New statistical methods in bioinformatics: for the analysis of quantitative trait loci (QTL), microarrays, and eQTLs

Rhonda DeCook
Iowa State University

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

Part of the Bioinformatics Commons, and the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1504

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
New statistical methods in Bioinformatics
for the analysis of
quantitative trait loci (QTL),
microarrays, and eQTLs

by

Rhonda DeCook

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dan Nettleton, Major Professor
   Alicia Carriquiry
   Ranjan Maitra
   Stephen Vardeman
   Stephen Howell

Iowa State University
Ames, Iowa
2006

Copyright © Rhonda DeCook, 2006. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI

UMI Microform 3229065
Copyright 2006 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346
This is to certify that the doctoral dissertation of

Rhonda DeCook

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................. vii

ABSTRACT ............................................................... 1

GENERAL INTRODUCTION ............................................. 3

1 QTL Analysis .......................................................... 4
   1.1 Background ....................................................... 4
   1.2 Genotyping Marker Loci ........................................ 7
   1.3 Experimental Populations in Interval Mapping ............... 8
   1.4 Expected Genotypes at Unobserved Loci ....................... 9

2 Affymetrix GeneChip Array Technology .......................... 15

3 Multiple Testing Adjustment Procedures ........................ 18
   3.1 For QTL Analysis ................................................ 19
   3.2 For Microarray Analysis ........................................ 20

4 Dissertation organization ......................................... 21

5 References ........................................................... 23

6 Glossary .............................................................. 25

CHAPTER 1. QTL DETECTION FOR ZERO-INFLATED POISSON

TRAITS ................................................................. 27

1 Introduction ......................................................... 28

2 Method for Detecting ZIP QTLs ................................... 31
   2.1 ZIP Distribution and Notation ................................. 31
vi

4 Discussion ......................................................... 126
5 References ......................................................... 129

CONCLUSIONS AND FUTURE WORK .............................. 133
1 Conclusions ......................................................... 133
2 Future Research .................................................... 135
ACKNOWLEDGEMENTS

For the many hours my major professor, Dr. Dan Nettleton, has spent giving me guidance and advice, I am incredibly grateful. Through years of weekly meetings and talks with Dan, I have not only broadened my statistical knowledge, but I feel I have observed how to be an encouraging and productive professor, something I hope to carry with me into my own professional career.

I would also like to thank Dr. Alicia Carriquiry for her endless optimism and energy that she gives so freely to her students. I had the pleasure of working with Alicia during my time in the Preparing Future Faculty program, and I hope that our friendship will continue to grow in the years to come.

Finally, I thank my family and friends for all their support. I feel incredibly lucky to have had such a wonderful group of people to lean on while finishing my graduate work. But above all, I thank my husband Howard for the patience, understanding, and generosity that he has shown me through the years.
ABSTRACT

This thesis focuses on new statistical methods in the area of bioinformatics which uses computers and statistics to solve biological problems. The first study discusses a method for detecting a quantitative trait locus (QTL) when the trait of interest has a zero-inflated Poisson (ZIP) distribution. Though existing methods based on normality may be reasonably applied to some ZIP distributions, the characteristics of other ZIP distributions make such an application inappropriate. In this study, we propose a QTL detection method, appropriate for any ZIP trait, that utilizes the EM algorithm to compute maximum likelihood estimates for the ZIP parameters. We compare our method to an existing non-parametric approach using simulation. The method is illustrated using QTL data collected on two ecotypes of the Arabidopsis thaliana plant where the trait of interest is shoot count.

The second study discusses a method to detect differentially expressed genes in an unreplicated multiple-treatment microarray timecourse experiment. In a two-sample setting, differential expression is well defined as non-equal means, but in the present setting, there are numerous expression patterns that may qualify as differential expression. By defining differential expression as any pattern other than a concurrent flat line over time for all treatment groups, we propose a method that allows the researcher to test the null hypothesis of no differential expression at every gene. This method provides the researcher with a list of significant genes, an associated false discovery rate for that list, and a 'best model' choice for every gene. The model choice component is relevant because the alternative hypothesis of differential expression does not dictate
one specific alternative expression pattern. In fact, in this type of experiment, there are many possible expression patterns of interest to the researcher. Using simulations, we provide information on the specificity and sensitivity of detection under a variety of true expression patterns using receiver operating characteristic curves. The method is illustrated using an *Arabidopsis thaliana* microarray experiment with five time points and three treatment groups.

The third study discusses a new type of analysis, called eQTL analysis. This analysis brings together the methods of microarray and QTL analyses in order to detect locations on the genome that control gene expression. These controlling loci are called expression QTL, or eQTL. Locating eQTL can help researchers uncover complex networks in biological systems. For data sets containing thousands of genes and hundreds of markers, there are potentially millions of tests of interest. Besides the difficulty involved in sifting through millions of tests, the issues previously discussed in QTL analysis and microarray analysis are also present here. For each of these types of analysis, a different multiple-testing adjustment is utilized. The adjustment for a QTL analysis accounts for the strong correlation between tests at consecutive markers, while the adjustment for a microarray experiment accounts for the block-structure correlation between gene expression values in an individual arising from gene coregulation and other gene-to-gene relationships. Both of these types of multiple testing must be considered when determining statistical significance of eQTLs. The method is illustrated using an *Arabidopsis thaliana* eQTL experiment with 22,787 genes and 288 markers.
GENERAL INTRODUCTION

The field of statistics is dynamic in that new methods are continually being developed. These new methods arise from a variety of motivations. For example, the discovery of better estimators, more efficient algorithms, or advances in technology can all motivate new methods. The new methods in this thesis are motivated by the new technology of microarrays, and the search for a more appropriate method in the area of quantitative trait locus (QTL) analysis. Specifically, we propose a method to locate a QTL when the trait of interest follows a zero-inflated Poisson (ZIP) distribution. Much research has been done for normally distributed QTL traits, but traits following other distributions have received much less attention. Applying existing QTL analysis methods, based on normality, to a ZIP trait provides less information than our proposed method, and in many cases, is even inappropriate. Second, we propose a method to detect differentially expressed genes in an unreplicated multiple-treatment microarray timecourse experiment. The analysis of this type of microarray data challenges the statistician because it must incorporate a model selection procedure and a differential expression testing procedure at each of thousands of genes simultaneously. Finally, QTL analysis and microarray analysis are joined together to detect eQTL, or expression-level quantitative trait loci. In this emerging area of research, gene expression is considered the quantitative trait of interest. Thus, there are thousands of traits on which to perform QTL analysis. Because any gene may be controlled by any location, there are potentially millions of gene-to-locus tests of interest. Just sifting through the vast number of tests is difficult, but another layer of difficulty is added to the analysis due to the strong
correlation in tests performed on consecutive chromosome locations on the genome.

The new methods proposed in this dissertation are all applied to data from the fields of genetics and biology. This introduction includes some background information on QTL analysis, microarrays, and related terminology. We also include some background on statistical procedures to adjust for multiple testing because procedures of this type are utilized for both QTL and microarray analyses.

1 QTL Analysis

1.1 Background

The intentional breeding of organisms to produce offspring with desirable characteristics, or traits, is an old and common practice. Successful breeding relies on the existence of an association between an observed trait and the genetic composition of the organism. The simplest association occurs when a trait is linked to a single locus on the genome. This describes a trait with single-locus control, otherwise known as a simple trait. For such a trait, we often find that much of the variability in the observed trait can be explained by the controlling locus genotype. More complex associations exist when a trait is linked to many loci on the genome. For these associations, the genotypes at the given loci work together to control the observed trait. This type of trait is called a complex trait. It is the discovery of either type of these associations that is the goal when performing quantitative trait locus (QTL) analysis.

With advances in technology, we can now determine the genotype (i.e. specific DNA sequence) an individual possesses at particular locations throughout the genome. QTL analysis applies statistical modeling to a data set containing information on location genotypes and an observed trait to search for potential controlling loci of the trait. The term locus in QTL refers to a section of double-stranded deoxyribonucleic acid (DNA),
containing the gene controlling the trait of interest. A gene is defined here as DNA that encodes for a protein or any RNA used in an organism's biological system. Many organisms, including humans, are diploid and their nucleus contains two copies of each gene, except for the genes in human males residing on the sex chromosomes. Each copy of a gene can take on any one of a number of DNA codings called alleles. Thus, at any locus on the genome, a pair of alleles defines the genotype at a gene. Letters are often used to represent alleles, such as B or b, and the combination of alleles define the genotype, such as BB, Bb, or bb at a given locus. When alleles differ at a given locus in a population, the locus is said to be polymorphic. Such loci are the subject of population genetic studies (Lynch and Walsh, 1998). We can also extend the concept of polymorphisms to DNA not in genes, sometimes referred to as non-coding regions. Thus, with infinite time and money, every DNA polymorphism in the genome could be genotyped. In reality, the number of known polymorphisms along the genome of an organism may be small, often in the hundreds. However, in some organisms, including man, thousands of polymorphisms have been identified. We refer to the DNA polymorphisms used in our QTL analysis as markers (see next section).

As the ability to genotype individuals progresses, so does our ability to associate traits of interest with specific genes on the genome. A QTL is a section of DNA containing the gene associated with a quantitative trait. When a QTL exists, the trait of interest is associated with the genotype at the controlling locus. As each QTL genotype is associated with a distinct trait distribution, we model the marginal distribution of observed traits as a finite mixture distribution. In this mixture model, each QTL genotype is associated with one component of the mixture. QTL experiments are often designed using experimental organisms or populations with very few possible genotypes at each genome position. For example, by backcrossing the offspring of parental inbred lines with one of the parents we can produce organisms with only two possible genotypes at each locus. Fortunately, the methods applied to such experimental populations
also apply to populations with more than two genotypes at each locus, which is more common in nature.

Traditional QTL analysis provides us with a likelihood-ratio (LR) test statistic for each genome position tested. The null hypothesis is that the tested location is not a QTL. The number of locations tested depends on how many locations the researcher chooses to test. If the researcher chooses to test for a QTL only at observed locations (markers), they can perform single marker analysis. But if the researcher wants to test for QTL at more locations, they can use interval mapping to test for QTL at unobserved locations between markers. A researcher may choose to apply interval mapping when the data set has a relatively sparse set of markers. Testing at more locations provides greater precision in detecting the QTL, but increases computation time. The information in the set of LR tests, one test for each location, is usually summarized using a plot showing genome location on the horizontal axis versus the LR test statistic on the vertical axis. The position with the largest LR test statistic shows the strongest evidence for being a QTL, but determining whether this finding is statistically significant requires applying a multiple testing adjustment.

As a final note in this QTL background section, even when a genome location tests as statistically significant for being a QTL, there is usually more work to be done. The testing position showing the greatest evidence for the presence of a QTL is not actually associated with an exact physical location on the genome. The location of the testing position is commonly defined in terms of centiMorgans, which is a genetic distance based on recombination fractions (see Section 1.4), rather than a physical distance measured in kilobases. This centiMorgan position does not necessarily translate directly into a specific physical location, rather it is associated with a region on the genome. Therefore, once a statistical analysis finds evidence for the existence of a QTL, the researcher must then search the region associated with the given testing position for the actual QTL.

In searching for the QTL, the researcher is looking for a location where a polymor-
such a polymorphism is expected to be present because the DNA sequence at the QTL determines trait group membership, and therefore the sequence should contain a categorizing feature (i.e. a polymorphism). Though we know each marker locus is polymorphic, we do not know ahead of time about the polymorphic state of locations between markers. To locate the QTL, the biologist must undertake the labor-intensive process of searching the genome for a candidate gene containing a polymorphism using available databases and sequencing information. A better statistical estimate for the QTL position (i.e. closer proximity to the true QTL) can equate to reduced time in follow-up work for the researcher. Unfortunately, translating the behavior of the QTL estimate into a confidence interval for the true QTL without applying strict, and perhaps questionable, assumptions is difficult. See Manichaikul et al. (2006) for a recent comparison of commonly used QTL confidence interval methods.

1.2 Genotyping Marker Loci

In order to genotype marker locations on the genome, we need to identify locations where a polymorphism exists, and then genotype the given location for each individual. A common approach to genotyping utilizes restriction enzymes that cut DNA when a specific sequence of nucleotides is present. A site that gets cut is called a restriction site and is often 4 to 6 nucleotide-bases long. Two restriction enzymes associated with two restriction sites in close proximity can cut the DNA and create a relatively short DNA fragment composed of the DNA that was between sites. When a polymorphism exists between two DNA strands, application of these two enzymes to the DNA strands can produce fragments of differing lengths. For instance, if one strand has an insertion of nucleotides between restriction sites, its fragment length will be longer than the strand without the insertion. This difference is called a Restriction Fragment Length Polymorphism (RFLP) and it allows us to detect differences in nucleotide sequences between
organisms. An RFLP can also arise when a mutation creates or destroys a restriction site (Lynch and Walsh, 1998).

In practice, DNA is digested with a variety of restriction enzymes at one time. When the digested DNA is run on a gel under an electric current, fragments of differing lengths will travel different distances. Thus, the fragments will group by size. With so many RFLPs represented on the gel, the groupings are not clear and rather uninformative. Thus, labeled DNA probes are used to identify particular regions of the DNA for marker analysis. Each probe represents a marker, and the differing alleles for the marker appear at distinct locations on the gel. This procedure can be done for each individual. There are also other techniques that can highlight several DNA fragments at a time such as Randomly Amplified Polymorphic DNAs (RAPDs). See Lynch and Walsh (1998) for more background on molecular markers.

1.3 Experimental Populations in Interval Mapping

Various mapping populations commonly used in QTL studies are backcross populations, intercross populations, and recombinant inbred line populations. These populations are valuable for QTL studies because we know the breeding structure under which they were propagated, and this structure provides linkage disequilibrium within the propagated individuals (Liu, 1989). Linkage disequilibrium (LD) exists when certain combinations of alleles at numerous genome locations occur more frequently than others. LD occurs because genetic material from each chromosome tends to be inherited in sections, so the alleles at locations in close proximity will tend to be inherited together. The existence of linkage disequilibrium is what allows us to search for QTL locations between observed markers. The small number of possible genotypes at each genome locus in these populations also make them easy to deal with in terms of modeling. In order to generate such populations, one starts with two inbred lines (homozygous
at every loci) of a particular species that differ for the trait of interest. We will refer
to these as the parental lines and refer to their respective genotypes as AA and BB.
Crossing the two parental lines generates an F$_1$ population that has an AB genotype
at every locus. By crossing this F$_1$ population with one of the parental lines we can
produce a backcross population. For example, crossing an F$_1$ organism with the AA
parent produces an organism with either an AA or an AB genotype at each locus. The
determining factor on whether an AA or AB genotype is present depends on the fre­
quency and location of crossover events during the meiosis phase of reproduction (see
next section). Intercross populations are developed by crossing the F$_1$ population with
itself. An organism in an intercross population has one of three possible genotypes at
each loci, AA, AB, or BB in this example. Finally, a recombinant inbred line population
is developed by first producing an F$_1$ population from the parental lines, then repeatedly
self-crossing the F$_1$ organisms until eventually, new homozygous lines are created that
have either an AA or BB at each locus. These new lines are called recombinant inbred
lines (RILs). In general, RIL populations are advantageous because they tend to have
a large frequency of recombination events across the genome compared to the backcross
or intercross populations. This provides more precision for detecting the location of a
QTL on the genome. RILs are also advantageous in genetics studies because they are
inbred and we can obtain many individuals with the same genotype.

1.4 Expected Genotypes at Unobserved Loci

In interval mapping for experimental organisms, we use the genotypes at observed
loci to place a probability distribution on the genotype at an unobserved location on the
same chromosome. Besides conditioning on observed loci, this probability distribution
is conditional on the probability of a crossover event occurring between the observed
loci and the unobserved location. A crossover event can occur during the meiosis phase
of sexual reproduction. During meiosis, the group of four chromatids composed of the two sister chromatids from each parent become close enough in proximity that they can actually exchange genetic material. Crossovers lead to recombinant gametes that have a genetic sequence different from that found in the parental chromatids (provided parents were not inbred and homozygous before mating). Such recombination of gametes contributes to genetic variability in a population as a whole.

The occurrence of crossover events on the genome can be modeled as a Poisson process. Consider the full length of the genome formed by sequentially placing the chromosomes end to end. We define one end of the genome to be positioned at the origin of a 1-dimensional axis. As one moves along the genome, the distance from the origin increases. We define \( x_d \) as the number of crossover events occurring between the origin and the position located at a distance \( d \) from the origin. Then, \( \{x_d : d \in D\} \) for the set of increasing genome positions \( D = \{d_1, d_2, \ldots\} \) is a stochastic process. The distance between two crossover events can modeled as an exponential(\( \lambda \)) random variable (parameterized such that the expected distance between events is \( 1/\lambda \)). Assuming there is no crossover interference\(^1\) (i.e. the occurrence of a crossover doesn’t inhibit another crossover occurring nearby), the distances between events across the full genome are all independently and identically distributed with this exponential(\( \lambda \)) distribution. Given this independence, the number of crossover events occurring at a distance \( d \) from the origin is modeled as a Poisson random variable with an expected number of \( d\lambda \) events. This Poisson(\( d\lambda \)) distribution arises out of the relationship between the gamma distribution (from the relevant sum of independent exponential random variables) and the Poisson distribution. The parameter \( \lambda \) itself is associated with the expected number of crossover events between the origin and the position at 1 distance unit from the origin.

The unit of measurement for distance along the chromosome is called the Morgan (M),

\(^{1}\text{No crossover interference is commonly assumed, but there is some evidence that crossover events are not uniformly distributed along a chromosome. For example, centromeres and telomeres tend to have lower frequencies of crossover events}\)
and is defined as the distance in which the expected number of crossovers is 1. Thus, the rate parameter $\lambda$ is 1 when using the Morgan as the distance measure, as we do in this paper. Finally, as the exponential distribution has the ‘memoryless’ property, the number of crossover events occurring between locations at a distance $d_i$ and $d_j$ with $j > i$ is Poisson distributed with an expected number of $(d_j - d_i)$ crossovers.

The Poisson process model now gives us a connection between the number of crossover events occurring between two loci on the genome and the length (in Morgans) of the interval formed by the loci. This implies that if we could observe crossover events in any interval, we could also estimate the distance between the loci forming the interval. Unfortunately, crossover events are not directly observable in experimental organisms because we do not genotype the full genome, only certain markers. The genotypes present at markers do give us some information on the frequency of crossovers. For example, consider the backcross population described in Section 1.3 above where each genome locus can be coded as either 0 or 1. Two markers on a chromosome forming an interval can be coded as (0,0), (0,1), (1,0), or (1,1). If we observe a (0,1) or (1,0), we say a recombination event has occurred between the loci. In this situation, we know that an odd number of crossover events have taken place between the loci. Similarly, if we observe a (0,0) or a (1,1) at the two locations, then we know an even number (including zero) of crossover events has occurred. In a genotyped sample, the fraction of the organisms showing a (0,1) or (1,0) marker configuration is defined as the estimated recombination fraction between the two markers. This recombination rate between markers is traditionally symbolized by $\theta$ and represents the probability of a recombination event occurring between the two loci.

Using the Poisson process model described above, we can model the probability of a recombination event occurring as the probability that an odd number of crossover events has occurred. Thus, letting $m = t - s$ represent the distance in Morgans between positions $t$ and $s$ with $t > s$, we have that the number of crossover events between $t$ and
s has a Poisson(m) distribution, and

\[ \theta = P(\text{recombination event}) = P(\text{an odd number of crossover events}) = 1 - P(\text{an even number of crossover events}) = 1 - \sum_{y=0}^{\infty} \frac{e^{-m}m^{2y}}{(2y)!} = 1 - e^{-m} \left( \frac{e^m + e^{-m}}{2} \right) = \frac{1}{2} - \frac{e^{-2m}}{2} = \frac{1}{2} \left(1 - e^{-2m}\right). \]

This relationship between \( \theta \) and \( m \) is known as Haldane's mapping function, and it allows us to convert from a Morgan distance to a recombination rate. By inverting the above equation, we form the function converting a recombination rate (estimated through observed marker genotypes) to a Morgan distance, shown as

\[ m = -0.5 \log(1 - 2\theta). \]

In the context of QTL interval mapping, we focus our attention on three particular locations. These locations represent a left marker L, a right marker R, and a putative QTL location Q between the markers. We let \( m_{LR} \) represent the genetic distance between L and R, \( m_{LQ} \) represent the genetic distance between L and Q, and \( m_{QR} \) represent the genetic distance between Q and R. We also let \( \theta \) represent the recombination rate between markers L and R, \( r_L \) represent the recombination rate between L and Q, and \( r_R \) represent the recombination rate between Q and R. As additivity of genetic distance holds, we have

\[ m_{LR} = m_{LQ} + m_{QR} \]

\[ -0.5 \log(1 - 2\theta) = -0.5 \log(1 - 2r_L) + -0.5 \log(1 - 2r_R) \]

\[ -0.5 \log(1 - 2\theta) = -0.5 \log(1 - 2r_L - 2r_R + 4r_Lr_R) \]
and after reducing the equation, we get

$$\theta = r_L + r_R - 2r_Lr_R.$$  

This relationship between the three recombination rates is sometimes utilized to simplify formulas as any one of the three can be written in terms of the other two.

As is apparent from the Poisson process for modeling, there is a higher probability of a recombination between loci that are farther apart than those that are closer together. This behavior plays a role in trying to predict the genotype of an unobserved locus that falls between two observed loci. For example, consider again a backcross organism that has two loci very close together with the first locus coded as a 0 and the second locus also coded as a 0. Because the chance for a crossover event occurring between the loci is very small, any location between the two loci is probably also a 0. If the two observed loci had been 0 and 1, we know that a crossover event occurred somewhere in the interval, and so we are less certain about the predicted genotype for any locus between the loci.

Continuing with the backcross scenario, we now look at estimating genotypes at unobserved loci. Let $x_{\text{left}}$ and $x_{\text{right}}$ be the observed marker genotypes at markers flanking a given genome interval. A putative QTL location between these markers is specified using the parameters $\theta, r_L,$ and $r_R$ as defined above. Each of these recombination fractions can be converted into a genetic distance, thus specifying the genome location (in Morgans) of the putative QTL. Using these recombination fractions, we can compute the probability that the putative QTL location has a genotype of 0 or 1 conditional on the observed flanking markers (see Table 1).
Table 1

Conditional probabilities of putative QTL genotype given observed flanking markers in a backcross population.

<table>
<thead>
<tr>
<th>Marker genotype</th>
<th>Conditional probability of QTL genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{left}$ $x_{right}$</td>
<td>0</td>
</tr>
<tr>
<td>0 0</td>
<td>$\frac{(1-r_L)(1-r_R)}{1-\theta}$</td>
</tr>
<tr>
<td>0 1</td>
<td>$\frac{(1-r_L)r_R}{\theta}$</td>
</tr>
<tr>
<td>1 0</td>
<td>$\frac{r_L(1-r_R)}{\theta}$</td>
</tr>
<tr>
<td>1 1</td>
<td>$\frac{r_{LR}}{1-\theta}$</td>
</tr>
</tbody>
</table>

The denominators in the Table 1 probabilities all contain $\theta$ and reflect the conditioning on what has been observed in the markers. The recombination fractions used in practice tend to be small, and markers are usually measured in centiMorgans (cM) rather than Morgans. Using Haldane’s mapping function, we see that two loci at 1cM apart have a recombination fraction of $\approx 0.01$. This suggests that when we observe the genotypes at these markers in 100 individuals, we expect, on average, 1 of the individuals to exhibit a recombination event. When markers are close together (i.e. $r_L$ and $r_R$ are very small) and no recombination has been observed, we can see in Table 1 that it is much more likely that any locus between the markers is the same genotype as the markers themselves. For such a locus to have a different genotype than the markers, an even number of crossover events greater than 0 would have had to occur between the markers.

In QTL analysis, we utilize the probability that a putative QTL is a specific genotype given the observed flanking markers. When there are only two possible genotypes, as in the backcross scenario, we can consider the QTL genotype as a binary random variable $\in \{0,1\}$, and we need only be concerned with the probability that the ran-
dom variable equals 1. Using the information in Table 1, we now write the probability that the putative QTL=1 for individual \( i \) as a function \( g \) of marker genotypes \( x_i = (x_{iL}, x_{iR}) = (x_{i\text{left}}, x_{i\text{right}}) \), and QTL location:

\[
\pi(x_i; \theta, r_L, r_R) = P(QTL = 1|x_i, \theta, r_L, r_R) = \frac{r_L \cdot r_R}{1 - \theta} + \left( \frac{1 - r_L}{\theta} \frac{r_R}{1 - \theta} \right) \cdot x_{iL} + \left( \frac{r_L \cdot (1 - r_R)}{\theta} \frac{r_R}{1 - \theta} \right) \cdot x_{iR} \tag{1}
\]

As this thesis focuses on populations with two possible genotypes, we will use this function \( \pi(x_i; \theta, r_L, r_R) \) in subsequent chapters as part of our mixture modeling notation and associated likelihoods.

2 Affymetrix GeneChip Array Technology

Microarray technology is used for measuring gene expression in an organism. After a gene is ‘turned-on’, the gene expresses itself by producing protein. The production of a protein is carried out in two distinct stages called transcription and translation. During transcription, the genetic information in a DNA sequence, or a gene, is transcribed into a complementary sequence called mRNA. The translation process follows next by translating this mRNA sequence into the protein the original DNA sequence was coded to produce. Base pairing characteristics of mRNA make it easier to measure than protein, and microarrays are designed to measure the steady state levels of specific RNAs at the time of sampling. In essence, we measure RNAs are a surrogate for protein measurement.

The key to measuring mRNA levels on a microarray is the complementary base-pair structure inherent in every DNA sequence. The four nitrogen bases that combine to form DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). Base-pairing occurs in the double helix structure as G pairs with C, and A pairs with T. mRNA
is a single-stranded complementary copy of one side of an ‘unzipped’ DNA sequence composed of similar bases, except that the T is replaced with a U, for uracil. In general, we can consider each mRNA sequence to be associated with one specific location on the genome, namely, the location of the gene that produced it. Microarrays are designed to take advantage of the uniqueness of these mRNA sequences.

To measure expression, we collect a sample from a biological organism that contains thousands of distinct mRNA sequences. The number of copies of each distinct mRNA sequence is related to the transcriptional output of the cognate gene and the turnover rate of the mRNA. When using Affymetrix GeneChip technology, collected mRNA is converted into a fluorescent labeled complementary RNA, or cRNA, and it is this labeled cRNA sample that is actually placed on the array and measured to determine gene expression.

A common characteristic for all arrays is that short sequences of DNA representing segments of the genome are attached to the array. Millions of copies of each short oligonucleotide sequence are attached in one location called a spot or a probe on the array. When the array is incubated with a solution containing the labeled cRNA, the cRNA hybridizes to the spot or probe that is complementary to its sequence. The fluorescent label allows measurement of the quantity of cRNA present at each spot or probe by its fluorescence intensity. The procedure provides an intensity measure at each probe allowing comparison of the relative levels of expression at the same gene across multiple samples (i.e. across multiple arrays).

The probe sequences can be printed, spotted, or synthesized on the arrays. Probes can be composed of a few bases (oligonucleotides), or they can be composed of hundreds of bases (cDNAs). In this thesis, we focus on expression data collected using Affymetrix GeneChips. In these microarrays, millions of copies of each sequence of interest (25 bases in length called 25-mers) are first covalently attached to the GeneChip through a process of photolithography and combinatorial chemistry. Each of these 25-mers is
called a probe. Affymetrix uses a grouping of 11 pairs of probes to represent a gene, and together these probes form a probe set. Probe synthesis for each probe set on a chip occurs in parallel, resulting in the addition of an A, C, T, or G nucleotide to multiple growing chains simultaneously across the whole array (Affymetrix, 2001). A probe pair has both a 25-mer called the perfect match probe (PM) designed to be complementary to the reference RNA, and a second 25-mer positioned next to the PM probe on the chip called the mismatch probe (MM) that is an exact replica of the first except that the middle base has been intentionally switched with its complementary base, therefore producing a sequence not naturally occurring in the organism. In theory, the presence of this MM probe on the microarray allows for a measure of noise in the data (called non-specific binding) because no cRNA sequence would be expected to correctly attach itself to a MM probe.

