2009

Estimating the number of true null hypotheses and the false discovery rate from multiple discrete non-uniform permutation p-values

Timothy John Bancroft
Iowa State University, timmyb@iastate.edu

Estimating the number of true null hypotheses and the false discovery rate from multiple discrete non-uniform permutation p-values

by

Timothy J. Bancroft

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
Ranjan Maitra
Dan Nordman
Stephen Vardeman
Roger Wise

Iowa State University

Ames, Iowa

2009

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

To this point in my academic life, there have been three major influences that have lead me to where I am today.

My parents, Dave and Diane, have been there from day one. I can still remember my grade school days when my mother would always ask me if I had my homework done, helping me with book reports, and overall stressing the importance of an education while my father sacrificed his own interests to provide for his family.

The constant encouragement of my parents landed me at Augsburg College where I majored in Mathematics and completed an independent study project with Professor Su Dorée. She pushed my intellectual boundaries and instilled in me academic fervor and integrity. She was the main impetus in my decision to attend graduate school and if not for her, I would not have been put in the position to pursue a graduate degree in statistics.

I chose the statistics graduate program at Iowa State University where I met Dr. Dan Nettleton at the beginning of my second year and had, what I now think of as, a golden opportunity to work under him as a research assistant. After working with him for the last four years, I can see his influence not only in my academic career, but in my everyday life as well. His kind, patient, and encouraging approach is not common in competitive academic settings. There is only so much a student can learn from a textbook or a lecture, and I feel I learned just as much from him in our hourly meetings than I did in all my classes. I may have made it through graduate school with a different major professor, but I know it wouldn't have been nearly as fulfilling and enjoyable.

ABSTRACT

This dissertation addresses statistical issues that arise in a multiple testing framework when each of m hypotheses is tested via permutation methods. A standard error rate to control in multiple testing situations (especially when $m \sim 10^4$) is the false discovery rate which describes the expected ratio of type I errors to the total number of rejections. An adaptive approach to controlling the false discovery rate is to estimate the number of type I errors using a data-based estimate of m_0, the number of true null hypotheses. Estimation of m_0 has received much interest in recent years. Existing methods assume each of the m p-values has a continuous uniform (0,1) null distribution. This dissertation discusses numerous ways in which p-values may not have continuous uniform (0,1) null distributions and proposes how to estimate m_0 and the false discovery rate in these scenarios. The first scenario involves a sequential permutation testing procedure that can substantially reduce computational expense when the test statistic is computationally intensive. The method is demonstrated via an application involving the genetic mapping of expression quantitative trait loci (eQTL). Other scenarios are motivated by problems that arise in genomics and proteomics.

CHAPTER 1. General Introduction

1.1 Introduction

This overview briefly describes the main components of this dissertation, including multiple testing, the false discovery rate, permutation testing, and how certain permutation testing scenarios produce null p-values (i.e., the set of p-values whose null hypotheses are true) that are not equally likely. This last concept is the central motivation of this dissertation.

1.2 Multiple Testing

A modern multiple testing scenario is a micrroarray experiment. Microarrays simultaneously measure the expression levels of thousands genes. Given two treatment conditions, the objective is to find genes whose expression distribution differs between the two treatment groups. Such genes are said to be differentially expressed. For example, say interest lies in determining which genes are involved in muscle hypertrophy (growth). Given a treatment that induces muscle growth, suppose five mice are randomly assigned to the control group and five mice are randomly assigned to the treatment group. After the treatment is applied, a tissue sample is taken from each mouse and from this sample, the expression of each of thousands of genes is measured. A subset of some hypothetical data is given in Table 1.1.

Once the data are collected, the idea is to test each gene for equivalent expression between the treatment and the control group. To accomplish this, the hypothesis $H_0 : \psi_j^c = \psi_j^t$ could be tested where ψ_j^c is the distribution function of expression values for gene j from mice given the control treatment and ψ_j^t is the distribution function of expression values for gene j from mice given the muscle growth treatment. The decision to reject or fail to reject each of $H_{01}, \ldots, H_{0m}$ is based on the corresponding p-values $p_1, \ldots, p_m$. Given independence among

Gene j	H_{0j}	Control					Treatment				
1	H_{01}	3.7	4.1	3.9	5.1	5.4	6.0	5.5	4.0	4.6	4.6
2	H_{02}	8.2	6.2	7.3	7.6	6.0	8.1	6.4	5.6	7.6	6.6
3	H_{03}	6.9	4.1	5.1	3.3	5.4	6.0	4.9	5.7	9.3	7.4
4	H_{04}	8.6	8.8	9.1	9.8	7.9	6.2	6.8	6.6	6.8	5.5
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
m	H_{0m}	3.5	1.5	2.9	4.5	0.9	3.0	3.9	3.8	3.1	3.9

Table 1.1 Hypothetical data from a microarray experiment.

mice and normally distributed data with constant variance, p-values from simple two-sample t-tests would each have a continuous uniform (0,1) null distribution. For now, suppose that is the case, and suppose that all genes whose corresponding p-value is less than some cutoff c are declared to be non-null. In theory, biological researchers can then make some genetic connection from these significant genes to muscle development. Now, an efficacious microarray experiment typically produces a gene list in the hundreds using a p-value cutoff of $c = 0.01$ when $m = 20,000$ genes are investigated. Not all genes present on the gene list will be truly differentially expressed as some false discoveries have most likely been made. For example, if all $m = 20,000$ genes are truly equivalently expressed and a p-value cutoff of $c = 0.01$ is used, then we would expect to make $20,000 \cdot 0.01 = 200$ type I errors. Controlling the number, or rate, of type I errors made in multiple testing scenarios is introduced in the following section.

1.3 False Discovery Rate

When m null hypothesis are tested, we choose to reject or fail to reject each null hypothesis based on its corresponding p-value. Each hypothesis p-value pair can be placed in one of the four cells of Table 1.2.

Multiple testing adjustments attempt to control quantities related to the unobservable random variable V. The family wise error rate (FWER) has historically been the error rate chosen to control, which amounts to choosing a significance cutoff c so that $\mathrm{P}(V \geq 1) \leq \alpha$, for

	Accept Null No Discovery	Reject Null Declare discovery	Total
True Nulls	U	V	m_0
False Nulls	T	S	$m - m_0$
Total	W	R	m

Table 1.2 Table of outcomes in a multiple testing scenario.

some $\alpha \in (0,1)$. The common Bonferonni adjustment chooses $c = \alpha/m$, which controls FWER at level α. Holm (1979) proposed choosing $c = p_{(k)}$, where $p_{(k)}$ is the k^{th} smallest p-value among the m p-values and $k = \max\left\{k^* : p_{(j)} \leq \frac{\alpha}{m-j+1} \text{ for all } j = 1, \ldots, k^*\right\}$. This method also controls FWER at level α. However, FWER is a conservative error rate in many multiple testing situations such as microarray experiments. FWER is the probability of making at least one type I error, and in microarray experiments, researchers will concede some type I errors as long as they are only a small proportion of all genes identified as differentially expressed. For example, suppose a gene list of 1,000 genes are declared to be differentially expressed. If just one of those 1,000 genes is a type I error, FWER considers the whole list of 1,000 genes to be an error. In contrast, a researcher would consider such a list to be very valuable. Therefore, focus has shifted to controlling the proportion of rejected null hypotheses that are false discoveries. Benjamini and Hochberg (1995) proposed the false discovery rate (FDR). They defined FDR as the expectation of the random variable Q, where $Q = V/\max\{1, R\}$. FDR is essentially the expected ratio of the number of type I errors to the total number of rejections. Benjamini and Hochberg (1995) show that under certain conditions, choosing $c = p_{(k)}$ where $k = \max\left\{k^* : \frac{p_{(k^*)} \cdot m}{k^*} \leq \alpha\right\}$, controls FDR at level $\frac{m_0}{m}\alpha$. Benjamini and Hochberg's (1995) approach is used in practice to control FDR at level α, and thus is unnecessarily conservative when $m_0 \leq m$. Replacing m with m_0 reduces the conservative bias but requires estimation as m_0 is unknown. Many methods exist to estimate m_0, but most assume each null p-value has a continuous uniform (0,1) distribution. The next section previews situations where each null p-value has a discrete distribution where the null probabilities of each possible p-value are not equally likely. Estimating m_0 in these cases, and therefore FDR, is one of the main aspects of this dissertation.

