}}_{OR} \) and \( Q \) under different values of \( \sigma_\lambda \)

larger probability on such EE distributional shapes, the magnitude of the negative correlation between FDR and \( Q \) will coincidentally increase. We also used truncated normal distributions on the \( \alpha_i \) and \( \beta_i \) parameters, where \( \mu_\beta \) is set equal to 8 and \( \sigma_\beta = 0.1 \), with truncation at +/- 0.5 from \( \mu_\beta \). Since \( \alpha \) is set equal to 1, \( f_{DE}(p) \) is ensured to be monotone decreasing.

Table 3.2 contains the Pearson correlation coefficients between \( \hat{\text{FDR}}_{OR} \) and \( Q \) using the threshold condition \( G = 100 \), and the corresponding values of \( \sigma_\lambda \) used for each simulation. We ran a total of 1000 replications in these simulations.

As \( \sigma_\lambda \) diminishes in magnitude, so does the magnitude of the negative correlation. Once \( \lambda_i = 0 \) for all the replications, correlation reduces to nearly zero. This is due to the fact that \( Q \) and \( \hat{\text{FDR}} \) are consistently close in value, with random noise the only source of variation of
the two. However, even when the standard deviation is relatively small (0.10), there is still a negatively correlated relationship between FDR and Q, suggesting that even slight deviations from uniformity can induce the negative correlation effect.

3.6 The Empirical Null Method

Given the evident relation between the extremity of the potential liberal bias of FDR control estimation and the size of $m_0$, an alternative approach in cases where the null proportion is expected to be very large would be the use of the empirical null procedure proposed by Efron (2007). By using a measure of large-scale testing correlation, Efron’s method estimates an appropriate empirical null distribution and offers corrective applications to FDR estimation. By assuming the empirical distribution of test statistics to be composed almost entirely of null $t$-test statistics, Efron argues that transformed values can be compared to a central $t$-distribution and the disproportionate representation of values at the tails or mode can be attributed to correlation effects. Note that there is a direct correspondence between the shape of the test statistic null distribution and the resulting null $p$-value distribution. A “narrow” empirical null test statistic distribution would result in an underrepresentation of small null $p$-values, while a “wide” null test statistic distribution would correspond to an overrepresentation of null $p$-values. Use of this method would then appropriately correct for the variety of null distributions encountered in the simulated microarray experiments, yielding more accurate estimates of FDR.

Let us contrast the $p$-value based FDR estimators described in Section 3.2.1, which assume theoretical null distributions, with the empirical null approach using the locfdr package (Efron et al., 2011) in R. Instead of the traditional “tail-area” approaches to FDR, Efron makes use of the local false discovery rate, or lFDR, which is defined as the probability that a given individual hypothesis test is null. In order to apply this method, we transform the calculated $t$-statistics from the analysis so that they are Normal(0,1) distributed under the theoretical null and employ the maximum likelihood estimator of the empirical null distribution. Since this method operates on the test statistic level, we order the hypothesis tests by their lFDR estimates and, in similar fashion to the $p$-value based estimators, declared the corresponding hypotheses with
Calculating the mean of the $\hat{\pi}$ with $\hat{\text{FDR}}_{EF}$ for each of the five significance thresholds.

Table 3.3  Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.90$ using $\hat{\text{FDR}}_{EF}$ for each of the five significance thresholds.

<table>
<thead>
<tr>
<th>G</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF</td>
<td>0.8786</td>
<td>0.8556</td>
<td>0.8413</td>
<td>0.8270</td>
<td>0.8193</td>
</tr>
</tbody>
</table>

Figure 3.5  A plot of $\hat{\text{FDR}}_{EF}$ vs. $Q$ for the simulation $\pi_0 = 0.90$, RND DE selection, and $G = 500$.

The $G$ smallest values of $l\text{FDR}$ to be DE. We then estimated the tail-based FDR estimate by calculating the mean of the $l\text{FDR}$ of all rejected hypotheses, such that $\hat{\text{FDR}} = \frac{1}{G} \sum_{i=1}^{G} l\text{FDR}_{(i)}$, for purposes of comparison. If we apply this approach to the same set of observations that were used in Table 3.1, we see that $\hat{\text{FDR}}_{EF}$ exhibits a very strong positively correlated relationship with $Q$ (Table 3.3). Similarly, we can plot $Q$ and $\hat{\text{FDR}}_{EF}$ to see an entirely different relationship (Figure 3.5). Note, however, that this relationship is very conservative, such that $\hat{\text{FDR}}_{EF}$ tends to overestimate $Q$, particularly when $Q$ is small.
3.7 Discussion

The initial screening process of microarray analysis is an exercise in dimensional reduction. Researchers using microarrays desire to limit the total possible number of genes they wish to declare as significant in order to curtail the expense of follow-up laboratory investigation. Consequently, common applications of FDR involve paring down the total list of genes to a manageable subset in correspondence to an acceptable false discovery rate, relative to the unique requirements of the analysis. Proper control of FDR at the screening stage of exploratory microarray analysis is critical for the identification of a list of candidate genes for secondary investigation. The existence of a negative correlation between $Q$ and estimated FDR is of serious concern since results which indicate FDR is controlled at a low value may actually have a much higher proportion of false positives, degrading the reliability of the data. Equally disquieting is when the estimate of FDR is too high when $Q$ is actually much smaller, leading to overly conservative elimination of candidate genes. While the simulation results show that overall the means of $Q$ and estimates of FDR are approximately equal or conservatively biased, supporting theoretical arguments in place that the estimators exhibit strong control, the research publication bias associated with publishable findings where $\hat{FDR}$ is very low would imply that this is even greater an issue, since it would be these very cases where we may see a large overrepresentation of false positives.

We believe that the principal mechanism behind the deviation of the empirical EE $p$-value distribution from the assumed uniform is genetic co-regulation. Whereas the statistical methods used in our simulation analyses rely upon the independence (or very weak dependence) of the individual hypothesis tests, independence is rarely, if ever, satisfied in multiple testing of the scale exhibited in microarray data. Moreover, genes are biologically related through intricate gene network pathways, leading to highly correlated subgroups of gene expression values, which clearly violate independence assumptions.

While the existence of the correlated relationship may be troubling, it is of note that for a large majority of the simulations, the mean of the estimates of FDR was greater than the mean of the $Q$ proportions, suggesting that conservative estimation of the expected FDR is
Table 3.4  Mean $Q$ and $\widehat{\text{FDR}}$ for all five estimators, using RND DE selection, $\pi_0 = 0.90$ and $G = 500$

<table>
<thead>
<tr>
<th></th>
<th>$Q$</th>
<th>NT</th>
<th>BH</th>
<th>OR</th>
<th>ST</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2024</td>
<td>0.3982</td>
<td>0.4248</td>
<td>0.3841</td>
<td>0.3959</td>
<td>0.5378</td>
</tr>
</tbody>
</table>

maintained, on average, by the proposed methods. For example, in examining the simulation results for $\pi_0 = 0.90$, DE method RND, and $t = p_{(500)} Q = 0.2024$, while the means of the FDR estimates for all five of the methods are $\geq 0.30$ (Table 3.4). The cause for concern instead lies in this phenomenon in conjunction with the reality of publication bias. Researchers tend to report results which lead to statistically significant findings. Consider the results of the same simulation in which all experiments with estimates of FDR greater than 0.20 are tossed out. For the OR estimator, this yields a mean $Q$ of 0.2511. If we lower this censoring FDR criterion to 0.15, the mean $Q$ of the experiment subset is even greater at 0.2642. Thus the consequences of introducing publication bias via censoring of experiments with FDR estimates above a particular threshold yield liberally biased estimates of FDR. This phenomenon is especially concerning in the context of meta-analyses, and would compound the already existing issues involved in the analysis of multiple published findings (Ramasamy et al., 2008). The extremity of the liberal bias was strongly related to the size of $\pi_0$ used in simulation, as only FDR threshold values $\leq 0.025$ began to exhibit bias for simulated cases where $\pi_0 = 0.75$.