After hybridization, the GeneChip is washed to remove excess cRNA that did not hybridize to any probes. Then, a staining reaction is performed on the chip to allow a scanner to quantify the amount of cRNA that hybridized to each probe set. The expression of any gene is represented by values in a 2x11 matrix with one row representing the PM probe fluorescence values, the other row representing the MM probe fluorescence values, and the columns corresponding to probe pairs. Usually, a summary statistic for these 22 values is computed in order to perform gene-based statistical analyses, such as identification of differentially genes.

Numerous probe set summary statistics have been suggested, and much controversy has arisen in choosing one as a 'best' summary (Choe et al., 2005). Affymetrix provides a commonly used summary statistic called MAS5.0 (Affymetrix, 2001) that incorporates Tukey’s bi-weight and the values of the MM probes. Another commonly used summary statistic that does not incorporate information from the MM probes is the Robust Multi-chip Average (RMA) suggested by Irizarry et al. (2003). Similarly, other procedures needed to soundly analyze expression data, such as normalization or background
correction, have generated much discussion in the research community. Because this thesis focuses on statistical issues that arise after the application of these procedures, we do not attempt to compare these procedures in this paper. Instead, we simply mention that our proposed methods are applicable to probe-set summary statistics computed after the researcher applies the relevant procedures of his or her choice.

3 Multiple Testing Adjustment Procedures

Both QTL and microarray analyses generate hundreds, maybe even thousands, of test statistics from one data set. The process for determining significance for such a situation is different than when there is only one test of interest. One goal of multiple testing procedures is to filter a large pool of tests down to a relatively short list of statistically significant tests that are associated with a low false positive rate. A second goal is to maintain a low false negative rate with respect to the tests that were not chosen for the significance list. Controlling the family-wise error rate (FWER) is a common approach to control the number of false positives in a multiple-testing scenario. After choosing a threshold that controls the FWER at $\alpha = 0.05$, any generated list of significant tests contains no false positives 95% of the time (i.e. if this procedure were repeated 100 times, only 5 of the 100 lists would be expected to contain one or more false positives). The Bonferroni adjustment can be easily applied to a multiple-testing scenario to achieve a chosen FWER, but this adjustment tends to be fairly conservative and is often associated with a high rate of false negatives. For example, when the number of tests performed is in the thousands, it is very feasible that no test is declared significant after applying the Bonferroni adjustment. The chosen FWER has been achieved, but an empty list of significant tests provides no information for further study. Perhaps applying the Bonferroni adjustment is appropriate when the cost for a false positive is very
high, or when very few of the tests are expected to be true positives. But in the case of QTL and microarray analyses, it is usually thought that at least some, and maybe even thousands, of the tests are true positives. Also, the significance list in a QTL or microarray analysis is usually not an end in itself, but rather a springboard for further biological investigation and verification. Thus, procedures other than the conservative Bonferroni adjustment, and error rates other than the FWER are often utilized for QTL and microarray analyses.

\subsection{For QTL Analysis}

As mentioned in Section 1.1, a researcher performing a QTL analysis chooses $\beta$ locations along the genome at which to test for a QTL. The null hypothesis for each test is the same, and it states that the the given location is not a QTL. Consecutive tests along the genome are often strongly correlated. When a genome location coincides with a large LR test statistic, the locations near it often show evidence of being a QTL as well due to linkage disequilibrium. This strong dependence in consecutive tests along the genome must be accounted for in determining significance.

Churchill and Doerge (1994) developed a method for determining significance in QTL studies that has become widely used. Their method uses permutation to develop an empirical distribution for the maximum of the $\beta$ test statistics that would be seen under the null. This empirical null distribution is then compared with the original test statistics for determining significance at a given FWER. Specifically, to create one null data set, the trait data is permuted among the individuals while the marker data remains unchanged (i.e. the genotype for each individual remains unaltered). This null data set has the same marker dependence structure as the original data, but has no association between the markers and the trait values. Instead of keeping track of all $\beta$ test statistics associated with this null data set, only the maximum test statistic is recorded. After
repeating this permutation process $N \geq 1000$ times, we have an empirical distribution of the maximum test statistic under the null. By summarizing the $\beta$ tests into one value, the issue of dependence between tests has been resolved. Choosing a significance threshold as the $100(1-\alpha)$ percentile of the maximum test statistic null distribution controls the FWER at the $\alpha$ level. The type I error rate holds under minimal assumptions about the distribution of traits, specifically, that the trait values are exchangeable under the null. This procedure essentially tests the null hypothesis of no association between the trait and the full genome map, an experimentwise statement. Rejecting this null suggests that there is at least one location on the genome that is associated with the trait.

### 3.2 For Microarray Analysis

The dependence between tests in a microarray analysis arises from the complex relationships between genes in a biological system. This type of dependence is inherently different than the dependence that exists between consecutive markers, and is dealt with in a different manner. Specifically, the adjustment we use for multiple testing in microarrays is appropriate for a block-structure correlation, which is a reasonable structure for modeling gene correlation. A common microarray analysis is to perform a hypothesis test for differential expression at each of $G$ genes, with $G$ usually in the thousands or tens of thousands. By applying a Bonferroni multiple-testing adjustment, we could achieve a FWER of 0.05 by using a $p$-value threshold of $0.05/G$. But because this method is quite conservative and researchers are often willing to risk getting a few false positives in order to decrease the number of false negatives, other methods are usually employed.

Benjamini and Hochberg (1995) introduced a new approach to multiple testing that advocated controlling the expected proportion of false positives among all rejected hypotheses. The motivation for their approach is expressed in their statement:

In many multiplicity problems the number of erroneous rejections should be
taken into account and not only the question whether any error was made.

(Benjamini and Hochberg, 1995)

Letting $R$ be the total number of hypotheses rejected and $V$ be the number of false rejections, they define the false discovery rate (FDR) to be $E\left(\frac{V}{R} \mid R > 0\right) \cdot P(R > 0)$. The inclusion of $P(R > 0)$ is needed because it is not possible to control the conditional expected value of $\frac{V}{R}$ alone. Specifically, in the case when all tests are true null tests, the expected value of the proportion, given that any rejections are made, is always 1. In their paper, they provide a sequential $p$-value procedure for choosing an appropriate number of rejections to control this FDR at the $\alpha$ level.

Storey (2002) extended the work of Benjamini and Hochberg by using a different approach. Instead of using a sequential $p$-value procedure, he described how to estimate the FDR once a rejection region $\Gamma$ has been chosen. By incrementally changing the rejection region and recomputing the estimated FDR, a $\Gamma_0$ rejection region can be chosen that coincides with a specific number of significant tests and an acceptable expected number of false positives. Storey et al. (2003) showed that this method conservatively and consistently estimates the FDR for all rejection regions simultaneously. In general, as the number of significant tests increases, so will the FDR. The possible pairs of values from which a researcher can choose depend on the data, and one hopes for a small increase in the FDR as more tests are designated as significant.

4 Dissertation organization

The three papers in this thesis all relate to the field of bioinformatics. The focus of the first paper is on the detection of quantitative trait loci. The second paper discusses the detection of differentially expressed genes in microarray analyses. The third paper brings these two methods together for a new analysis called eQTL analysis that seeks
to detect locations of genetic control of gene expression.

The sections of this thesis include an introduction, three research papers, and a conclusion discussing future work. Chapter 1 discusses the first research paper where we propose a method to detect a QTL in the case of a zero-inflated Poisson trait. We show how this method outperforms existing methods based on normality in many situations using simulations. The EM algorithm is utilized to estimate parameters in the related likelihood. In Chapter 2, we propose a method to detect differentially expressed genes in a multiple-treatment microarray timecourse experiment. This type of experiment presents challenges to the statistician because it incorporates both multiple-testing and model choice issues. Our method provides both a ‘best’ model choice, and a p-value for determining differential expression significance at each of thousands of genes. We provide information on the specificity and sensitivity of detection under a variety of true expression patterns using receiver operating characteristic curves. In Chapter 3, we discuss a new type of analysis called eQTL analysis. The goal of this analysis is to detect loci controlling gene expression at other locations on the genome. These controlling loci are called expression QTL, or eQTL. For data sets containing thousands of genes and hundreds of markers, there are potentially millions of tests of interest. Besides the difficulty involved in sifting through millions of tests, the issues seen in QTL analysis and microarray analysis are both present. For example, in multiple-testing adjustments, the type of dependence between tests must be considered. The adjustment for a QTL analysis accounts for the strong correlation between tests at consecutive markers, while the adjustment for a microarray experiment accounts for the block-structure correlation between gene expression values in an individual. Both of these types of multiple testing must be considered when determining statistical significance of eQTLs. We conclude this thesis with a section discussing our conclusions and future work.
5 References


6 Glossary

**allele** any one of a number of viable DNA codings of the same gene occupying a given locus on a chromosome.

**backcross population** offspring generated through a cross between an inbred parent and an F1 organism.

**centiMorgan (cM)** genetic distance unit of the genome. One crossover event is expected in a genetic distance of 1 Morgan, and 1 Morgan is 100 cM.

**centromere** a region of chromosome at which sister chromatids are attached during cell division.

**crossover** exchange of chromosomal genetic material between nonsister chromatids of homologs during meiosis.

**cQTL** a section of DNA containing the gene associated with a gene expression trait.

**F1 population** offspring generated through a cross between two inbred parental lines.

**F2 population** offspring generated through a cross between F1 organisms.

**gene** DNA that encodes for a protein or any RNA used in an organism’s biological system.

**genotype** the specific genetic makeup of an individual, in the form of DNA.

**hybridization** the process in which a labeled cRNA sample adhere to a microarray according to base-pair complementation.

**inbred line** population of organisms that are genetically uniform. Breeding within the population produces another genetically identical organism.

**interval mapping** a procedure used to test for QTL within the interval bound by two marker locations.
interval mapping  a procedure used to test for QTL within the interval bound by two marker locations.

kilobase (kb)  physical distance unit of the genome. Unit of length for DNA fragments equal to 1000 nucleotides.

linkage disequilibrium  describes a situation in which some combinations of alleles or genetic markers on a chromosome occur more frequently in a population than would be expected under independence between the locations.

locus  a section of double-stranded deoxyribonucleic acid (DNA).

marker  a known DNA sequence that can be identified in an organism by a simple assay.

Morgan (cM)  genetic distance unit of the genome. One crossover event is expected in a genetic distance of 1 Morgan, and 1 Morgan is 100 centiMorgans.

mRNA  RNA that encodes and carries information from DNA during transcription to sites of protein synthesis to undergo translation in order to yield a gene product.

oligonucleotide  short sequences of nucleotides (RNA or DNA), typically with twenty or fewer bases.

polymorphism  different, detectable alleles for a gene or marker in a population.

QTL  a section of DNA containing the gene associated with a quantitative trait.

restriction fragment length polymorphisms (RFLP)  an assay in which organisms may be differentiated by analysis of patterns derived from cleavage of their DNA. Used to genotype markers.

telomere  DNA at ends of chromosome.
CHAPTER 1. QTL DETECTION FOR ZERO-INFLATED POISSON TRAITS

A paper to be submitted to *Biometrics*

Rhonda DeCook and Dan Nettleton

Abstract

Much work has been done on detecting QTLs when the trait of interest is normally distributed. This paper presents a method for detecting QTLs for zero-inflated Poisson (ZIP) traits. A few non-normal trait distributions have received some attention in the area of QTL analysis. Methods for binary, ordinal, and Poisson traits have been proposed, but many other non-normal traits have yet to be investigated. Though existing methods based on normality may be reasonably applied to some ZIP distributions, the characteristics of other ZIP distributions make such an application inappropriate. The method proposed in this paper is appropriate for any ZIP distribution. Using simulation, we compare our method to two existing applicable approaches. The method is illustrated using QTL data collected on two ecotypes of the *Arabidopsis thaliana* plant where the trait of interest is shoot count.
1 Introduction

A quantitative trait locus (QTL) is a locus on the genome that contributes to a phenotype that varies quantitatively. In general, the alleles present at such a locus are thought to play a role in determining the level of the quantitative trait. The genetic mapping of QTLs gives insight into the genetics behind observed traits and can help advance genetic research on the trait of interest. For example, detecting QTLs associated with a disease can provide useful information for developing therapeutic drugs. For desirable traits, such as high yield in corn plants, detecting QTLs can provide useful information for selective breeding.

In order to quantify the association between a trait and a putative QTL location, we need information on the alleles, or genotypes, at the locus of interest. In QTL statistical modeling, observed genotype values are the explanatory variables $x$ used to model the trait of interest $y$. Because a single genome can be composed of billions of basepairs, we presently do not genotype the genome completely. Instead, we genotype particular locations called markers that are dispersed throughout the genome. These markers are characterized as short basepair sequences containing a polymorphism in the population (see Section 1.2 of the General Introduction). Essentially, genotypes are observed at marker locations, and genotypes at all other locations on the genome are unobserved. Fortunately, the observed marker genotypes provide useful information for predicting the genotypes at unobserved loci based on simple genetics principles.

Simply due to the vast length of the genome and the relatively small number of observed loci, QTLs are not expected to fall directly on a marker. Instead, we expect QTLs to be located somewhere between observed locations. A single marker QTL analysis that does not investigate unobserved locations may be sufficient when the genome is densely genotyped, but most QTL analysis methods use observed markers to investigate the unobserved genome regions. Specifically, these methods utilize the information at
observed markers that flank an interval of unobserved locations to predict the genotype at any unobserved location in the interval. This allows the researcher to compute a test statistic for QTL existence at any location, observed or unobserved. Lander and Botstein (1989) introduced this method called interval mapping in their QTL analysis seminal paper.

Interval mapping models the marginal trait distribution as a mixture model with each mixture component representing a unique QTL genotype. For each putative QTL location along the genome, this mixture model is fitted and the maximum likelihood estimates of the parameters are used to compute a likelihood ratio test statistic. The location with the highest test statistic shows the greatest evidence for being a QTL. The conditional probability distributions of unobserved genotypes at putative QTL locations given the available marker information are incorporated into the mixing proportions of the mixture model.

Since the introduction of interval mapping by Lander and Botstein (1989), much work has been done on interval mapping for QTL traits that follow a normal distribution. Carbonell et al. (1992) investigated normal traits in the case of nonadditivity. Zeng (1994) introduced composite interval mapping that tests for a QTL after accounting for variability in the quantitative trait due to other marker locations on the genome. Kao et al. (1999) proposed an multiple interval mapping method used for detecting QTL when multiple QTL are present on the genome.

For trait distributions that are not approximately normal, other methods must be applied. Methods for binary response traits, such as diseased versus not diseased, have received some attention in recent years. Xu and Atchley (1996) proposed a method for detecting a binary trait generated from an underlying normal trait in conjunction with a threshold model. McIntyre et al. (2001) investigated mapping a single binary trait, while Xu et al. (2005) investigated jointly mapping multiple binary traits. Ordinal traits have also been explored by Hackett and Weller (1995). A method for detecting QTL
associated with a Poisson trait has been proposed by Kayis et al. (1998), and Thomson (2003) extended the Poisson method by using a generalized estimating equations approach in order to include random effects. For detecting non-normal traits in general, a non-parametric approach has been proposed by Kruglyak and Lander (1995).

In this paper, we are interested in mapping a zero-inflated Poisson (ZIP) trait. Though methods based on normality may be reasonably applied to some ZIP distributions, the characteristics of other ZIP distributions make such an application inappropriate. For example, when the Poisson parameter is relatively large and there is a non-ignorable amount of mass at zero, a normal distribution will approximate the ZIP distribution poorly. Broman (2003) discusses QTL analysis in the case of a spike in the quantitative trait distribution. The mixture scenario he describes is closely related to the ZIP trait scenario, but differs in some important aspects. Broman's method assumes that for each QTL genotype, knowledge of which mixture component (either the spike or the smooth part of the distribution) generated the observation is apparent in the observation itself. He illustrates this situation using a mixture between a normal distribution with $\mu \gg 0$ and a spike at 0. If $y_i = 0$, then he essentially knows the observation $i$ was generated from the point mass component of the mixture distribution. In the situation of a ZIP trait, we can not determine which distribution generated the observation simply from the observation itself. The $P(Y_i = 0)$ is potentially non-ignorable in both components of the mixture. Thus, if we consider $Z$ as the random variable specifying which mixture component generated the observation from a given QTL genotype, $Z$ is unobserved when $y_i = 0$. For this reason, Broman's method is not applicable in all ZIP scenarios.

In this paper, we develop a method for detecting QTL for ZIP traits that incorporates the ZIP distribution and can be appropriately applied to any ZIP distribution. In the next section, we describe the ZIP distribution and develop the likelihood for a mixture of ZIPS in the context of QTL mapping. We describe the procedure for using the EM
algorithm to compute maximum likelihood estimates for the ZIP parameters, and we discuss convergence of the algorithm. The asymptotic behavior of the estimates is also discussed. Section 3 provides results from a simulation study comparing our proposed method to applicable existing methods. In Section 4, we illustrate the method using QTL data collected on two ecotypes of the *Arabidopsis thaliana* plant where the trait of interest is shoot count. The final section is the Appendix and provides verification on some EM convergence conditions and maximum likelihood regularity conditions.

2 Method for Detecting ZIP QTLs

2.1 ZIP Distribution and Notation

In this section we provide notation for a single ZIP distribution and the related likelihood function. We also discuss the missing data issue involved in the ZIP distribution, and describe the complete-data likelihood. The same information is also provided for a mixture of ZIP distributions which is the relevant distribution in the case of a ZIP QTL trait.

2.1.1 Single ZIP Distribution

The probability mass function (PMF) for a ZIP random variable can be seen as a mixture between a Poisson PMF and a point mass at zero. Two parameters are required to completely specify the distribution. The first parameter, $p \in (0,1)$, specifies the proportion of total mass coming from the point mass component and is the mixing parameter. Similarly, $1 - p$ is the proportion of total mass coming from the Poisson component of the mixture. The second parameter, $\lambda > 0$, is the mean of the Poisson
distribution. If \( Y_i \) is distributed as a ZIP random variable, then

\[
f(y_i|p, \lambda) = p \cdot I_{\{0\}}(y_i) + (1 - p) \cdot \left( \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right) \quad \text{for } y_i = 0, 1, 2, \ldots
\]

After rewriting the PMF as

\[
f(y_i|p, \lambda) = \left\{ p + (1 - p)e^{-\lambda} \right\} I_{\{0\}}(y_i) \left\{ (1 - p) \left( \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right) \right\}^{1-I_{\{0\}}(y_i)} \quad \text{for } y_i = 0, 1, 2, \ldots
\]

we can write the log-likelihood function for \( n \) independent observations as

\[
L(p, \lambda|y) = \sum_{y_i=0} \log \left\{ \left( \frac{p}{1 - p} \right) + e^{-\lambda} \right\} + \sum_{y_i>0} (y_i \log \lambda - \lambda) + \sum_{i=1}^{n} \log(1 - p) - \sum_{y_i>0} \log(y_i!).
\]

Taking the derivatives of \( L(p, \lambda|y) \) in order to determine the parameter values for maximizing the function does not lead to closed-form solutions for \( \hat{p} \) and \( \hat{\lambda} \). For this reason, an alternative approach utilizing the complete-data likelihood and the EM algorithm (Dempster et al., 1977) is usually taken to find the maximum likelihood estimates (MLEs) for \( p \) and \( \lambda \).

### 2.1.2 Complete-Data Likelihood for Single ZIP

In a complete-data framework, we conceptualize some part of the data as missing. For the ZIP, the missing information is associated with an observed zero value. When \( y_i = 0 \), we can not directly determine whether the point mass or the Poisson distribution generated the observation. If we had complete information on which distribution generated each zero, we could easily solve the related likelihood equations for the two MLEs. Dempster et al. (1977) developed an interative process called the EM algorithm that utilizes this desirable feature of the complete-data likelihood and the information in the observed data to compute the MLEs for the incomplete-data likelihood. It is this algorithm that we employ to compute MLEs in our method. In order to develop the complete-data likelihood for the ZIP distribution, we must introduce a random variable.
that coincides with the missing information. We let $z_i = 0$ when the observation was
generated from the Poisson. Similarly, we let $z_i = 1$ when the observation was gen-
erated from the point mass. An observation is expressed as $(y_i, z_i)$ and $z_i$ is unobserved
when $y_i = 0$. We note that $z_i$ is observed when $y_i > 0$ because such an observation must
have been generated from the Poisson distribution. Thus, it is the zero observations
that motivate the transition into the complete-data framework for the ZIP distribution.

The complete-data joint PMF, or complete-data likelihood, for the ZIP distribution
can be written in exponential-family form as

$$f(y, z | p, \lambda) = (1 - p)^n e^{-n\lambda} \left( \prod_{i=1}^{n} \frac{y_i}{y_i!} \right) e^{\{ (1 - p) \lambda \sum_{i=1}^{n} z_i + \log(1 - p) \sum_{i=1}^{n} y_i \}}$$

The two minimal sufficient statistics are $\sum_{i=1}^{n} z_i$, the number of observations from the
point mass, and $\sum_{i=1}^{n} y_i (1 - z_i)$, the sum of the observations generated from the Pois-
son. The complete-data log-likelihood is easily maximized by computing derivatives,
but unfortunately we cannot use this method directly as it requires observation of the
unobserved z values. Instead, we use the EM algorithm to iterate between two steps
involving i) replacing the unobserved z by its conditional expectation given the observed
data and present estimates for $\lambda$ and $p$, and ii) estimating $\lambda$ and $p$ by maximizing the re-
sulting complete-data likelihood. At convergence, the EM algorithm provides parameter
estimates that are a solution to the score functions of the incomplete-data log-likelihood.

2.1.3 Mixture of ZIP Distributions

In a QTL analysis, the marginal distribution of traits is represented as a mixture
of distributions. Each QTL genotype is associated with one component of the mixture.
For purposes of illustration in this paper, we will limit our modeling to populations with
only two possible genotypes at any location on the genome. Thus, our marginal trait
distribution is a two-component mixture. This situation occurs in backcross populations
and recombinant inbred line populations derived from two inbred parental lines. With minor modifications, this method can be extended to populations with more than two possible QTL genotypes.

When there are two possible genotypes at each genome location, there are four possible genotype combinations for the observed markers flanking each unobserved location. Let $x_i = (x_{iL}, x_{iR})$ denote the vector holding the left flanking marker genotype and right flanking marker genotype, respectively, for individual $i$ for a given interval on the genome. Then, $x_i \in \{(0,0), (0,1), (1,0), (1,1)\}$. Though the QTL genotype can be coded similarly as 0 or 1, we instead code the QTL as $q$ or $Q$, respectively, to distinguish it from marker locations.

In a backcross QTL scenario, we can statistically model the trait of individual $i$ given $\beta_q, \beta_Q$, and $x_i$ as

$$f_{y|x}(y_i|\beta_q, \beta_Q, x_i) = (1 - \pi(x_i))f_q(y_i|\beta_q) + \pi(x_i)f_Q(y_i|\beta_q) \quad (1)$$

where $\beta_q$ contains the parameters needed to specify the trait distribution when the QTL is a $q$, $\beta_Q$ contains the parameters needed to specify the trait distribution when the QTL is a $Q$, and $\pi(x_i)$ is the probability that individual $i$ has a QTL genotype of $Q$. The mixing proportion $\pi(x_i)$ for individual $i$ is a function of the genotypes at observed markers flanking the QTL, the distance between markers, and the location of the putative QTL.

The location of the putative QTL is usually specified in terms of genetic distance from the right and left flanking markers. The genetic distance between any two of the three locations of interest (i.e. the left marker, the QTL, and the right marker) is determined by the probability of a recombination event occurring during reproduction. Genetic distance, measured in Morgans or centiMorgans (cM), is determined from a mapping function that maps a recombination rate to a distance value. The three probabilities of interest are the probability $r_L$ of a recombination event occurring between the left
marker and the QTL, the probability \( r_r \) of a recombination event occurring between the QTL and the right marker, and the probability \( \theta \) of a recombination event occurring between the markers. Together, the trio define a putative QTL location and are taken as known fixed quantities when computing a LR test statistic.

For a given testing location, both the distance between markers, represented as \( \theta \), and the putative QTL location are considered known and fixed, but the marker genotypes are randomly generated (the modeling of marker genotypes was discussed in Section 1.4 of the General Introduction). Letting \( w(x_i) \) represent the PMF for the randomly generated markers, we have

\[
w(x_i) = \begin{cases} 
0.5(1 - \theta) & \text{if } x_i = (0,0) \\
0.5\theta & \text{if } x_i = (0,1) \\
0.5\theta & \text{if } x_i = (1,0) \\
0.5(1 - \theta) & \text{if } x_i = (1,1).
\end{cases}
\]

We also showed in Section 1.4 of the General Introduction that the mixing proportion \( \pi(x_i) = P(\text{QTL}=Q | x_i) \) could conveniently be written as a function \( \pi(x_i; \theta, r_L, r_R) \) of the relevant observed markers and fixed values. Hereafter, we replace \( \pi(x_i) \) with \( \pi(x_i; \theta, r_L, r_R) \) to emphasize that the probability of getting QTL genotype of Q is a function of the observed flanking markers, the distance between markers, and the location of the QTL. By including this notation, the joint likelihood for the observed trait \( y_i \) and observed marker genotypes \( x_i \) can be written as

\[
f(y_i, x_i | \beta_q, \beta_Q) = f_{y|x}(y_i | \beta_q, \beta_Q, x_i)w(x_i) \\
= \{ (1 - \pi(x_i; \theta, r_L, r_R)) f_q(y_i | \beta_q) + \pi(x_i; \theta, r_L, r_R) f_Q(y_i | \beta_Q) \} w(x_i)
\]

In the case of a ZIP trait, there are four parameters required to specify the model in Equation 2. We let the parameters \( p_q \) and \( \lambda_q \) specify the ZIP distribution given the QTL genotype is \( q \). Similarly, \( p_Q \) and \( \lambda_Q \) specify the ZIP distribution given the QTL genotype is \( Q \). As defined before, \( p_q \) and \( p_Q \) represent the probability that an observation is gener-
ated by the point mass component of the ZIP, respectively. Letting \( \phi = (p_q, \lambda_q, p_Q, \lambda_Q)' \), the joint log-likelihood \( L(\phi|y, x) \) for \( n \) independent observations can be written as

\[
L(\phi|y, x) = \sum_{i=1}^{n} \log \left( (1 - \pi(x_i; \theta, r_L, r_R)) \cdot \left\{ (p_q + e^{-\lambda_q}(1 - p_q))^{I(y_i|Q)} \left( 1 - p_q \frac{e^{-\lambda_q} \lambda_q^{y_i}}{y_i!} \right)^{(1-I(y_i|Q))} \right\} + \right.
\]
\[
\pi(x_i; \theta, r_L, r_R) \left\{ (p_Q + e^{-\lambda_Q}(1 - p_Q))^{I(y_i|Q)} \left( 1 - p_Q \frac{e^{-\lambda_Q} \lambda_Q^{y_i}}{y_i!} \right)^{(1-I(y_i|Q))} \right\} \left. \right\} + \sum_{i=1}^{n} \log w(x_i)
\]

The form of the log likelihood reflects the mixture of ZIP distributions. This form makes direct computation of the MLE for \( \phi \) difficult. As discussed earlier, the common approach for finding the MLE for this type of likelihood is to apply the EM algorithm to the related complete-data likelihood. Though the incomplete-data log-likelihood function in Equation 3 is not used in the EM algorithm directly, the convergence of the algorithm does depend on its behavior. Thus, we will return to Equation 3 upon investigating convergence properties in Section 2.3.