1.4 Permutation Testing

Permutation testing is a popular alternative testing procedure used when the assumptions of parametric testing are in question. Permutation testing builds a reference distribution by computing the chosen test statistic for every possible assignment of observed responses to observational units. The p-value is the proportion of test statistics computed from all possible assignments of responses to observational units that are as extreme or more extreme than the test statistic corresponding to the observed pairing of responses to observational units. This procedure produces a valid p-value when the null hypothesis implies that the w total observations have the same joint distribution under any random assignment of the w responses to the w observational units, i.e. the w observations are exchangeable under the null hypothesis. This null hypothesis implies that each of the $w!$ assignments of the w responses to the w observational units is equally likely, and thus, for continuously distributed random variables, the observed test statistic is equally likely to be any rank between 1 and $w!$ under the null hypothesis. This implies, that under the null hypothesis, $\mathrm{P}(p\text{-value} \leq i/w!) = i/w!$ for all $i = 1, \ldots, w!$.

For example, suppose $Y_{11}, \ldots, Y_{1w_1}$ are iid ψ_1 and $Y_{21}, \ldots, Y_{2w_2}$ are iid ψ_2 for any two continuous distribution functions ψ_1 and ψ_2 and let $w = w_1 + w_2$. Suppose we wish to test $H_0 : \psi_1 = \psi_2$ versus $H_A : \psi_1 \neq \psi_2$, a simple test of distributional equality. Under H_0, the joint distribution of the combined $w_1 + w_2 = w$ observations, say, $Y_1^*, \ldots, Y_w^*$, is equivalent to the joint distribution of $Y^*_{\pi(1,\ldots,w)_1}, \ldots, Y^*_{\pi(1,\ldots,w)_w}$ for any permutation $\pi(1, \ldots, w)$ of the indices $(1, \ldots, w)$, where $\pi(1, \ldots, w)_i$ is the i^{th} of the w elements of the random permutation of $(1, \ldots, w)$. This property is known as exchangeability and is required whenever a null hypothesis is tested using permutation techniques. The next step is to define a test statistic that discriminates between H_0 and H_A. Here, the usual pooled variance two sample t-statistic, t, could be chosen as the test statistic. Hence, we can calculate $t(\text{obs})$, the test statistic corresponding to the observed assignment and we can calculate $t_1, \ldots, t_{w!}$, where t_i is the value of the test statistic for the i^{th} of the $w!$ assignments of the w responses to the w observational units. The two-sided p-value is $\sum_{i=1}^{w!} \frac{I(|t(obs)| \leq |t_i|)}{w!}$, where $I(\cdot) = 1$ if the argument is true and 0 otherwise. It can be shown that under H_0, this p-value is equally likely to be any element in the set $\left\{\frac{1}{w!}, \frac{2}{w!}, \ldots, 1\right\}$ since the test statistics are equally likely to be any rank between 1 and $w!$ under H_0 (again, for continuously distributed random variables).

Advantages to permutation testing are the distribution free assumptions; only exchangeability is required under the null hypothesis. Also, a permutation test exists for any test statistic regardless of whether or not its null distribution is known. Therefore, one is always free to choose a test statistic that best discriminates between the null and alternative hypotheses. One drawback to permutation testing is the computational expense, which can happen when the number of observations per group is large and/or the test statistic is difficult to compute. When the number of observations in each group is large, asymptotic results can often be utilized and the computational expense of permutation tests in these cases can be avoided by using parametric methods. To alleviate some computational expense when the test statistic is computationally expensive, this dissertation employs the sequential procedure of Besag and Clifford (1991). This procedure is introduced in the following subsection.

1.4.1 Sequential P-values

Suppose a hypothesis H_0 is to be tested using a test statistic Z with a null distribution function ψ and that small values of Z are consistent with H_0. Given the observed test statistic z, suppose the quantity $\mathrm{P}(Z \geq z) \overset{H_0}{=} 1 - \psi(z)$ cannot analytically be determined, but we can sample iid values from ψ, namely, $z_1, z_2, \ldots, z_{n-1}$. Then, $\frac{\sum_{i=1}^{n-1} I(z \leq z_i) + 1}{n}$ can used to approximate the true p-value $1 - \psi(z)$. These approximations are commonly used in permutation testing settings, especially when the number of possible permutations is large and/or the test statistic has a non-trivial computational expense. In that case, ψ is the distribution of the test statistic computed for each permutation of the data, (i.e., $t_1, \ldots, t_{w!}$). Typically, the number of values, $n-1$, drawn from ψ is taken to be large when computation time is not an issue. Regardless, a lack of evidence against H_0 can be observed in a sample much smaller than size

n saving unnecessary computation time. Besag and Clifford (1991) propose sampling until one of the following occurs: 1) a fixed number, h, of sampled values are larger than z or 2) a fixed number, $n-1$, of values have been sampled from ψ. The p-value under this scheme is defined as

$$p = \left\{ \begin{array}{ll} h/L & \text{if } G = h, \\ (G+1)/n & \text{if } G < h \end{array} \right\},$$

where L denotes the number of values sampled at termination and G denotes the number of sampled values that are strictly larger than the observed test statistic z.

The set $S(h,n) = \left\{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{h-1}{n}, \frac{h}{n}, \frac{h}{n-1}, \ldots, \frac{h}{h+1}, 1 \right\}$ describes the possible p-values that are obtainable under this procedure for given values of h and n. We will show that each sequential permutation p-value p^*, has a null distribution that is similar to a continuous uniform (0,1) distribution in the $\mathrm{P}(p^* \leq s) \overset{H_0}{=} s$ for all $s \in S(h,n)$. However, because the elements of $S(h,n)$ are not equally spaced, the null probabilities $\mathrm{P}(p^* = s)$ are not equal for all $s \in S(h,n)$. Hence, if we tested m hypotheses using this sequential approach with a fixed h and n, the resulting null p-values would each have the same non-uniform discrete distribution. One main focus of this dissertation is how to estimate m_0, the number of true null hypotheses out of m total tests, and the false discovery rate when m hypotheses are tested using p-values that are not continuous uniform under their null hypotheses. Sequential permutation p-values serve as one example where the collection of p-values do not have a continuous uniform distribution under the null hypothesis. Next, a second scenario is described where the null distribution of the p-values is a mixture of non-uniform discrete distributions.

1.4.2 Permutation Testing of Data with many Ties

1.4.2.1 Completely Randomized Design

Consider a two treatment completely randomized design with three observations per group. Suppose the observed values are $\{0,0,0\}$ and $\{0,1,2\}$ for the two treatment groups, respectively. The null hypothesis of equal distributions across the two treatment groups is tested using permutation. The test statistic chosen is the absolute difference in treatment sums.

There are a total of 6! assignments of responses to experimental units, each of which is equally likely under the null hypothesis. Table 1.3 displays an example arrangement for each of the possible values of the test statistic and the p-value permutation distribution. The cardinality of the p-value support in this case is two and is related to the number of ties in the data set. To see this, consider another simple example with fewer ties.

Example Arrangement	Test Statistic	Probability under H_0	p-value
$\{2, 0, 1\}$ $\{0, 0, 0\}$	3	$6 \cdot 2 \cdot 4!/6! = 0.4.$	0.4
$\{0, 1, 0\}$ $\{0, 2, 0\}$	1	$6 \cdot 3 \cdot 4!/6! = 0.6.$	1

Table 1.3 One example arrangement for each value of the test statistic and the p-value permutation distribution for some example data with four tied observations from a completely randomized design with two treatments.