Given the evident relation between the extremity of the potential liberal bias of FDR control estimation and the size of $\pi_0$, an alternative approach in cases where the null proportion is expected to be very large would be the use of the empirical null procedure proposed by Efron (2007). By using a measure of large-scale testing correlation, Efron’s method estimates an appropriate empirical null distribution and offers corrective applications to FDR estimation. Note that there is a direct correspondence between the shape of the test statistic null distribution and the resulting null $p$-value distribution. A “narrow” empirical null test statistic distribution would result in an underrepresentation of small null $p$-values, while a “wide” null test statistic distribution would correspond to an overrepresentation of null $p$-values. Use of this method would then appropriately correct for the variety of null distributions encountered in the simulated microarray experiments, yielding more accurate estimates of FDR. If we em-
ploy our model of publication bias and eliminate all results with $\hat{\text{FDR}} \leq 0.20$, the resultant $\bar{Q} = 0.1282$, restoring conservative estimation of FDR. It is clear that via the inclusion of the estimation of the empirical null distribution, we circumvent liberally biased estimates of the other estimators under the impacts of publication bias. However, as seen in Figure 3.5, this approach is highly conservative, and $\hat{\text{FDR}}$ is typically much larger than $Q$. This itself presents a different problem, since the inflation of $\hat{\text{FDR}}$ could lead to a false sense of a lack of result integrity when significance is determined by $t = p(G)$.

### 3.8 Conclusions

Much work has gone into the theoretical properties of various statistical methods of estimating FDR in the presence of mathematically simulated dependency structures (Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2003; Dabney and Storey, 2006), however the implications of actual biological dependence structures among genes in gene expression data appear to be far-reaching. This would indicate that ignoring the effects of gene co-regulation would be inappropriate for practical exploratory gene expression analysis, and investigators should be cautious in how they approach threshold selection in estimating FDR without acknowledging the existence of dependency in expression data and employing estimators which take such dependency into account.
### APPENDIX

#### 3.A Pearson Correlations from Data Simulations

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0232</td>
<td>-0.0235</td>
<td>-0.0371</td>
<td>-0.0433</td>
<td>-0.0522</td>
</tr>
<tr>
<td>BH</td>
<td>0.0056</td>
<td>0.0130</td>
<td>0.0064</td>
<td>0.0088</td>
<td>0.0069</td>
</tr>
<tr>
<td>OR</td>
<td>0.0064</td>
<td>0.0140</td>
<td>0.0077</td>
<td>0.0100</td>
<td>0.0082</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0323</td>
<td>-0.0354</td>
<td>-0.0509</td>
<td>-0.0593</td>
<td>-0.0701</td>
</tr>
<tr>
<td>PC</td>
<td>-0.0728</td>
<td>-0.0933</td>
<td>-0.1032</td>
<td>-0.1108</td>
<td>-0.1226</td>
</tr>
<tr>
<td>EF</td>
<td>0.9111</td>
<td>0.9069</td>
<td>0.9024</td>
<td>0.8969</td>
<td>0.8910</td>
</tr>
</tbody>
</table>

Table 3.5 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.75$ and RND gene selection

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0012</td>
<td>-0.0373</td>
<td>-0.0277</td>
<td>-0.0301</td>
<td>-0.0299</td>
</tr>
<tr>
<td>BH</td>
<td>0.0260</td>
<td>0.0002</td>
<td>0.0175</td>
<td>0.0241</td>
<td>0.0282</td>
</tr>
<tr>
<td>OR</td>
<td>0.0297</td>
<td>0.0062</td>
<td>0.0250</td>
<td>0.0326</td>
<td>0.0379</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0102</td>
<td>-0.0493</td>
<td>-0.0427</td>
<td>-0.0470</td>
<td>-0.0475</td>
</tr>
<tr>
<td>PC</td>
<td>-0.0712</td>
<td>-0.1073</td>
<td>-0.1191</td>
<td>-0.1281</td>
<td>-0.1251</td>
</tr>
<tr>
<td>EF</td>
<td>0.9065</td>
<td>0.9000</td>
<td>0.8908</td>
<td>0.8814</td>
<td>0.8723</td>
</tr>
</tbody>
</table>

Table 3.6 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.80$ and RND gene selection
### Table 3.7 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.85$ and RND gene selection

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0414</td>
<td>-0.1062</td>
<td>-0.1446</td>
<td>-0.1839</td>
<td>-0.2116</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0093</td>
<td>-0.0632</td>
<td>-0.0945</td>
<td>-0.1291</td>
<td>-0.1544</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0085</td>
<td>-0.0621</td>
<td>-0.0934</td>
<td>-0.1282</td>
<td>-0.1539</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0481</td>
<td>-0.1144</td>
<td>-0.1544</td>
<td>-0.1947</td>
<td>-0.2237</td>
</tr>
<tr>
<td>PC</td>
<td>-0.1961</td>
<td>-0.2473</td>
<td>-0.2782</td>
<td>-0.3101</td>
<td>-0.3274</td>
</tr>
<tr>
<td>EF</td>
<td>0.8867</td>
<td>0.8632</td>
<td>0.8409</td>
<td>0.8199</td>
<td>0.8063</td>
</tr>
</tbody>
</table>

### Table 3.8 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.95$ and RND gene selection

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.2263</td>
<td>-0.3750</td>
<td>-0.4819</td>
<td>-0.5500</td>
<td>-0.5904</td>
</tr>
<tr>
<td>BH</td>
<td>-0.1952</td>
<td>-0.3474</td>
<td>-0.4611</td>
<td>-0.5360</td>
<td>-0.5801</td>
</tr>
<tr>
<td>OR</td>
<td>-0.1948</td>
<td>-0.3475</td>
<td>-0.4615</td>
<td>-0.5367</td>
<td>-0.5810</td>
</tr>
<tr>
<td>ST</td>
<td>-0.2342</td>
<td>-0.3824</td>
<td>-0.4881</td>
<td>-0.5551</td>
<td>-0.5954</td>
</tr>
<tr>
<td>PC</td>
<td>-0.3641</td>
<td>-0.4136</td>
<td>-0.4428</td>
<td>-0.4698</td>
<td>-0.4904</td>
</tr>
<tr>
<td>EF</td>
<td>0.8126</td>
<td>0.7845</td>
<td>0.7616</td>
<td>0.7409</td>
<td>0.7237</td>
</tr>
</tbody>
</table>
\[ G \]

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0376</td>
<td>-0.0342</td>
<td>-0.0547</td>
<td>-0.0578</td>
<td>-0.0770</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0164</td>
<td>-0.0066</td>
<td>-0.0207</td>
<td>-0.0194</td>
<td>-0.0342</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0165</td>
<td>-0.0068</td>
<td>-0.0207</td>
<td>-0.0195</td>
<td>-0.0342</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0447</td>
<td>-0.0454</td>
<td>-0.0685</td>
<td>-0.0719</td>
<td>-0.0917</td>
</tr>
<tr>
<td>PC</td>
<td>-0.0810</td>
<td>-0.1007</td>
<td>-0.1169</td>
<td>-0.1220</td>
<td>-0.1295</td>
</tr>
<tr>
<td>EF</td>
<td>0.9159</td>
<td>0.9030</td>
<td>0.8917</td>
<td>0.8820</td>
<td>0.8736</td>
</tr>
</tbody>
</table>