### 2.1.4 Complete-Data Likelihood for Mixture of ZIPS

Complete information for an observation generated from a mixture of two ZIP distributions includes the observed count, knowledge of which ZIP generated the observation (i.e. knowledge of the QTL genotype), and knowledge of which component of the given ZIP generated the observation. In order to develop the complete-data likelihood we introduce the random variables \( Z_1 \), to designate the QTL genotype, and \( Z_2 \), to designate the component of the ZIP generating the observation. For observation \( i \), we have the
complete-data vector \((y_i, z_{1i}, z_{2i}, x_i')\). We let \(z_{1i} = 0\) when the QTL genotype is \(q\), and \(z_{1i} = 1\) when the QTL genotype is \(Q\). Similarly, we let \(z_{2i} = 0\) when observation \(i\) was generated by the Poisson component of the ZIP, and \(z_{2i} = 1\) when it was generated by the point mass component. As mentioned earlier, for \(y_i > 0\), \(z_{2i}\) is actually observed because such an observation must have been generated by the Poisson distribution.

The complete-data joint PMF can be written in exponential-family form as

\[
\prod_{i=1}^{n} f_{Y,Z,X} (y_i, z_{1i}, z_{2i}, x_i | \phi) = \prod_{i=1}^{n} f_{Y,Z|X} (y_i, z_{1i}, z_{2i} | \phi, x_i) w(x_i)
\]

\[
= \left[ \prod_{i=1}^{n} \left( \frac{1}{y_i!} \right) \pi(x_i; \theta, r_L, r_R)^{z_{1i}} \cdot (1 - \pi(x_i; \theta, r_L, r_R))^{(1-z_{1i})} \cdot w(x_i) \right] \cdot p_q^{z_{1i}} \cdot e^{\left\{ \log\left( \frac{\lambda Q}{p Q} \right) \sum_{i=1}^{n} z_{1i} \right\}} \cdot e^{\left\{ \log\left( \frac{e^{-\lambda Q (1-p Q)}}{p Q} \right) \sum_{i=1}^{n} z_{1i} (1-z_{2i}) \right\}} \cdot e^{\left\{ \log\left( \frac{e^{-\lambda Q (1-p Q)}}{p Q} \right) \sum_{i=1}^{n} z_{1i} (1-z_{2i}) \right\}} \cdot e^{\left\{ \log\left( \frac{e^{-\lambda Q (1-p Q)}}{p Q} \right) \sum_{i=1}^{n} z_{1i} (1-z_{2i}) \right\}}.
\]

As the complete-data joint PMF shows, there are five linearly independent natural parameters, each a function of the parameters in the original parameterization contained in \(\mathbb{R}^4\). This type of exponential family, where the natural parameters are a function of a lower-dimensional parameter set, is called a \textit{curved exponential family}. The natural parameters \(\eta = (\log \left( \frac{p Q}{p Q} \right), \log \lambda Q, \log \left( \frac{e^{-\lambda Q (1-p Q)}}{p Q} \right), \log \lambda q, \log \left( \frac{e^{-\lambda Q (1-p Q)}}{p Q} \right))'\) in this curved exponential family are restricted to lie in a curved subset of 5-dimensional Euclidean space.

Describing the five minimal sufficient statistics \(t = (\sum z_{1i}, \sum y_i z_{1i}, \sum z_{1i} (1-z_{2i}), \sum y_i (1-z_{1i})(1-z_{2i}), \sum (1-z_{1i})(1-z_{2i})))'\), into terms of the ZIP distributions gives us insight as to why each statistic is essential for parameter estimation. The five statistics provide information on \(i\) the number of observations with a QTL genotype of \(Q\), \(ii\) the sum of the counts generated from the Poisson when the QTL was \(Q\), \(iii\) the number of observations generated from the Poisson when the QTL was \(Q\), \(iv\) the sum.
of the counts generated from the Poisson when the QTL was \( q \), and \( v \) the number of observations generated from the Poisson when the QTL was \( q \).

The likelihood equations for the complete-data log-likelihood are easily solvable. As the EM algorithm is applicable to curved exponential families (Dempster et al., 1977), we now discuss its application to compute the MLE for the incomplete-data likelihood function associated with Equation 3.

### 2.2 EM Algorithm

In this subsection, we first provide general background information on the EM algorithm when the complete-data likelihood is an exponential or curved exponential family. Then we provide the EM steps for the specific case of a mixture of two ZIP distributions arising from a backcross QTL study.

#### 2.2.1 Background

The EM algorithm introduced by Dempster et al. (1977) is commonly used to compute maximum likelihood estimates (MLEs) in the case of incomplete data. Data generated from a mixture distribution is often used as an example of this type of data. The framework of the EM algorithm includes two sample spaces that we will denote as \( S \) and \( U \). The observed data \( s \) is a realization from the sample space \( S \). For ZIP QTL data, \( s \) includes observed counts and marker genotype information. We also refer to \( s \) as the incomplete data. The sample space \( U \) involves the unobserved data. We assume there exists a mapping, not 1-to-1, from \( U \) to \( S \), that allows us to infer information about the unobserved \( u \) through the observed \( s \). For ZIP QTL data, \( u \) includes information on the QTL genotype for all individuals, and information on which component of the given ZIP generated each observation. Information about these unobserved values lies in the observed count and observed marker information. We refer to the pair \((s, u)\) as
the complete data.

The goal of the EM algorithm is to compute maximum likelihood estimates for the parameters in the data generating mechanism for \( s \), we shall presently denote these as the natural parameters \( \eta \). We define the density for the incomplete data as \( h(s|\eta) \). As we assume the same natural parameters \( \eta \) are also involved in generating the joint \((s, u)\), we define the complete-data joint density as \( f(s, u|\eta) \). When we integrate the joint density with respect to the distribution of the unobserved data, we are left with the density for the observed data:

\[
h(s|\eta) = \int_{\mathcal{U}} f(s, u|\eta) du
\]

The results presented in Dempster et al. apply to a variety of complete-data density forms, but we will focus on the results that apply to curved exponential families as this is the relevant form for modeling a mixture of ZIP traits in the QTL setting when utilizing marker information. Writing the complete-data density in exponential family form (or curved exponential family form) using Dempster et al. notation, we have

\[
f(s, u|\eta) = b(s, u) \exp(\eta^T \cdot t(s, u))/\alpha(\eta) \tag{5}
\]

where \( \eta \) is the \( r \times 1 \) natural parameter vector, and \( t(s, u) \) is the \( r \times 1 \) vector of sufficient statistics for estimating the natural parameters. For an exponential family form, \( \eta \) is restricted to lie in an \( r \)-dimensional convex set \( \Omega \) such that any \( \eta \in \Omega \) defines a viable density for \( f(s, u|\eta) \). For a curved exponential family form, \( \eta \) is restricted to lie in a subset of \( \Omega \) denoted as \( \Omega_0 \). This subset \( \Omega_0 \) is a curved submanifold of the \( r \)-dimensional \( \Omega \) space. In the case of a ZIP QTL, the five natural parameters are all functions of the four parameters \( p_q, \lambda_q, p_Q, \) and \( \lambda_Q \). Thus, the natural parameters are restricted to lie in a curved submanifold of 5-dimensional Euclidean space. It is assumed that the support set of \( f(s, u|\eta) \) is the same for all \( \eta \in \Omega_0 \).

A fundamental concept used in the EM algorithm is that the parameter values
that maximize the complete-data log-likelihood also maximize the incomplete-data log-likelihood. This concept arises out of the relationship between the complete-data joint density $f(s, u|\eta)$, the incomplete data density $h(s|\eta)$, and the associated conditional density $g(u|s, \eta)$. Dempster et al. showed that the derivative of the incomplete-data log-likelihood $L(\eta)$ is equal to the expected value of the sufficient statistics under the complete-data density subtracted from the expected value of the sufficient statistics under the conditional density. Letting $D$ represent differentiation with respect to the natural parameters, we can write this statement as

$$ DL(\eta) = -E(t(s, u)|\eta) + E(t(s, u)|s, \eta). $$

Thus, if $\eta^*$ provides the equality

$$ E(t(s, u)|\eta^*) = E(t(s, u)|s, \eta^*), $$

then $DL(\eta^*) = 0$. Whether or not $\eta^*$ is associated with a global maximum for $L(\eta)$ is discussed in Section 2.3.

To find an $\eta^*$ for which Equation 7 holds, the EM algorithm uses an iterative procedure that updates an estimated $\eta^{(k)}$ with each iteration $k$. The first step of each iteration, called the E-step, focuses on the right side of Equation 7. In this step, we use the parameter estimates from the previous iteration $\eta^{(k)}$ to compute the expected values of the sufficient statistics represented in the right side of the equation. The vector of expected values can be expressed as $E(t(s, u)|s, \eta^{(k)})$. These resulting expected values of the sufficient statistics serve as a prediction for the unobservable sufficient statistics in the $k$-th iteration and are denoted as $t^{(k)}$.

In the second step of each iteration, called the M-step, we take $t^{(k)}$ from the E-step and consider it to be the observed $t$. This allows us to solve the score functions of the complete-data log-likelihood given the observed values as $t^{(k)}$. Thus, in the M-step, we are maximizing the complete-data log-likelihood based on the observed $t^{(k)}$ and computing the next iteration of MLEs denoted as $\eta^{(k+1)} \in \Omega_0$. These updated parameter
estimates are chosen to make what we observed, or $t^{(k)}$, be the expected observation. This can be written as $E(t(s, u)|\eta^{(k+1)}) = t^{(k)}$. The EM algorithm moves back and forth between these two steps providing a sequence of estimates for the parameters $\{\eta^{(k)}\}$. If the algorithm converges and some regularity conditions are fulfilled, then the parameter estimates at convergence $\eta^*$ satisfy Equation 7, and are at least a stationary point for $L(\eta)$. We discuss whether or not $\eta^*$ is associated with a global maximum for $L(\eta)$ in Section 2.3.

We now define the specific E-step and M-step used to compute parameter estimates in the case of a mixture of two ZIP distributions in the QTL scenario. As there is a known one-to-one mapping between the natural parameters $\eta$ and the parameters $\phi = (p, \lambda, pQ, \lambda Q)$ used in the original conventional parameterization, we proceed by writing formulas and equations in terms of the conventional parameters $\phi$, with $\phi \in \Theta = (0,1) \times R^+ \times (0,1) \times R^+$, for convenience and ease of interpretation. For example, we use $L(\phi)$ and $L_c(\phi)$ to represent the log-likelihood and complete-data log-likelihood as functions of the conventional parameters.

### 2.2.2 E-step

In the E-step, we compute the expected value of the sufficient statistics with respect to the conditional density of $(z_1, z_2|y, \phi^{(k)}, x)$, where $\phi^{(k)}$ represents the present parameter estimates from the previous iteration, $y$ represents the vector of observed counts, and $x$ represents the observed marker information. Given $n$ independent observations, the conditional expected value $E$ of the five sufficient statistics in $t(y, z_1, z_2)$ can be
computed as:

\[ E(t_1) = \sum E(z_{1i}) = \sum E(z_{1i}) \]

\[ E(t_2) = E \sum (1 - z_{1i})(1 - z_{2i}) = n - \sum E(z_{1i}) - \sum E(z_{2i}) + \sum E(z_{1i}z_{2i}) \]

\[ E(t_3) = E \sum y_i (1 - z_{1i})(1 - z_{2i}) = \sum y_i - \sum y_i E(z_{1i}) - \sum y_i E(z_{2i}) + \sum y_i E(z_{1i}z_{2i}) \]

\[ E(t_4) = E \sum z_{1i} (1 - z_{2i}) = \sum E(z_{1i}) - \sum E(z_{1i}z_{2i}) \]

\[ E(t_5) = E \sum y_i z_{1i} (1 - z_{2i}) = \sum y_i E(z_{1i}) - \sum y_i E(z_{1i}z_{2i}) \]

Thus, the required conditional expected values for E-step computation are

\[ E(z_{1i}|y_i, \phi^{(k)}, x_i), \]

\[ E(z_{2i}|y_i, \phi^{(k)}, x_i), \]

\[ E(z_{1i}z_{2i}|y_i, \phi^{(k)}, x_i). \]

Recall that \( z_{1i} = 0 \) when the QTL genotype is \( q \) and \( z_{1i} = 1 \) when the QTL genotype is \( Q \). Also, \( z_{2i} = 0 \) when the observation was generated by the Poisson component of the ZIP and \( z_{2i} = 1 \) when the observation was generated by the point mass component. As \( z_{1i} \) and \( z_{2i} \) are not independent, we must use the joint conditional density, denoted as \( p(z_{1i}, z_{2i}|y_i, \phi^{(k)}, x_i) \), to compute the above conditional expected values. In the case of a mixture of two ZIP distributions, the pair \((z_{1i}, z_{2i})\) can take on four possible values, namely \{(0,0), (0,1), (1,0), (1,1)\}. Taking the present parameter estimates to be \( \phi^* \), the joint conditional density can be written for each pair of values as:

\[
p(z_{1i} = 0, z_{2i} = 0|y_i, \phi^*, x_i) = \begin{cases} 
\frac{e^{-\lambda \nu}(1-p^*_Q)(1-\pi(x_i;\theta, rL, rR))}{e^{-\lambda \nu}(1-p^*_Q)(1-\pi(x_i;\theta, rL, rR)) + p^*_Q(1-\pi(x_i;\theta, rL, rR)) + e^{-\lambda \nu}(1-p^*_Q)\pi(x_i;\theta, rL, rR) + p^*_Q\pi(x_i;\theta, rL, rR)} & \text{if } y_i = 0 \\
\frac{e^{-\lambda \nu_y}(1-\pi(y_i;\theta, rL, rR))(1-p^*_Q)(1-\pi(x_i;\theta, rL, rR))}{(1-p^*_Q)(1-\pi(x_i;\theta, rL, rR)) + \left(\frac{e^{-\lambda \nu_y}}{y_i!}\right)(1-p^*_Q)(1-\pi(x_i;\theta, rL, rR)) + \left(\frac{e^{-\lambda \nu_y}}{y_i!}\right)(1-p^*_Q)\pi(x_i;\theta, rL, rR)} & \text{if } y_i > 0 
\end{cases}
\]
In the conditional density, we can see that when we observe \( y_i > 0 \), there is no chance that the observation was generated by the point mass component of the ZIP (represented as \( z_{2i} = 1 \)).

Once the four probabilities in the joint conditional density are computed for observation \( i \), the marginal conditional densities for \( z_{1i} \) and \( z_{2i} \) are easily attained through summation. The three conditional expected values can be stated as:
$E(z_{i1}|y_i, \phi^{(k)}, x_i) = \sum_{j=0}^{1} j \cdot p(z_{i1} = j|y_i, \phi^{(k)}, x_i)$

$= \sum_{m=0}^{1} p(z_{i1} = 1, z_{2i} = m|y_i, \phi^{(k)}, x_i)$

$E(z_{i2}|y_i, \phi^{(k)}, x_i) = \sum_{m=0}^{1} m \cdot p(z_{i2} = m|y_i, \phi^{(k)}, x_i)$

$= \sum_{j=0}^{1} p(z_{i1} = j, z_{i2} = 1|y_i, \phi^{(k)}, x_i)$

$E(z_{i1}z_{i2}|y_i, \phi^{(k)}, x_i) = \sum_{j=0}^{1} \sum_{m=0}^{1} j \cdot m \cdot p(z_{i1} = j, z_{i2} = m|y_i, \phi^{(k)}, x_i)$

$= p(z_{i1} = 1, z_{i2} = 1|y_i, \phi^{(k)}, x_i)$.

These conditional expected values are used to compute the conditional expected values of the sufficient statistics, which serve as a prediction for the unobserved sufficient statistics at the present iteration. In the next step of the EM algorithm, we consider these expected values of the sufficient statistics, denoted as $t(y, z_1, z_2)^{(k)}$, to be the observed values for $t(y, z_1, z_2)$, which allows us to compute the MLEs for the complete-data log-likelihood.

### 2.2.3 M-step

Replacing the unobservable sufficient statistics in the complete-data likelihood in Equation 4 with their expected values $t(y, z_1, z_2)^{(k)}$ computed in the E-step allows us to solve the related score functions relatively easily. The subsequent solutions, which are the MLEs for the complete-data likelihood, provide the next parameter estimates $\phi^{(k+1)}$ in the iterative procedure. As the formulas for the parameter estimates are more easily interpreted in terms of $(y, z_1, z_2)$ than in terms of $t$, we have shown them in both forms.
Estimates in terms of parameter \((y, z_1, z_2)\)

- \(p_q\): \[
\frac{\sum_{i=1}^{n} [(1 - z_{1i})z_{2i}]}{\sum_{i=1}^{n} (1 - z_{1i})} = \frac{n - t_1 - t_2}{n - t_1}
\]

- \(\lambda_q\): \[
\frac{\sum_{i=1}^{n} [(1 - z_{1i})(1 - z_{2i})y_i]}{\sum_{i=1}^{n} [(1 - z_{1i})(1 - z_{2i})]} = \frac{t_3}{t_2}
\]

- \(pQ\): \[
\frac{\sum_{i=1}^{n} z_{1i}z_{2i}}{\sum_{i=1}^{n} z_{1i}} = \frac{t_1 - t_4}{t_1}
\]

- \(\lambda_Q\): \[
\frac{\sum_{i=1}^{n} z_{1i}(1 - z_{2i})y_i}{\sum_{i=1}^{n} z_{1i}(1 - z_{2i})} = \frac{t_5}{t_4}
\]

After plugging in the expected values of the sufficient statistics in the far right column, the updated parameter estimates are then used in the next iteration of the E-step. This procedure continues until convergence of the sequence \(\{\phi^{(k)}\}\) to some \(\phi^*\) for which we have the equality

\[
E(t(y, z_1, z_2)|\phi^*, x) = E(t(y, z_1, z_2)|y, \phi^*, x).
\]  

We now investigate the convergence behavior of the EM algorithm in the case of a mixture of two ZIP distributions.

### 2.3 EM Convergence

The EM algorithm definition in Section 2.2 applies to exponential and curved exponential families, but this definition is just a special case of a more general definition of the EM algorithm. In order to investigate the convergence of the algorithm for a mix-
ture of ZIP distributions, we will now utilize the more general definition using notation specific to the ZIP QTL scenario. We use $L_c(\phi)$ to denote the complete-data log-likelihood. Earlier, iteration $k$ of the EM algorithm was described in two steps. First, a prediction of the unobservable sufficient statistics in the $k$-th step was computed as $t(y, z_1, z_2)^{(k)} = E \left[ t(y, z_1, z_2) \mid y, \phi^{(k)}, x \right]$. Then, this expected value for the sufficient statistics was taken to be the observed values for $t(y, z_1, z_2)$, and the subsequent $L_c(\phi)$ was maximized, producing the next parameter estimates $\phi^{(k+1)}$. In the more general definition of steps, the first step is defined as computing the conditional expected value of $L_c(\phi)$ with respect to the conditional density of $(z_1, z_2 \mid y, \phi^{(k)}, x)$. In this step, the unobserved variables in $L_c(\phi)$ are integrated out according to our present estimate of their conditional density based on our present estimate $\phi^{(k)}$ for $\phi$. In the second step, we maximize this conditional expected value (a function free of $z_1, z_2$) producing the next parameter estimates $\phi^{(k+1)}$. When $L_c(\phi)$ is linear in the sufficient statistics, as in the case of an exponential family, these two definitions are equivalent. We denote the conditional expected value of $L_c(\phi)$ computed in the first step as $Q(\phi \mid \phi^{(k)})$, and define it as

$$Q(\phi \mid \phi^{(k)}) = E \left[ L_c(\phi) \mid y, z_1, z_2, x \right] \mid y, \phi^{(k)}, x.$$  

This function computes the conditional expected value of $L_c(\phi)$ based on the conditional density of the unknown variables $z_1$ and $z_2$ given the current parameter estimates $\phi^{(k)}$ and the observed values of $y$ and $x$. In the case of a mixture of two ZIP distributions, the pair $(z_{1i}, z_{2i})$ can take on four possible values, namely $\{(0,0), (0,1), (1,0), (1,1)\}$. For $n$ independent observations, we can write $Q(\phi \mid \phi^{(k)})$ as

$$Q(\phi \mid \phi^{(k)}) = \sum_{j=0}^1 \sum_{m=0}^1 \sum_{i=1}^n \log \left\{ f(y_i, z_{1i} = j, z_{2i} = m \mid \phi, x_i) \right\} p(z_{1i} = j, z_{2i} = m \mid y_i, \phi^{(k)}, x_i)$$

where $f(\cdot)$ is the complete-data density, $p(z_{1i}, z_{2i} \mid y_i, \phi^{(k)}, x_i)$ is the conditional joint density of $(z_{1i}, z_{2i})$ under the present parameter estimates of $\phi^{(k)}$, and $y_i$ and $x_i$ are the
observed count and marker information, respectively, for observation \( i \).

The convergence of the EM algorithm depends on the characteristics of both the log-likelihood \( L(\phi) \), and the complete-data log-likelihood \( L_c(\phi) \) through \( Q(\phi|\phi^{(k)}) \). As the likelihood is a PMF in the ZIP QTL backcross scenario, the \( L(\phi) \) is bounded above. We also know that iterative evaluations of \( L(\phi) \) are monotone in the sense that \( L(\phi^{(k+1)}) \geq L(\phi^{(k)}) \) simply by the EM algorithm procedure itself. These two elements provide convergence of the sequence \( L(\phi^{(k)}) \), which can be stated as \( L(\phi^{(k)}) \rightarrow L^* \) for some \( L^* \). Though we wish \( L^* \) to be a global maximum of \( L(\phi) \), this is not a given from only the convergence of the sequence. Wu (Theorem 2, 1983) showed that \( L^* \) is at least a stationary point of \( L(\phi) \) under certain regularity conditions. These conditions can be summarized in this case as:

1. \( \Omega_0 \) is a subset in the \( r \)-dimensional Euclidean space \( \mathbb{R}^r \),
2. \( L(\phi) \) is bounded above,
3. \( L(\phi) \) is continuous in \( \Theta \) and differentiable in the interior of \( \Theta \),
4. \( Q(\phi|\phi^{(k)}) \) is continuous in both \( \phi \) and \( \phi^{(k)} \).

Conditions (i) and (ii) have already been discussed. For any \( \phi \in \Theta \), the \( L(\phi) \) is continuous and all first derivatives exist. These characteristics are apparent from the form of the log-likelihood in Equation 3 which shows smoothness in \( L(\phi) \) for any \( \phi \) in the interior of \( \Theta \). Thus, Condition (iii) holds. Condition (iv) holds in the case of a mixture of two ZIPS and is discussed in Section 5, the Appendix.

Given these regularity conditions, Wu (Corollary 1, 1983) showed that \( L^* \) is at least a local maximizer if the derivative of \( Q(\phi|\phi^*) \) w.r.t. \( \phi^* \) is continuous in \( \phi \) and \( \phi^* \). Again, for \( \phi \) and \( \phi^* \) in the interior of \( \Theta \), this continuity holds. The step from local maximum to global maximum is more difficult. Looking at the form of \( L(\phi) \), it is not apparent that \( L(\phi) \) is unimodal, which implies that the \( L^* \) found at convergence may not be a global maximizer. For this reason, exploring sensitivity to different starting points is a good practice.
2.4 Asymptotic Distribution of MLE

At convergence, the EM algorithm provides a solution to the likelihood equations. As stated above, we have shown that the EM algorithm, when applied to the likelihood function for a mixture of two ZIPs in the QTL scenario, will converge to a solution that is at least a local maximizer of $L(\phi)$. At the testing position associated with the true QTL (provided one exists), the MLE for $\phi$, or $\hat{\phi}_{MLE}$, for $n$ independent observations has the asymptotic distribution of

$$\sqrt{n}(\hat{\phi}_{MLE} - \phi_0) \xrightarrow{d} N(0, I^{-1}_1(\phi_0))$$

where $\phi_0$ denotes the true value of the parameter, and $I_1(\phi_0)$ denotes the Fisher expected information for one observation (see Section 5 for regularity conditions). For $(Y_1, X_1), (Y_2, X_2), (Y_3, X_3), \ldots, (Y_n, X_n) \overset{i.i.d.}{\sim} f(y, x|\phi)$, we define the Fisher expected information in the data as

$$I_n(\phi_0) = -E_{\phi_0} \left[ \frac{\partial^2}{\partial \phi \partial \phi'} \log \prod_{i=1}^{n} f(y_i, x_i|\phi) \right]_{\phi = \phi_0}$$

$$= -n E_{\phi_0} \left[ \frac{\partial^2}{\partial \phi \partial \phi'} \log f(y, x|\phi) \right]_{\phi = \phi_0}$$

$$= n I_1(\phi_0)$$

2.5 Estimating the Asymptotic Variance-Covariance Matrix When Using the EM Algorithm

When the derivatives of the log-likelihood $L(\phi)$ are easily attainable, we can approximate the asymptotic variance-covariance matrix for the parameter estimates by replacing the unknown $\phi_0$ in $I_n(\phi_0)$ with $\hat{\phi}$. This approximation for the information is denoted as $I_n(\hat{\phi})$. Therefore, the approximated variance-covariance matrix for $\hat{\phi}$ is
Another approach to estimating $I_n(\phi_0)$, that does not require taking the expected values of the derivatives, uses the observed information in the data rather than the expected information. We define the observed information as

$$I_{\text{obs}}(\hat{\phi}) = - \left( \sum_{i=1}^{n} \frac{\partial^2}{\partial \phi \partial \phi'} \log f(y_i, x_i | \phi) \bigg|_{\phi=\hat{\phi}} \right)$$

where $(Y_1, X_1), (Y_2, X_2), (Y_3, X_3), \ldots, (Y_n, X_n)$ are i.i.d. with joint density $f(y, x | \phi)$.

The estimated variance-covariance matrix based on $I_{\text{obs}}(\hat{\phi})$ has been found to be superior in performance to the former estimate of $I_n(\phi)$ (Efron and Hinkley, 1978).

In both of the above estimates for $I_n(\phi_0)$, second derivatives of $L(\phi)$ are utilized. In cases where the EM algorithm is applied, direct maximization of $L(\phi)$ through derivatives is a difficult task due to the form of the equation. For the same reason, use of either of the approaches described above for estimating the variability in the parameter estimates is also difficult. Fortunately, we can again utilize the relationship between the complete-data density, the incomplete-data density, and the joint conditional density denoted as $\prod_{i=1}^{n} g(z_{1i}, z_{2i} | y_i, \phi, x_i)$ in order to attain an estimate for $I_{\text{obs}}(\hat{\phi})$ at EM convergence without directly computing the second derivatives of $L(\phi)$. Recall,

$$L(\phi) = L_c(\phi) - \sum_{i=1}^{n} \log g(z_{1i}, z_{2i} | y_i, \phi, x_i).$$

Taking the second derivative of both sides of the equation

$$\frac{\partial^2}{\partial \phi \partial \phi'} L(\phi) = \frac{\partial^2}{\partial \phi \partial \phi'} L_c(\phi) - \frac{\partial^2}{\partial \phi \partial \phi'} \left\{ \sum_{i=1}^{n} \log g(z_{1i}, z_{2i} | y_i, \phi, x_i) \right\},$$

then computing the expected value of both sides with respect to the joint conditional density of $z_1$ and $z_2$ with $\phi$ replaced by the parameter estimates $\hat{\phi}_{\text{MLE}}$ at EM convergence, we have

$$\frac{\partial^2}{\partial \phi \partial \phi'} L(\phi) = -E \left[ \frac{\partial^2}{\partial \phi \partial \phi'} L(\phi) \bigg| y, \hat{\phi}_{\text{MLE}}, x \right]$$

$$= -E \left[ \frac{\partial^2}{\partial \phi \partial \phi'} L_c(\phi) \bigg| y, \hat{\phi}_{\text{MLE}}, x \right]$$

$$- E \left[ \frac{\partial^2}{\partial \phi \partial \phi'} \left\{ \sum_{i=1}^{n} \log g(z_{1i}, z_{2i} | y_i, \phi, x_i) \right\} \bigg| y, \hat{\phi}_{\text{MLE}}, x \right].$$
Substituting $\hat{\phi}_{MLE}$ for $\phi$ and inserting the observed values into the left side of Equation 16 provides us with the observed information, or $I_{\text{obs}}(\hat{\phi})$, as defined above. By instead utilizing the right side of the equation, we can easily estimate $I_{\text{obs}}(\hat{\phi})$ using the conditional expected values computed in the final E-step of the EM algorithm at convergence.