For this example suppose the responses are now $\{0, 0, 0\}$ and $\{1, 2, 3\}$. For these data, there are only four values of the test statsitic, namely zero, two, four, six. Table 1.4 displays the same information as Table 1.3 for this example dataset which contains only three tied observations. The cardinality of the p-value support in this case is four and larger than in the previous example since there are less ties in the dataset.

Example Arrangement	Test Statistic	Probability under H_0	p-value
$\{2, 3, 1\}$ $\{0, 0, 0\}$	6	$6 \cdot 2 \cdot 3!/6! = 0.1.$	0.1
$\{3, 0, 2\}$ $\{1, 0, 0\}$	4	$6 \cdot 2 \cdot 3 \cdot 3!/6! = 0.3.$	0.4
$\{0, 2, 0\}$ $\{0, 1, 3\}$	2	$6 \cdot 2 \cdot 3 \cdot 3!/6! = 0.3.$	0.7
$\{0, 1, 2\}$ $\{3, 0, 0\}$	0	$6 \cdot 2 \cdot 3 \cdot 3!/6! = 0.3.$	1

Table 1.4 One arrangement for each value of the test statistic and the p-value permutation distribution for some example data with three tied observations from a completely randomized design with two treatments.

These examples show that when permutation testing is applied to data with many tied observations, the resulting p-value null distribution can be discrete non-uniform. Another example of data with many tied observations is given next.

1.4.2.2 Fisher's Exact Test

Consider two categories (A and B), each with two levels (1 and 2), measured on $n_{..}$ subjects. This data can be summarized in a two-by-two table, where each cell corresponds to one of the four combinations of category levels. Suppose the data, given in Table 1.5 for $n_{..} = 5$, are $\{A1,\ B1\}$, $\{A1,\ B1\}$, $\{A1,\ B2\}$, $\{A2,\ B2\}$, and $\{A2,\ B2\}$ where $A1$ represents that the subject has level 1 of category A. This kind of data inherently contains many ties as there are only two possible responses to each of the two categories. Fisher's exact test (Fisher 1934) can be used to test for independence between the two categories. This test is a classic example of a test that yields p-values that are discrete and non-uniform under the null hypothesis as this test is based on exact small sample distributions in contrast to the chi-square test based on asymptotic arguments.

	$B1$	$B2$	Total
$A1$	$n_{11} = 2$	$n_{12} = 1$	$n_{1.} = 3$
$A2$	$n_{21} = 0$	$n_{22} = 2$	$n_{2.} = 2$
Total	$n_{.1} = 2$	$n_{.2} = 3$	$n_{..} = 5$

Table 1.5 Generic data from $n_{..} = 5$ observations measured on two categories (A and B), each with two levels (1 and 2).

Fisher's exact test is explained in more detail in Chapter 3. Briefly, Fisher's exact test p-value is used to test for independence between category A and category B. Under independence, n_{11} has a hypergeometric distribution. Fisher's exact test p-value is calculated by summing all hypergeometric probabilities that are less than or equal to the hypergeometric probability for the observed data. These probabilities are computed for the range of values that n_{11} could take on given the marginal totals. These probabilities are small when n_{11} is extreme, and therefore, large and small values of n_{11} provide evidence against independence. Since the range of possible values for n_{11} depends only on the marginal totals, the p-value is discrete and can take on only one of finitely many values, each of which are not equally likely under independence.

This test can also be viewed from a permutation testing perspective. Under the null

Subject	Responses	
1	$A1$	$B1$
2	$A1$	$B1$
3	$A1$	$B2$
4	$A2$	$B2$
5	$A2$	$B2$

	$B1$	$B2$	Total
$A1$	2	1	3
$A2$	0	2	2
Total	2	3	5

Figure 1.1 One of 5! example arrangements and the corresponding two-by-two table for responses on two categories (A and B), each with two levels (1 and 2) on 5 subjects.

hypothesis of independence, each assignment of the category B responses to the category A responses is equally likely. Consider the test statistic as the value of n_{11}. Then, we can record the value of n_{11} for each of the 5! assignments. Once all 5! assignments have been made, the proportion of times each possible value of n_{11} has occurred is recorded. Next, the p-value is obtained by sum all proportions that are less than or equal to the proportion corresponding to the observed data. It turns out, this process is exactly constructing the hypergeometric null distribution given the observed marginal counts used by Fisher. Here, given the marginal totals, there are only three possible values of the test statistic, namely zero, one, and two. The arrangement for a test statistic equal to two is given in Figure 1.1. The probability for this arrangement is $3!2!\binom{2}{2}\binom{3}{1}/5! = \binom{2}{2}\binom{3}{1}/\binom{5}{3} = 0.3$. Arrangments for the other two values of the test statistic are given in Figure 1.2 and Figure 1.3. These arrangements occur with probability $3!2!\binom{2}{1}\binom{3}{2}/5! = 0.6$ and $3!2!\binom{2}{0}\binom{3}{3}/5! = 0.1$, respectively. Together, the proportion of arrangements that yield $n_{11} = 0,\ 1,$ and 2 is 0.1, 0.6, and 0.3, respectively. Since the observed value of $n_{11} = 2$, the p-value for the observed arrangement is $0.3 + 0.1 = 0.4$.

Subject	Responses	
1	$A1$	$B1$
2	$A1$	$B2$
3	$A1$	$B2$
4	$A2$	$B1$
5	$A2$	$B2$

	$B1$	$B2$	Total
$A1$	1	2	3
$A2$	1	1	2
Total	2	3	5

Figure 1.2 One of 5! example arrangements and the corresponding two-by-two table for responses on two categories (A and B), each with two levels (1 and 2) on 5 subjects.

Subject	Responses	
1	$A1$	$B2$
2	$A1$	$B2$
3	$A1$	$B2$
4	$A2$	$B1$
5	$A2$	$B1$

	$B1$	$B2$	Total
$A1$	0	3	3
$A2$	2	0	2
Total	2	3	5

Figure 1.3 One of 5! example arrangements and the corresponding two-by-two table for responses on two categories (A and B), each with two levels (1 and 2) on 5 subjects.

Regardless of how the test is interpreted, the p-value has a discrete support where the elements are not equally likely under the null hypothesis. Here, only one p-value has been obtained. Next generation sequencing data, to be described in Chapter 3, is one application where the analysis of one dataset may involve analysis of m two-by-two tables. In that case, the marginal counts may not all be the same across the two-by-two tables, and thus, the null p-values are from a mixture of discrete non-uniform distributions. Another application of data with many ties includes some proteomics experiments, which will also be described in Chapter 3, where data for each of m proteins contains many ties. Applying permutation testing to each of the typically hundreds of proteins results in a collections of m p-values that do not all follow the same discrete non-uniform null distribution.

1.5 Organization

The core ideas of estimating the number of true null hypotheses and the false discovery rate when using permutation to test multiple hypotheses are seen throughout this dissertation.

The second chapter develops sequential permutation p-values for application to multiple testing problems. One case study will show the dramatic reduction in the total number of test statistics computed across the m tests by using the sequential approach.

The third chapter introduces methods to estimate the number of true null hypotheses and the false discovery rate when the collection of null p-values arise from a mixture of discrete non-uniform distributions. Applications to proteomics, gene set testing, and next generation sequencing data are presented.

The fourth and final piece of this dissertation applies the ideas of the second chapter to expression quantitative trait loci mapping studies. These studies are intense computationally as the genome is scanned along small increments to determine which position on the genome has the most association with expression for each of thousands of genes. Finding this candidate position involves employing the EM algorithm at each of the hundreds of testing positions along the genome, and assessing the significance of the candidate position requires permutation as the null distribution of the maximum association (across testing positions) between the expression of the given gene and the candidate position is not analytically tractable due to dependence along the genome. Given the computational intensity of the test statistic, using sequential permutation p-values can greatly reduce the computational burden in this scenario.

1.6 References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

[2] Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*, **Vol. 78, No. 2**, 301-304.