Table 3.9 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.75$ and GOT gene selection

\[ G \]

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0365</td>
<td>-0.0706</td>
<td>-0.1132</td>
<td>-0.1419</td>
<td>-0.1635</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0106</td>
<td>-0.0319</td>
<td>-0.0677</td>
<td>-0.0901</td>
<td>-0.1078</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0096</td>
<td>-0.0307</td>
<td>-0.0664</td>
<td>-0.0889</td>
<td>-0.1069</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0432</td>
<td>-0.0802</td>
<td>-0.1240</td>
<td>-0.1555</td>
<td>-0.1787</td>
</tr>
<tr>
<td>PC</td>
<td>-0.1256</td>
<td>-0.1773</td>
<td>-0.2159</td>
<td>-0.2405</td>
<td>-0.2585</td>
</tr>
<tr>
<td>EF</td>
<td>0.8942</td>
<td>0.8794</td>
<td>0.8649</td>
<td>0.8509</td>
<td>0.8385</td>
</tr>
</tbody>
</table>

Table 3.10 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.80$ and GOT gene selection
### Table 3.11  Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.85$ and GOT gene selection

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0851</td>
<td>-0.1072</td>
<td>-0.1412</td>
<td>-0.1690</td>
<td>-0.2037</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0551</td>
<td>-0.0658</td>
<td>-0.0933</td>
<td>-0.1154</td>
<td>-0.1467</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0539</td>
<td>-0.0647</td>
<td>-0.0924</td>
<td>-0.1148</td>
<td>-0.1465</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0909</td>
<td>-0.1156</td>
<td>-0.1510</td>
<td>-0.1802</td>
<td>-0.2157</td>
</tr>
<tr>
<td>PC</td>
<td>-0.1808</td>
<td>-0.2180</td>
<td>-0.2544</td>
<td>-0.2829</td>
<td>-0.3088</td>
</tr>
<tr>
<td>EF</td>
<td>0.9014</td>
<td>0.8809</td>
<td>0.8679</td>
<td>0.8574</td>
<td>0.8473</td>
</tr>
</tbody>
</table>

### Table 3.12  Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.90$ and GOT gene selection

<table>
<thead>
<tr>
<th>Estimator</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>-0.0665</td>
<td>-0.1332</td>
<td>-0.1999</td>
<td>-0.2672</td>
<td>-0.3189</td>
</tr>
<tr>
<td>BH</td>
<td>-0.0313</td>
<td>-0.0899</td>
<td>-0.1552</td>
<td>-0.2244</td>
<td>-0.2782</td>
</tr>
<tr>
<td>OR</td>
<td>-0.0300</td>
<td>-0.0890</td>
<td>-0.1548</td>
<td>-0.2246</td>
<td>-0.2789</td>
</tr>
<tr>
<td>ST</td>
<td>-0.0706</td>
<td>-0.1399</td>
<td>-0.2078</td>
<td>-0.2763</td>
<td>-0.3285</td>
</tr>
<tr>
<td>PC</td>
<td>-0.2153</td>
<td>-0.2882</td>
<td>-0.3312</td>
<td>-0.3724</td>
<td>-0.3990</td>
</tr>
<tr>
<td>EF</td>
<td>0.8377</td>
<td>0.8131</td>
<td>0.7942</td>
<td>0.7938</td>
<td>0.7912</td>
</tr>
</tbody>
</table>
\begin{tabular}{|c|cccccc|}
\hline
Estimator & 100 & 200 & 300 & 400 & 500 \\
\hline
MN & -0.1304 & -0.3077 & -0.4208 & -0.4938 & -0.5315 \\
BH & -0.0923 & -0.2717 & -0.3933 & -0.4736 & -0.5159 \\
OR & -0.0921 & -0.2719 & -0.3939 & -0.4745 & -0.5170 \\
ST & -0.1413 & -0.3173 & -0.4291 & -0.5016 & -0.5395 \\
PC & -0.3223 & -0.3745 & -0.4053 & -0.4270 & -0.4440 \\
EF & 0.8091 & 0.7913 & 0.7666 & 0.7380 & 0.7146 \\
\hline
\end{tabular}

Table 3.13 Pearson correlation coefficients for the data simulation using the $\pi_0 = 0.95$ and GOT gene selection

\textbf{Bibliography}


CHAPTER 4. METHODOLOGY FOR EVALUATING THE COMPLEMENTATION MODEL OF HETEROSIS USING HIERARCHICAL LATENT-CLASS MODELING OF RNA-SEQ DATA

Nick Larson and Dan Nettleton

Abstract

Heterosis, otherwise known as “hybrid vigor,” is a well documented phenomenon in crop breeding in which the genetic cross of two diverse inbred varieties tends to exhibit traits which are agriculturally superior to either of the two parents. While this topic has been researched for over a century (Bruce, 1910), the molecular mechanisms behind the phenomenon are still poorly understood. One proposed explanation is referred to as the complementation model, which argues that parent inbred lines whose progeny exhibit heterosis possess genotypes which are complementary for deleterious alleles. Recent advances in genomics have yielded new technologies which can further contribute to the ongoing debate of how heterosis operates. In this paper, we propose an analysis which aims to determine which genes are being transcribed in a given genotype using next-generation sequencing data for mRNA quantification (RNA-Seq). We develop a hierarchical latent class model that can be used to predict whether or not each gene is transcriptionally active in each of three genotypes (the two parents and the F\textsubscript{1} cross). This enables us to identify genes which exhibit possible expression complementation under a dominance model, where F\textsubscript{1} progeny with a transcriptionally active allele from one parent and a transcriptionally inactive allele from the other parent for a given gene exhibits expression. We evaluate the efficacy of this approach using simulated data, and then fit the model to real data involving the B73 and Mo17 inbred lines of maize (Zea mays). We show this model leads to reasonable classifications of transcriptional activity for specific genes.
4.1 Introduction

In recent years, high-throughput next-generation sequencing (NGS) technologies for DNA have been developed using massively parallel sequencing technology. The application of NGS for quantifying mRNA transcripts, referred to as RNA-Seq, can be used to evaluate the relative expression levels of particular genes by mapping mRNA fragments to a target genome (Wang et al., 2009). Unlike microarray technology, which quantifies expression via continuous-valued fluorescence intensities, RNA-Seq processing protocols result in an integer-type output, which reflects the total read count mapped to each transcript. As such, unique statistical applications must be developed to address the resulting digital nature of RNA-Seq data.

One possible application of this new technology is in the field of plant breeding, where the molecular mechanisms of heterosis, or hybrid vigor, are of great interest. Inbred cultivars of particular species, such as maize (Zea mays), are traditionally crossbred to create F₁ hybrid cultivars. These hybrids often possess phenotypes which are more agriculturally advantageous than either inbred parent line. While currently the mechanisms of heterosis are poorly understood (Birchler et al., 2003), one possible explanation at the molecular level is genetic complementation. First proposed in 1910 (Bruce), the model argues that since completely inbred lines possess genome-wide homozygosity, they are susceptible to the effects of deleterious allele homozygosity. If we consider a metabolic pathway, homozygosity of deleterious alleles may disrupt the production of the end product. Parents which are homozygous for deleterious alleles in genes within the pathway can yield offspring which recover the functionality of the pathway if the two sets of properly working genes are complementary. Figure 4.1 illustrates this principle in a simple fashion, where each parent has a functioning set of alleles in one gene that the opposing parent does not. If two particular inbred lines have highly complementary genomes in this respect, it is possible that the resulting F₁ generation exhibits hybrid vigor.