Louis (1982), Meilijson (1989), and Meng and Rubin (1991) all refer to the first term on the right side of Equation 16 as the complete information, or $I_{\text{complete}}$, and the subtracted term as the missing information, or $I_{\text{missing}}$. They also note that the relationship

$$I_{\text{obs}}(\hat{\phi}) = I_{\text{complete}} - I_{\text{missing}}$$

is an application of the ‘missing information principle’ described by Orchard and Woodbury (1972).

### 2.6 Genome-wide Test for Existence of QTL

In practice, due to the strong correlation that exists between test statistics at consecutive locations, we do not look at an individual test statistic to determine if there is sufficient evidence for the existence of a QTL. Instead, we consider the full set of test statistics across the genome simultaneously to test for the existence of a QTL anywhere along the genome. If a QTL exists, we expect the test statistic at the QTL position to be large, but we also expect the test statistics at nearby locations to be inflated even though no QTL exists at these locations (the reason for this dependence among test statistics was discussed in Section 1.4 of the General Introduction). The test statistic computed at each location compares a full model allowing for the presence of two distinct ZIPs requiring 4 parameters (i.e. the model when a QTL is present) and a reduced model allowing for the presence of only one ZIP requiring 2 parameters. The test statistic $T_j$
at each location \( j \) for \( j \in \{1, 2, \ldots, M\} \) takes on the familiar form of \( T_j = -2 \log \lambda_j \) where \( \lambda_j \) is a likelihood ratio defined as

\[
\lambda_j = \frac{\sup_{\phi \in \Theta_0} \prod_{i=1}^{n} f(y_i, x_{ij} | \phi)}{\sup_{\phi \in \Theta_0} \prod_{i=1}^{n} f(y_i, x_{ij} | \phi)},
\]

and \( \Theta_0 = \{ \phi \in \Theta : p_q = p_Q \text{ and } \lambda_q = \lambda_Q \} \). We have added the subscript \( j \) to the previously denoted marker information vector \( x_i \) in order to specify which markers are used in the likelihood equations when testing for a QTL at position \( j \). Specifically, \( x_{ij} \) contains the genotypes of the nearest observed markers flanking the test position \( j \) for individual \( i \).

The actual set of hypotheses related to testing for a QTL can be expressed as

\[
H_0 : \text{no QTL is present} \\
H_A : \text{a QTL is present.}
\]

To test this set of hypothesis we use the test statistic \( T^* \) which represents the maximum test statistic among all test statistics \( T_j \) computed across the genome, or

\[
T^* = \sup_j T_j.
\]

We use a permutation procedure that accounts for the correlation between consecutive testing locations to establish an empirical null distribution for \( T^* \) (see Section 3.1 of the General Introduction). The \((1 - \alpha)^{th}\) quantile of the null distribution of maximum test statistics is used for a significance threshold providing a genome-wide error rate no larger than \( \alpha \).

3 Simulation Study

We conducted a Monte Carlo simulation study to compare the power and accuracy for QTL detection between our method and two other existing methods that are applicable in backcross studies involving a ZIP trait. We limited the parameter configurations
investigated for \((\lambda_q, \lambda_Q)\) to those having at least one of the two ZIP Poisson components providing an appreciable proportion of mass at zero. In other words, for \(Y \sim \text{Poisson}(\lambda)\), we chose the configurations of \((\lambda_q, \lambda_Q)\) from \(\{(\lambda_q, \lambda_Q) : P(Y = 0|\lambda_q) \gg 0 \text{ and/or } P(Y = 0|\lambda_Q) \gg 0\}\). For \((\lambda_q, \lambda_Q)\) parameter configurations in the complement of this set, any observed zero is extremely likely to have been generated by a point mass ZIP component. For such distributions, this essentially equates to being able to observe \(z_{2i}\), the variable that designates whether the observation was generated from the point mass or Poisson component. When \(z_{2i}\) is observable, the method of Broman (2003) could be extended to address our Poisson case. In his paper, Broman proposed a method to detect a QTL when the trait of interest is a mixture between a spike at zero and a normal distribution.

### 3.1 Design

We simulated data for a backcross population with a genome composed of 2 chromosomes. Each chromosome was 100 cM in length, and we simulated a single QTL near the middle of chromosome 2 (at 55 cM from the left end). Genotypes for the markers and the QTL were simulated based on the Poisson process modeling of crossover event occurrence on the genome (see Section 1.4 of the General Introduction for a more detailed description). Markers were genotyped every 10 cM across the genome, and we randomly removed 10% of the observed markers to simulate missing marker data that is commonly seen in practice.

All investigated parameter configurations provided a lower expected count for an observation from an individual with a QTL genotype of \(q\), compared to an observation from an individual with a QTL genotype of \(Q\). If a high count was seen as advantageous, then this would imply that the QTL genotype of \(Q\) was the favorable allele. With four parameters, there are many configurations that fulfill this criterion, but we have limited our investigated parameter space to the region where \(p_q \geq p_Q\) and \(\lambda_q \leq \lambda_Q\).
This parameter subspace represents parameter configurations that are believed to be the most relevant from the biological perspective. For this subspace, an observation from an organism with a favorable allele is both less likely to have been generated from the point mass component, and more likely to have a high positive count from the Poisson component.

For each parameter configuration, we simulated 500 data sets each with \( n = 160 \) individuals. The observed trait \( y_i \) for each individual in each data set was generated based on the given parameters and simulated QTL genotype. We tested for the presence of a QTL every 2 cM across the genome. The significance threshold for QTL detection for each data set was based on 100 permutations of the observed traits, and provides a genome-wide type I error rate no larger than \( \alpha = 0.05 \).

3.2 Results

For comparison, we applied our proposed method and two other existing methods for QTL detection to each simulated data set. The first method for comparison was the classical method, based on normality, applied to transformed simulated data. The transformation was the square-root transformation recommended for count data by Anscombe (1948), and it was applied in order to better achieve the assumptions of equal variance and approximate normality. The second method was a rank-based nonparametric method for mapping a quantitative trait proposed by Kruglyak and Lander (1995). Their method is an extension of the Wilcoxon rank-sum test. QTL were classified as detected when the LR test statistic exceeded the significance threshold and the estimated location of the QTL was on chromosome 2. The power of each procedure was computed as the proportion of simulations with a detected QTL. To compare the performance of estimating the QTL location, we computed the average absolute deviation (AAD) of the estimated location for all detected QTL.
Figure 1  Estimated power (+/- 2 SD) to detect QTL based on 500 simulations for all methods at each parameter configuration. Each box is associated with a fixed parameter setting for \( p_q \) and \( P_Q \) (see top of box), and a variety of parameter setting for \( \lambda_q \) and \( \lambda_Q \) (see horizontal axis). Movement to the right on the horizontal axis is associated with larger values for both \( \lambda_Q \) and \( \lambda_q \) while maintaining the constant ratio \( \lambda_Q/\lambda_q = 1.5 \).

### 3.2.1 Power and Estimated Location of QTL

Figure 1 is a plot of the estimated power for each method at each parameter configuration, and and Table 1 provides the power estimates and estimated standard deviations. As Figure 1 shows, there is little distinction in the performance of the methods when \( \lambda_q \) and \( \lambda_Q \) are relatively close to zero. These locations are represented at the far left of each of the three boxes in Figure 1. Using McNemar’s test for paired data, we tested the null hypothesis of no difference in power between each pairwise comparison of methods at \( \alpha = 0.05 \) level. When \((\lambda_q, \lambda_Q) = (2, 3)\) and \((p_q, p_Q) = (0.20, 0.20)\), there was no significant difference between the ZIP method and the normal method, but the NP method tested significantly lower in power than both the ZIP method and the normal method. For the same \((\lambda_q, \lambda_Q)\) configuration with \((p_q, p_Q) = (0.20, 0.15)\), there was no significant difference in the methods. For \((p_q, p_Q) = (0.30, 0.15)\), the ZIP method tested
Table 1
Power estimates and estimated (SD) for 500 simulations

<table>
<thead>
<tr>
<th>$(\lambda_q, \lambda_Q)$</th>
<th>$(0.20,0.20)$</th>
<th>$(p_q, p_Q)$</th>
<th>$(0.20,0.15)$</th>
<th>$(0.30,0.15)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIP</td>
<td>Normal</td>
<td>NP</td>
<td>ZIP</td>
</tr>
<tr>
<td>(2,3)</td>
<td>0.522</td>
<td>0.508</td>
<td>0.446</td>
<td>0.600</td>
</tr>
<tr>
<td>(0.022) (0.022) (0.022)</td>
<td>0.021</td>
<td>0.021</td>
<td>0.021</td>
<td>(0.017)</td>
</tr>
<tr>
<td>(3,4.5)</td>
<td>0.814</td>
<td>0.558</td>
<td>0.572</td>
<td>0.878</td>
</tr>
<tr>
<td>(0.017) (0.022) (0.022)</td>
<td>0.015</td>
<td>0.020</td>
<td>0.017</td>
<td>(0.010)</td>
</tr>
<tr>
<td>(4,5)</td>
<td>0.918</td>
<td>0.498</td>
<td>0.674</td>
<td>0.946</td>
</tr>
<tr>
<td>(0.012) (0.022) (0.021)</td>
<td>0.010</td>
<td>0.021</td>
<td>0.014</td>
<td>(0.005)</td>
</tr>
</tbody>
</table>

significantly lower in power than the other two methods. Though some of the methods were significantly different when $\lambda_q$ and $\lambda_Q$ were near zero, the magnitude of the power estimates were all relatively close in proximity with respect to the possible range from 0 to 1.

As $\lambda_q$ and $\lambda_Q$ move farther away from zero, the ZIP and nonparametric (NP) method behave as expected, in that they gain power as $\lambda_Q - \lambda_q$ increases for a fixed $(p_q, p_Q)$ configuration. The normal model analysis is not always predictable in this respect, especially for $(\lambda_q, \lambda_Q)$ far from zero, such as $(\lambda_q, \lambda_Q) = (6, 9)$ seen in earlier simulations not shown here, where it becomes less likely that the normal approximation is reasonable. When $(\lambda_q, \lambda_Q) = (3, 4.5)$ and $(p_q, p_Q) = (0.20, 0.20)$, the ZIP method performed significantly better than both the normal and NP method, which were not significantly different from each other. For the same $(\lambda_q, \lambda_Q)$ with $(p_q, p_Q) = (0.20, 0.15)$ all methods were significantly different with the ZIP method exhibiting the highest power and the normal method exhibiting the lowest power. When $(p_q, p_Q) = (0.30, 0.15)$, the ZIP method was not significantly different than either of the other two methods, but the NP was significantly better than the normal method.

In the final configuration with $(\lambda_q, \lambda_Q) = (4, 6)$ in which the normal approximation
becomes less reasonable, the power for the normal method was significantly lower than the power for both the ZIP and NP method for all \((p_q, p_Q)\) configurations. At this same \((\lambda_q, \lambda_Q)\) configuration, the ZIP method was significantly higher than the NP when \((p_q, p_Q) = (0.20, 0.20)\) and \((0.20, 0.15)\). There was no significant difference between the ZIP and NP methods when \((p_q, p_Q) = (0.30, 0.15)\).

The ZIP method exhibits the greatest advantage over other approaches when \(p_q = p_Q\) (see Figure 1A). This advantage was also seen for proportion parameter settings of \((0.15, 0.15)\) and \((0.30, 0.30)\) in earlier simulations (not shown here). This suggests that the ZIP method does better compared to the other methods at recognizing the difference between \(\lambda_q\) and \(\lambda_Q\) that manifests itself in the positive observed counts. When a difference in the proportions is included along with a difference in the Poisson parameters, the ZIP approach loses some advantage (see Figure 1B). When more and more distinguishing power is placed on the proportion parameters of the ZIP distributions (movement from the left box to the right box), the normal and nonparametric seem to catch up to the ZIP method. Though, again, we see the normal method start to drop in performance after the Poisson parameters are so far from zero that the gain in power from the proportion differential can no longer overcome the drop in power due to the non-normality of the data. Overall, the ZIP method does as well or better than the other methods at the configurations investigated.

To compare the accuracy in locating the QTL among methods, we computed the average absolute deviation (AAD) of the estimated location for all detected QTL (see Table 2). Figure 2 provides a plot of the AAD for each method at each configuration. As Figure 2 shows, the ZIP method tends to have the smallest AAD for all configurations. There is also a general trend across all \((p_q, p_Q)\) configurations for the AAD to decrease as the Poisson parameters move farther away from zero. This trend roughly coincides with an increase in power that was seen in Figure 1.
3.2.2 Estimated Parameters at Detected QTL

Summaries of the parameter estimates from the ZIP method at detected locations and their sampling variability are shown in Tables 3, 4, and 5. The empirical standard deviations (SD) represent the square root of the variance of the parameter estimates at all detected QTL. The asymptotic standard deviations (ASD) represent the average of the asymptotic standard deviations computed at each detected QTL. Each table coincides with a fixed \((p,q, pQ)\) parameter configuration. Tables 3, 4, and 5, coincide with \((p,q, pQ)\) parameter configurations of \((0.20, 0.20)\), \((0.20, 0.15)\), and \((0.30, 0.15)\), respectively.

Though the MLEs are asymptotically unbiased, there does appear to be some selection bias that manifests in the estimated \(\lambda\) values. This can be seen when we subset the estimates based on detection. Estimated parameter configurations with \(\hat{\lambda}_q < \lambda_q\) and \(\hat{\lambda}_Q > \lambda_Q\) were more likely to be detected than were estimated configurations with \(\hat{\lambda}_q > \lambda_q\) and \(\hat{\lambda}_Q < \lambda_Q\). As all tables show, the average estimated \(\lambda_q\) was slightly deflated and the average estimated \(\lambda_Q\) was slightly inflated compared to their true values.
Table 2
Estimated average absolute deviation (AAD) of the estimated QTL location and estimated (SD) at detected QTL for 500 simulations.

<table>
<thead>
<tr>
<th>(λ_q, λ_Q)</th>
<th>(0.20,0.20)</th>
<th>(p_q, p_Q)</th>
<th>(0.20,0.15)</th>
<th>(0.30,0.15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIP Normal</td>
<td>NP</td>
<td>ZIP Normal</td>
<td>NP</td>
</tr>
<tr>
<td>(2,3)</td>
<td>9.31 9.76</td>
<td>10.97</td>
<td>8.34 8.54</td>
<td>9.01</td>
</tr>
<tr>
<td></td>
<td>(0.36) (0.42)</td>
<td>(0.56)</td>
<td>(0.24) (0.24)</td>
<td>(0.26) (0.13) (0.13)</td>
</tr>
<tr>
<td>(3,4,5)</td>
<td>7.77 9.32</td>
<td>9.71</td>
<td>7.20 8.26</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>(0.19) (0.36)</td>
<td>(0.36)</td>
<td>(0.13) (0.22)</td>
<td>(0.19) (0.09) (0.12)</td>
</tr>
<tr>
<td>(4,5)</td>
<td>6.23 8.00</td>
<td>8.74</td>
<td>5.44 7.91</td>
<td>6.92</td>
</tr>
<tr>
<td></td>
<td>(0.11) (0.31)</td>
<td>(0.28)</td>
<td>(0.08) (0.21)</td>
<td>(0.14) (0.06) (0.11)</td>
</tr>
</tbody>
</table>

This phenomenon was not seen when considering the full set of estimated parameters, at detected and not detected QTL locations. The effect of selection bias was not as prominent in the proportion parameter estimates.

We now discuss the estimated variability in the parameter estimates by comparing the estimates to their complete-data likelihood counterparts. The expected asymptotic SD under the complete-data scenario are provided in the last column of Tables 3, 4, and 5. With complete information (i.e. knowledge of the QTL genotype and knowledge of whether the Poisson or point mass generated each observation), the asymptotic variances of the proportion parameter estimates would be the traditional variances of \( \frac{p_i(1-p_i)}{n_i} \) computed within each of the two QTL groups with \( i \in \{q, Q\} \). Because we do not know which part of the ZIP generated each observed zero, and because QTL group membership is not observed, we do not have complete information, and the asymptotic variances for \( \hat{p}_q \) and \( \hat{p}_Q \) are higher than in the complete-information case. In each table, though the ASD estimates for \( \hat{p}_q \) and \( \hat{p}_Q \) are both larger than would be expected in the complete-data case, we do see these values decrease and get closer to those expected in the complete-data case as the \( \lambda \) values get larger, i.e. as we move down a given table. This is because missing information is associated with observed zeroes that can be
Table 3
Summary of parameter estimates for 500 simulations
Configuration: $p_q = p_Q = 0.20$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical mean of estimates</th>
<th>Complete data mean of estimates</th>
<th>Complete data asym. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_q = 2 &lt; \lambda_Q = 3$</td>
<td>2.00</td>
<td>1.878</td>
<td>0.1927</td>
<td>0.2200</td>
<td>0.1768</td>
</tr>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.192</td>
<td>0.0749</td>
<td>0.0709</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>0.20</td>
<td>0.187</td>
<td>0.0608</td>
<td>0.0510</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>3.00</td>
<td>3.122</td>
<td>0.2218</td>
<td>0.2455</td>
<td>0.2165</td>
</tr>
<tr>
<td>$\lambda_q = 3 &lt; \lambda_Q = 4.5$</td>
<td>3.00</td>
<td>2.929</td>
<td>0.2337</td>
<td>0.2455</td>
<td>0.2165</td>
</tr>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.200</td>
<td>0.0537</td>
<td>0.0466</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>0.20</td>
<td>0.197</td>
<td>0.0539</td>
<td>0.0477</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>4.50</td>
<td>4.579</td>
<td>0.2586</td>
<td>0.2820</td>
<td>0.2652</td>
</tr>
<tr>
<td>$\lambda_q = 4 &lt; \lambda_Q = 6$</td>
<td>4.00</td>
<td>3.976</td>
<td>0.2640</td>
<td>0.2704</td>
<td>0.2500</td>
</tr>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.199</td>
<td>0.0472</td>
<td>0.0459</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>0.20</td>
<td>0.199</td>
<td>0.0472</td>
<td>0.0459</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>6.00</td>
<td>6.040</td>
<td>0.3093</td>
<td>0.3189</td>
<td>0.3062</td>
</tr>
</tbody>
</table>

Based on $259^a$, $407^b$, $459^c$ significant simulations

generated from either the Poisson or the point mass. As the $\lambda$ values move away from zero and we move down on a table, there is less missing information. Thus, we see the ASD estimates get smaller even though the proportion parameter values have remained fixed. As another comment related to the proportion parameter ASD estimates, the ASD estimates for $\hat{p}_q$ are consistently larger than the ASD estimates for $\hat{p}_Q$. This result stems from both the fact that $p_q \geq p_Q$, and also that configurations of $p_q$ was always associated with a $\lambda_q$ that was smaller than $\lambda_Q$. Thus, the ZIP associated with a QTL of $q$ was always associated with more missing information in the configurations we investigated. Similar trends were seen in the ASD estimates for $\hat{\lambda}_q$ and $\hat{\lambda}_Q$. All estimated ASD values were larger than those expected under the complete-data scenario, and all ASD estimates moved closer to those expected in the complete-data scenario as the true
Table 4
Summary of parameter estimates for 500 simulations
Configuration: $p_q = 0.20$ and $p_Q = 0.15$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asym. SD</th>
<th>Complete data asym. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.190</td>
<td>0.0780</td>
<td>0.0701</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>2.00</td>
<td>1.903</td>
<td>0.2001</td>
<td>0.2217</td>
<td>0.1768</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.139</td>
<td>0.0544</td>
<td>0.0475</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>3.00</td>
<td>3.090</td>
<td>0.2313</td>
<td>0.2374</td>
<td>0.2100</td>
</tr>
</tbody>
</table>

$\lambda_q = 2 < \lambda_Q = 3^a$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asym. SD</th>
<th>Complete data asym. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.200</td>
<td>0.0541</td>
<td>0.0528</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>3.00</td>
<td>2.936</td>
<td>0.2149</td>
<td>0.2463</td>
<td>.02165</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.146</td>
<td>0.0460</td>
<td>0.0421</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>4.50</td>
<td>4.547</td>
<td>0.2628</td>
<td>0.2722</td>
<td>0.2572</td>
</tr>
</tbody>
</table>

$\lambda_q = 3 < \lambda_Q = 4.5^b$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asym. SD</th>
<th>Complete data asym. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.20</td>
<td>0.201</td>
<td>0.0504</td>
<td>0.0478</td>
<td>0.0450</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>4.00</td>
<td>3.962</td>
<td>0.2662</td>
<td>0.2699</td>
<td>0.2500</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.150</td>
<td>0.0446</td>
<td>0.0415</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>6.00</td>
<td>6.023</td>
<td>0.3038</td>
<td>0.3100</td>
<td>0.2970</td>
</tr>
</tbody>
</table>

$\lambda_q = 4 < \lambda_Q = 6^c$

Based on 330$^a$, 439$^b$, 473$^c$ significant simulations

$\lambda$ values moved away from zero.

We now compare the ASD estimates to the empirical SD values. Comparing the mean of the estimated ASD to the SD for all the tables, we see a general trend that the ASD estimates for $\hat{p}_q$ and $\hat{p}_Q$ tended to underestimate the observed sampling variance in these parameter estimates. In contrast, we see that the ASD estimates for $\hat{\lambda}_q$ and $\hat{\lambda}_Q$ tended to overestimate the observed sampling variance. By considering the variability in our ASD estimates, we can see if the empirical SD is within the range of mean ASD +/- 2 standard deviations of the ASD (not shown in tables). Only one of the nine empirical SD values for $\hat{p}_q$ in Tables 3-5 was within this range. Similarly, none of the empirical SD values for $\hat{p}_Q$ were within the estimated ASD interval. In contrast to these rates, the coverage for the empirical SD values for $\hat{\lambda}_q$ and $\hat{\lambda}_Q$ were much higher. Seven out of
Table 5
Summary of parameter estimates for 500 simulations
Configuration: $p_q = 0.30$ and $p_Q = 0.15$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asympt. SD</th>
<th>Complete data asympt. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.30</td>
<td>0.300</td>
<td>0.0730</td>
<td>0.0717</td>
<td>0.0510</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>2.00</td>
<td>1.955</td>
<td>0.2450</td>
<td>0.2414</td>
<td>0.1890</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.137</td>
<td>0.0517</td>
<td>0.0480</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>3.00</td>
<td>3.027</td>
<td>0.2148</td>
<td>0.2348</td>
<td>0.2100</td>
</tr>
</tbody>
</table>

$\lambda_q = 2 < \lambda_Q = 3^a$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asympt. SD</th>
<th>Complete data asympt. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.30</td>
<td>0.305</td>
<td>0.0561</td>
<td>0.0578</td>
<td>0.0510</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>3.00</td>
<td>2.963</td>
<td>0.2532</td>
<td>0.2663</td>
<td>0.2315</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.146</td>
<td>0.0447</td>
<td>0.0423</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>4.50</td>
<td>4.526</td>
<td>0.2627</td>
<td>0.2709</td>
<td>0.2572</td>
</tr>
</tbody>
</table>

$\lambda_q = 3 < \lambda_Q = 4.5^b$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Mean of estimates</th>
<th>Empirical SD</th>
<th>Mean of asympt. SD</th>
<th>Complete data asympt. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_q$</td>
<td>0.30</td>
<td>0.296</td>
<td>0.0556</td>
<td>0.0536</td>
<td>0.0510</td>
</tr>
<tr>
<td>$\lambda_q$</td>
<td>4.00</td>
<td>3.972</td>
<td>0.2869</td>
<td>0.2910</td>
<td>0.2673</td>
</tr>
<tr>
<td>$p_Q$</td>
<td>0.15</td>
<td>0.150</td>
<td>0.0441</td>
<td>0.0417</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\lambda_Q$</td>
<td>6.00</td>
<td>6.017</td>
<td>0.3095</td>
<td>0.3082</td>
<td>0.2970</td>
</tr>
</tbody>
</table>

$\lambda_q = 4 < \lambda_Q = 6^c$

Based on 420$^a$, 471$^b$, 493$^c$ significant simulations

the nine intervals covered the empirical SD for both $\hat{\lambda}_q$ and $\hat{\lambda}_Q$. This suggests that the sample size required for asymptotic approximation to be effective is perhaps less for $\lambda$ than for $p$. In general, when we look at which parameter configurations were associated with coverage, we see that larger $\lambda$ values and larger $p$ values were associated with a higher rate of coverage compared to smaller $\lambda$ values and smaller $p$ values. Though not all empirical SD values are covered by the estimated ASD intervals, the ASD still seems to estimate its respective empirical value reasonably well for use in practice with this sample size.
4 Example: Shoot Regeneration Study

The commercial production of some ornamental plants can be achieved through a method called micropropagation. This is the process of cloning offspring from the tissue culture of a single plant. It has been shown that the ability to micropropogate varies widely within a plant species. In this study, researchers were interested in studying the genetic basis for differences in shoot regeneration between ecotypes of the *Arabidopsis thaliana* plant (Lall et al., 2004). The Columbia ecotype, an inbred line, shows a strong capacity for shoot regeneration compared to the inbred Landsberg ecotype. Recombinant inbred lines (RILs) of these two ecotypes were analyzed in order to detect QTLs affecting shoot regeneration. Shoot regeneration was initially quantified as a count of shoots on a single root segment after going through tissue culture procedures. After transforming the shoot count observations, Lall et al. (2004) applied a QTL mapping method based on normality. In this example, we apply our QTL method based on the ZIP distribution to the untransformed shoot count data. Because the distribution of shoot counts conditional on the major QTL found by Lall et al. (2004) appears to look more like a ZIP distribution than either a transformed normal or a Poisson distribution (see Figure 3), we expect our method to be a more appropriate application.

4.1 Design

Root segments from each of 100 RILs and parental ecotypes went through a tissue culture procedure to promote shoot formation. Segments were placed in a callus induction medium for four days to promote cell growth, and then placed in a shoot induction medium for 15 days to promote shoot growth. The number of shoots forming on each root segment after the 15 days was documented. The process provided either five or six independent shoot count observations for each RIL.
In this QTL model, the variables used to predict the shoot count trait are the loci along the *Arabidopsis* genome, which is composed of five chromosomes and is approximately 600cM in length. A set of 62 markers was used providing us with an observed marker at approximately every 10cM across the genome. The two possible genotypes present at each marker are \( L \), for the Landsberg allele, or \( C \), for the Columbia allele. We used interval mapping to test for a QTL every 1cM.