[3] Fisher, R.A. (1934). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Loyd.

[4] Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, **Vol. 6, No. 2**, 65-70.

CHAPTER 2. Computationally Efficient Estimation of False Discovery Rate Using Sequential Permutation *P*-Values

Abstract

We consider the problem of testing each of m null hypotheses with a sequential permutation procedure in which the number of draws from the permutation distribution of each test statistic is a random variable. Each sequential permutation p-value has a null distribution that is non-uniform on a discrete support. We show how to use a collection of such p-values to estimate the number of true null hypotheses m_0 among the m null hypotheses tested and how to estimate the false discovery rate (FDR) associated with p-value significant thresholds. The properties of our estimates of m_0 and FDR are evaluated through simulation and illustrated through the analysis of a microarray dataset. An additional application to the problem of mapping quantitative trait loci for thousands of gene expression traits is discussed.

Key Words: Expression quantitative trait locus; Microarray; Monte Carlo testing; Multiple testing; Sequential analysis

2.1 Introduction

For most Monte Carlo testing procedures, it can be clear well before a large sample from the null distribution is taken that there is little evidence against H_0. For example, consider a simple two group comparison where a permutation test is used to test for a difference between two continuous distributions. If the sample size for each group is 10 and the absolute difference between sample means is used as the test statistic, then the total number of distinct values in the support of the test statistic's permutation distribution is almost surely $C\left(20,10\right)/2 = 92,378$. An exact permutation p-value is given by the fraction of all values in the permutation distribution as extreme or more extreme than the value of the test statistic computed from the observed data. Imagine that the observed test statistic falls near the 50^{th} percentile of 100 random draws from the permutation distribution. Even though only a small fraction of the total number of values from the permutation distribution has been examined, there is very little evidence against H_0 and little relevant information to be gained by examining additional values from the permutation distribution if our goal is to test H_0 at traditional significance levels.

Besag and Clifford (1991) recognized this issue and proposed a sequential procedure that stops sampling from a test statistic's null distribution as soon as (1) the number of values more extreme than the observed statistic reaches a prespecified value or (2) the total number of draws from the null distribution reaches a prespecified value. We will demonstrate in Section 2 that a sequential permutation p-value resulting from this procedure has a non-uniform null distribution with discrete support in (0,1] and that such a p-value can be used to obtain exact tests of significance at levels contained in the p-value's support and conservative tests for other levels. Depending on the choice of prespecified values and the status of the null hypothesis, the sequential permutation p-value can be computed at far less computational expense than the standard permutation p-value and with little loss of relevant information.

If there is only a single hypothesis to be tested, generating a large sample from the test statistic's permutation distribution – or even generating the entire permutation distribution – is not necessarily a serious computational burden. However, when multiple hypotheses are each to be tested with a permutation approach, the overall computational expense is the sum of the expenses of the individual tests so that when the number of tests is large, the computational burden of traditional permutation testing may be substantial. Thus, we find Besag and Clifford's (1991) approach to be particularly relevant for the case of multiple permutation tests. Extending Besag and Clifford's approach to the multiple testing scenario is the main focus of this paper.

Our work is primarily motivated by applications in microarray data analysis where tests number in the thousands and false discovery rate (FDR) is the error rate of interest. Much of the research on FDR since Benjamini and Hochberg's (1995) seminal paper has focused on obtaining less conservative estimates of FDR by incorporating estimates of the number of true null hypotheses m_0 in FDR calculation (see, for example, Benjamini and Hochberg, 2000; Storey, 2000a,b; Mosig et al., 2001; Storey and Tibshirani, 2003; Langaas, Lindqvist, and Ferkingstad, 2005; Nettleton et al., 2006; Ruppert, Nettleton, and Hwang, 2007). Nearly all existing methods rely heavily on the assumption that a p-value from a test with a true null hypothesis is continuous and uniformly distributed on the interval (0,1). Because the sequential permutation p-values do not have this uniformity property, a new approach for estimating m_0 and FDR from sequential permutation p-values is needed. To address this problem, we extend the histogram-based estimator of Nettleton et al. (2006) to obtain an estimator of m_0 and use a variation of Benjamini and Hochberg's (1995) procedure to estimate FDR from discrete, non-uniform null p-values (i.e., the set of p-values whose null hypothesis is true).

In Section 2 of this paper, we review the sequential procedure of Besag and Clifford (1991) with emphasis on the null distribution of sequential permutation test p-values. The procedure of Benjamini and Hochberg (1995) to control FDR is described in Section 3. In Section 4, we review the histogram-based estimator of m_0 presented by Nettleton et al. (2006) and extend this estimator to address the case of discrete, non-uniform p-values. In Section 5, we apply the proposed procedure to the Acute Lymphoma Leukemia (ALL) dataset and discuss the results and computational savings compared to testing using other permutation schemes. We also discuss other natural applications of the proposed procedure in Section 5. Three simulation studies that compare the estimated FDR with the observed false positive fraction for cases of independent and dependent p-values are presented in Section 6.

2.2 Sequential Permutation P-Values

Suppose a permutation distribution ψ places equal probability on every element in a discrete, finite support. Suppose that large values of a test statistic Z provide evidence against a null hypothesis H_0. For an observed test statistic z, a Monte Carlo approximation to the exact permutation p-value $1 - \psi(z)$ is given by $\frac{\sum_{i=1}^{n-1} I(z \le Z_i)+1}{n}$, where $Z_1, \ldots, Z_{n-1}$ is a simple random sample from ψ and $I(\cdot) = 1$ if the argument is true and 0 otherwise. Such Monte Carlo p-values are often used when conducting a permutation test – particularly when the number of distinct elements in the support of ψ is large. Henceforth, we will assume that sampling from ψ is without replacement although the distinction between sampling with or without replacement is practically unimportant when the cardinality of the support of ψ is large.

Usually, the Monte Carlo sample size $n - 1$ is taken to be large when computation time is not an issue. However, regardless of computational expense, it can be clear in a sample much smaller than $n - 1$ whether there is evidence against H_0. In the general context of Monte Carlo testing, Besag and Clifford (1991) propose sampling until one of the following occurs: 1) a fixed number, h, of sampled values are larger than z or 2) a fixed number, $n - 1$, of values have been sampled from the null distribution. The p-value under this scheme is defined as

$$p = \left\{ \begin{array}{ll} h/L & \text{if } G = h, \\ (G+1)/n & \text{if } G < h \end{array} \right\},$$

where L denotes the number of values sampled at termination and G denotes the number of sampled values that are strictly larger than the observed test statistic z.

The p-values that can result from this procedure depend on h and n and lie in the set $S(h, n) = \left\{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{h-1}{n}, \frac{h}{n}, \frac{h}{n-1}, \ldots, \frac{h}{h+1}, 1 \right\}$. In this paper, we consider the case of sequential permutation p-values where ψ serves as the null distribution in Besag and Clifford's sequential procedure. Under H_0, the cumulative distribution function of the sequential permutation p-

value coincides with the cumulative distribution function of a continuous uniform$(0,1)$ random variable at the points in $S(h,n)$; that is,

$$P\left(p^* \leq s\right) = s \text{ for all } s \in S(h,n) \text{ when } H_0 \text{ is true.} \tag{2.1}$$

To see that (2.1) holds, first consider the case where $p^* = \frac{j}{n}$ for some $j \in \{1,\ldots,h\}$. Note that

$$\begin{aligned}
P\left(p \leq p^*\right) &= P\left(p \leq \frac{j}{n}\right) \\
&= P\left(p = \frac{1}{n}\right) + \ldots + P\left(p = \frac{j}{n}\right) \\
&= P\left(G = 0 \text{ in a sample of size } n-1 \text{ from } \psi\right) + \ldots + \\
&\quad\; P\left(G = j-1 \text{ in a sample of size } n-1 \text{ from } \psi\right) \\
&= P\left(Z \text{ is largest in a sample of total size } n \text{ from } \psi\right) + \ldots + \\
&\quad\; P\left(Z \text{ is } j^{\text{th}} \text{ largest in a sample of total size } n \text{ from } \psi\right) \\
&\overset{H_0}{=} \frac{1}{n} + \ldots + \frac{1}{n} = \frac{j}{n} = p^*.
\end{aligned}$$