While there are many reasons an allele may be nonfunctional, one possible explanation is that the gene is simply not transcribed. Let us define the “transcriptional state” of a gene for a given genotype to be the binary condition of whether or not a gene is transcribed. This leads to an absence/presence model for gene expression, where we consider the transcriptional state
Figure 4.1  Basic representation of the genetic concept of complementation
of a gene to be “on” or “off.” RNA-Seq is a natural tool to use in investigating this hypothesis, since transcripts are directly sequenced and aligned to known exon sequences. However, there are many complications that arise in trying to ascribe probability statements about the transcriptional activity of a gene using RNA-Seq data. The number of sequenced reads obtained from each observation is a subset, and does not represent an exhaustive identification of every mRNA transcript. Referred to as the “sequencing depth” of a given sample, this subsampling can result in issues of nondetection for genes that generate a relatively rare number of transcripts. Thus, a gene which is actually transcriptionally active can appear to be off. Similarly, reads can also be misappropriated to a specific gene due to improper sequencing or misalignment with the reference genome. If such misappropriation occurs for a transcriptionally inactive gene, it will appear to be on when it is not. These complications make the classification of transcriptional activity for genes a nontrivial procedure.

By analyzing RNA-Seq data of two inbred cultivars and the F\textsubscript{1} generation and developing a probabilistic model of gene activity, we can estimate the subset of genes which are transcriptionally active in each genotype. The same method could be applied to a collection of inbred lines to identify pairs of lines with complementary alleles that may be expected to exhibit heterosis when crossed. Such a molecular diagnostic for heterosis would be greatly beneficial to the plant breeding community because it has been found that genetically divergent breeding pairs often present greater heterosis than those that are genetically more similar (Hallauer and Miranda Filho, 1981).

Current methods for predicting whether or not a gene is transcribed use some type of count thresholding criterion (Ramsköld et al., 2009) or test each gene for expression significantly greater than an estimated background expression level due to sequencing error (Li et al., 2010). As an alternative to these approaches, we aim to develop a fully defined probabilistic model for classifying genes as transcriptionally active or inactive.

In this paper, we assume a negative binomial log-linear model for RNA-Seq read counts to develop a statistical model for simultaneously classifying the transcriptional activity of two inbred lines and their reciprocal F\textsubscript{1} crosses. In Section 4.2, we describe a motivating example for our research involving a maize crop breeding experiment, and formally define our data model.
for the RNA-Seq read counts. We also describe a Bayesian approach to fitting our model using Markov Chain Monte Carlo methods. In Section 4.3, we apply our model to the example RNA-Seq data, and evaluate the performance of this method using simulated read counts based on these data in Section 4.4. We discuss the results of these analyses in Section 4.5, which show that our method produces reasonable results in spite of the small sample sizes used in the experiment. Our findings show that the vast majority of the classifications are consistent with our assumptions, and we identify a subset of genes which may contribute to complementation. Concluding thoughts and future research directions are discussed in Section 4.6.

4.2 Methods

4.2.1 Latent Class Definitions

Consider an experiment involving two inbred parental genotypes \((j = 1, 2)\) and a hybrid genotype \((j = 3)\) formed by crossing the two inbred parental genotypes. Suppose we have RNA-Seq data from \(K_j\) independent samples of genotype \(j\) \((j = 1, 2, 3)\). Let \(y_{ijk}\) represent the total number of mapped reads for the \(k^{th}\) replication of the \(i^{th}\) gene and \(j^{th}\) genotype, where \(i = 1\ldots g, j = 1, 2, 3, \) and \(k = 1, \ldots, K_j\). We denote the transcriptional state of gene \(i\) in genotype \(j\) by \(z_{ij}\), such that \(z_{ij} \in \{0, 1\}\) where 0 indicates that the gene is not being transcribed and 1 that the gene is transcriptionally active. With respect to the three unique genotypes, there are 8 possible combinations of the 3 binary states, which we refer to as the “transcriptional state vector” for gene \(i\), denoted by \(z_i = (z_{i1}, z_{i2}, z_{i3})\). The possible values for \(z_i\) are outlined in Table 4.2.1.

Through the mechanics of genetic heritability, the transcriptional state of the gene within an offspring genotype may be determined by those of its parent genotypes. Thus, \(z_{i3}\) may depend on \(z_{i1}\) and \(z_{i2}\). We can address this dependence by considering each state vector as its own unique class. Because \(z_i\) is not directly observable, we have a latent class (LC) model with 8 possible classes. This results in a notational simplification, which we can characterize by replacing the state vectors \(z_i\) with the numerical class numbers defined in Table 4.2.1. Let \(S_i\) denote the class number for gene \(i\), and let the marginal probability that any given gene exhibits
transcriptional state vector $s$ to be represented by $p_s$, where $s = 1, \ldots, 8$, and $p = (p_1, \ldots, p_8)$.

Let $Y$ represent all the observed count data. Then, via Bayes’ Theorem, it follows that

$$P(S_i = s \mid Y) = \frac{p_s P(Y \mid S_i = s)}{\sum_{t=1}^{8} p_t P(Y \mid S_i = t)}.$$  \hspace{1cm} (4.1)

Using (4.1), we can determine the conditional probability that a given gene exhibits the behavior associated with a particular state vector given the observed data, and use these probabilities to make inferences as to the nature of transcriptional activity across genotypes for each gene. In the simplest dominance model of heritability, it is assumed that any gene that is active in at least one parent will be active in offspring. Under this model, we would expect most genes to have one of four state vectors: $(0,0,0)$, $(0,1,1)$, $(1,0,1)$ and $(1,1,1)$. However, other state vectors may be possible.

### 4.2.2 Data Model

From (4.1), it is necessary that we define a model that allows the conditional probability $P(Y \mid S_i = s)$ to be determined. While the first data models developed for RNA-Seq data utilized the Poisson distribution to represent the read counts (Wang et al., 2010), research has shown that the strict mean-variance relationship of this probability distribution is insufficient to capture the overdispersion present in biological data (Langmead et al., 2010; Robinson and Smyth, 2007). A solution that has been developed is the use of the negative binomial%}

<table>
<thead>
<tr>
<th>State $S_i$</th>
<th>$z_i$</th>
<th>Inbred 1</th>
<th>Inbred 2</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0,0,0)</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>2</td>
<td>(1,0,0)</td>
<td>ON</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>3</td>
<td>(0,0,1)</td>
<td>OFF</td>
<td>OFF</td>
<td>ON</td>
</tr>
<tr>
<td>4</td>
<td>(1,0,1)</td>
<td>ON</td>
<td>OFF</td>
<td>ON</td>
</tr>
<tr>
<td>5</td>
<td>(0,1,0)</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>6</td>
<td>(0,1,1)</td>
<td>OFF</td>
<td>ON</td>
<td>ON</td>
</tr>
<tr>
<td>7</td>
<td>(1,1,0)</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
</tr>
<tr>
<td>8</td>
<td>(1,1,1)</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
</tr>
</tbody>
</table>

Table 4.1 Definition of all 8 possible state vectors for binary classification of transcriptional activity for the three genotypes.
(NB) distribution. A common data model for biological count data (Bliss and Fisher, 1953; Cox, 1989), the NB distribution can be derived by integrating over the Poisson with respect to a mixture distribution, whereby the rate parameter is modeled via the conjugate gamma distribution. This yields a count data model that allows for a flexible mean-variance relationship due to the inclusion of a dispersion parameter. The NB model is currently used by many RNA-Seq analysis methods, such as EdgeR (Robinson et al., 2010) and DESeq (Anders, 2010).