### 4.2 Results

The plot of the likelihood ratio (LR) test statistic against genomic position is shown in Figure 4. Using a permutation method described in Churchill and Doerge (1994), we computed a significance threshold at the 0.01 significance level (shown as horizontal line in Figure 3). At this level, two loci were found to have significant association with shoot formation. A major QTL occurs on chromosome V, and a minor QTL occurs on chromosome IV. These locations were close in proximity to a major and a minor QTL found using the method based on normality. We note that a third QTL was detected on
chromosome I using the normality method that was not detected using the ZIP method. As is apparent from Figure 4, no locations on chromosome I appear to be near the significance threshold when using the ZIP method.

The location with the highest LR test statistic serves as a point estimate for the

![Plot of likelihood ratio versus genomic position. The horizontal line represents a significance threshold at the 0.01 level based on 500 permutations. A major QTL is located on chromosome V, and a minor QTL is located on chromosome IV, shown with vertical dotted lines.](image)

**Figure 4** Plot of likelihood ratio versus genomic position. The horizontal line represents a significance threshold at the 0.01 level based on 500 permutations. A major QTL is located on chromosome V, and a minor QTL is located on chromosome IV, shown with vertical dotted lines.

major QTL. Given this is the QTL location and this is the only QTL on the genome, we could estimate the ZIP parameters at the QTL using the MLEs from the EM algorithm. Because two locations tested as significant, a minor QTL on chromosome IV and a major QTL on chromosome V, we instead chose to extend the method described earlier for fitting a mixture of two ZIPS (one for each of two QTL groups) to fit a mixture of four ZIPS (one for each of four QTL groups). The extension changes the QTL indicator variable $Z_1$ from a random variable in $\{0, 1\}$, to a random variable in $\{1, 2, 3, 4\}$. In this extension, we used the markers flanking the estimated minor QTL and the markers flanking
the estimated major QTL and the EM algorithm to compute MLEs for the parameters in a mixture of four ZIP distributions based on the genotypes at both the major and minor QTL. Each distribution was associated with one of the four QTL groups in the set \( \{LL, LC, CL, CC\} \). The QTL group \( LC \) represents organisms with a Landsberg allele at the minor QTL, and a Columbia allele at the major QTL. The parameter estimates for the four ZIP distributions are shown in Table 6, and the density of shoot counts for each QTL group with the expected values given the estimated parameters is shown in Figure 5.

The most favorable QTL genotype configuration was a Columbia allele at the major QTL and a Landsberg allele at the minor QTL. This configuration had the lowest chance of an observation being generated from the point mass component, and it had the highest mean parameter for the Poisson component for all four ZIPs. The least favorable configuration was a Landsberg allele at the major QTL and a Columbia allele at the minor QTL. This configuration had the highest chance of an observation being generated from the point mass component, and it had the lowest mean parameter for the Poisson component.

Similar to the ZIP parameter interpretation in the context of manufacturing provided by Lambert (1992), we can interpret the \( p \) parameters in the context of biological shoot production. The point mass parameter for a given QTL genotype configuration represents the probability that a root segment is in a state of dormancy, or a state that

---

**Table 6**

Parameter estimates for a mixture of four ZIPs. Each ZIP is associated with one of the QTL groups in \( \{LL, LC, CL, CC\} \).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( LL )</th>
<th>( LC )</th>
<th>( CL )</th>
<th>( CC )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>0.578</td>
<td>0.018</td>
<td>0.754</td>
<td>0.168</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>2.918</td>
<td>4.774</td>
<td>2.587</td>
<td>2.641</td>
</tr>
</tbody>
</table>
does not allow for shoot formation. This state of being is most likely a result of both genetics and the shoot regeneration environment. Root segments that are not in this state have the capacity to develop shoots, but there is no guarantee that a shoot will form on all such segments. In this example, there is a much higher chance that a root segment with a QTL genotype configuration of $CL$ is in the dormant state, and much smaller chance that a root segment with a QTL genotype configuration of $LC$ is in the dormant state.
5 Appendix

5.1 Asymptotic Distribution of MLE

The maximum likelihood estimator for $\phi$ at any given testing location, denoted here as $\hat{\phi}_n$, is defined as

$$\hat{\phi}_n = \arg \max_{\phi \in \Theta} \prod_{i=1}^{n} f(y_i, x_i; \phi).$$

We use the following explicit form for $f(y_i, x_i; \phi)$ in the case of a mixture of two ZIPs in the backcross QTL scenario:

$$f(y, x; \phi) = f_{\gamma|x}(y|x; \phi) \cdot w(x)$$

where $f(y, x; \phi) = f_{\gamma|x}(y|x; \phi) \cdot w(x)$ is the PMF for $x_i$. $\pi(x)$ is a function of $x$ and known values for a given testing location, and $w(x)$ is the PMF for $x_i$.

In order to verify the asymptotic distribution of the maximum likelihood estimator above, we utilize the following theorem and lemma from Ferguson(1996)(see pages 121 and 124) modified to apply to a multivariate observation. Using notation from Ferguson, we first define $\Psi$ and $\hat{\Psi}$ as

$$\Psi(y, x; \phi) = \frac{\partial}{\partial \phi} \log f(y, x; \phi)$$

$$\hat{\Psi}(y, x; \phi) = \frac{\partial^2}{\partial \phi \partial \phi'} \log f(y, x; \phi).$$

**Theorem 1** (Cramér). Let $(Y_1, X_1), (Y_2, X_2), (Y_3, X_3), \ldots$ be i.i.d with density $f(y, x; \phi)$ (with respect to $dv$), and let $\phi_0$ denote the true value of the parameter. Suppose
1. $\Theta$ is an open subset of $\mathbb{R}^k$;

2. second partial derivatives of $f(y, x; \phi)$ with respect to $\phi$ exist and are continuous for all $(y, x)$, and may be passed under the integral sign in $\int f(y, x; \phi) dv(y, x)$;

3. there exists a function $K(y, x)$ such that $E_{\phi_0} K(Y, X) < \infty$ and each component of $\dot{\Psi}(y, x, \phi)$ is bounded in absolute value by $K(y, x)$ uniformly in some neighborhood of $\phi_0$;

4. $I(\phi_0) = -E_{\phi_0} \dot{\Psi}(Y, X, \phi_0)$ is positive definite; and

5. $f(y, x; \phi) = f(y, x; \phi_0)$ a.e. $dv \Rightarrow \phi = \phi_0$.

Then there exists a strongly consistent sequence $\hat{\phi}_n$ of roots of the likelihood equation such that

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N(0, I_1(\phi_0)^{-1}).$$

**Lemma 1.** If $\frac{\partial}{\partial \phi} g(y, x, \phi)$ exists and is continuous in $\phi$ for all $(y, x)$ and all $\phi$ in an open interval $S$, and if $|\frac{\partial}{\partial \phi} g(y, x, \phi)| \leq K(y, x)$ on $S$ where $\int K(y, x) dv(y, x) < \infty$, and if $\int g(y, x, \phi) dv(y, x)$ exists on $S$, then

$$\frac{d}{d\phi} \int g(y, x, \phi) dv(y, x) = \int \frac{\partial}{\partial \phi} g(y, x, \phi) dv(y, x).$$

Condition 1 is met as $\Theta = (0, 1) \times \mathbb{R}^+ \times (0, 1) \times \mathbb{R}^+$. Condition 2 requires both the existence and continuity of second partial derivatives, and the ability to exchange integration and differentiation for the second partial derivatives. It is easy to verify that the first and second partial derivatives with respect to each $\phi_j, j \in \{1, 2, 3, 4\}$, exist and are continuous for all $\phi$ in the interior of $\Theta$. To verify the exchangeability of integration and differentiation, we utilize Lemma 1 to first verify this exchangeability for the first
partial derivatives, and then verify the exchangeability for the second partial derivatives.

We first show that the absolute value of each first partial derivative with respect to \( \phi_j \) is bounded by a function \( K(y, x) \), with \( \int K(y, x)dv(y, x) \) being finite. For the parameter \( \phi_1 = p_q \), we have

\[
\left| \frac{\partial}{\partial p_q} f(y, x; \phi) \right| = \left| I_{(0)}(y) \left[ (1 - \pi(x))(1 - e^{-\lambda_q}) \right] 
- (1 - I_{(0)}(y)) \left[ (1 - \pi(x)) \left( \frac{e^{-\lambda_q \lambda_q^y}}{y!} \right) \right] \cdot w(x) 
\right|
\]

\[
< I_{(0)}(y) + (1 - I_{(0)}(y)) \frac{e^{-\lambda_q \lambda_q^y}}{y!}
< I_{(0)}(y) + \frac{e^{-\lambda_q \lambda_q^y}}{y!}.
\]

Now, for any \( \phi_0 \in \Theta \), define \( S = (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{q_0} - \epsilon, \lambda_{q_0} + \epsilon) \times (p_{Q_0} - \epsilon, p_{Q_0} + \epsilon) \times (\lambda_{Q_0} - \epsilon, \lambda_{Q_0} + \epsilon) \) where \( \epsilon \) is a positive constant small enough that \( S \in \Theta \). We define the bounding function \( K(y, x) \) for any \( \phi \) in \( S \) as

\[
K(y, x) = \left( 1 - \pi(x) \right)(1 - e^{-\lambda_q})
+ \frac{e^{-\lambda_{q_0} - \epsilon} \lambda_{q_0}^y}{y!}
+ \frac{e^{-(\lambda_{q_0} + \epsilon)} \lambda_{q_0}^y}{y!}
= I_{(0)}(y) + \frac{e^{-\lambda_{q_0} - \epsilon} \lambda_{q_0}^y}{y!} \cdot \frac{e^{-(\lambda_{q_0} + \epsilon)}}{y!}
\]

Thus, for any \( \phi \in S \), \( \left| \frac{\partial}{\partial \lambda_q} f(y, x; \phi) \right| \leq K(y, x) \), and \( \int K(y, x)dv(y, x) = \sum_{y=0}^{\infty} K(y, x) = 1 + e^{2\epsilon} < \infty. \)

For the parameter \( \phi_2 = \lambda_q \), we have

\[
\left| \frac{\partial}{\partial \lambda_q} f(y, x; \phi) \right| = \left| I_{(0)}(y) \left[ (1 - \pi(x))(1 - p_q)(-e^{-\lambda_q}) \right] 
+ (1 - I_{(0)}(y)) \left[ (1 - \pi(x))(1 - p_q) \frac{1}{\lambda_q} \left( \frac{e^{-\lambda_q \lambda_q^y}}{y!} \right)(y - \lambda_q) \right] \cdot w(x) 
\right|
\]

\[
< I_{(y:y<\lambda_q)}(y) + \frac{1}{\lambda_q} \frac{e^{-\lambda_q \lambda_q^y}}{y!}.
\]
Now, for any $\phi_0 \in \Theta$, define $S = (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{q_0} - \epsilon, \lambda_{q_0} + \epsilon) \times (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{Q_0} - \epsilon, \lambda_{Q_0} + \epsilon)$ where $\epsilon$ is a positive constant small enough that $S \in \Theta$. We define the bounding function $K(y, x)$ for any $\phi$ in $S$ as

$$K(y, x) = I_{\{y : y \leq (\lambda_{q_0} + \epsilon)\}}(y) + \frac{1}{(\lambda_{q_0} - \epsilon)} y \left( \frac{e^{-(\lambda_{q_0} - \epsilon)(\lambda_{q_0} + \epsilon)y}}{y!} \right).$$

Thus, for any $\phi \in S$, $\left| \frac{\partial}{\partial \lambda_q} f(y, x; \phi) \right| \leq K(y, x)$, and

$$\int K(y, x) dv(y, x) = \sum_{y=0}^{\infty} K(y, x) = \#\{y : y \leq (\lambda_{q_0} + \epsilon)\} + \frac{e^{2(\lambda_{q_0} + \epsilon)}}{\lambda_{q_0} - \epsilon} < \infty.$$

Similar $K(y, x)$ exist for $\phi_3 = p_Q$ and $\phi_4 = \lambda_Q$. As $f(y, x; \phi)$ is a density, we have $\int f(y, x; \phi) dv(y, x) = 1 < \infty$ and Lemma 1 provides, for each $j \in \{1, 2, 3, 4\}$, that

$$\frac{d}{d\phi_j} \int f(y, x; \phi) dv(y, x) = \int \frac{\partial}{\partial \phi_j} f(y, x; \phi) dv(y, x).$$

To verify exchangeability for the second partial derivatives, we show similar results as above. For the second partial $\frac{\partial^2}{\partial \lambda_q \partial p_q} = \frac{\partial^2}{\partial p_q \partial \lambda_q}$, we have

$$\left| \frac{\partial^2}{\partial \lambda_q \partial p_q} f(y, x; \phi) \right| = \left| I_{\{0\}}(y)(1 - \pi(x)) e^{-\lambda_q} \right.$$

$$- (1 - I_{\{0\}}(y))(1 - \pi(x)) \frac{1}{\lambda_q} \left( \frac{e^{-\lambda_q \lambda_q y}}{y!} \right) (y - \lambda_q) \cdot w(x)$$

$$< I_{\{y, y \leq \lambda_q\}}(y) + \frac{1}{\lambda_q} y \left( \frac{e^{-\lambda_q \lambda_q y}}{y!} \right).$$

Now, for any $\phi_0 \in \Theta$, define $S = (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{q_0} - \epsilon, \lambda_{q_0} + \epsilon) \times (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{Q_0} - \epsilon, \lambda_{Q_0} + \epsilon)$ where $\epsilon$ is a positive constant small enough that $S \in \Theta$. We define the bounding function $K(y, x)$ for any $\phi$ in $S$ as

$$K(y, x) = I_{\{y : y \leq (\lambda_{q_0} + \epsilon)\}}(y) + \frac{e^{2\epsilon}}{(\lambda_{q_0} - \epsilon) y} \left( \frac{e^{-(\lambda_{q_0} + \epsilon)(\lambda_{q_0} + \epsilon)y}}{y!} \right).$$
Thus, for any \( \phi \in S \), \( \left| \frac{\partial^2}{\partial \lambda_q \partial \phi} f(y, \phi) \right| \leq K(y, x) \), and \( \int K(y, x) dv(y, x) = \sum_{y=0}^{\infty} K(y, x) = \#\{y : y \leq (\lambda_{q_0} + \epsilon)\} + \frac{e^{2\epsilon(\lambda_{q_0} + \epsilon)}}{\lambda_{q_0} - \epsilon} < \infty. \)

For the second partial \( \frac{\partial^2}{\partial \lambda_q^2} \), we have

\[
\left| \frac{\partial^2}{\partial \lambda_q^2} f(y, x; \phi) \right| = \left| I_{(0)}(y) \left[ (1 - \pi(x))(1 - p_q)(e^{-\lambda_q}) \right] \\
+ (1 - I_{(0)}(y)) \left[ (1 - \pi(x))(1 - p_q) \left( \frac{e^{-\lambda_q} \lambda_q^y}{y!} \right) \right] \\
+ (1 - I_{(0)}(y)) \left[ (1 - \pi(x))(1 - p_q) \frac{1}{\lambda_q} (y^2 - y(1 + 2\lambda_q)) \left( \frac{e^{-\lambda_q} \lambda_q^y}{y!} \right) \right] \cdot w(x) \\
< I_{(0)}(y) + \frac{e^{-\lambda_q} \lambda_q^y}{y!} + \frac{1}{\lambda_q} y^2 \left( \frac{e^{-\lambda_q} \lambda_q^y}{y!} \right). \]

Now, for any \( \phi_0 \in \Theta \), define \( S = (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{q_0} - \epsilon, \lambda_{q_0} + \epsilon) \times (p_{Q_0} - \epsilon, p_{Q_0} + \epsilon) \times (\lambda_{Q_0} - \epsilon, \lambda_{Q_0} + \epsilon) \) where \( \epsilon \) is a positive constant small enough that \( S \in \Theta \). We define the bounding function \( K(y, x) \) for any \( \phi \in S \) as

\[
K(y, x) = I_{(0)}(y) + \frac{e^{-\lambda_{q_0} - \epsilon} (\lambda_{q_0} + \epsilon)^y}{y!} + \frac{1}{(\lambda_{q_0} - \epsilon)^2 y^2} \left( \frac{e^{-\lambda_{q_0} - \epsilon} (\lambda_{q_0} + \epsilon)^y}{y!} \right) \\
= I_{(0)}(y) + e^{2\epsilon} \left( \frac{e^{-\lambda_{q_0} - \epsilon} (\lambda_{q_0} + \epsilon)^y}{y!} \right) + \frac{e^{2\epsilon}}{(\lambda_{q_0} - \epsilon)^2 y^2} \left( \frac{e^{-\lambda_{q_0} - \epsilon} (\lambda_{q_0} + \epsilon)^y}{y!} \right).
\]

Thus, for any \( \phi \in S \), \( \left| \frac{\partial^2}{\partial \lambda_q \partial \phi} f(y, x; \phi) \right| \leq K(y, x) \), and \( \int K(y, x) dv(y, x) = \sum_{y=0}^{\infty} K(y, x) = 1 + e^{2\epsilon} + \frac{e^{2\epsilon}}{(\lambda_{q_0} - \epsilon)^2} [(\lambda_{q_0} + \epsilon) + (\lambda_{q_0} + \epsilon)^2] < \infty. \)

Similar \( K(y, x) \) exist for the second partial derivatives involving the parameters \( p_Q \) and \( \lambda_Q \). As \( f(y, x; \phi) \) is a density, and we’ve already verified the exchangeability of integration and differentiation for first partial derivatives, we have \( \int \frac{\partial}{\partial \phi_j} f(y, x; \phi) dv(y, x) = 0 < \infty \) for all \( \phi_j \) and Lemma 1 provides

\[
\frac{d^2}{d\phi_i d\phi_j} \int f(y, x; \phi) dv(y, x) = \int \frac{\partial^2}{\partial \phi_i \partial \phi_j} f(y, x; \phi) dv(y, x).
\]
For condition 3, let \( S = (p_{q_0} - \epsilon, p_{q_0} + \epsilon) \times (\lambda_{q_0} - \epsilon, \lambda_{q_0} + \epsilon) \times (p_{Q_0} - \epsilon, p_{Q_0} + \epsilon) \times (\lambda_{Q_0} - \epsilon, \lambda_{Q_0} + \epsilon) \) for any \( \phi_0 \in \Theta \) where \( \epsilon \) is a positive constant small enough that \( S \subseteq \Theta \).

It is straightforward to show that

\[
K(0, x) = \frac{3}{\left[1 - c_0(p_{q_0} - \epsilon + \epsilon(\lambda_{q_0} - \epsilon)(1 - (p_{q_0} + \epsilon))) + c_0(p_{Q_0} - \epsilon + \epsilon(\lambda_{Q_0} - \epsilon)(1 - (p_{Q_0} + \epsilon)))\right]^2}
\]

bounds all ten derivatives for all \( \phi \in S \) when \( y = 0 \).

For \( y > 0 \), we define \( K(y, x) \) as \( \max_{i=1, \ldots, 10} K_i(y, x) \) where \( K_i(y, x) \) are defined as follows:

\[
\begin{align*}
\left| \frac{\partial^2}{\partial p_q \partial p_q} \log f(y, x; \phi) \right| &< K_1(y, x) = \left( \frac{1}{1 - (p_{q_0} + \epsilon)} \right)^2 \\
\left| \frac{\partial^2}{\partial \lambda_q \partial p_q} \log f(y, x; \phi) \right| &< K_2(y, x) = \left( \frac{1}{\lambda_{q_0} - \epsilon} \left( \frac{1}{1 - (p_{q_0} + \epsilon)} \right) \right) y \\
\left| \frac{\partial^2}{\partial p_Q \partial p_Q} \log f(y, x; \phi) \right| &< K_3(y, x) = \left( \frac{1}{1 - (p_{Q_0} + \epsilon)} \right)^2 \left( \frac{1}{1 - (p_{Q_0} + \epsilon)} \right) \\
\left| \frac{\partial^2}{\partial \lambda_Q \partial \lambda_Q} \log f(y, x; \phi) \right| &< K_4(y, x) = \left( \frac{1}{\lambda_{Q_0} - \epsilon} \left( \frac{1}{1 - (p_{Q_0} + \epsilon)} \right) \right) y \\
\left| \frac{\partial^2}{\partial p_Q \partial \lambda_Q} \log f(y, x; \phi) \right| &< K_5(y, x) = 2 \left( \frac{1}{\lambda_{q_0} - \epsilon} \right)^2 \max \{ y^2, (\lambda_{q_0} + \epsilon)^2, 1 - \frac{1}{4} \} \\
\left| \frac{\partial^2}{\partial \lambda_Q \partial \lambda_Q} \log f(y, x; \phi) \right| &< K_6(y, x) = \left( \frac{1}{\lambda_{q_0} - \epsilon} \left( \frac{1}{\lambda_{q_0} - \epsilon} \right) \right) y \\
\left| \frac{\partial^2}{\partial \lambda_Q \partial \lambda_Q} \log f(y, x; \phi) \right| &< K_7(y, x) = \frac{1}{(\lambda_{q_0} - \epsilon)(\lambda_{q_0} - \epsilon)} (y^2 + \max \{ \lambda_{q_0} + \epsilon, \lambda_{Q_0} + \epsilon \}^2) \\
\left| \frac{\partial^2}{\partial p_Q \partial p_Q} \log f(y, x; \phi) \right| &< K_8(y, x) = \left( \frac{1}{1 - (p_{Q_0} + \epsilon)} \right)^2 \\
\left| \frac{\partial^2}{\partial \lambda_Q \partial p_Q} \log f(y, x; \phi) \right| &< K_9(y, x) = \left( \frac{1}{\lambda_{Q_0} - \epsilon} \left( \frac{1}{1 - (p_{Q_0} + \epsilon)} \right) \right) y \\
\left| \frac{\partial^2}{\partial \lambda_Q \partial \lambda_Q} \log f(y, x; \phi) \right| &< K_{10}(y, x) = 2 \left( \frac{1}{(\lambda_{Q_0} - \epsilon)} \right)^2 \max \{ y^2, (\lambda_{Q_0} + \epsilon)^2, 1 - \frac{1}{4} \},
\end{align*}
\]

where the inequalities hold for all \( \phi \in S \).

It follows that the absolute value of each component of \( \dot{\Psi} \) is bounded above by \( K(y, x) \) uniformly on \( S \). Furthermore, \( E_{\phi_0} \dot{K}_i(y, x) < \infty \) for all \( i \), and it follows that \( E_{\phi_0} K(y, x) < \infty \).
For condition 4, we note that we have already verified the exchangeability of differ-
entiation and integration. Therefore, we have

\[ I(\phi_0) = -E_{\phi_0} [\hat{\Psi}(Y, X, \phi) = E_{\phi_0} \left[ \Psi(Y, X, \phi)^T \right]] . \]

For any vector \( a \in \mathbb{R}^d \) we have

\[
a^T I(\phi_0) a = a^T \left[ E_{\phi_0} [\Psi(Y, X, \phi)] \right] a \\
= E_{\phi_0} [a^T \Psi(Y, X, \phi)] [a^T \Psi(Y, X, \phi)]^T \\
= E_{\phi_0} (a^T \Psi(Y, X, \phi))^2 \\
= E_{\phi_0} \left[ E_{\phi_0} \left\{ (a^T \Psi(Y, X, \phi))^2 \right\} \right] , \tag{17} \]

From the above, \( I(\phi_0) \) is at least non-negative definite.

Suppose \( \exists a \) such that \( a^T I(\phi_0) a = 0 \). Then by Equation (17), we have

\[
E_{\phi_0} \left[ (a^T \Psi(Y, X, \phi))^2 \right| X = x = 0 \quad \forall x \text{ such that } w(x) > 0 ,
\]

which implies \( a^T \Psi(y, x, \phi) = 0 \quad \forall y \in \{0, 1, 2, 3, \ldots\} \) and any \( x \) such that \( w(x) > 0 \).

Now, note that \( \forall x \) such that \( w(x) > 0 \),

\[
a^T \Psi(1, x, \phi) = a_1 \left[ -\left(1 - \pi(x)\right)e^{-\lambda_q} \lambda_q \right] + a_2 \left[ \left(1 - \pi(x)\right)(1 - p_q)e^{-\lambda_q}(1 - \lambda_q) \right] \\
+ a_3 \left[ -\pi(x)e^{-\lambda_q} \lambda_q \right] + a_4 \left[ \pi(x)(1 - p_q)e^{-\lambda_q}(1 - \lambda_q) \right] = 0
\]

and

\[
a^T \Psi(2, x, \phi) = a_1 \left[ -\left(1 - \pi(x)\right)e^{-\lambda_q} \lambda_q^2 \right] + a_2 \left[ \left(1 - \pi(x)\right)(1 - p_q)e^{-\lambda_q} \lambda_q(2 - \lambda_q) \right] \\
+ a_3 \left[ -\pi(x)e^{-\lambda_q} \lambda_q^2 \right] + a_4 \left[ \pi(x)(1 - p_q)e^{-\lambda_q} \lambda_q(2 - \lambda_q) \right] = 0.
\]

The above equations can be written in matrix form as

\[
[(1 - \pi(x))A_q \pi(x)A_q] a = 0
\]

where \( A_q \) is the nonsingular \( 2 \times 2 \) matrix defined as

\[
A_q = \begin{bmatrix}
-\lambda_q e^{-\lambda_q} & (1 - p_q)(1 - \lambda_q)e^{-\lambda_q} \\
-\lambda_q^2 e^{-\lambda_q} & (1 - p_q)(2 - \lambda_q)e^{-\lambda_q}
\end{bmatrix}
\]
and $A_Q$ is defined analogously. Now let $x_1 = (0,0)'$ and $x_2 = (1,1)'$. Then, $w(x_1) > 0$, $w(x_2) > 0$, and $\pi(x_1) \neq \pi(x_2)$, and we have
\[
[(1 - \pi(x_1))A_q \quad \pi(x_1)A_Q]a = 0
\]
and
\[
[(1 - \pi(x_2))A_q \quad \pi(x_2)A_Q]a = 0.
\]
Thus, for
\[
A = \begin{bmatrix}
(1 - \pi(x_1))A_q & \pi(x_1)A_Q \\
(1 - \pi(x_2))A_q & \pi(x_2)A_Q
\end{bmatrix}
\]
we have
\[
Aa = 0
\]
which can be written as
\[
Aa = \begin{bmatrix}
(1 - \pi(x_1))I & \pi(x_1)I \\
(1 - \pi(x_2))I & \pi(x_2)I
\end{bmatrix} \begin{bmatrix}
A_q & 0 \\
0 & A_Q
\end{bmatrix}a = 0.
\]
As $A_q, A_Q$ are nonsingular, the right matrix in the product above is rank 4. The left matrix is also nonsingular as its inverse can be stated as
\[
\frac{1}{\pi(x_2) - \pi(x_1)} \begin{bmatrix}
\pi(x_2)I & -\pi(x_1)I \\
-(1 - \pi(x_2))I & (1 - \pi(x_1))I
\end{bmatrix}
\]
Thus, $A$ is full rank and $Aa = 0 \Rightarrow a = 0$. Hence, $a^T I(\phi_0)a = 0 \Rightarrow a = 0$, and we have established that $I(\phi_0)$ is positive definite.