If $p^* = \frac{h}{j}$ for any $j \in \{n-1,\ldots,h\}$, then,

$$\begin{aligned}
P\left(p \leq p^*\right) &= P\left(p \leq \frac{h}{j}\right) \\
&= P(Z \text{ is among the } h \text{ largest values when included} \\
&\quad\; \text{with the first } j-1 \text{ sampled from } \psi) \\
&= P\left(Z \text{ is largest in a sample of total size } j \text{ from } \psi\right) + \ldots + \\
&\quad\; P\left(Z \text{ is } h^{\text{th}} \text{ largest in a sample of total size } j \text{ from } \psi\right) \\
&\overset{H_0}{=} \frac{1}{j} + \ldots + \frac{1}{j} = \frac{h}{j} = p^*.
\end{aligned}$$

Thus, under H_0, the distribution of the sequential permutation p-value of Besag and Clifford (1991) is given by

$$\mathrm{P}(p = S_j(h,n)) = S_j(h,n) - S_{j-1}(h,n),$$

where $S_j(h,n)$ denotes the j^{th} smallest element of $S(h,n)$ and $S_0(h,n) \equiv 0$. For example, for $h = 2$ and $n = 4$, the distribution places probabilities 1/4, 1/4, 1/6, and 1/3 on support points 1/4, 1/2, 2/3, and 1, respectively.

As mentioned in the Introduction, the main focus of this paper is to use sequential permutation p-values in a multiple testing framework. In particular, we are motivated by problems where the number of tests is in the thousands. FDR has become the error rate of choice for such problems. In the next section, we review the basics of FDR and discuss the challenges that arise when estimating FDR with sequential permutation p-values.

2.3 Estimation of False Discovery Rate with Sequential Permutation *P*-Values

Suppose $p_{(1)}, \ldots, p_{(m)}$ are the m ordered p-values used to test the corresponding m null hypothesis $H_{(01)}, \ldots, H_{(0m)}$. Let m_0 denote the unknown number of true null hypotheses among the m hypotheses tested. Consider a decision procedure that results in rejection of R of the m null hypotheses. Let V denote the number of true null hypotheses among the R rejected null hypotheses. Benjamini and Hochberg (1995) defined FDR as the expectation of the random variable $V/\max\{1, R\}$. They showed that, under certain conditions, FDR will be controlled at level $\frac{m_0}{m}\alpha$ for any $\alpha \in (0,1)$ if $H_{(01)}, \ldots, H_{(0k)}$ are rejected, where

$$k = \max\left\{k^* : \frac{p_{(k^*)}m}{k^*} \leq \alpha\right\}. \tag{2.2}$$

When control of FDR at level α is desired, Benjamini and Hochberg's procedure will be conservative whenever $m_0 < m$. If m_0 were known, the inherent conservativeness could be removed by substituting m_0 for m in (2.2) to obtain control at level α rather than $\frac{m_0}{m}\alpha$. Because m_0 is unknown, a natural strategy is to replace m in (2.2) with an estimate of m_0. This strategy was first proposed by Benjamini and Hochberg (2000) and by Storey (2002, 2003) and Storey et. al (2004).

As noted in the introduction, estimation of m_0 has become a central issue in estimation of FDR, but the case in which null p-values are uniformly and continuously distributed on

(0,1) has been the focus of past research. Here we address the previously unconsidered case in which the m hypothesis $H_{(01)}, \ldots, H_{(0m)}$ are tested using the sequential permutation procedure introduced in Section 2. The corresponding m p-values each have a discrete non-uniform null distribution as described in (2.1). The next section focuses on estimating m_0 from a collection of such p-values.

Given an estimate of m_0 denoted $\widehat{m}_0$, we define for any $c \in S(h, n)$,

$$\widehat{\mathrm{FDR}}(c) = \min\left\{\frac{p_{(k)}\widehat{m}_0}{k} \,:\, p_{(k)} \geq c\right\}. \tag{2.3}$$

If (2.3) is calculated for each value of $c \in \{p_{(1)}, \ldots, p_{(m)}\}$, the resulting values of $\widehat{\mathrm{FDR}}(c)$ are essentially the q-values of Storey (2003). As noted by Storey (2002), the decision rule that rejects $H_{(0k)}$ if and only if $\widehat{\mathrm{FDR}}\left(p_{(k)}\right) \leq \alpha$ is equivalent to the decision rule obtained by replacing m with $\widehat{m}_0$ in (2.2).

2.4 A Histogram-Based Estimator of the Number of True Null Hypotheses

Mosig et al. (2001) proposed an iterative histogram-based algorithm that estimates m_0 when each p-value has a continuous uniform$(0, 1)$ null distribution. The first part of this section describes, both informally and formally, an extension of Mosig et al. (2001) to the case where each p-value has a discrete non-uniform null distribution. The second part of this section illustrates the extension of the algorithm with a simple example. Nettleton et al. (2006) proved the existence of and characterized the limit of the iterative procedure of Mosig et al. (2001). The third and final part of this section states the extension of the convergence result of Nettleton et al. (2006) for the case where each p-value has a discrete non-uniform distribution under its null hypothesis. The result is proved in the appendix.

2.4.1 The Iterative Algorithm for Estimating m_0 from Sequential Permutation P-Values

The basic idea of the iterative algorithm of Mosig et al. (2001) is the following. Given a histogram (with any bin size) of p-values $p_{(1)}, \ldots, p_{(m)}$, start by assuming that the null hypothesis is true for all m tests. Then, the expectation of the number of p-values in each bin is known if we assume that each p-value is has a continuous uniform$(0,1)$ null distribution. Next, find the leftmost bin where the number of p-values fails to exceed expectation, and for each bin to the left of this bin, compute the observed number of p-values minus the expected. The sum of these differences is an estimate of $m - m_0$, and therefore m minus the sum yields an estimate of m_0. Now, recalculate the number of null p-values expected in each bin using the new estimate of m_0 as the number of true null hypotheses. Again, find the leftmost bin where the number of p-values fails to exceed the new expectation, and for each bin to the left of this bin, compute the observed number of p-values minus the expected. As before, m minus the sum of these differences provides an updated estimate of m_0. The algorithm continues in this fashion until convergence.

To adapt this algorithm to the case of discrete, non-uniformly distributed sequential permutation p-values, let n_j denote the number of p-values that equal $S_j(h,n)$, the j^{th} element of $S(h,n)$; $j = 1,\ldots,n$. Let $s_j = S_j(h,n) - S_{j-1}(h,n)$, where $S_0(h,n) \equiv 0$. By (2.1), s_j is the probability under the null hypothesis that a sequential permutation p-value equals $S_j(h,n)$. Let

$$\tilde{n}_{j:n} = \frac{\sum_{i=j}^{n} n_i}{\sum_{i=j}^{n} s_i}.$$

Let $m_0^{(0)} = \sum_{j=1}^{n} n_j$ and define

$$m_0^{(k)} = \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \tilde{n}_{j^{(k)}:n} \text{ for all } k \geq 1, \tag{2.4}$$

where

$$j^{(k)} \equiv \min\left\{j : n_j \leq s_j \cdot m_0^{(k-1)}\right\}. \tag{2.5}$$

Then, $m_0^{(k)}$ is the estimated number of true nulls at iteration k.

2.4.2 A Simple Example

Suppose we have a collection of $m = 100$ sequential permutation p-values produced using $h = 4$ and $n = 10$. Table 2.1 below displays the possible p-values and summarizes the frequency of the observed p-values and the expected frequency under the null hypothesis for the first two iterations of the algorithm. The expected null frequencies have been rounded to the nearest whole number for the first iteration and to the nearest tenth for the second iteration.