For our model, we use the parametric form \( NB(\mu, \phi) \), where \( \mu \) represents the mean and \( \phi \) is the dispersion parameter, with the resulting probability mass function:

\[
f(x \mid \mu, \phi) = \frac{\Gamma(x + \phi)}{x!\Gamma(\phi)} \left( \frac{\phi}{\phi + \mu} \right)^\phi \left( \frac{\mu}{\phi + \mu} \right)^x, x = 0, 1, 2, \ldots
\]  

(4.2)

where \( \Gamma \) is the Gamma function. Under this parameterization, the variance is given as \( \mu(1+\mu/\phi) \) such that the distribution reduces to a Poisson when \( \phi \) approaches infinity. We then assume the data model

\[
y_{ijk} \mid z_i, M_{jk}, \mu_{ij}, \lambda, \phi_i \sim NB(M_{jk} \ast (\mu_{ij} \ast z_{ij} + \lambda), \phi_i)
\]  

(4.3)

where \( M_{jk} \) represents the sequencing depth for observation \( k \) of genotype \( j \), \( \mu_{ij} \) is a relative expression level parameter for gene \( i \) and genotype \( j \), \( \lambda \) is a background false mapping rate parameter, and \( \phi_i \) is the dispersion parameter for gene \( i \). Furthermore, we assume \( \log(\mu_{ij}) = \eta_i + \epsilon_{ij} \), where the parameters \( \eta_i \) and \( \epsilon_{ij} \) respectively represent overall gene effects and gene-specific genotype effects for the mean read rate of genotype \( j \) and gene \( i \).

The sequencing depth \( M_{jk} \) represents the value by which all read counts from the \( k^{th} \) sample of the \( j^{th} \) genotype are proportional. Although the total number of reads for sample \( k \) of genotype \( j \) would be a natural choice for \( M_{jk} \), the total number of reads for a sample is often dominated by a few heavily expressed genes (Anders, 2010). Thus, for our purposes, we use the median count of reads within an expression profile.

We assume that \( \eta_i \) and \( \epsilon_{ij} \) terms are independent and random, such that \( \eta_i \sim N(\mu_\eta, \sigma_\eta) \) and \( \epsilon_{ij} \sim N(0, \sigma_{\epsilon}) \). We also consider an underlying distribution for the dispersion parameters \( \phi_i \), which we assume follows a gamma\( (\alpha_{\phi}, \beta_{\phi}) \), where the probability density of the gamma
distribution is defined as \( f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\alpha \beta} \) for \( x > 0 \). This follows previous work which has found that the application of negative binomial data models using a single-value dispersion approach may be too limiting (Anders, 2010).

### 4.2.3 Bayesian Approach

The negative binomial random effects model described above translates well to a Bayesian hierarchical model. Let \( Z = \{ z_1, \ldots, z_g \} \) represent the unobserved latent data, and \( \theta = \{ \eta_1, \ldots, \eta_g, \epsilon_1, \ldots, \epsilon_g, p_1, \ldots, p_8, \mu_\eta, \sigma_\eta, \sigma_\epsilon, \alpha_\phi, \beta_\phi, \lambda_e \} \) represent the data model parameters. Thus, \( \{ Y, Z \} \) represents the \emph{complete} augmented data. Note that \( p(Z \mid Y) \) is proportional to

\[
\int_{\Theta} p(Z \mid \omega, Y) p(\omega \mid Y) d\omega
\]  

where \( \Theta \) indicates the parameter space of \( \theta \). This integration in (4.4) is intractable, and an analytic solution to obtaining \( p(Z \mid Y) \) is not obtainable. However, \( p(Z \mid Y) \) and \( p(\theta \mid Y) \) can be simultaneously approximated through the use of data-augmented MCMC to obtain posterior samples from \( p(\theta, Z \mid Y) \). Similar in application to the EM algorithm (Dempster et al., 1977), data-augmented MCMC approaches to LC modeling rely upon a two-stage iterative algorithm which alternates between (1) imputing the missing latent variables and (2) sampling the rest of the model parameters. As Gelfand and Smith (1990) show, this iterative procedure results in a Markov chain with a stationary distribution which is equivalent to the posterior distribution of interest. For sufficiently large chain length, draws from \( p(\theta, Z \mid Y) \) approximate \( p(Z \mid Y) \) and \( p(\theta Y) \) (Feigelson et al., 2003). Thus, this approach will allow us to simultaneously characterize the posterior distribution of the parameters and transcriptional state vector status for each gene.

This method then requires specification of the distributions \( p(\theta \mid Y, Z) \) \( p(Z \mid Y, \theta) \). The former is proportional to

\[
\prod_{i=1}^{M} \prod_{j=1}^{3} \prod_{k=1}^{K_j} NB(y_{ijk} ; M_{ij}(z_{ij} + \lambda_e), \phi_i) \varphi(\eta_i ; \mu_\eta, \sigma_\eta) \varphi(\epsilon_{ij} ; 0, \sigma_\epsilon) \gamma(\phi_i ; \alpha_\phi, \beta_\phi) \pi(\theta^*) \quad (4.5)
\]

where \( \varphi(x ; \mu, \sigma) \) and \( \gamma(x ; \alpha, \beta) \) represent the probability density functions of the normal and gamma distributions, respectively, and \( \pi(\theta^*) \) represents the joint prior distribution on \( \theta^* = \)
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_\eta$</td>
<td>Normal(0,1000)</td>
</tr>
<tr>
<td>$\sigma_\eta$</td>
<td>uniform(0,100)</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>uniform(0,100)</td>
</tr>
<tr>
<td>$\alpha_\eta$</td>
<td>gamma(1,0.1)</td>
</tr>
<tr>
<td>$\beta_\eta$</td>
<td>gamma(1,0.1)</td>
</tr>
<tr>
<td>$p$</td>
<td>Dirichlet(1)</td>
</tr>
</tbody>
</table>

Table 4.2  Prior distribution assignments for hyperparameters of the read count model. Normal distribution is presented using the $N(\mu, \sigma^2)$ parameterization.

To obtain the conditional predictive posterior, $p(S_i = s \mid y_i, \theta)$, where $y_i$ is the vector of observed read counts for gene $i$ for all genotypes and replications. Let $z_{ij}(s)$ be the binary state of $z_{ij}$ defined by state vector assignment $S_i = s$. Then, it follows that

$$P(S_i = s \mid y_i, \theta) \propto \prod_{j=1}^{3} \prod_{k=1}^{K_i} f(y_{ijk}; M_{jk}(z_{ij}(s)\mu_{ij} + \lambda_{ij}), \phi_k).$$

(4.6)

Then, define the normalized conditional probability to be

$$\kappa_i(s) = \frac{P(S_i = s \mid y_i, \theta)p_s}{\sum_{t=1}^{8} P(S_i = t \mid y_i, \theta)p_t}.$$  

(4.7)

Let $\kappa_i = (\kappa_i(1), \ldots, \kappa_i(8))$. State vector assignment values from the posterior predictive distribution $p(S_i \mid y_i, \theta)$ can then be imputed using Multinomial(1, $\kappa_i$).

4.2.4 Model Fitting and Inference

We opt for an MCMC approach in sampling $Z$ and $\theta$ from their respective posterior distributions, using the data augmentation approach similar to that described by Tanner and Wong (1987). Table 4.2, specifies the prior distributions for all model hyperparameters. As per Gelman (2006), we adopt diffuse uniform priors on the variance parameters of the log-linear model of the mean ($\sigma_\eta$ and $\sigma_\epsilon$). The remaining choices for prior distributions were made to either take advantage of conjugacy where possible and/or impart non-informative prior distributions on the model parameters.
With respect to $\lambda_e$, we let this be a fixed, a priori specified value, which gives the investigator the ability to fit the data under differing assumptions of read count integrity. This is necessary in order to avoid issues of identifiability and label-switching, a common problem in MCMC applications to models with hidden states (Jasra et al., 2005). By fixing $\lambda_e$ a priori to a sufficiently small value, we ensure that it does not become problematically large within the sampling process to the point that convergence of the MCMC iterations to $p(\theta \mid Y)$ becomes prohibitively difficult.