To show condition 5, suppose $f(y, x; \phi) = f(y, x; \phi_0)$ a.e. $dv$. Then,
\[
f_{Y|X}(y|x; \phi)w(x) = f_{Y|X}(y|x; \phi_0)w(x) \quad a.e. \ dv
\]
\[
\Rightarrow f_{Y|X}(y|x; \phi) = f_{Y|X}(y|x; \phi_0) \quad \forall y \text{ and } \forall x \text{ s.t. } w(x) > 0.
\]
In the present situation, $X$ represents random marker genotypes, and $w(x)$ represents the PMF for $X$. The conditional density $f_{Y|X}(y|x; \phi)$ is a mixture of two ZIP distributions for which the mixing proportion depends on the marker genotypes $x$. The
support set for \( w(x) \) is the set of all possible marker genotypes in a backcross population, which is the set \{ (0, 0), (0, 1), (1, 0), (1, 1) \}. When markers are separated by a distance of \( \theta \), \( w(x) \) places a positive mass of \( 0.5(1 - \theta) \) on \((0, 0)\) and \((1, 1)\), and a positive mass of \( 0.5\theta \) on \((0, 1)\) and \((1, 0)\). The support set for \( f_{Y|X}(y|x; \phi) \) is the set \{0, 1, 2, 3, \ldots\}. Thus, the conditional densities above that are equal a.e. both represent mixtures on \( y \in \{0, 1, 2, 3, \ldots\} \) with mixing proportions depending on \( x \).

**Lemma 1.** Consider \( \pi_1, \pi_2 \in [0, 1] \) where \( \pi_1 \neq \pi_2 \). Now, suppose \( h, g, h_0, \) and \( g_0 \) are any four distributions on \{0, 1, 2, 3, \ldots\} with
\[
\pi_1 h + (1 - \pi_1) g = \pi_1 h_0 + (1 - \pi_1) g_0 \text{ a.e.}
\]
and
\[
\pi_2 h + (1 - \pi_2) g = \pi_2 h_0 + (1 - \pi_2) g_0 \text{ a.e.}
\]
Then,
\[
g = g_0 \text{ and } h = h_0.
\]

**Pf:** Subtracting the equations in (18) yields
\[
(\pi_1 - \pi_2) h + (\pi_2 - \pi_1) g = (\pi_1 - \pi_2) h_0 + (\pi_2 - \pi_1) g_0 \text{ a.e.}
\]
\[
\Rightarrow (\pi_1 - \pi_2)(h - h_0) = (\pi_1 - \pi_2)(g - g_0) \text{ a.e.}
\]
\[
\Rightarrow h - h_0 = g - g_0 \text{ a.e.} \quad (19)
\]
Thus, the difference in the first component of each mixture must equal the difference in the second component of each mixture. From the first equation in (18),
\[
\pi_1 h + (1 - \pi_1) g = \pi_1 h_0 + (1 - \pi_1) g_0
\]
\[
\Rightarrow \pi_1 (h - h_0) + (1 - \pi_1)(g - g_0) = 0.
\]
This together with (19) yields
\[
\pi_1(g - g_0) + (1 - \pi_1)(g - g_0) = 0
\]
\[
\Rightarrow g = g_0 \text{ and } h = h_0
\]
In the context of the mixture distributions \( f_{Y|X}(y|x; \phi) \) and \( f_{Y|X}(y|x; \phi_0) \), the mixing proportions \( \pi_1 \) and \( \pi_2 \) represent the functions \( \pi(x_j) \) and \( \pi(x_k) \) for any \( x_i, x_j \in \{(0,0), (0,1), (1,0), (1,1)\} \) on which \( w(x) \) places positive mass. As \( w(x) \) always places positive mass on at least two of the four marker configurations, Lemma 1 provides equality of the first component of \( f_{Y|X}(y|x; \phi) \) and the first component of \( f_{Y|X}(y|x; \phi_0) \), as well as equality of the second component of \( f_{Y|X}(y|x; \phi) \) and the second component of \( f_{Y|X}(y|x; \phi_0) \).

The first component of \( f_{Y|X}(y|x; \phi) \) is a ZIP distribution defined by the parameters \( p_q \) and \( \lambda_q \), denoted as ZIP\((p_q, \lambda_q)\). The second component of \( f_{Y|X}(y|x; \phi) \) is also a ZIP distribution defined as ZIP\((p_{q_0}, \lambda_{q_0})\). Thus, it remains to show that if ZIP\((p_q, \lambda_q)\)=ZIP\((p_{q_0}, \lambda_{q_0})\), then \((p_q, \lambda_q) = (p_{q_0}, \lambda_{q_0})\).

**Lemma 2.** If the PMF for the two ZIP distributions ZIP\((p_q, \lambda_q)\) and ZIP\((p_{q_0}, \lambda_{q_0})\) place equal mass at every \( y \in \{0,1,2,3,\ldots\} \), then \((p_q, \lambda_q) = (p_{q_0}, \lambda_{q_0})\).

**Pf:** To verify this relationship, we use the probability of observing \( Y = 1 \) and \( Y = 2 \).

For \( Y = 1 \), we have
\[
P(Y = 1|\phi) = P(Y = 1|\phi_0)\\
\implies (1 - p_q)e^{-\lambda_q} = (1 - p_{q_0})e^{-\lambda_{q_0}}\\
\implies \frac{(1 - p_q)}{(1 - p_{q_0})} = \frac{\lambda_{q_0}}{\lambda_q}e^{\lambda_q - \lambda_{q_0}}
\]

For \( Y = 2 \), we have
\[
P(Y = 2|\phi) = P(Y = 2|\phi_0)\\
\implies (1 - p_q)\frac{e^{-\lambda_q} \lambda_q^2}{2} = (1 - p_{q_0})\frac{e^{-\lambda_{q_0}} \lambda_{q_0}^2}{2}\\
\implies \frac{(1 - p_q)}{(1 - p_{q_0})} = \left(\frac{\lambda_{q_0}}{\lambda_q}\right)^2 e^{\lambda_q - \lambda_{q_0}}
\]

Solving the system of equations for \( \phi, \phi_0 \in \Theta \) provides the equalities \( \lambda_q = \lambda_{q_0} \) and \( p_q = p_{q_0} \). \(\square\)
Applying Lemma 2 to the first components of the mixtures provides \((p_q, \lambda_q) = (p_{q0}, \lambda_{q0})\).

Similarly, applying Lemma 2 to the second components provides \((p_Q, \lambda_Q) = (p_{Q0}, \lambda_{Q0})\)

Thus, if \(f(y, x; \phi) = f(y, x; \phi_0)\) a.e. \(dv\), then \(\phi = \phi_0\).

### 5.2 Continuity of \(Q(\phi|\phi^*)\) in \(\phi\) and \(\phi^*

Note that

\[
P(Y_i = 0|\phi^*, x_i) \equiv e^{-\lambda_i^*}(1 - p^*_q)(1 - \pi(x_i; \theta, r_L, r_R)) + e^{-\lambda_Q^*}(1 - p^*_Q)\pi(x_i; \theta, r_L, r_R)
\]

\[+ p^*_q(1 - \pi(x_i; \theta, r_L, r_R)) + p^*_Q\pi(x_i; \theta, r_L, r_R)
\]

and

\[
P(Y_i = y_i > 0|\phi^*, x_i) \equiv \frac{e^{-\lambda_i^*} \lambda_i^{y_i}}{y_i!} (1 - p^*_q)(1 - \pi(x_i; \theta, r_L, r_R))
\]

\[+ \frac{e^{-\lambda_Q^*} \lambda_Q^{y_i}}{y_i!} (1 - p^*_Q)\pi(x_i; \theta, r_L, r_R)
\]

Then, we can write \(Q(\phi|\phi^*)\) as:

\[
Q(\phi|\phi^*) = \sum_{i=1}^{n} \left[ \log \left( \frac{e^{-\lambda_i^*} \lambda_i^{y_i}}{y_i!} (1 - p^*_q)(1 - \pi(x_i; \theta, r_L, r_R)) \right) \right.
\]

\[
\left. \left( \frac{e^{-\lambda_i^*} (1 - p^*_q)(1 - \pi(x_i; \theta, r_L, r_R))}{P(Y_i = 0|\phi^*, x_i)} \right)^{I_{\{0\}}(y_i)} \cdot \left( \frac{e^{-\lambda_Q^*} (1 - p^*_Q)\pi(x_i; \theta, r_L, r_R)}{P(Y_i = y_i > 0|\phi^*, x_i)} \right)^{1 - I_{\{0\}}(y_i)} \right.
\]

\[+ \log (1 - \pi(x_i; \theta, r_L, r_R)) \cdot \left( \frac{p^*_q(1 - \pi(x_i; \theta, r_L, r_R))}{P(Y_i = 0|\phi^*, x_i)} \right) \cdot I_{\{0\}}(y_i)
\]

\[+ \log \left( \frac{e^{-\lambda_Q^*} \lambda_Q^{y_i}}{y_i!} (1 - p^*_Q)\pi(x_i; \theta, r_L, r_R) \right).
\]

\[
\left( \frac{e^{-\lambda_i^*} (1 - p^*_Q)\pi(x_i; \theta, r_L, r_R)}{P(Y_i = 0|\phi^*, x_i)} \right)^{I_{\{0\}}(y_i)} \cdot \left( \frac{e^{-\lambda_Q^*} \lambda_Q^{y_i} (1 - p^*_Q)\pi(x_i; \theta, r_L, r_R)}{P(Y_i = y_i > 0|\phi^*, x_i)} \right)^{1 - I_{\{0\}}(y_i)}
\]

\[+ \log (p_Q \pi(x_i; \theta, r_L, r_R)) \cdot \left\{ \left( \frac{p^*_Q\pi(x_i; \theta, r_L, r_R)}{P(Y_i = 0|\phi^*, x_i)} \right) \cdot I_{\{0\}}(y_i) \right\}
\]
For continuity of $\phi$, we note that each of the four summand components for each observation $i$ of $Q(\phi|\phi^*)$ are all differentiable in $\phi$. Thus, $Q(\phi|\phi^*)$ is continuous in $\phi$. Continuity of $\phi^*$ depends on the functions $P(Y_i = 0|\phi^*, x_i)$ and $P(Y_i = y_i > 0|\phi^*, x_i)$ which contain $\phi^*$ and are in the denominators of the four summand components. Because a zero can be generated from either QTL distribution, and from either component of each ZIP distribution, there are four summands in the conditional probability of generating a zero $P(Y_i = 0|\phi^*, x_i)$. The first two summands, associated with the probability of generating a zero from a Poisson component, can not simultaneously be zero for $\phi^* \in \Theta$. This is because $\pi(x_i; \theta, r_L, r_R)$ and $(1 - \pi(x_i; \theta, r_L, r_R))$, the probabilities that individual $i$ has a QTL genotype of $Q$ or $q$, respectively, can not simultaneously be zero. A similar argument holds for the third and fourth summands which also can not simultaneously be zero. Thus, division by zero will not occur in the interior of $\Theta$, and $Q(\phi|\phi^*)$ is continuous in both $\phi$ and $\phi^*$.

6 References


CHAPTER 2. IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN UNREPLICATED MULTIPLE-TREATMENT MICROARRAY TIMECOURSE EXPERIMENTS

A paper published in *Computational Statistics and Data Analysis*

Rhonda DeCook, Dan Nettleton, Carol Foster, and Eve S. Wurtele

Abstract

Microarray technology has become widespread as a means to investigate gene function and metabolic pathways in an organism. A common experiment involves probing, at each of several time points, the gene expression of experimental units subjected to different treatments. Due to the high cost of microarrays, such experiments may be performed without replication and therefore provide a gene expression measurement of only one experimental unit for each combination of treatment and time point. Though an experiment with replication would provide more powerful conclusions, it is still possible to identify differentially expressed genes and to estimate the number of false positives for a specified rejection region when the data are unreplicated. We present a method for identifying differentially expressed genes in this situation that utilizes polynomial regression models to approximate underlying expression patterns. In the first stage of a
two-stage permutation approach, we choose a 'best' model at each gene after considering all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms. In the second stage, we identify genes whose 'best' model differs significantly from the overall mean model as differentially expressed. The number of expected false positives in the chosen rejection region and the overall proportion of differentially expressed genes are both estimated using a method presented by Storey and Tibshirani (2003). For illustration, the proposed method is applied to an Arabidopsis thaliana microarray data set.

1 Introduction

Present microarray technology allows a researcher to simultaneously measure the mRNA expression level of thousands of genes from a single experimental unit. This ability serves as a powerful genomics tool for uncovering gene function. By applying this technology to a multiple-treatment timecourse experiment, a researcher can locate genes whose expression patterns over time are biologically interesting as defined in a number of ways. For example, genes of biological interest may include genes whose expression changes significantly over time in an identical fashion for all treatments (time main effects), genes whose expression differs significantly among treatments but remains constant over time (treatment main effects), or genes whose pattern of expression over time differs significantly among treatments (time-by-treatment interaction).

Though the cost of microarrays has greatly decreased since their emergence, a multiple-treatment microarray timecourse experiment can still be quite costly. Therefore, a researcher interested in such an experiment may choose to perform it with no replications in hopes of finding interesting expression patterns that will be studied more intensively in follow-up experiments. In this situation, a statistical analysis must rely on only one
expression measurement per treatment-time combination for each gene, making separation of signal and noise a difficult task.

Even when replication is present, a microarray data set that holds information on both treatment and time effects poses challenges to the statistical researcher that go beyond the one-way ANOVA methods or the cluster analysis methods that are often applied to microarray data. Issues such as normalization, signal-to-noise ratios, and multiple testing are still present, but the fact that numerous alternative hypotheses may be of interest at each of thousands of genes adds a layer of complexity to the analysis.

In a microarray experiment that compares only two groups, the same test statistic, namely a standard \( t \)-test or one of its modifications, can be used to test for differential expression at every gene. Storey and Tibshirani (2001) provide a method in this scenario for identifying differentially expressed genes that estimates the positive False Discovery Rate (pFDR) for the chosen rejection region. To apply this method in a multiple-treatment timecourse experiment, a researcher could use the \( F \)-statistic comparing the most complex polynomial regression model with positive degrees of freedom for error to the model that assumes a common mean for all conditions to test for the existence of any effects at each gene. Employing this overall \( F \)-test for all genes, however, would waste degrees of freedom for genes whose true expression patterns can be adequately described with a relatively simple alternative polynomial regression model. An \( F \)-statistic comparing a simpler polynomial regression model to the overall mean model would have more power for detecting genes with simpler expression patterns.

We propose a method for locating genes whose expression patterns over time differ in any way from the overall mean model that provides more power for simpler alternatives than the overall \( F \)-test described above. A concurrent flat line profile for all treatment groups represents the expected expression pattern at any gene with a true null hypothesis. This null reflects the reasoning that any gene represented by a pattern other than the overall mean model is potentially biologically interesting. To test this
null hypothesis, we first choose a 'best' model at each gene from the pool of candidate models including all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms. The model choice criterion is based on the vector of $F$-statistics comparing each candidate model to the overall mean model. The model with the most extreme $F$-statistic when compared to the distribution of similar $F$-statistics generated from permuted data sets is considered the 'best' model. In the second step, we identify genes whose 'best' model differs significantly from the overall mean model by using the same permuted data sets as those used in the model-selection step to generate a relevant $p$-value for each gene. A $p$-value threshold is chosen that coincides with both a list of significant genes and an estimated number of expected false positives among the genes declared to be significant.

There are several benefits of our proposed procedure. First, it allows us to associate a gene with a 'best' model requiring fewer degrees of freedom than the most complex model possible. The $F$-statistic comparing this 'best' model with the overall mean model will potentially have more degrees of freedom for error and more power for detecting differential expression than the overall $F$-test. Furthermore, differentially expressed genes will be automatically sorted into groups of genes exhibiting similar expression patterns. Lastly, considering time as a quantitative factor allows us to detect genes exhibiting treatment-by-time interaction in unreplicated data because the pool of candidate models includes those that fit a separate polynomial to each treatment group while still allowing for degrees of freedom for error.

In the next section, we discuss the form of the data, the model selection process, the test statistic, and the multiple-comparison adjustment. Section 3 provides a simulation study comparing sensitivity and specificity for our method and the overall $F$-test method when six particular alternative expression patterns are present in the data. In Section 4, we apply our method to data generated from a multiple-treatment timecourse experiment that exposed three genetic lines of the *Arabidopsis thaliana* plant to five
different durations of ultra-violet light. In Section 5 we discuss recent related work in this area and in Section 6 we provide some final discussion.

2 Method for Detecting Differential Expression

2.1 Notation and Hypotheses

Suppose an unreplicated multiple-treatment timecourse experiment includes \( J \) treatments denoted by \( 1, 2, \cdots, J \) and \( T \) time points denoted by \( 1, 2, \cdots, T \). This experiment would require \( M = J \cdot T \) microarrays each providing an expression level measure on \( G \) distinct genes. Though this experiment provides only one expression measurement for a particular \( \text{gene-treatment-time} \) combination, \( Y_{gjt} \), it does provide \( M \) expression measurements per gene.

The expression at each gene can be described by a model allowing for a unique expected value at each \( \text{treatment-time} \) combination and a random error term:

\[
Y_{gjt} = \mu_{gjt} + \epsilon_{gjt} \quad \text{with} \quad \epsilon_{gjt} \sim iid(0, \sigma^2_g).
\]  

(1)

As stated in Section 1, we consider any gene with an expression pattern different from the overall mean pattern to be potentially biologically interesting. Using this reasoning, we propose a set of hypotheses to be tested at every gene that has the ability to detect a variety of interesting expression patterns. The mean structure at a gene with a true null hypothesis depends neither on treatment group nor time and can therefore be fully described by the single parameter \( \mu_g \). This leads to the following set of hypotheses:

\[
H_{0g} : \mu_{gjt} = \mu_g \quad \text{versus} \quad H_{1g} : \text{not} \ H_{0g}
\]  

(2)

Hereafter, we use the terms \textit{overall mean model} and \textit{null model} interchangeably and use both as a reference to model (1) under \( H_{0g} \).
2.2 Model Selection

In a multiple-treatment microarray timecourse experiment, there are many expression patterns that may be of interest to a researcher. Exploring the data with traditional model selection methods such as BIC or AIC could provide useful information about the existence of particular patterns, but they neither incorporate a formal hypothesis test nor provide a relevant $p$-value for differential expression (as defined by $H_{0g}$ being false).

To provide such a $p$-value, we first select a ‘best’ model at each gene from a given candidate pool using a permutation method described below. This selection process provides us with a test statistic for testing the hypotheses in (2). We compare this test statistic to a permutation distribution of similar test statistics generated under $H_{0g}$ to compute a relevant permutation $p$-value to be used for determining significance of differential expression.

The pool of candidate models consists of all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms that still allow for degrees of freedom for error. We exclude the overall mean model from the candidate pool, and when a particular term is included in a candidate model, we assume that all lower order terms are also included to coincide with the hierarchical order principle described in Wu and Hamada (2000). Thus, a given experiment with $M$ observations per gene will have $S$ models in the candidate pool. For example, in an experiment with three treatments and five time points there are fifteen observations at each gene. The most complex model in the candidate pool fits a unique cubic polynomial to each treatment group. This experiment will have 13 models in its candidate pool (see Section 3.2.2 for models). In general, the most complex model considered in an experiment with $J$ treatments and $T$ time points will fit a unique polynomial of order $(T - 2)$ to each of the $J$ treatment groups.

In the selection process, we first compute the vector of observed $F$-statistics $F_o =$
(F_{10}, F_{20}, \ldots, F_{S0})$ comparing each possible model $s$ to the null model at each gene in the original data set. (For simplicity, we will suppress the subscript $g$ in our notation because the same process applies to all genes.) To generate a possible realized vector of $F$-statistics under $H_0$, we permute the labels on the $M$ microarrays and then re-compute the previously described vector of $F$-statistics $F_s^* = (F_{1s}^*, F_{2s}^*, \ldots, F_{Ss}^*)$ at each gene. Creating a permuted data set in this manner preserves the unknown covariance structure between genes in a single experimental unit, while providing a possible realized data set generated under $H_0$. Repeating this process for a total of $B$ permuted data sets randomly chosen from the $M!$ possible sets provides us with a permutation distribution of $F$-statistic vectors $F_1^*, F_2^*, \ldots, F_B^*$ at every gene.

Our model selection method chooses the model whose $F$-statistic is most extreme when compared to the permutation distribution of $F$-statistics computed for the same model. To make this decision, we compute an approximate permutation $p$-value $p_{so}^*$ for each model $s$,

$$p_{so}^* = \frac{\sum_{b=0}^{B} I(F_{sb}^* \geq F_{so}^*)}{B + 1}$$

where $F_{so}^* = F_{so}$, and choose the 'best' model $m_o \in \{1, \ldots, S\}$ as the model with the smallest permutation $p$-value, i.e. choose $m_o$ such that $p_{m_o o}^* = \min_s p_{so}^*$. Under the null-hypothesis, each $p_{so}^*$ behaves as a conservative $p$-value in that $P_{H_0}(p_{so}^* \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

### 2.3 Test Statistic and $P$-values

To test the set of hypotheses in (2), we need (i) a relevant test statistic, (ii) knowledge of the test statistic's distribution under $H_0$, and (iii) some general knowledge of the test statistic behavior when $H_0$ is false. For (i), we propose using the minimum permutation $p$-value $p_{m_o o}^* \in (0, 1]$. For (ii), we generate an empirical distribution of test statistics under $H_0$ denoted as $p_{m_{11}}^*, p_{m_{22}}^*, \ldots, p_{m_{BB}}^*$ by sending each permuted data
set through the same model selection process as the original data. In essence, when
the null hypothesis is true, $p^*_{m,o}$ acts as the minimum order statistic computed from $S$
dependent Uniform$(0,1)$ random variables. Because the dependence structure in the $S$
random variables is unknown, we use a permutation distribution of the test statistic to
test the hypotheses of interest. With respect to (iii), when the null hypothesis is false we
expect at least one of the values in $F_a$ to be more extreme than most of its $F$-statistic
counterparts generated from the permuted data. This would equate to a relatively small
test statistic $p^*_{m,o}$ for the original data when compared to the permutation distribution
of test statistics $p^*_{m_1}, p^*_{m_2}, \ldots, p^*_{m_B}$.

By comparing the observed test statistic to the permutation distribution of test
statistics we compute an approximate permutation $p$-value $p^{**}$ appropriate for testing
the set of hypotheses in (2) at each gene using the following expression:

$$p^{**} = \frac{\sum_{b=0}^{B} I(p^*_{m,b} \leq p^*_{m,o})}{B + 1}$$

where $p^*_{m,0} \equiv p^*_{m,o}$.

Under the null-hypothesis, each $p^{**}$ behaves as a conservative $p$-value in that $P_{H_0}(p^{**} \leq
\alpha) \leq \alpha$ for any $\alpha \in [0,1]$. This behavior is demonstrated empirically by the distribution
of $p^{**}$ values taken from 114,000 null genes in our simulated data study (Section 3).
The proportion of $p^{**}$ values that fell below 0.001, 0.01, 0.05, 0.10 were 0.0010, 0.0099,
0.0510, and 0.1010, respectively.

The set of empirical $p$-values can be adjusted for multiple comparisons and used to
determine significance by using one of numerous proposed methods. We have opted to
use the method proposed by Storey and Tibshirani (2003). This method estimates the
proportion of genes with a true null hypothesis using the fact that the $p$-values generated
from null genes are uniformly distributed. Using this estimate, the expected proportion
of false positives among all significant genes, also known as the $q$-value, is estimated for
each gene (i.e. for each potential significance threshold). A threshold is chosen that
coincides with an acceptable number of genes declared to be differentially expressed and
an acceptable \( q \)-value.

Up to this point, we have described two possible methods which could be used for identifying differentially expressed genes in a multiple-treatment timecourse experiment. For ease of comparison in the simulation study, we will refer to our proposed method as the \textit{CALM} method to emphasize that its candidate pool considers all models. The second method, related to the overall \( F \)-test, computes only one \( F \)-statistic at each gene comparing the most complex model possible to the overall mean model and we will refer to this method as the \textit{Overall-\( F \)} method. We now introduce a third method that follows the same procedure as the CALM method, but considers only those models that include an interaction term in its candidate pool. This method would be most useful for the researcher who is only interested in locating genes whose expression patterns over time differ across treatment groups. We refer to this method as the \textit{COIM} method to emphasize that its candidate pool includes only interaction models. Specifically, the COIM candidate pool is the subset of the CALM candidate pool that includes only interaction models.

3 Simulation Study

We conducted a Monte Carlo simulation study to compare the sensitivity and specificity of the CALM method, the COIM method, and the Overall-\( F \) method when six particular alternative gene expression patterns are present in the data. We used 2500 permutations for both the CALM and COIM methods.
3.1 Design

The design of the simulated data is based on the motivating experiment described in Section 4 that explored the gene expression of three treatment groups over five time points with one microarray assigned to each treatment-time combination. For this simulation study, each of the hypothetical microarrays contains 5000 probe sets (or genes), and the five time points are assigned at uniform intervals of 0, 6, 12, 18, and 24 hours.

We chose to include the given six alternative patterns (Table 1) primarily because their mean structures represented potentially biologically interesting phenomena. Beyond this initial criterion for inclusion, we were also interested in including both simple and complex as well as polynomial and non-polynomial alternatives to permit a comparison of methods in a variety of situations. The mean structures of the first two alternatives (i & ii) are considered simple and can be fully described by a polynomial regression model in the CALM candidate pool that does not include a treatment-by-time interaction. The mean structures of the next two alternatives (iii & iv) can be described by a candidate model considered by both the CALM and COIM methods. However, these mean structures are more complex and require an interaction term to fully describe their pattern. The last two alternatives (v & vi) have mean structures that can not be described by a model in any candidate pool, but may still be of interest to the researcher. Table 1 describes the specific expression pattern used for each alternative and lists the number of parameters required to fully describe the mean structure when a candidate polynomial regression model is applicable.
Table 1
The six alternative patterns included in the simulated data, the number of parameters required to fully describe the mean structure, and an example of the pattern.

<table>
<thead>
<tr>
<th>Alternative pattern description</th>
<th>Parameters in model</th>
<th>Graph of pattern (e.g.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i Linear in time</td>
<td>2</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
<tr>
<td>ii Linear in time with treatment main effects</td>
<td>4</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
<tr>
<td>iii Non-parallel lines for treatment groups</td>
<td>6</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
<tr>
<td>iv Non-parallel quadratics for treatment groups</td>
<td>9</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
<tr>
<td>v Concurrent jump at time midpoint</td>
<td>na</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
<tr>
<td>vi Non-concurrent logistic growth curves</td>
<td>na</td>
<td><img src="image" alt="Graph of pattern" /></td>
</tr>
</tbody>
</table>

To generate each data set, we simulated each alternative at 200 genes by adding a random error term $\epsilon \sim N(0,1)$ to each of the 15 expected values dictated by the given alternative. For the first four alternatives, we chose mean structures that would provide common Type I and Type II error rates when using the relevant $F$-statistic to test for differential expression. This relevant $F$-statistic uses the simplest model describing the alternative as the full model and the the overall mean model as the reduced model. The chosen mean structures coincide with a comparison-wise Type I error rate of 0.005 and
Type II error rate of 0.20 for all four alternatives. With respect to the non-polynomials, we simulated alternative \( v \) by including a jump of 2.5 standard deviations at the time midpoint. The horizontal asymptotes of the logistic growth curves in alternative \( v_i \) have heights of 3, 4, and 5 standard deviations, respectively, relative to time point 0, and each logistic curve is within 0.01 standard deviations of its asymptote by the 18 hour time point. We simulated the remaining 3800 genes in each data set as a null model genes by randomly generating 15 values from the \( N(0,1) \) distribution. Thus, the null hypothesis is false at 24% of the genes in each simulated data set.