Iter.	p-value	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{4}{9}$	$\frac{4}{8}$	$\frac{4}{7}$	$\frac{4}{6}$	$\frac{4}{5}$	1
	Obs. Freq.	23	13	9	1	1	8	7	8	13	17
1	Exp. Freq.	10	10	10	10	4	6	7	10	13	20
2	Exp. Freq.	8.4	8.4	8.4	8.4	3.7	4.7	6.0	8.0	11.2	16.8

Table 2.1 Observed frequencies and (rounded) expected frequencies under the null hypothesis for two steps of the iterative algorithm of Mosig et al. (2001) extended to $m = 100$ sequential permutation p-values with $h = 4$ and $n = 10$.

For example, the expected null frequency at iteration 1 for a p-value of 4/9 is given by $m_0^{(0)} \cdot s_5 = 100 \cdot (4/9 - 4/10) = 4.\bar{4}$, which has been rounded to 4 in Table 2. To determine the expected null frequencies in the second line of the table, we first compute $m_0^{(1)}$ as follows. At iteration 1, the first observed frequency that does not exceed its expected null frequency occurs at the p-value $\frac{3}{10}$. Thus, we add the differences between the observed number of p-values and the expected number of null p-values for p-values $\frac{1}{10}$ and $\frac{2}{10}$ to obtain $(23 - 10) + (13 - 10) = 16$ as an estimate of $m - m_0$. Therefore, $m_0^{(1)}$ is set equal to $100 - 16 = 84$.

Given that the current estimate of m_0 after iteration 1 has been updated from 100 to 84, we then recalculate the number of each of the possible p-values that would be expected to originate from tests with a true null hypothesis. These expectations are given in the bottom line of the table by $84 \cdot s_j$ for $j = 1, \ldots, 10$. Next, we find the smallest p-value whose observed frequency is less than its updated expected null frequency. In this case, the relevant p-value is $\frac{4}{10}$. The excess for p-values less than $\frac{4}{10}$ is $23 - 8.4 = 14.6$, $13 - 8.4 = 4.6$, and $9 - 8.4 = 0.6$, respectively. The iterative algorithm next updates m_0 to $100 - (14.6 + 4.6 + 0.6) = 80.2$. Continuing to iterate, the estimated number of true nulls at iteration k decreases monotonically to approximately

78.57.

Using the formal definitions of the algorithm given by (3) and (4) to obtain $m_0^{(1)}$, we have $n_1 = 23 > 10 = \frac{1}{10}100 = s_1 \cdot m_0^{(0)}$, $n_2 = 13 > 10 = \frac{1}{10}100 = s_2 \cdot m_0^{(0)}$, and $n_3 = 9 \leq 10 = \frac{1}{10}100 = s_3 \cdot m_0^{(0)}$, which implies $j^{(1)} = 3$ and

$$\begin{aligned} m_0^{(1)} &= \left(\sum_{i=1}^{3-1} s_i\right) m_0^{(0)} + \left(1 - \sum_{i=1}^{3-1} s_i\right) \tilde{n}_{3:10} \\ &= \left(\frac{1}{10} + \frac{1}{10}\right) 100 + \left(1 - \left(\frac{1}{10} + \frac{1}{10}\right)\right) \frac{9+1+1+8+7+8+13+17}{1 - \left(\frac{1}{10} + \frac{1}{10}\right)} \\ &= 20 + 64 = 84. \end{aligned}$$

Continuing to iterate using the formal definitions yields $m_0^{(2)} = 80.2$, $m_0^{(3)} = 79.06$, $m_0^{(4)} = 78.718$, $m_0^{(5)} = 78.6154$, $m_0^{(6)} = 78.58462$, $m_0^{(7)} = 78.57539$, $m_0^{(8)} = 78.57262$, $m_0^{(9)} = 78.57178$, etc.

2.4.3 Convergence Extension

Nettleton et al. (2006) showed the existence of and characterized the limit of the iterative algorithm when the p-values have a continuous uniform$(0, 1)$ null distribution. We now state an analogous result for discrete p-values that satisfy (2.1). A proof is provided in the Appendix.

Convergence Result: Let $J = \min\{j : n_j \leq s_j \cdot \tilde{n}_{j:n}\}$. Then,

$$\widehat{m}_0 = \lim_{k\to\infty} m_0^{(k)} = \tilde{n}_{J:n} = \frac{\sum_{i=J}^{n} n_i}{\sum_{i=J}^{n} s_i}. \tag{2.6}$$

For the simple example in Subsection 4.2, note that $n_1 = 23 > 10 = \tilde{n}_{1:n} \cdot s_1$, $n_2 = 13 > 8.\bar{5} = \tilde{n}_{2:n} \cdot s_2$, $n_3 = 9 > 8 = \tilde{n}_{3:n} \cdot s_3$, and $n_4 = 1 \leq 55/7 = \tilde{n}_{4:n} \cdot s_4$. Thus $J = 4$ and

$$\widehat{m}_0 = \frac{\sum_{i=4}^{10} n_i}{\sum_{i=4}^{10} s_i} = \frac{1+1+8+7+8+13+17}{1 - 3/10} = 55/0.7 \approx 78.57143.$$

Henceforth, we propose to use $\widehat{m}_0 = \frac{\sum_{i=J}^{n} n_i}{\sum_{i=J}^{n} s_i}$ as the estimate of m_0 in (2.3) for FDR estimation. The next section applies this proposed procedure to a subset of the Acute Lymphoma Leukemia (ALL) dataset.

2.5 Acute Lymphoma Leukemia Data Analysis

In this section, we present the results of the proposed procedure applied to a subset of the B- and T-cell Acute Lymphocyctic Leukemia (ALL) dataset. This dataset can be accessed via the Bioconductor ALL package at www.bioconductor.org. Measures of messenger ribonucleic acid (mRNA) – commonly referred to as expression values – are available for 12,625 probesets (which we refer to here as genes) in 128 ALL patients. Of these 128 patients, we focus on the 21 males who have been classified as having a translocation between chromosomes 9 and 22 (BCR/ABL) and the 5 males who have a translocation between chromosomes 4 and 11 (ALL1/AF4). We treat these 21 males and 5 males as independent random samples of males with the BCR/ABL translocation or the ALL1/AF4 translocation, respectively. For each gene, we wish to test whether the population expression distributions corresponding to the two translocations are identical or whether the gene is differentially expressed across translocation type. The proposed procedure with $h = 10$ and $n = 1,000$ was employed to find genes that are differentially expressed. Given $n = 1,000$, the smallest possible p-value is $1/1,000 = 0.001$. Using this as the significance cutoff c, the analysis identifies 210 genes as differentially expressed between the two translocation groups. Of the 12,625 null hypotheses tested, the extension of Nettleton et al. (2006) estimated that approximately 12,005 of those hypotheses were true. Hence, the estimated FDR associated with a significance threshold of 0.001 is less than or equal to $\frac{0.001*12,005}{210} = 0.0572$. Using equations (2.3) and (2.6), the results for several significance thresholds are given below in Table 2.2.

The total number of test statistics computed using our procedure to produce Table 2.2 was 1,616,148. In contrast, non-Monte Carlo permutation tests would require the computation of $C(26, 5) = 65,780$ test statistics for each of 12,625 genes which requires computing $12,625 \cdot 65,870 = 831,608,750$ total test statistics. Alternatively, a random sample of, say, 1,000 permutations could be taken to reduce computation. Even then, the number of test statistics computed would be $12,625 \cdot 1,000 = 12,625,000$. In either case, for some of these 12,625 genes there will not be significant statistical evidence to declare differential expression, and this lack of evidence surely can be seen prior to computing the entire 65,780 element support of the

cutoff c	0.001	0.002	0.003	0.004	0.005
$R(c)$	210	304	383	440	503
$\widehat{\text{FDR}}(c)$	0.0572	0.0790	0.0940	0.1091	0.1193

Table 2.2 The number of rejected hypotheses ($R(c)$) and the estimated FDR ($\widehat{\text{FDR}}(c)$) for each of the five smallest significance thresholds resulting from analyzing a subset of the Acute Lymphoma Leukemia (ALL) dataset using sequential permutation p-values with $h = 10$ and $n = 1,000$. FDR was estimated using a variation of Benjamini and Hochberg (1995) defined by equations (2.3) and (2.6).

permutation distribution or 1,000 randomly sampled elements.