The updating algorithm of the model consists of two stages: imputing the hidden latent classes $S = \{S_1, \ldots, S_g\}$, and sampling the rest of the model parameters $\theta$ via either Gibbs if conjugacy is present ($p, \mu_\eta, \alpha_\phi$) or via slice sampling (Neal, 2003). The $(n + 1)^{th}$ sample is derived as follows

- Conditional on the $n^{th}$ sample of $\theta$, $\theta^{(n)}$, sample each $S^{(n+1)}_i$ from a Multinomial(1, $\kappa_i^{(n)}$) to obtain $S^{(n+1)}$, which is equivalent to $Z^{(n+1)}$
- Given $Z^{(n+1)}$, sample each parameter in $\theta$ from $p(\theta \mid Z, Y)$ to obtain $\theta^{(n+1)}$
- Calculate $\kappa_i^{(n+1)}$ from $\theta^{(n+1)}$
- Repeat

These steps are simulated via JAGS (Plummer, 2011), an implementation of the BUGS statistical language (Gilks et al., 1994), in conjunction the R2jags R library (Su and Yajima, 2011). For latent class models, the appropriate selection of starting values is prudent in facilitating expedient convergence of the chains to the joint posterior distribution. For the purposes of generating initial values for the latent data $Z$, we used the naive classifier

$$z_{ij} = \begin{cases} 1 & \text{if } \sum_{k=1}^{K_j} (I(y_{ijk} > 0)) > K_j/2 \\ 0 & \text{otherwise} \end{cases} \quad (4.8)$$

The determination of the $z_{ij}$ simultaneously determines $S_i$. The observations which were classified as transcriptionally active were then used to generate initial values for the $\eta_i$ and $\epsilon_{ij}$. 

We used the Markov chain traces of the state vector for each gene to make inferential statements about its class membership, such that \( \hat{\kappa}_i(s) = \frac{1}{N} \sum_{n=1}^{N} I(S_i^{(n)} = s) \), where \( N \) is the number of iterations used for posterior inference after sufficient burn-in and thinning have been applied to the chains and \( I(S_i^{(n)} = s) \) is the indicator function that returns a value of 1 if the \( n^{th} \) value in the trace of \( S_i \) is equal to \( s \) and 0 otherwise. For the purposes of classification, we define \( \hat{S}_i = \arg\max_{s \in S} \hat{\kappa}_i(s) \).

4.3 Application

A recently conducted experiment examined the expression profiles of maize (Zea mays) for the purposes of exploring the molecular components of heterosis. A total of four genotypes were examined: inbred lines B73 and Mo17, and their reciprocal F\(_1\) crosses, B73\times Mo17 and Mo17\times B73, where the first line indicated in the crosses indicates maternal lineage. Four biological replicate samples were separately measured for each genotype. Samples of RNA were obtained from the primary root tissues. The samples were then sequenced using an Illumina Genome Analyzer II (www.illumina.com).

We apply our method to this maize RNA-Seq dataset. Let \( j = 1 \) indicate the B73 inbred line, \( j = 2 \) indicate the Mo17 inbred line, and \( j = 3 \) indicate the F\(_1\) intercross, regardless of maternal lineage. This provides sample sizes of \( K_1 = K_2 = 4 \) and \( K_3 = 8 \).

Mapping of the sequenced sample reads was conducted using the B73 reference genome (Schnable et al., 2009), allowing for a maximum of two mismatches per read. Reads which were redundantly “stacked”, such that map location and direction were identical, were removed from the analysis. Alignment of the reads to the reference genome was determined using NOVOALIGN software (www.novocraft.com). This process resulted 16 RNA-Seq read count profiles, each consisting of counts for 32540 unique targets ranging in value from 0 to 6117.

To select an appropriate value \( \lambda_e \) in our model fit, we used the naive classification algorithm defined previously to generate starting state vectors for each gene, and approximated \( \lambda_e \) using all of the observations that were initially classified as corresponding to transcriptionally inactive genes. For our model fit, we select 0.001 as our value for \( \lambda_e \) based upon the observed maize data.
Table 4.3 Posterior means and 95% credible intervals for $\theta^*$ applied to the experimental data setting $\lambda_c = 0.001$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>2.5 Percentile</th>
<th>97.5 Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_\eta$</td>
<td>0.14787</td>
<td>0.12160</td>
<td>0.17305</td>
</tr>
<tr>
<td>$\sigma_\eta$</td>
<td>0.22610</td>
<td>0.22164</td>
<td>0.23081</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>0.18252</td>
<td>0.17983</td>
<td>0.18520</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0.17500</td>
<td>0.17080</td>
<td>0.17934</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.00004</td>
<td>&lt;0.00001</td>
<td>0.00014</td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.00006</td>
<td>&gt;0.00001</td>
<td>0.00021</td>
</tr>
<tr>
<td>$p_4$</td>
<td>0.01422</td>
<td>0.01286</td>
<td>0.01560</td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.00176</td>
<td>0.00121</td>
<td>0.00240</td>
</tr>
<tr>
<td>$p_6$</td>
<td>0.02402</td>
<td>0.02320</td>
<td>0.02593</td>
</tr>
<tr>
<td>$p_7$</td>
<td>0.00004</td>
<td>&lt;0.00001</td>
<td>0.00014</td>
</tr>
<tr>
<td>$p_8$</td>
<td>0.78487</td>
<td>0.78031</td>
<td>0.78930</td>
</tr>
<tr>
<td>$\alpha_\phi$</td>
<td>0.79352</td>
<td>0.77913</td>
<td>0.80904</td>
</tr>
<tr>
<td>$\beta_\phi$</td>
<td>0.01805</td>
<td>0.01753</td>
<td>0.01855</td>
</tr>
</tbody>
</table>

Table 4.4 Predicted number of genes exhibiting each state vector based upon $\hat{Z}_i$.

<table>
<thead>
<tr>
<th>State</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>5792</td>
<td>0</td>
<td>1</td>
<td>409</td>
<td>47</td>
<td>708</td>
<td>0</td>
<td>25583</td>
</tr>
</tbody>
</table>

To obtain posterior estimates, we generated two chains at a total of 15,000 iterations each, burning the first 10,000 and thinning the remaining iterations by saving every 5th draw. Convergence of the Markov chains was determined using the coda package (Plummer et al., 2010) in R. The summary statistics of the posterior distributions of the hyperparameters in $\theta$ can be found in Table 4.3.

The quantity of classifications via use of $\hat{Z}_i$ can be found in Table 4.4. A total of 1117 genes exhibit the state vectors of interest mentioned in the beginning of this paper, whereby one parent and the hybrid exhibit transcriptional activity for a gene which is not transcribed by the remaining parent, breaking down into 409 for (1,0,1) and 708 for (0,1,1). Of the 32540 transcript sequences in the reference genome, 8484 involved non-zero estimated probabilities for multiple state vectors, the vast majority of which were classified as (0,0,0).

To examine the assumption that B73×Mo17 and Mo17×B73 can be represented by a single
genotype, we also fit a model where we relaxed this restriction and treated the reciprocal F₁ crosses as two unique genotypes. This leads to a total of four genotypes with four observations each, as well as 16 possible state vectors. Fitting the expanded model in the same way as the three genotype approach, 32440/32540 of the state classifications were consistent with those of the three genotype model. Of these classifications, only 49 genes were classified in a manner that the transcriptional status of one reciprocal F₁ cross did not match that of the other. However, if the identification of such cases is a research goal, this expanded model approach would be useful. Otherwise, the three genotype model performs quite well and yields reasonable results.