3.2 Results

3.2.1 Receiver Operating Characteristic (ROC) curves

Figure 1 shows the ROC curves for each method and each alternative computed as an average of 30 simulated data sets. We generated each ROC curve using test statistics from the genes simulated as the given alternative and the full set of null model genes. The vertical axis for each plot represents sensitivity, which is the proportion of truly differentially expressed genes that were declared to be significant. The horizontal axis is plotted as 1-specificity, which is the proportion of null model genes declared as differentially expressed. Each point on the curve represents a possible rejection region that coincides with a specific number of genes declared to be significant. ROC curves are commonly plotted from 0 to 1.0 on both axes, but we have graphed the plots with a horizontal axis from 0 to 0.10 for a more discerning view in the rejection regions likely to be of interest in practice.

By comparing the ROC curves in Figure 1, we see that the CALM method performs best for the simpler alternatives \( i \) & \( ii \), and it is followed by the COIM method and Overall-\( F \) method, respectively. The curves for these two alternatives also illustrate a
Figure 1 ROC curves for each method at each of the six alternatives $i$ through $vi$ included in the simulated data based on the average of 30 simulated data sets. The solid line, dotted line, and dashed line represent the CALM method, COIM method, and Overall-$F$ method, respectively.

general trend that as the alternative polynomial patterns move from simple to complex (i.e. more parameters are required to describe the alternative’s mean structure), the distance between the CALM and COIM method ROC curves decreases until the alternative complexity becomes large enough to require an interaction term for full description, as in alternatives $iii$ & $iv$. Once this level of alternative complexity is reached, the COIM method overtakes the CALM method as best shown by the COIM method curves being above the other curves in these two plots. We also see for alternative $iv$, the most complex polynomial alternative included in the simulated data, that the Overall-$F$ method comes much closer in performance to the CALM and COIM methods. This coincides with our earlier statements that the Overall-$F$ method may have little power in detecting
simple alternative patterns because it focuses its attention on more complex patterns. The ROC curves for the non-polynomial alternatives $v$ & $vi$ show a comparative performance that is similar to that found in the simpler alternatives of $i$ & $ii$.

In practice, a researcher will rank the pooled set of test statistics from all genes and reject in order starting with the most extreme and stopping when the rejection region coincides with an acceptable number of significant genes and estimated number of false positives. For example, if we apply the CALM method to our simulated data and choose a $p$-value threshold of 0.005, then on average we reject 345 genes of which 329 are truly differentially expressed. This leads to an overall proportion of false positives of 4.6%. When the most extreme 345 test statistics are rejected for the COIM method, the false positive proportion is 8.4% and for the Overall-$F$ method the proportion is 28.1%.

The set of correctly rejected genes includes all types of alternative expressions. Though we generated the alternatives to be equally represented in each simulated data set, their representation in the set of rejected genes is not equal and depends on the type of alternative, the method applied, and the chosen significance threshold. In Table 2, we show the composition of the average rejection set for each method when the most extreme 345 test statistics are rejected in each simulation. The composition of the rejection region for the Overall-$F$ method coincides with earlier statements that this method focuses on detecting more complex alternatives. Similarly, alternative $iii$, which includes interaction, represents the largest portion of the COIM method rejection set, and the simplest polynomial alternative represents the largest portion of the CALM rejection set.

We computed the overall sensitivity for each method as the proportion of alternative genes that appeared in the aforementioned rejection region by considering the full set of 1200 alternative genes in every data set. On average, the sensitivity was 0.275, 0.264, and 0.206 for the CALM, COIM, and Overall-$F$ method, respectively. The related overall 1-specificity values were 0.0042, 0.0076, and 0.0255.
Table 2
Composition of the average rejection set for each method when the most extreme 345 test statistics are rejected (provided as a count and as a percentage of correct rejections).

<table>
<thead>
<tr>
<th>Alternative</th>
<th>CALM Count</th>
<th>%</th>
<th>COIM Count</th>
<th>%</th>
<th>Overall- ( F ) Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>81</td>
<td>24.6</td>
<td>30</td>
<td>9.4</td>
<td>19</td>
<td>7.8</td>
</tr>
<tr>
<td>( ii )</td>
<td>76</td>
<td>23.0</td>
<td>68</td>
<td>21.6</td>
<td>31</td>
<td>12.6</td>
</tr>
<tr>
<td>( iii )</td>
<td>50</td>
<td>15.3</td>
<td>96</td>
<td>30.3</td>
<td>50</td>
<td>20.2</td>
</tr>
<tr>
<td>( iv )</td>
<td>23</td>
<td>7.1</td>
<td>59</td>
<td>18.8</td>
<td>105</td>
<td>42.1</td>
</tr>
<tr>
<td>( v )</td>
<td>37</td>
<td>11.3</td>
<td>10</td>
<td>3.2</td>
<td>8</td>
<td>3.1</td>
</tr>
<tr>
<td>( vi )</td>
<td>62</td>
<td>18.7</td>
<td>53</td>
<td>16.7</td>
<td>35</td>
<td>14.2</td>
</tr>
<tr>
<td>Total</td>
<td>329</td>
<td>100.0</td>
<td>316</td>
<td>100.0</td>
<td>248</td>
<td>100.0</td>
</tr>
</tbody>
</table>

3.2.2 Model choice

The model choice candidate pool for each simulated data set included 13 polynomial regression models (Table 3). We labeled the models from 1 to 13 to coincide with the ordered level of model complexity. We quantified complexity as the number of parameters required to fully describe a model’s mean structure. For models of equal complexity, we arbitrarily labeled the models using the appropriate consecutive numbers in the rank. Recalling that the null model is not included as a candidate, the simplest candidate, Model 1, included only a linear time effect. The most complex candidate, Model 13, included enough terms to describe an expression pattern represented by separate cubic polynomials for each treatment group.

The true models used to generate the mean structures of alternatives \( i \) through \( iv \) were present in the candidate pool. For this reason, we recorded the CALM method model choice at the genes where alternatives \( i \) through \( iv \) were simulated in order to estimate how often the CALM method chose the correct model. The models used to generate alternatives \( i \) through \( iv \) were Models 1, 5, 8, and 11, respectively, and Figure 2 shows the normalized histograms representing the frequency of model choice. In gen-
Table 3
Terms included in each of the 13 candidate models for the CALM method.
The COIM method candidate pool included only models 8 - 13.

<table>
<thead>
<tr>
<th>Model term</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>time</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>time²</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>time³</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>group * time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>group * time²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>group * time³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

In general, the CALM method chose the correct model a majority of the time for each of these alternatives. For Model 1, the CALM method chose the correct model more than 80% of the time. The histograms also show that when an incorrect model choice is made, the chosen model is usually close in complexity to the correct model.

Figure 2  Normalized histograms based on all 30 simulated data sets showing CALM method model choice frequencies at genes where alternatives i through iv were simulated. The model numbers representing the true models used to simulate alternatives i through iv are 1, 5, 8, and 11, respectively.
4 Example: Arabidopsis Experiment

We applied the CALM method to a data set generated from a microarray experiment designed to explore how different genetic lines of the *Arabidopsis thaliana* plant react to varying durations of ultra-violet light exposure. In this experiment, researchers harvested RNA from wildtype and two mutant lines of the plant at time points 0, 1, 4, 10, and 24 hours representing the duration of exposure. Fourteen plants were allocated to each line-time combination, but their RNA was pooled and hybridized to one array, leaving one expression measurement for each gene-line-time combination. Performance of this experiment required fifteen microarrays. Though there was no true replication, we were able to provide the researchers with a list of significantly differentially expressed genes and to estimate the expected number of false positives using our proposed method.

Each of the fifteen microarrays contained 8297 probe sets (or genes) and the expression at each gene was originally represented by the Affymetrix MAS 5.0 signal intensity (Affymetrix, 2001). To make measurements comparable across arrays, we logged and mean centered the MAS 5.0 values within each array. Finally, before applying our method to this data we replaced each of the original five time points with its respective square root value in order to reduce the influence of the data point collected at 24 hours on the fitted polynomials.

The empirical distribution of p-values generated after applying the CALM method is shown in Figure 3a. Using the procedure presented by Storey and Tibshirani, we approximated the proportion of null genes to be 43% and computed a q-value for each gene. Figure 3b shows the relationship between the number of significant genes in the rejection region and the associated number of expected false positives to fall within that region.

Choosing a q-value of 5.5% as the significance threshold provides us with a list of 500 significant genes. We chose three of these genes labeled as 4949, 8247, and 247
and plotted their observed expression patterns in Figure 4. The observed points are labeled with a 1, 2, or 3 coinciding with wildtype, mutant 1, and mutant 2 genetic lines, respectively. We then overlaid each gene’s fitted expression pattern based on the CALM method model choice. The fitted model for Gene 4949 contains only a linear time effect. The fitted model for Gene 8247 is represented by parallel lines for each genotype and the fitted model for Gene 247 includes interaction and is represented by non-parallel lines for each genotype.

5 Related Work

The two-stage approach we are proposing has some similarities to the two-stage approach proposed by Peddada et al. (2003) used to cluster and select genes according to their timecourse expression. In their first step, the researcher chooses candidate profiles.
Figure 4  Observed expression patterns for Genes 4949, 8247, and 247. The labels of 1, 2, and 3 represent observations from the wildtype, mutant 1, and mutant 2 genetic lines, respectively. The overlaid black lines represent the CALM method fitted model for each gene.

of interest represented by order-restricted means along the given time points. Thus, the shape of a candidate profile can be described as moving up, down, or staying the same for sequential observations in time. The choice of candidates appears to be biologically driven. In the second step, a best fitting profile is chosen for each gene from the pool of candidates, essentially clustering the genes into as many categories as there are candidate profiles. The authors then select genes only if their fit to the best candidate is stronger than the fit that would be seen when there was no correlation in the data. The strength of the fit is judged by comparing the observed fit statistic with a bootstrap null distribution of fit statistics.

Both our method and the method by Peddada et al. involve a model choice component followed by a hypothesis test based on the chosen model. The global null models for both methods are similar in that the expected expression at a gene with a true null hypothesis is represented as a constant across all time points, but because our analysis includes a treatment effect, there is a definite difference in how these null models
are interpreted. Accordingly, Peddada et al. focuses on the timecourse aspect by referring to the flat line expression as the null profile model and we focus on the concurrent movement of treatment groups as well as the time component by referring to the flat line expression as the overall mean model. We should point out, as Peddada et al. mentioned, that though our method is applied to a timecourse study, it could also be applied to a dose-response study, or to any experiment involving one categorical and one quantitative factor.

Park et al. (2003) present a method applied to multiple-treatment timecourse cDNA microarray data to identify genes with different expression profiles between treatment groups. As their first step, Park et al. dichotomizes the data into genes with and without significant treatment-by-time interaction using the relevant traditional $F$-test if the normality assumption is met or a permutation test if it is not. Genes in the first category are associated with the $p$-value testing for an interaction effect and genes in the second category are associated with the $p$-value testing for the main effect of treatment. The $p$-values are then adjusted for multiple comparisons and a $p$-value threshold is chosen to determine significance.

Similar to the overall $F$-test mentioned in Section 1, Park et al. focus on specific alternative patterns in the data by itemizing only two possible alternative hypotheses, whereas our method allows for numerous alternative hypotheses of interest. For example, our method has the ability to identify genes with concurrently upregulated or downregulated expression levels across all treatments, which is not an alternative pattern of interest in the method proposed by Park et al. Also, though they mentioned using their method for an unreplicated experiment, to do so would require making the assumption of no treatment-by-time interaction, which is an assumption our method does not require.
6 Discussion

We have proposed a general method for identifying differentially expressed genes in a multiple-treatment timecourse microarray experiment where differential expression is defined as any expression profile that differs from the simple overall mean profile. Our method utilizes a permutation approach to first choose a best fitting model for each gene from a candidate pool of possible polynomial regression models and then generate a related permutation p-value to be used for testing the hypothesis of differential expression at each gene. Once the distribution of empirical p-values is generated, we suggest using the method proposed by Storey and Tibshirani (2003) to determine significance. This method uses the distribution of p-values to approximate the overall proportion of genes with a true null hypothesis. This subsequent estimate can then be used to help guide in the choice of a rejection region by providing an estimate for the number of expected false positives for any chosen significance threshold.

The candidate pool for model choice plays a large part in our method. By considering time as a continuous variable and including only models allowing for degrees of freedom for error in the candidate pool, we are able to apply our method to relevant unreplicated as well as replicated experiments. Also, after narrowing the candidate pool to a subset of the regression models, a researcher can apply our same proposed process, but focus more detection power on specific expression profiles. We exemplified this flexibility in our method by providing an example when a researcher would be most interested in detecting profiles that included interaction.

We chose the six alternatives included in our simulations to represent a variety of expression pattern complexities. As predicted, simulations showed our method to be more powerful in detecting relatively simple alternative patterns when compared to the method that uses the F-statistic from the overall F-test to detect differential expression. The results suggested that the advantage in our method diminishes as the alternative of
interest becomes more complex. Further research may be conducted to determine the specific level of polynomial complexity at which the methods crossover in performance.

Undoubtedly, to gain power for detecting the simpler alternatives, our method must sacrifice some power for the more complex alternatives. To the researcher who is equally interested in a variety of expression patterns, this trade-off would seem worthwhile. Plus, as the number of treatments or time points in the design gets larger, the potential for net gain also increases. Regardless of which method is more powerful for a particular alternative, one advantage inherent in our method is that it offers a 'best' fitting model for every gene and as the simulations showed, its frequency of choosing the correct model for significant genes with a polynomial expression pattern appears to be high.

Pan (2003) warned that permutation techniques can lead to overly conservative testing procedures for microarray experiments. Pan compared the true null distribution of a standard two-sample $t$-statistic with a permutation distribution obtained by permuting data for 1000 genes 50 times and pooling the permutation replications of the $t$-statistics for all genes. In the simulation, 500 of the genes were generated as differentially expressed. The combined permutation distribution was found to be heavier tailed than the true null distribution, and the heavy tails were attributed to the permutation replications of the $t$-statistics from the 500 differentially expressed genes.

Although the two-sample setting considered by Pan (2003) is quite different from the scenario we consider in this paper, the findings of Pan (2003) do suggest a potential loss of power for the use of permutation procedures in microarray experiments. To investigate this phenomenon in our regression framework, we simulated normally distributed data designed to mimic our three-treatment, five-time-point structure, and compared the true null distribution of our $F$-statistics to their permutation distributions. We found no statistically significant differences between the distributions for simpler alternative patterns and slight differences in the distributions for more complex alternative patterns. The disparity between the theoretical and permutation distributions tended to increase
as the complexity of the alternative increased, i.e. as the degrees of freedom for error decreased. The same phenomenon holds true in the setting considered by Pan (2003). Pan (2003) considered a two-sample scenario with 6 degrees of freedom for error. Simulations similar to those conducted by Pan (2003) indicated no noticeable difference between the permutation distribution and the true null distribution when the error degrees of freedom was increased from 6 to 10. Differences between the true null distribution and the permutation distribution may lead to a loss of power for a gene with a complex pattern, but this inefficiency will not compromise the power for detecting differential expression at other genes because we do not pool our null statistics before computing p-values. Though the permutation procedure may be associated with loss of power in some scenarios, the lack of replication in the data and the desire to make no parametric assumptions suggests a permutation test is presently the best option for this type of data.

The computation time required to apply our method to a microarray data set depends on the experimental design and the chosen number of permutations. Using a 2.66 GHz dual processor machine, 2500 permutations, and the R environment (Ihaka and Gentleman, 1996), it took approximately 12 hours to complete a run on a single simulated data set. After switching to parallel R processing as described by Rossini, Tierney, and Li (2003) on a local cluster composed of four similar machines, a single run was completed in approximately 3 hours. We should also note the we were able to complete a run on a 3 GHz single processor machine using the open source programming language Python (van Rossum and Drake, 2002) in conjunction with Numerical Python (Ascher et al., 2001) in approximately 3.5 hours and this would be our preferred language for this method if restricted to a single processor. Related code is available from the first author upon request.
7 References


CHAPTER 3. GENETIC REGULATION OF GENE EXPRESSION DURING SHOOT DEVELOPMENT IN ARABIDOPSIS THALIANA

A paper published in Genetics

Rhonda DeCook, Sonia Lall, Dan Nettleton, and Stephen H. Howell

Abstract

The genetic control of gene expression during shoot development in Arabidopsis thaliana was analyzed by combining quantitative trait loci (QTL) and microarray analysis. Using oligonucleotide array data from thirty recombinant inbred lines derived from a cross of Columbia and Landsberg erecta ecotypes, the Arabidopsis genome was scanned for marker-by-gene linkages or so-called eQTLs. Single feature polymorphisms (SFPs) associated with sequence disparities between ecotypes were purged from the data. SFPs may alter the hybridization efficiency between cDNAs from one ecotype with probes of another ecotype. In genome scans, five eQTL hot spots were found with significant marker-by-gene linkages. Two of the hot spots coincided with classical QTLs conditioning shoot regeneration suggesting that some of the heritable gene expression changes observed in this study are related to differences in shoot regeneration efficiency between ecotypes. Some of the most significant eQTLs, particularly those at the shoot regeneration QTL sites, tended to show cis-chromosomal linkages in that the target genes were
located at or near markers to which their expression was linked. However, many linkages of lesser significance showed expected ‘trans-effects’, whereby a marker affects the expression of a target gene located elsewhere on the genome. Some of these eQTLs were significantly linked to numerous genes throughout the genome suggesting the occurrence of large groups of coregulated genes controlled by single markers.

1 Introduction

Shoots develop from shoot apical meristems (SAMs) formed during zygotic embryogenesis in plants (Takada and Tasaka, 2002; Baurle and Laux, 2003). Shoots can also be produced adventitiously or regenerated in tissue culture by organogenesis or through somatic embryogenesis. Shoot regeneration in tissue culture is a means by which plants can be propagated and transgenic plants generated (Preil, 2003). In addition, shoot regeneration in tissue culture makes possible the study of shoot development under controlled conditions.

Shoot regeneration in tissue culture is a trait that often varies between plant species and within a plant species among subspecies, varieties, cultivars or ecotypes. Variation in shoot regeneration can be troublesome especially when elite lines are difficult to regenerate. Shoot regeneration efficiency is a quantitative trait, and loci (QTLs) associated with variation in shoot regeneration efficiency have been identified in Arabidopsis (Schiantarelli et al., 2001) and other plants (Komatsuda et al., 1993; Taguchi-Shiobara et al., 1997; Holme et al., 2004). Shoot regeneration QTLs in Arabidopsis were identified in recombinant inbred (RI) lines that differ in shoot regeneration efficiency (Lall et al., 2004). Three significant QTLs associated with shoot regeneration efficiency were found – a major QTL on chromosome 5 in which the superior allele was derived from the parent of the Columbia ecotype, and two minor loci on chromosomes 1 and 4 in
which the Landsberg erecta ecotype parent contributed the superior alleles (Lall et al., 2004). Because superior alleles are distributed between the two parents, the RI population exhibits transgressive segregation of the shoot regeneration trait in that some lines are more or less robust shoot regenerators than either parent (Lall et al., 2004).

Gene expression profiling during shoot regeneration in Arabidopsis has revealed a complex gene expression program with hundreds of significant expression changes (Che et al., 2002). The most significant components of variation contributing to the overall pattern of gene expression changes during shoot development are waves of genes that turn on at one developmental stage and off at the next (Che et al., 2002). One stage with significant gene expression changes occurs about the time of shoot commitment when shoot formation in root explants becomes independent of added plant hormones (Cary et al., 2002) and when an abundance of genes encoding transcription factors and signaling components are upregulated (Che et al., 2002).

An effort has been undertaken by others to explore the genetic control of gene expression programs in a variety of organisms, such as yeast, maize, mouse, rat and man, by combining microarray and QTL analyses (Brem et al., 2002; Schadt et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; Hubner et al., 2005). In doing so, gene expression levels are considered as metric traits and genetic linkages between genes with heritable expression levels and so-called eQTLs (regulatory loci) have been sought (Jansen and Nap 2001; Doerge 2002; Jansen 2003). Both cis and trans-acting eQTLs have been described for regulatory loci that either do or do not colocalize with the regulated genes (Brem et al., 2002; Schadt et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; Hubner et al., 2005). True cis-acting eQTLs are thought to represent genes with polymorphisms that affect their own expression (Schadt et al., 2003). In addition, genome scans conducted in populations segregating for heritable gene expression variation in these organisms has revealed eQTL hot spots. Such hot spots are thought to represent key regulatory loci controlling multiple transcripts – hundreds of transcripts,
as in the case of mouse brain gene expression (Chesler et al., 2005). In several cases analyzed so far, some eQTLs with multiple linkages tended to locate at classical QTLs associated with traits segregating in the population under study (Schadt et al., 2003; Hubner et al., 2005).

In this study, we scanned the Arabidopsis genome for eQTLs that control gene expression at the time of shoot commitment. We attempted to distinguish cis- from trans-chromosomal effects and to determine whether the eQTLs were coincident with classical QTLs associated with shoot regeneration.

2 Materials and Methods

Plant materials and tissue culture procedures: Thirty recombinant inbred lines generated from a cross of Arabidopsis thaliana (L.) ecotypes Ler and Col, Landsberg erecta Columbia were used in this study (Lister and Dean, 1993). The RI lines were chosen as having the greatest number of recombination breakpoints across the genome. Seeds were obtained from Arabidopsis Biological Resource Centre (ABRC). Shoots were regenerated in tissue culture through a two-step regeneration procedure according to Valvekens et al. (1988) as described in Lall et al. (2004).

RNA extraction and DNA chip analysis: Plant material for RNA extraction was collected 6 days after transferring the root segments to shoot induction medium (Lall et al., 2004). Root explants from several hundred seedlings in each line (1 g total) were pooled for RNA extraction. RNA extraction and hybridization to Affymetrix ATH1 (Affymetrix, Inc, CA, USA) oligonucleotide arrays were carried out as described in Che et al. (2002) except, GeneChip Scanner 3000 was used for scanning the chips and the image data generated from the scans was converted to numerical data using GeneChip Operating System v 1.0 (Affymetrix, Inc, CA, USA).
SFP-affected probe pair removal and gene expression: The gene chip scans provided probe intensity readings for all probe pairs in the 22,810 probe sets on the Affymetrix ATH1 array (Affymetrix, 2002). Of the 22,775 non-control probe sets, almost all (99.78%) are composed of 11 probe pairs. The few remaining non-control probe sets are composed of 8, 9, or 10 probe pairs. We removed any probe pair from the data set identified as a SFP according to Borevitz et al. (2003) where the reference ecotype (Col) hybridized with significantly greater intensity than the Ler ecotype (false discovery rate (FDR) < 8%). Using data and scripts from Borevitz (http://www.naturalvariation.org/methods), R software (http://www.r-project.org), and the affy package available through Bioconductor (http://www.bioconductor.org), we investigated the presence of SFPs with a higher intensity in the Ler ecotype. These SFPs were not removed because we found them to be much less prevalent and of lesser effect than their counterparts, making them more difficult to detect at a high level of confidence. Removal of probe pairs was accomplished by defining an alternative CDF environment using the altcdfenvs package (http://www.bioconductor.org) and scripts available in Supplemental Materials. After SFP-affected probe pair removal, we were left with 22,787 probe sets (35 of which were control probe sets) because 23 probe sets had all probe pairs removed. Using the remaining probe pairs in each probe set, we computed MAS 5.0 signal intensities for all 22,787 probe sets via the affy Bioconductor package. The MAS 5.0 values were logged and mean centered for each of the 30 oligonucleotide arrays as a normalization procedure so that expression measures would be comparable across slides. Raw CEL files are available at the Plant Expression Database (PLEXdb) website (http://www.barleybase.org/plexdb/html/index.php)

Data analysis comparing expression and shoot regeneration phenotype: The shoot regeneration phenotype was determined from the number of shoots per root explant using data and methods described by Lall et al. (2004). Briefly, mixed linear model analysis of shoot counts on the square-root scale recommended by Anscombe (1948) was
used to obtain measures of shoot regeneration efficiency for each line.

At each gene (probe set), we tested the null hypothesis of no correlation between expression and phenotype using Spearman’s rank correlation coefficient, a nonparametric test. For sample size greater than 10, the null distribution of the test statistic can be approximated by a t distribution with n-2 degrees of freedom. The resulting P value was used to determine the significance of the relationship. For various thresholds, we estimated a FDR using the method described by Storey and Tibshirani (2003). For example, the threshold P value of $8.83 \times 10^{-5}$ was associated with a significance set containing 20 genes and a FDR of 7.8%.

**Data analysis identifying eQTL:** A filtered set of 288 markers positioned approximately every 2 cM was chosen to represent the genome. The majority of these markers were nonredundant in that at least 1 of the 30 lines had a recombination event between consecutive markers at 240 of the possible 287 consecutive-marker pairings. Markers with a Col allele were coded as 1 and markers with a Ler allele were coded as 0. Missing genotypes were replaced with the estimated probability of a Col allele based on flanking markers. For every marker-by-gene combination, a least squares linear regression was fitted using the coded genotype as the independent variable and expression as the dependent variable. The P value associated with testing the hypothesis of slope equal to zero was used to determine the significance of the relationship. Using a linear regression with the coded genotypes instead of a 2-sample t test with the original genotype groupings to evaluate the strength of the relationship allowed us to include genotype information from all 30 lines in every test. The linear regression is equivalent to the two-sample t test when no genotypes are missing.

Under the required assumptions for a t test, we could apply a Bonferroni adjustment to control the genome-wide error rate for each trait (gene expression) at the 0.05 level by choosing a threshold P value of $1.7361 \times 10^{-4}$, but this adjustment does not account for multiple testing over the 20k genes. In order to estimate an error rate for
the full experiment, we instead used a permutation approach to investigate the FDR associated with various significance thresholds for marker-by-gene linkages. We created 1000 permuted data sets by permuting the RIL labels on the 30 microarrays. For each of five P value thresholds coinciding with Bonferroni genome-wide error rates between 0.0035 and 0.05 (i.e. five P value thresholds between 1.2153 x 10^{-5} and 1.7361 x 10^{-4}), we counted the number of significant marker-by-gene linkages in the set of 6.56 x 10^6 tests for each permuted data set. A FDR for each threshold was estimated applying the method of Storey and Tibshirani (2001) using the region from 0.99 to 1.00 to estimate the proportion of true null hypotheses \( p_0 \) (estimated \( p_0 = 60748/65063.01 = 0.934 \)). The FDRs for these thresholds ranged from 2.3% to 10.2%. The density map of significant linkages was shown for the two extreme thresholds, and thresholds between these values showed density maps with similar patterns.

3 Results

3.1 Single Feature Polymorphisms

We used RI lines derived from a cross between the two standard ecotypes, Col x Ler (Lister and Dean, 1993) as a mapping population to identify eQTLs. A problem with RI lines derived from two different ecotypes is the presence of single feature polymorphisms (SFPs), small sequence differences between the ecotypes (Borevitz et al., 2003). They have the potential to confound oligonucleotide chip analysis because differences in hybridization efficiency caused by SFPs could be interpreted as differences in gene expression levels.

To circumvent this problem, we utilized information from Borevitz to identify and eliminate from our data analysis SFP-affected probe pairs (a probe pair being a set of two probes with one a perfect match to the gene sequence and the other a mismatch)
Table 1
Frequency of probe sets with SFP-free probe pairs.