The amount of computation savings in terms of time in this case may not be as marked as the savings in terms of the number of test statistics computed. The two-sample t-statistic used in this example is relatively inexpensive to compute. Furthermore, it can be shown that the permutation test based on the two-sample t-statistic is identical to the permutation test that uses the sum of the observations in one group as the test statistic, further reducing the computational expense. However, in other cases, the cost of computing each test statistic can be non-trivial. One such example is interval mapping of quantitative trait loci (QTL). Lander and Botstein (1989) wrote the seminal paper on the topic, and Churchill and Doerge (1994) first recommended the use of permutation testing in QTL mapping. The goal in QTL mapping studies is to find the locations of genes that affect quantitative characteristics. Hundreds or thousands of genetic positions (loci) are tested for association with a quantitative trait. Typically, a likelihood ratio test statistic is computed at each locus. Because the likelihood under the alternative hypothesis involves a mixture distribution, the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is used to compute each test statistic. To conduct a permutation test, this analysis is repeated for thousands of data permutations, and the largest likelihood ratio statistic across loci is computed for each permutation. Hence, obtaining a single test statistic value in the permutation distribution requires hundreds or thousands of calls to the EM algorithm, and each use of the EM algorithm may involve several iterations.

In expression QTL (eQTL) mapping, the computational costs are multiplied by a factor of tens of thousands as traditional QTL mapping is carried out for each of tens of thousands of genes represented on a microarray (see, e.g., Jansen and Nap, 2001; Brem et al., 2002; Schadt et al., 2003). The expression level associated with each gene is treated as a quantitative trait. The goal of eQTL analysis is to identify the loci that affect the expression of each gene. The total number of times the EM algorithm might be called in an eQTL permutation analysis is on the order of 10^9 or 10^{10} if 1,000 statistics are sampled from each gene's permutation distribution. Employing sequential permutation p-values can greatly reduce the number of test statistics that must be computed and lead to a considerable computational savings, in terms of both time and number of test statistics computed.

2.6 Simulation Studies

Three simulation studies are presented in this section. In each study, data are simulated in a multiple testing scenario, and the multiple hypotheses are tested using the proposed sequential permutation p-values. Each study compares the average of the true false positive fraction $V/\max\{1, R\}$ to the average estimated FDR as defined by (2.3) using (2.6) to estimate m_0. All three simulation studies use randomly generated effect sizes. The first simulation study examines the case of randomly generated independent normally distributed data. The second and third simulation studies randomly sample data from an actual microarray experiment where, in the second simulation study, the dependence structure is preserved, while in the third simulation study, the dependence structure is not preserved.

2.6.1 Independent Normally Distributed Data

2.6.1.1 Setup

For each run of the simulation, each of m independent null hypotheses was tested for a difference in location between two treatment groups of *eight* observations each. For m_0 of the tests, the data for each treatment group were simulated from the same normal distribution, while for the other $m - m_0$ of the tests, the data for each treatment group were simulated from normal distributions differing only in location. Specifically, for each simulation run, we simulated $Z_{1jw} \sim \mathrm{N}(0,1)$, $Z_{2jw} \sim \mathrm{N}(\delta_j, 1)$ for $j = 1, \ldots, m$ and $w = 1, \ldots, 8$, and

$$\delta_j \sim \left\{ \begin{array}{ll} 0, & j = 1, \ldots, m_0 \\ \Gamma(\lambda, 1), & j = m_0 + 1, \ldots, m \end{array} \right\},$$

where $\Gamma(\lambda, 1)$ denotes a gamma distribution with mean and variance λ. All random variables were generated independently.

Then, for $j = 1, \ldots, m$, $H_{0j} : \delta_j = 0$ was tested using the proposed sequential permutation test based on a two-sample t-statistic with $h = 10$ and $n = 1,000$. This resulted in a collection of m sequential permutation p-values for each simulated dataset. For each simulation run, the estimated $\mathrm{FDR}(c)$ and $V(c)/\max\{1, R(c)\}$ were recorded for several values of c. The simulation was executed for six different combinations of m_0 (7500 or 9000) and λ (1, 2, or 3) to assess the effects of the proportion of null genes and the size of non-null effects on performance. The results for each parameter combination were based on $N = 1,000$ simulation runs with $m = 10,000$.

2.6.1.2 Results

Table 2.3 displays summary statistics for $\widehat{m}_0$, the number of rejections $R(c)$, estimated $\mathrm{FDR}(c)$ and $V(c)/\max\{1, R(c)\}$ for the smallest possible p-value of 0.001 and two other significance cutoffs.

Table 2.3 shows that the average estimated FDR compares very well to the average true false positive fraction, $V/\max\{1, R\}$. The average estimated FDR is slightly conservative, which is preferred over being liberal, and the bias is reduced as the power increases. This is presumably due to the diminishing bias in the estimates of m_0 (as the power increases) since $\widehat{m}_0$ is used to estimate FDR as defined by (3). The performance of the m_0 estimator seen here is consistent with the performance of the histogram-based m_0 estimator for the case of uniform$(0, 1)$ p-values as discussed by Nettleton et al. (2006). When effect sizes are small, the distribution of p-values from tests with true alternatives is very similar to the null p-value distribution. Thus, many p-values from non-null tests are likely to fall in the right tail of the

		$\widehat{m}_0$			$R(c)$		$\widehat{\mathrm{FDR}}\,(c)$		$V(c)/\max\{1, R(c)\}$	
m_0	λ	mean	se	c	mean	se	mean	se	mean	se
7500	1	9351	.76	.001	333.73	0.54	.0281	.0000	.0193	.0002
				.005	594.32	0.69	.0788	.0001	.0613	.0003
				.010	742.32	0.75	.1261	.0001	.1000	.0003
7500	2	8569	.83	.001	912.90	0.79	.0094	.0000	.0069	.0000
				.005	1332.04	0.81	.0322	.0000	.0271	.0001
				.010	1516.18	0.83	.0565	.0000	.0488	.0002
7500	3	8012	.70	.001	1503.12	0.79	.0053	.0000	.0042	.0000
				.005	1915.97	0.71	.0209	.0000	.0189	.0001
				.010	2068.40	0.69	.0387	.0000	.0356	.0001
9000	1	9742	.58	.001	138.79	0.34	.0706	.0002	.0543	.0006
				.005	267.23	0.47	.1829	.0003	.1639	.0007
				.010	356.14	0.56	.2739	.0004	.2480	.0007
9000	2	9430	.58	.001	369.73	0.50	.0256	.0000	.0207	.0002
				.005	562.50	0.54	.0841	.0000	.0775	.0004
				.010	664.04	0.57	.1421	.0001	.1329	.0004
9000	3	9207	.51	.001	606.50	0.49	.0152	.0000	.0125	.0001
				.005	795.25	0.48	.0580	.0000	.0541	.0002
				.010	885.72	0.49	.1040	.0000	.0990	.0003

Table 2.3 Means and standard errors of the means computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\widehat{m}_0$), the number of rejected null hypotheses ($R(c)$), estimated FDR ($\widehat{\mathrm{FDR}}\,(c)$), and false positive fraction ($V(c)/\max\{1, R(c)\}$). The number of tests was $m = 1,000$ for each simulation run. The number of true null hypotheses (m_0) and the mean effect size of differentially expressed genes (λ) varied across simulation settings. All data were independent and effect sizes were generated from a gamma distribution as described in Subsection 6.1.1.

p-value distribution, and estimates of m_0 are inflated as a result. Ruppert et al. (2007) discuss strategies for reducing this bias for the case of p-values from t-tests, but these strategies are not easy to adapt for our nonparametric setting.