4.4 Simulation Analysis

The model defined in Section 4.2 is admittedly simple and clearly not an exact representation of the data generating mechanism it intends to model. While simulation studies where the data is generated from our model perform very well, the assumptions of our model may not be satisfied by the actual data. To evaluate the efficacy of this approach without any distributional assumptions, we simulated pseudo-data based upon the results from the fitting of the model to the Zea mays data.

To generate the pseudo-data, a total of $R = 1000$ genes were randomly selected from the 32540 total targets in the experimental gene profiles. Let $y_{rjk}$ represent the $k^{th}$ observation of the $j^{th}$ genotype for the $r^{th}$ selected gene, where $r = 1, \ldots, R$, and $\hat{z}_r = (\hat{z}_{r1}, \hat{z}_{r2}, \hat{z}_{r3})$ its corresponding state vector classification. In order to extend this approach for modeling falsely mapped reads, we tabulated all of the observations corresponding to $\hat{z}_{ij} = 0$ (Table 4.4), and defined $\Lambda$ to be the normalized vector of these counts such that its elements sum to 1. To simulate $\tilde{y}_{rjk}$, the pseudo-data observations, we use the following algorithm:

<table>
<thead>
<tr>
<th>Reads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>93388</td>
<td>3882</td>
<td>381</td>
<td>54</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>
1. If $\hat{z}_{rj} = 0$, simulate $y_{rjk}^*$ such that $y_{rjk}^* \sim \text{Multinomial}(1, \Lambda)$

2. If $\hat{z}_{rj} = 1$, calculate the empirical summary statistics $m_{rj} = \bar{y}_{rj}$ and $v_{rj} = \frac{1}{K_j-1} \sum_{k=1}^{K_j} K_j (y_{ijk} - \bar{y}_{rjk})^2$

3. Convert these estimates into gamma distribution parameters, such that $\alpha_{rj} = \frac{m_{rj}^2}{v_{rj}}$ and $\beta_{rj} = \frac{v_{rj}}{m_{rj}}$

4. Simulate $y_{rjk}^*$ such that $y_{rjk}^*$ is drawn from $\text{gamma}(\alpha_{rj}, \beta_{rj})$ and rounded to the nearest integer

To examine the impact of $\lambda_e$ selection, we fit the model using a variety of parameter values, conducting model fittings using six separate specifications of $\lambda_e$ (0.000001, 0.00001, 0.0001, 0.001, 0.01, and 0.1). Two chains of 10,000 iterations each were generated for each model fit, with the first 5000 iterations used as a burn-in sample. The remaining 5000 iterations were then thinned such that every fifth value was saved in the trace to avoid the effects of autocorrelation, yielding a total of 1000 iterations for posterior inference. Convergence was determined in the same fashion described in Section 4.3. The posterior means and standard deviations for model hyperparameters in $\theta$ for each of the six fits of the parameters can be found in Table 4.7.

Define $\hat{z}_r^*$ to be the state vector classifications for the pseudo-data model fits. From these simulations we provide the number of correct classifications (matching $\hat{z}_r$), as well as a breakdown of assignments for each state vector, in Table 4.8. We also provide a contingency table comparing the true versus predicted state classifications for the simulation using $\lambda_e = 0.01$ (Table 4.9).

Similar contingency tables for the remaining five simulations can be found in the appendix.

### 4.5 Discussion

Using MCMC, we have constructed a method for classifying transcriptional activity in a breeding experiment using RNA-Seq data. The results from our application indicate that a vast majority of the genes are classified as state vector categories consistent with a dominance genetic model, with only 48 total genes deviating from this family of classifications (patterns
Table 4.7 Table of posterior means and 95% credible intervals of the model parameters for separate model fits using a grid of $\lambda_e$ values. The columns of results corresponding to model fit using the true value of $\lambda_e$ are indicated in bold.
<table>
<thead>
<tr>
<th>$\lambda_e$</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-06</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>862</td>
<td>896</td>
</tr>
<tr>
<td>1e-05</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>32</td>
<td>0</td>
<td>860</td>
<td>894</td>
</tr>
<tr>
<td>1e-04</td>
<td>148</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>2</td>
<td>31</td>
<td>0</td>
<td>799</td>
<td>955</td>
</tr>
<tr>
<td>1e-03</td>
<td>186</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>6</td>
<td>33</td>
<td>0</td>
<td>752</td>
<td>962</td>
</tr>
<tr>
<td>1e-02</td>
<td>72</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>7</td>
<td>31</td>
<td>0</td>
<td>866</td>
<td>768</td>
</tr>
<tr>
<td>1e-01</td>
<td>143</td>
<td>1</td>
<td>0</td>
<td>26</td>
<td>6</td>
<td>19</td>
<td>0</td>
<td>805</td>
<td>668</td>
</tr>
<tr>
<td>TRUE</td>
<td>175</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>776</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.8 State vector classification results $\hat{Z}_g^*$ from simulation analysis along with true values derived from $\hat{Z}_g$

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>156</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>49</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>709</td>
</tr>
</tbody>
</table>

Table 4.9 Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.01$
(0,0,1) and (0,1,0)). We were also able to identify 409 (1,0,1) and 708 (0,1,1) candidate genes that could partly explain the heterotic progeny of B73 and Mo17 via the complementation model for heterosis.

From our simulations in Section 4.4, we see that the method performs quite well when $\lambda_e$ is in the vicinity of its true value, even when the sample sizes are as small as those used in our case study and the data generating mechanism differs from the assumed model. When the $\lambda_e$ value used in the model fit is identical to that used in the case study, 962 out of 1000 genes were correctly classified. The model begins to do poorly with respect to proper classification when the $\lambda_e$ parameter is set to be too high, with classification accuracy dropping to 668/1000 when $\lambda_e$ is 100 times too large. This is due to complications of identifiability, where $\lambda_e$ is set high enough that the probabilities that any relatively low expressed gene is on or off are very similar. We see this when we examine the number of genes which exhibit positive estimated probability for multiple state classifications relative to the value of $\lambda_e$ used in the model fit. The number of these cases increases as $\lambda_e$ increases, with 392 such observations when $\lambda_e = 0.1$ compared to 241 when $\lambda_e$ is set to 1e-06. Thus, it may be prudent to use a relatively small value for $\lambda_e$ when applying this method, since larger values of $\lambda_e$ can result in difficulty discerning (0,0,0) and (1,1,1) transcriptional activity vector patterns.

One concern we had using this approach was the implication of $z_{ij} = 0$ negating the roles of specific $\epsilon_{ij}$ and possibly $\eta_i$ in the model likelihood using this approach. Even though particular observations are classified as transcriptionally inactive during imputation of $Z$, the mean-read parameters that comprise $\mu_i$ ($\epsilon_{ij}$ and $\eta_i$) are still simulated during the MCMC process. This is particularly relevant to genes classified as (0,0,0), since none of the corresponding mean-read parameters are taken into account in the likelihood. We recreated our simulation in Section 4.4 by instead modeling the data directly from the probability model described in Section 4.2, thus requiring us to define every parameter value prior to simulation. We defined all mean-value parameters ($\eta_i$ and $\epsilon_{ij}$) prior to state vector assignment, generating them from their respective distributions. Our model fits show that all of our posterior distributions are centered very closely to their true values, indicating that biased estimates do not seem to be a great concern in the application of this model.
Another concern is the admittedly ad hoc nature of the selection of $\lambda_e$. In reality, this modeling approach is an expansion of the thresholding concept described at the beginning of this paper, in which a model parameter and probability distribution are used in lieu of any strict universal count threshold. While the use of the naive classifier to motivate the value selected for $\lambda_e$ involves a data-driven approach in our case-study, it is not a very statistically motivated strategy. It would be prudent to make use of additional biological information, such as the methods used in Li et al. (2010), to obtain a valid estimate for $\lambda_e$.