<table>
<thead>
<tr>
<th>Number of SFP-free probe pairs</th>
<th>Frequency out of 22,775 probe sets ¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>16,047</td>
</tr>
<tr>
<td>10</td>
<td>3,321</td>
</tr>
<tr>
<td>9</td>
<td>1,483</td>
</tr>
<tr>
<td>8</td>
<td>755</td>
</tr>
<tr>
<td>7</td>
<td>473</td>
</tr>
<tr>
<td>6</td>
<td>236</td>
</tr>
<tr>
<td>5</td>
<td>157</td>
</tr>
<tr>
<td>4</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>0</td>
<td>23</td>
</tr>
</tbody>
</table>

¹ Does not include 35 control probe sets.

where the reference ecotype (Col) hybridized with greater intensity than the Ler ecotype. Removing these probe pairs dispenses with the problem of hybridization artifacts due to ecotype-specific sequence differences within probe pairs. We did not, however, address other, less frequent issues such as possible hybridization differences due to ecotypic differences in gene copy numbers. Out of 22,775 non-control probe sets on the Affymetrix ATH1 chip, 16,047 had no significant SFPs, and at the other extreme, 23 had significant SFPs in all probe pairs of the probe set (Table 1). The elimination of SFP-affected probe pair data from our dataset is not without some impact on the reliability of our data. However, Gautier et al. (2004) suggested that the number of probe pairs needed for reliable gene expression is probably less than 11, the number used by Affymetrix for most non-control probes, but the minimal number is not known. To check the sensitivity of our results to the inclusion of less reliable probe sets, we repeated our analysis after removing all probe sets with fewer than 5 probe pairs. All aspects of the results showed very little change. For example, the 34 genes found to be linked to the major shoot
regeneration QTL remained the same, and only one gene containing more than 5 SFPs was removed from the list of 100 genes showing the most evidence of association between expression and phenotype.

### 3.2 Gene Expression Pattern Signatures

Thirty of the most informative Arabidopsis RI lines (with the most recombinant breakpoints) from Lister and Dean (1993) were analyzed in this study. RNA was extracted from root explants that had been preincubated on CIM for 4 days, transferred to SIM and incubated for 6 more days. There are no obvious morphological differences between the two ecotypes at this stage, which precedes shoot emergence. However, about this time, root explants become 'committed' to shoot development, that is, they continue to form shoots even when transferred to basal medium without hormones (Cary et al., 2002). This stage is characterized by abundant expression changes in genes encoding transcription factors and signaling pathway components (Che et al., 2002). Labeled cRNAs were generated from the 30 RNA samples and hybridized to Affymetrix Arabidopsis gene chips with 22,810 genes (probe sets).

Genes were identified with expression patterns that significantly correlate with the shoot regeneration efficiency phenotype. Expression levels for individual genes were plotted against a shoot regeneration phenotype computed from the number of shoots per root explant for each RI line. The genes with the strongest correlation between gene expression and shoot regeneration phenotype at a false discovery rate (FDR) of 7.8% were identified (Table 2). At5g48330, a putative regulator of chromosome condensation (cell cycle regulatory protein), showed the most significant correlation between gene expression and shoot regeneration phenotype (t-value = 6.6912) (Figure 1). Other genes with high correlation between gene expression and shoot regeneration phenotype include those encoding a VAMP membrane protein, RNA recognition motif protein,
Figure 1  Plot of gene expression vs. shoot regeneration phenotype (shoots per explant on square-root scale) for At5g48330 in the 30 RI lines used in this study. At5g48330, encoding a regulator of chromosome condensation family protein, shows the strongest relationship between gene expression and phenotype. Linear regression line and parental allele at the shoot regeneration QTL indicated by C (Col) or L (Ler). Expression values represent MAS 5.0 signals that were logged and mean centered for each gene chip. A negative expression reflects a gene with a logged MAS 5.0 value below the average logged expression of genes on a gene chip.

SWIB complex BAF60b domain-containing protein, two proteases (Clp and subtilisin-like), protodermal factor (located in the L1 embryonic layer) and so forth.

3.3 Genome Scan

To identify loci controlling heritable gene expression patterns, such as those described above, the Arabidopsis genome was scanned for marker-by-gene expression linkages. Expression signals (corrected for SFP-affected probe pairs) in the 30 RI line data set were treated as quantitative traits and subjected to linkage analysis using a filtered set of 288 markers that were uniformly positioned about every 2 cM. In doing so, a test statistic
Table 2
Genes with significant\(^2\) relationship between expression and shoot regeneration phenotype and an assessment of their association with the major QTL.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Probe</th>
<th>(t)-value</th>
<th>Gene function(^3)</th>
<th>Linkage (p)-value(^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At5g48330</td>
<td>248693_at</td>
<td>6.691211</td>
<td>Regulator of chromosome condensation family protein</td>
<td>0.0004187</td>
</tr>
<tr>
<td>At4g17870</td>
<td>254705_at</td>
<td>-5.9651</td>
<td>Expressed protein</td>
<td>0.0463669</td>
</tr>
<tr>
<td>At5g49840</td>
<td>248575_at</td>
<td>-5.9250</td>
<td>Clp protease</td>
<td>0.0000004</td>
</tr>
<tr>
<td>At5g47180</td>
<td>248796_at</td>
<td>-5.6123</td>
<td>VAMP membrane protein</td>
<td>0.0291548</td>
</tr>
<tr>
<td>At1g22910</td>
<td>257413_at</td>
<td>-5.5869</td>
<td>RNA recognition motif protein</td>
<td>0.0074835</td>
</tr>
<tr>
<td>At1g44800</td>
<td>261335_at</td>
<td>5.4033</td>
<td>Nodulin MtN21 family protein</td>
<td>0.0063793</td>
</tr>
<tr>
<td>At3g60390</td>
<td>251374_at</td>
<td>5.1479</td>
<td>Leucine zipper protein</td>
<td>0.0242114</td>
</tr>
<tr>
<td>At3g03150</td>
<td>258845_at</td>
<td>5.1256</td>
<td>Expressed protein</td>
<td>0.0000358</td>
</tr>
<tr>
<td>At5g54970</td>
<td>248139_at</td>
<td>5.1098</td>
<td>Expressed protein</td>
<td>0.0003867</td>
</tr>
<tr>
<td>At2g35605</td>
<td>266641_at</td>
<td>5.0940</td>
<td>SWIB complex BAF60b domain-containing protein</td>
<td>0.0018222</td>
</tr>
<tr>
<td>At5g48360</td>
<td>248696_at</td>
<td>5.0192</td>
<td>Formin homology 2 domain-containing protein</td>
<td>0.0002446</td>
</tr>
<tr>
<td>At5g47760</td>
<td>248780_at</td>
<td>-4.9368</td>
<td>Putative 4-nitrophenylphosphatase</td>
<td>0.0172426</td>
</tr>
<tr>
<td>At5g45650</td>
<td>248961_at</td>
<td>-4.8329</td>
<td>Subtilisin-like protease</td>
<td>0.0003827</td>
</tr>
<tr>
<td>At4g39860</td>
<td>252821_at</td>
<td>4.8270</td>
<td>Expressed protein</td>
<td>0.0001439</td>
</tr>
<tr>
<td>At5g56970</td>
<td>247956_at</td>
<td>4.7951</td>
<td>Cytokinin oxidase family protein</td>
<td>0.0000326</td>
</tr>
<tr>
<td>At5g53850</td>
<td>248234_at</td>
<td>-4.7576</td>
<td>Haloacid dehalogenase-like</td>
<td>0.0125426</td>
</tr>
<tr>
<td>At2g45510</td>
<td>267500_s_at</td>
<td>-4.6979</td>
<td>Cytochrome P450</td>
<td>0.0125426</td>
</tr>
<tr>
<td>At2g42840</td>
<td>263979_at</td>
<td>4.6615</td>
<td>Protodermal factor 1 (L1 layer protein)</td>
<td>0.0034888</td>
</tr>
<tr>
<td>At5g46510</td>
<td>248847_at</td>
<td>4.6337</td>
<td>TIR-NBS-LRR class protein</td>
<td>0.0010187</td>
</tr>
<tr>
<td>At5g58350</td>
<td>247819_at</td>
<td>-4.5761</td>
<td>WNK family protein kinase</td>
<td>0.0000059</td>
</tr>
</tbody>
</table>

\(^2\) List represents significant genes associated with a 7.8% FDR.

\(^3\) Gene functions according to TAIR.

\(^4\) \(p\)-value for testing eQTL linkage to major shoot regeneration QTL site (marker 270).
evaluating each marker-by-gene association was computed. The resulting $6.56 \times 10^6$ P values were subjected to a significance threshold and the proportion of significant linkages was plotted across the Arabidopsis genome (Figure 2). Using a permutation approach, a FDR was estimated for various significance thresholds coinciding with genome-wide error rates for a single trait (gene expression) between 0.0035 and 0.05. These thresholds equate to comparison-wise P values of $1.2153 \times 10^{-5}$ and $1.7361 \times 10^{-4}$. The more stringent threshold is associated with 3,525 significant linkages distributed over 958 genes and a FDR of 2.3%. The less stringent threshold is associated with 10,521 significant linkages distributed over 2,637 genes and a FDR of 10.2%. Due to correlation in markers, many of these linkages can be considered redundant in that they represent single marker-by-gene linkages. Limiting each gene to link to only one marker (chosen as the marker with the highest test statistic) would remove these redundancies, but the peak of a marker regulating a large number of genes (Figure 2) may appear falsely low due to a 'spreading-out' of its regulated genes to nearby markers. (For purposes of illustrating where target genes were located on the genome relative to their linked markers, we did remove these redundancies later in Figure 6 to produce a clearer plot while maintaining the existing relationship.)

At all threshold levels, the genome scan of markers with significant linkages revealed peaks with higher densities of significant linkages (Figure 2). These peaks are similar to linkage 'hot spots' found in the yeast or mouse genomes (Brem et al., 2002; Schadt et al., 2003; Chesler et al., 2005). It was of interest that two of the hot spots corresponded to two of three shoot regeneration QTLs identified in a prior study (Lall et al., 2004). The major shoot regeneration QTL is located on the lower arm of chromosome 5, centered on marker 270, and one of two minor QTLs is located on chromosome 4 and centered on marker 190 (Figure 2). At the threshold associated with a FDR of 2.3%, marker 270 links to 34 genes – 23 genes were upregulated in association with the Col allele at the marker site, and 11 were downregulated (Table 3). A sampling of the single
marker-by-gene associations at this site (selected for genes upregulated in association with the Col allele) clearly shows that the genes are expressed at a much higher level when Col alleles, rather than the Ler alleles, are present at the marker site (Figure 3A-C). Similar plots for genes downregulated in association with the Col allele show lower levels of expression when Col alleles are present (Figure 3D). In general, genes with a strong relationship between expression and shoot regeneration phenotype also showed a strong association with the major shoot regeneration QTL (Table 2), though declaration of linkage significance depends on the chosen threshold.

Could the linkage hot spots be artifacts due to allele frequency differences across the genome? To examine this, the proportion of Col and Ler at each of the 288 markers
Figure 3  Examples of the effect of the genotype at marker 270 on the expression of various genes to which the marker is significantly linked. Expression levels in the 30 RI lines are grouped according to presence of the Landsberg erecta (L) or Columbia (C) allele at marker 270. Horizontal lines represent QTL genotype group means.

was estimated. Known allele genotypes were coded as the probability of a Col allele, and unknown allele genotypes were also coded as the probability of a Col allele based on flanking markers. The estimated proportion of Col alleles was based on the expected count of such alleles and was computed as the average of the coded values at each marker. This estimate reduces to a straightforward proportion when all genotypes are known. The out-of-balance group size or proportion of alleles in the larger group (Col or Ler) was plotted across the genome (Figure 4A). If equal numbers of Col and Ler alleles were present at each marker, then a flat line centered on 0.5 would be expected. However, allele imbalances were observed across the genome for this set of RI lines. The comparison-wise power of detecting a 1.5 standard deviation difference in the average
gene expression for the Col and Ler groups at each marker was then calculated based on a type I error rate of 0.05 (Figure 4B). By fixing the difference in the means and the type I error rate, power becomes dependent on only the sample size of the two groups and on the accuracy of genotyping. Power is greatest when all genotypes are known and there are equal numbers in each group (15 in each group for this data set). In plotting out-of-balance group size and power across the genome, it can be seen that there are regions, especially in chromosome I where there are out-of-balance group sizes and low power markers.

However, the question is whether the linkage hot spots derive from markers with high power of detection. It can be seen in scans of the genome that clusters of significant linkages and marker power don’t necessarily align (Figure 4B and C). That becomes clearer in comparing numbers of significant linkages to power of detection on

Figure 4  Allele frequency distribution in a genome-wide scan. (A) Out-of-balance group size or the proportion of Col or Ler alleles in the larger group at each of the 288 markers used in genome scans. (B) The comparison-wise power of detecting a 1.5 standard deviation in the average gene expression at each marker based on a type I error rate of 0.05. Power is greatest when genotypes are known and equal numbers of the two different parental alleles are present at a given marker. The two spikes associated with lower power occur due to missing genotypes at the given markers. (C) Distribution of number of significant linkages at a threshold associated with a FDR of 2.3%.
a chromosome-by-chromosome basis. None of the chromosomes show a strong linear association between power at marker and number of significant associations (data not shown).

3.4 Cis- and Trans-Chromosomal Effects

What was unexpected was to find that the most significant marker-by-gene associations at the FDR of 2.3% involved linkages to genes located in the region of the markers. For example, all of the significant associations with marker 44 were linked to genes in the vicinity of the marker (Figure 5). Of the 23 upregulated genes linked to marker 270 at the shoot regeneration QTL site, all but one were located in the region of the major shoot regeneration QTL itself, and of the 11 downregulated genes, all were located near the QTL site. Thus, some of the most significant linkages were ‘neighborhood-effects’ in which the marker was in the vicinity of genes to which the marker was expression-linked.

Cis-effects have been reported in yeast and mouse, and markers linked to genes in cis have more significant associations than those in trans (Brem et al., 2002; Schadt et al., 2003). In these systems, cis-effects appear to have a simple genetic explanation in that they result from polymorphisms that affect the expression level of the genes in which the polymorphisms occur (Brem et al., 2002; Schadt et al., 2003). In this study, we eliminated from consideration polymorphisms (SFPs) that may give rise by artifact to apparent expression differences. These are SFPs within the probe sets that may alter the hybridization efficiency of cDNA made from the RNA of one ecotype to the probe of another ecotype. However, genes purged of SFPs in their probe sets are still included in our data (only the probe sets comprised completely of SFPs have been eliminated).

Therefore, we asked whether SFPs and cis-effects of the kind described in yeast and mice account for the observed variation in expression in the 23 upregulated and 11 downregulated genes linked to the major shoot regeneration marker 270. Out of
Marker 44 (24 upreg, 8 downreg)

Marker 124 (17 upreg, 14 downreg)

Marker 190 (32 upreg, 15 downreg)

Marker 242 (18 upreg, 16 downreg)

Marker 270 (23 upreg, 11 downreg)

Figure 5  Location of genes linked to markers at the eQTL hot spots in the Arabidopsis genome at a threshold associated with a FDR of 2.3%. Markers 190 and 270 are centered on the shoot regeneration QTLs. Ticks pointing upward show location of upregulated genes and downward pointing ticks are downregulated genes. Markers are located about every 2 cM and position of markers are indicated by an open oval.
23 upregulated genes linked to marker 270, 9 had reported single nucleotide (SNPs) or single feature polymorphisms (SFPs) (Table 3). (A genome-wide set of SNP markers distinguishing Col and Ler ecotypes is available through SeqViewer at The Arabidopsis Information Resource or TAIR http://www.arabidopsis.org/servlets/sv.) The remaining 14 upregulated genes had no reported SNPs or SFPs in comparing Col or Ler. Of the 11 downregulated genes, 3 had reported SNPs or SFPs, and 8 did not. Even in the target genes with SFPs or SNPs, there is a low probability that any polymorphism affects gene expression. Thus, the neighborhood effects associated with marker 270 are probably not attributed to ‘cis-effects’ of the sort proposed in yeast and mice, i.e., polymorphisms in the target gene (Brem et al., 2002; Schadt et al., 2003).

It is conceivable that the neighborhood effects we observe might not result from polymorphisms in genes acting on their own expression, but by polymorphisms in nearby genes. Such neighborhood effects might be revealed in higher resolution studies involving more RI lines. Nonetheless, it does appear that several genes within a neighborhood are commonly controlled in our study. For example, the four genes most strongly associated with the major shoot regeneration QTL site are located within 37 kb of each other and the correlation in expression between any two of the four genes is positive and above 0.81 suggesting that these genes might be commonly controlled.

Significant marker-by-gene linkages across the genome were also illustrated by plotting the position of markers against the position of their corresponding linked genes. Specifically, each gene in the list of significant linkages was plotted against its best controlling marker for the two thresholds associated with FDRs of 2.3% and 10.2% (Figure 6). As reported in previous studies (Schadt et al., 2003; Bystrykh et al., 2005) we found that the most significant marker-by-gene linkages tended to represent cis-chromosomal effects (a marker affects the expression of a target gene in close proximity) as seen in the plot by the many black dots falling along the diagonal. As the number of significant linkages increases, more dots appear off the diagonal suggesting trans-chromosomal
Table 3
Genes with the strongest linkage to marker 270, the effect of the Col allele on their expression, and the existence of SFPs\(^5\) or SNPs\(^6\) within their probe sets.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Effect of Col allele</th>
<th>SFP</th>
<th>SNP</th>
<th>Gene</th>
<th>Effect of Col allele</th>
<th>SFP</th>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>At2g18540</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g53070</td>
<td>up-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g47800</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g53120</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>At5g47940</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g53950</td>
<td>up-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g48240</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g58730</td>
<td>up-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50410</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g59140</td>
<td>up-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50550</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g59290</td>
<td>up-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50560</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g48110</td>
<td>down-regulated</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>At5g50565</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g49840</td>
<td>down-regulated</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>At5g50570</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g50230</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50580</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g51390</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50630</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g52540</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g51670</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g53360</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g51820</td>
<td>up-regulated</td>
<td></td>
<td></td>
<td>At5g53370</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g51960</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g53420</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g51980</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g53760</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g53000</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g53850</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g53050</td>
<td>up-regulated</td>
<td>X</td>
<td></td>
<td>At5g58350</td>
<td>down-regulated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^5\) As defined in Borevitz et al. (2003) using a FDR threshold of 8%.
\(^6\) As defined by The Arabidopsis Information Resource (TAIR).

effects (a marker affects the expression of a target gene elsewhere on the genome) are present, but not as strong as the cis-effects.

Also apparent in the plot are dense vertical bands indicating regions of the genome that regulate a large number of genes in trans during shoot development (Figure 6). Two of these regions are located on chromosome 5 and coincide with previously described hot spots for the 3,525 most significant linkages (Figure 2). Genes linked in trans to these hot spots show no obvious pattern in that the targeted genes are scattered throughout the genome. The trans-chromosomal hot spot located on the lower arm of chromosome 5 is correlated with the major shoot regeneration QTL. That finding was expected because
Figure 6  Significant marker-by-gene linkages plotted as (x, y) coordinates with the x-axis representing the genome location of the marker and the y-axis representing the genome location of the linked gene. Each dot represents a single gene plotted against its best controlling marker. Significant linkages at two thresholds (see text) are shown. Black dots represent significant linkages at a threshold associated with a FDR of 2.3%, and grey dots represent significant linkages at a less stringent threshold associated with a FDR of 10.2%. Chromosome endpoints are indicated by dashed lines.

This QTL likely controls many genes throughout the genome associated with shoot regeneration. However, trans-effects were not concentrated at the chromosome 4 hot spot described above suggesting most of the strong linkages in this region, associated with a minor shoot regeneration QTL, are cis-effects.
4 Discussion

In scanning the Arabidopsis genome for eQTLs associated with heritable changes in gene expression during shoot development, it was found that significant marker-by-gene linkages tended to cluster in ‘hot spots’ as they do in the yeast or mouse genomes (Brem et al., 2002; Schadt et al., 2003). Why they tend to do so is not clear. A hot spot could be due to a single gene at the hot spot that influences the expression of many other genes or it could be a cluster of several genes at the hot spot each of which act on a few genes. In any case, two of the eQTL hot spots coincided with two of the three QTLs associated with the efficiency of shoot regeneration - the major shoot regeneration QTL on chromosome five and a minor QTL on chromosome 4 (Lall et al., 2004). It was expected that eQTLs and shoot regeneration QTLs might coincide because the gene expression data were collected during the process of shoot regeneration. Furthermore, QTLs that condition the efficiency of shoot regeneration undoubtedly affect the expression of many other genes. It was noted in the other studies that eQTL hot spots correspond to QTLs or sites of marker genes involving a phenotype that segregated in the mapping population (Schadt et al., 2003; Hubner et al., 2005).

However, it was unexpected to find that the markers with the most significant associations are linked to the expression of genes in the same vicinity of the chromosome as the marker. We refer to these effects as ‘neighborhood effects’, and neighborhoods are very large in molecular terms. The genes with the most significant linkages to marker 270 at the shoot regeneration QTL site cover nearly 1.5 Mbp of DNA. Neighborhood effects might be due to mechanisms similar to those that regulate genes in operons. That, however, seems unlikely given the large size of the neighborhoods and the distant spacing of some of the affected genes.

Another possibility is that genes are regulated by epigenetic mechanisms acting at the chromosome level - such as chromatin effects. Chromatin structure is known to
influence gene regulation locally and globally and to specify functional differentiation of chromosomal domains during development in a number of organisms (Weiler and Wakimoto, 1995). Chromatin features have been described for some of the Arabidopsis chromosomes on which eQTL hot spots were found (CSHL/WUGSC/PEB, 2000; Lippman et al., 2004). In particular, one of the eQTL hot spots in our study was located in a euchromatic region on the long arm of chromosome 4, coincident with a shoot regeneration QTL. This chromosome is thought to be organized in euchromatic loops which emanate from condensed heterochromatic chromacenters (Fransz et al., 2002). The chromacenters are composed of heterochromatin from pericentric and nucleolus organizing regions (NOR) and the loops extending from these chromacenters are estimated to be 0.2-2.0 Mbp (Fransz et al., 2002), consistent with the dimensions of the neighborhood effects. It is possible that some of cis-chromosomal gene regulation effects we see on chromosome 4 may involve chromatin or chromosome loops. It would be of interest to compare chromatin structure around some of the up or downregulated genes in the parental ecotypes.

Two other unexpected findings in our study are also of note. One is that the eQTL hot spot on chromosome 4 is the site of a minor shoot regeneration QTL that has significant epistatic effects on the major shoot regeneration QTL on chromosome 5 (Lall et al., 2004). Given this interaction one might expect significant trans-chromosomal linkages between markers and target genes at the QTLs on chromosome 4 and 5. Such linkages were found for the significance thresholds we investigated at the QTL on chromosome 5, but not at the QTL on chromosome 4. Another unexpected finding was the presence of eQTL hot spots at sites other than the QTL sites. The eQTL hot spot on the upper arm chromosome 5 is particularly prominent and is not associated with a shoot regeneration QTL. That must mean that although the locus is associated with ecotype-specific gene expression changes, those expression changes have little impact on shoot regeneration.

This study would not have been possible without information on single feature poly-
morphisms (SFPs). Probe set SFPs resulting in a higher hybridization affinity for the Col ecotype had the potential to alter this analysis significantly. Performing similar analyses on the data before and after SFP probe pair removal allowed us to determine the impact of SFPs on our results. After SFP removal, 34 genes were significantly linked to marker 270 in the eQTL analysis (Table 3). Using the same P value threshold applied to the data before SFP removal, we found 40 significant linkages to marker 270. The number of genes significantly upregulated in Col decreased from 31 to 23 after SFP removal, while the number significantly downregulated in Col increased from 9 to 11. This implies many apparent strong marker-by-gene expression linkages were due to SFP probes. All genes that were eliminated from the significance list contained at least 1 SFP probe pair and most contained numerous SFPs. The genes that were dropped may still have strong relationships with marker 270, but the relationship was not strong enough to be considered significant at the given level once SFPs were removed. No genes were dropped from the downregulated gene list, but 2 genes that contained SFPs in the original data were added. Although many of the same genes appeared on both significance lists (before and after SFP removal), the fact that some did not suggests that it’s important to consider SFPs in data analysis. Study results often guide ongoing research and the removal of SFP probe pairs in data analysis may help researchers avoid inefficient use of resources.

The genetic basis for the major shoot regeneration QTL on chromosome 5 has not yet been determined, however, a number of candidate genes are under study. It will be interesting to know whether the genetic entity that conditions shoot regeneration at this site is also responsible for controlling the target genes in the eQTL analysis.
5 References


The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems Arabidopsis Sequencing Consortium (2000). The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* 100, 377-386.


1 Conclusions

In this first research chapter of this dissertation, we proposed a method that extends the classical QTL analysis for normal traits to the detection of zero-inflated Poisson (ZIP) traits. QTL methods search for locations along the genome that are associated a quantitative trait. For example, a genome location that is associated with tumor size in a mouse or crop yield for a plant would be considered a QTL. Though the classical QTL analysis based on normality can be applied to traits that are approximately normal, not all traits can be transformed to attain approximate normality. When the normal method is not applicable, one option is to use a nonparametric method which has been proposed in the past, but nonparametric methods tend to have lower power than appropriately applied parametric methods. We proposed a parametric QTL detection method for ZIP traits. The classical normal method fits a mixture of normal distributions, but in the case of a ZIP trait, the underlying distribution is itself a mixture. Therefore, a distinguishing characteristic of the ZIP QTL method is that it uses the EM algorithm to fit a mixture of mixtures in order to determine parameter estimates. Our simulation results showed that the ZIP method performed as well as or better than the nonparametric in terms of power, and tended to have more accuracy in detecting the QTL location. When the ZIP trait distribution could be transformed to be approximately normal, we found that the ZIP method performed as well as or better than the normal method in terms
of power. The ZIP method also tended to have more accuracy in detecting the QTL location than the normal method.

In the second research chapter, we proposed a method to analyze microarray data that has been generated from a more complicated design than the common two-sample scenario. The method can be applied when sample sizes are very small, and even when there is only one observation for each treatment. Specifically, we proposed a method for detecting differentially expressed genes in an unreplicated multiple-treatment time-course experiment. Besides the issues with multiple testing that arise in gene expression data, this type of design must deal with a model selection issue as well. The method we proposed provides the researcher with a list of significant genes, an estimated false discovery rate for that list, and a 'best model' choice for every gene. The model choice component is relevant because the alternative hypothesis in our hypothesis test of differential expression does not dictate one specific alternative expression pattern. Using simulations, we showed that this method performed better at detecting truly differentially expressed genes when compared to a competing method. Simulations also showed that the proposed method chose the correct model for a given differentially expressed gene in more than half of the simulated data sets for the parameter configurations investigated.

In the final research chapter, we brought together QTL analysis and microarray analysis in a new analysis called eQTL analysis. In this type of analysis, the classical QTL detection method described in the background of the first chapter is applied to gene expression data. As there are thousands of genes in a gene expression data set, there are also thousands of traits of interest. The researcher is interested in finding any genome location that is associated with the expression at any gene. Because there are usually hundreds of testing positions along the genome in a QTL study, there are essentially millions of tests in an eQTL study. Accounting for the correlation between consecutive marker genotypes and the correlation between genes, we determined a sig-
nificance threshold providing the researcher with a list of locations on the genome that controlled gene expression, a list of genes that were being controlled, and an estimated false discovery for the given list. The strongest associations were found between a gene (i.e. a locus) that controlled its own expression, but there were also many significant associations where a locus controlled a gene other than itself.

2 Future Research

In the eQTL analysis method described in this dissertation, a test for association was performed between each marker and each gene spotted on a microarray. The resulting pool of tests was then adjusted for multiple testing. We would like to investigate a unified approach for finding eQTL and for clustering genes that are coregulated. The dimension reduction capability of clustering could greatly reduce the number of relevant tests needed to perform an eQTL analysis.

The method proposed in this dissertation for detecting ZIP traits models the marginal trait distribution as a mixture of ZIPs. We are interested in extending this method to incorporate random effects introduced by experimental design. The parameters of interest, the proportion $p$ and the Poisson parameter $\lambda$, could be estimated using a generalized linear model approach that incorporates the random effects, and could also incorporate multiple covariates of interest.