While the design of this model is specific to the research question at hand, it can easily be modified to address the overall issue of transcriptionally inactive genes in RNA-Seq data analysis and improve existing approaches. For example, in popular methods used for the purposes of detecting differential expression (Anders, 2010; Robinson et al., 2010), the issue of nontranscribed genes is not taken into account. Consequently, estimates of various sources of variance may be destabilized by transcriptional inactivity, affecting detection of differential expression.

### 4.6 Conclusions

We have produced a novel statistical model for the classification of transcriptional activity across multiple genotypes for the purposes of detecting possible genes involved in complementation. This model can be applied to a variety of breeding experiments to fully explore whether or not this hypothesis is viable, tracing the functions of these particular genes and comparing the results of heterotic against non-heterotic breeding pairs. Moreover, the probability model can be applied to a variety of research questions outside the scope of original purpose.

While we have shown via simulation that the model performs well under small samples and adequate approximation of $\lambda_e$, there is significant room for improvement in the use of additional information. Further research should be done to determine how to make use of such information as exon length, gene function, and other biological information that assist in fully characterizing the gene-specific contributions to mean read counts.
APPENDIX

4.A JAGS Model Code

model{
    tau.eta <- sig.eta^-2
    tau.eps <- sig.eps^-2
    for(i in 1:N){
        state[i] ~ dcat(p[])
        phi[i] ~ dgamma(alpha.phi,beta.phi)
        eta[i] ~ dnorm(mu.eta,tau.eta)
        for(j in 1:G){
            eps[i,j] ~ dnorm(0,tau.eps) #Genotype random effect
            l.state[i,j]<-state.mat[state[i],j]
            log(mu[i,j]) <- eta[i] + eps[i,j] #Log-linear regression
        }
        for(k in 1:16){
            y[i,k] ~ dnegbin((phi[i]/(phi[i]+M[k]*
            (mu[i,R[k]]*l.state[i,R[k]]+lambda.e))),phi[i])
        }
    }
    #Priors
    p[1:8] ~ ddirch(a[])
    mu.eta ~ dnorm(0,0.001)
    sig.eta ~ dunif(0,100)
    sig.eps ~ dunif(0,100)
    alpha.phi ~ dgamma(1,0.1)
    beta.phi ~ dgamma(1,0.1)
}
### 4.B Simulation Contingency Tables

#### Table 4.10
Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.000001$

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>767</td>
</tr>
</tbody>
</table>

#### Table 4.11
Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.00001$

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>765</td>
</tr>
</tbody>
</table>
Table 4.12  Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.0001$

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>765</td>
</tr>
</tbody>
</table>

Table 4.13  Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.001$

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>174</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>745</td>
</tr>
</tbody>
</table>
### State Vector Assignments

<table>
<thead>
<tr>
<th>TRUE</th>
<th>(0,0,0)</th>
<th>(1,0,0)</th>
<th>(0,0,1)</th>
<th>(1,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>167</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>108</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>6</td>
<td>12</td>
<td>10</td>
<td>638</td>
</tr>
</tbody>
</table>

Table 4.14  Two-way contingency table comparing the true transcriptional vector states (rows) against the predicted states (columns) for simulation using $\lambda_e = 0.1$

### Bibliography


G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R.,
Wing, R. A., and Wilson, R. K. (2009). The b73 maize genome: Complexity, diversity, and

0.02-17.


CHAPTER 5. GENERAL CONCLUSIONS

The research in this dissertation has sought to advance statistical knowledge and methodology regarding estimation of FDR in microarray experiments (Chapters 2 and 3) and develop novel statistical models for applications involving RNA-Seq data analysis (Chapter 4). These methods were evaluated with sophisticated simulation techniques that aimed to preserve biologically realistic correlation structures and accurately portray gene expression data.

We specifically have shown that the underlying assumptions of popular approaches to multiple testing procedures may be inappropriate for application to microarray experiments, resulting in biased estimates related to FDR. In Chapter 2, we addressed the non-normality of expression values and the small sample sizes that are typical of microarray experiments and used nonparametric approaches in the adaptation of an existing procedure. This showed marked improvement over the original method, leading to more accurate and less variable estimates of the true null proportion, $\pi_0$. This work not only argues for the use of permutation tests for exploratory microarray data analysis under small samples, but exposes a distinct distributional pattern for the empirical null $p$-values when $t$-tests are applied for gene-wise hypothesis testing. Other multiple testing procedures with similar conditions should take this finding into consideration if the assumption of a uniform empirical null is necessary.

The work in Chapter 3 identified a serious issue in certain significance threshold selection strategies with respect to violations of the assumption of the independence of hypothesis tests. We determined that modern approaches to estimating FDR actually result in these estimates being negatively correlated with the true false discovery proportion when a fixed number of genes is repeatedly rejected in exploratory microarray data analysis. This result could lead to these estimators yielding very liberal estimates of FDR, especially when publication bias is taken into consideration. As such, we advise great caution in any type of meta-analysis that is
dependent upon the findings of microarray analysis studies.

In Chapter 4, we presented a unique approach to simultaneously determine the transcriptional activity of genes across multiple genotypes using MCMC. This model may become quite useful in exploring the molecular components of heterosis, identifying genes of interest in this regard. It may also be useful for models designed for the detection of differential expression which currently do not take transcriptional activity into account in their application.

5.1 Future Research Directions

An obvious expansion of the work completed in Chapter 2 would be the extension of the adapted estimator of $\pi_0$ to include a data-driven shape for the null $p$-values rather than relying upon the assumption of uniformity. Addressed in Chapter 3, our findings led us to argue that dependence among genes in expression analysis distorts the empirical null $p$-value distribution. This phenomenon could lead to poor estimation of the true null proportion if the uniformity assumption is retained.

It would also be prudent to explore the role of sample size on the impact of the original semi-parametric estimator with an additional simulation study using a base data set with a much larger number of observed expression profiles. In this manner, a set of simulations scaled upwards in sample size could isolate the impact of the number of expression profiles in a sample on the performance differences between the semi-parametric and adapted nonparametric estimators of $\pi_0$. The use of multiple data sets on various platforms would also help validate the findings of this research.

The research in Chapter 3 exposed a negatively correlated relationship with common estimators of FDR and $Q$, the true realized proportion of Type I errors. Despite the investigation of one procedure that addresses the complications of dependence in microarray data, examining the behavior of alternative methods would also be of use. It would also be of interest to explore the possibility of taking advantage of the distributional patterns observed in the empirical null $p$-value distributions from the simulation data in the development of a new estimator of FDR or $\pi_0$. Similar in application to Efron’s empirical null approach, we could use a histogram of the $p$-values to infer the shape of the null distribution based upon the shape of the empirical
density near 1. This could be a source for expanding the estimator of \( \pi_0 \) discussed in Chapter 2.

In Chapter 4, we developed a simple model for RNA-Seq read count data. There are multiple factors that affect the rate of transcription of a gene, including exon length and gene function. Additional information could also be used to address gene-specific false read rates, such as genes which exhibit high sequence-relatedness with other genes. Careful distillation of this information could strengthen the model and result in more accurate inferential statements about transcriptional activity.

For the purposes of exploring the complementation theory of heterosis, application of this model to several other breeding experiments would assist in determining the genetic components of this phenomenon.