2.6.2 Microarray-Derived Dependent Data

2.6.2.1 Setup

In many modern multiple testing problems, tests have an unknown and presumably complex dependence structure. Testing for differential expression in microarray analysis is a common example where dependence across genes within experimental units leads to dependence across tests. To evaluate the performance of the sequential permutation p-values when tests are dependent, we revisit the ALL dataset discussed in Section 5. Recall that expression values were recorded for 12,625 genes on two groups of 21 males and 5 males. In this section, we focus on the data from the 21 males with a translocation between chromosomes 9 and 22 to obtain a gene expression matrix A with 12,625 rows (one for each gene) and 21 columns (one for each subject). This data is inherently dependent across genes as the 12,625 gene expression values for each subject are correlated.

To be consistent across simulation schemes, $m = 10,000$ genes were randomly selected from all 12,625 genes and used as the entire dataset for every run of this simulation and the simulation to follow in Section 6.3. For each run of the simulation, 16 of the 21 males were randomly selected yielding a gene expression sub-matrix B of 10,000 rows and 16 columns. At the gene specific level, this subset can be thought of as a simple random sample from a finite homogeneous population. Next, if we further divide these 16 randomly selected males into two sub-groups of the first eight males and last eight males, respectively, we create two sub-groups for which the null hypothesis for each gene is true.

To generate data where some null hypotheses are true and some null hypotheses are false, we alter the gene expression sub-matrix B as follows. First, we randomly select $m - m_0$ genes from the subset of the $m = 10,000$ genes. Next, for each of those $m - m_0$ genes, we add to each of the eight males in the second sub-group a random effect generated from a $\Gamma(\lambda, 1)$ with mean λ multiplied by the gene-specific standard deviation of all 21 males. This multiplier will yield a noncentrality parameter similar to the previous simulation study. In this new 10,000 by 16 sub-matrix (denoted C), $m - m_0$ genes are differentially expressed across sub-groups (first eight columns vs. the last eight columns) and dependencies among genes within experimental units mimic those in the original data.

For each simulation run, a sub-matrix C was generated using the strategy described above. Genes were tested for differential expression using the sequential permutation p-values with $h = 10$ and $n = 1,000$. A total of $N = 1,000$ runs were simulated for the same six combinations of m_0 (7,500 or 9,000) and λ (1, 2, or 3) as the previous simulation study. The number of rejections, estimated FDR, and actual false positive fraction were calculated for significance thresholds of $c = 0.001,\ 0.005$, and 0.01. The estimated $\widehat{m}_0$ were also recorded.

2.6.2.2 Results

Just as in the independent data simulation study, Table 2.4 shows that the average estimated FDR compares very well to the average true false positive fraction $V/\max\{1, R\}$. Here, the conservative bias is larger than that seen in the first simulation study for both $\widehat{m}_0$ and $\widehat{\text{FDR}}$. Also, the standard error of $\widehat{m}_0$ is much larger in this simulation study compared to the first simulation study. This phenomena has been seen in other simulation studies with dependent data (e.g., Langaas et al., 2005; Nettleton et al., 2006). Of course, the marginal distribution of each gene was changed along with the dependence structure. Thus, we conducted one additional simulation study to determine if the differences between the first two simulations were due to dependence or the change in marginal distributions.

2.6.3 Microarray-Derived Independent Data

2.6.3.1 Setup

The third simulation was conducted as in the dependent data simulation with one key difference. Each row of sub-matrix B was formed by independently randomly sampling 16 of the 21 expression values from the corresponding row of matrix A. Because selections were made

		$\widehat{m}_0$			$R(c)$		$\widehat{\text{FDR}}\,(c)$		$V(c)/\max\{1, R(c)\}$	
m_0	λ	mean	se	c	mean	se	mean	se	mean	se
7500	1	9374	4.61	.001	324.77	1.87	.0298	.0002	.0172	.0008
				.005	572.52	3.28	.0842	.0004	.0514	.0018
				.010	719.76	4.56	.1345	.0007	.0840	.0024
7500	2	8596	4.27	.001	891.59	3.08	.0098	.0000	.0067	.0003
				.005	1301.18	3.53	.0333	.0001	.0244	.0009
				.010	1489.96	4.23	.0582	.0002	.0440	.0014
7500	3	8026	3.81	.001	1482.70	3.07	.0054	.0000	.0047	.0003
				.005	1895.99	3.04	.0212	.0000	.0180	.0008
				.010	2054.28	3.77	.0392	.0000	.0334	.0012
9000	1	9752	4.46	.001	136.15	1.13	.0751	.0005	.0501	.0022
				.005	258.32	2.85	.2016	.0014	.1310	.0037
				.010	345.78	4.42	.3074	.0023	.2007	.0044
9000	2	9442	3.90	.001	364.01	1.42	.0263	.0001	.0198	.0001
				.005	548.79	2.54	.0875	.0003	.0661	.0022
				.010	652.65	3.85	.1484	.0007	.1133	.0031
9000	3	9217	4.00	.001	598.81	1.48	.0155	.0000	.0126	.0008
				.005	785.41	2.55	.0592	.0002	.0465	.0019
				.010	875.92	3.96	.1068	.0004	.0832	.0027

Table 2.4 Means and standard errors of the means computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\widehat{m}_0$), the number of rejected null hypotheses ($R(c)$), estimated FDR ($\widehat{\text{FDR}}\,(c)$), and false positive fraction ($V(c)/\max\{1, R(c)\}$). The number of tests was $m = 10,000$ for each simulation run. The number of true null hypotheses (m_0) and the effect sizes for differentially expressed genes (λ) varied across simulation settings. Dependence across tests was generated from the dependence in actual microarray data as described in Subsection 6.2.1.

independently from row to row, dependence among genes was removed from the dataset. The next subsection summarizes the results of this simulation study.

2.6.3.2 Results

The results displayed in Table 2.5 address questions raised by contrasting the results of our first two simulation studies. The estimates and standard errors of all quantities are strikingly similar to the first simulation study. This suggests that, because the noncentrality parameter distribution is nearly the same in all simulation studies, dependent data can result in much more variable estimates of m_0 and an increase in the conservative bias of the FDR estimates.

		$\widehat{m}_0$			$R(c)$		$\widehat{\text{FDR}}\,(c)$		$V(c)/\max\{1, R(c)\}$	
m_0	λ	mean	se	c	mean	se	mean	se	mean	se
7500	1	9352	.77	.001	351.14	0.56	.0267	.0000	.0226	.0003
				.005	597.70	0.69	.0783	.0001	.0642	.0003
				.010	741.51	0.76	.1261	.0001	.1023	.0003
7500	2	8572	.84	.001	940.20	0.77	.0091	.0000	.0084	.0000
				.005	1335.30	0.82	.0321	.0000	.0290	.0001
				.010	1514.01	0.83	.0566	.0000	.0504	.0002
7500	3	8007	.69	.001	1537.50	0.75	.0052	.0000	.0052	.0000
				.005	1924.44	0.70	.0208	.0000	.0200	.0001
				.010	2072.63	0.68	.0386	.0000	.0366	.0001
9000	1	9741	.57	.001	146.56	0.36	.0669	.0002	.0634	.0006
				.005	269.47	0.46	.1813	.0003	.1695	.0007
				.010	357.15	0.53	.2724	.0004	.2530	.0007
9000	2	9428	.58	.001	381.65	0.48	.0247	.0000	.0245	.0002
				.005	564.49	0.53	.0836	.0000	.0817	.0003
				.010	665.92	0.57	.1416	.0001	.1368	.0004
9000	3	9204	.54	.001	620.84	0.52	.0148	.0000	.0151	.0002
				.005	799.41	0.47	.0576	.0000	.0572	.0003
				.010	889.05	0.51	.1035	.0000	.1021	.0003

Table 2.5 Means and standard errors of the means computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\widehat{m}_0$), the number of rejected null hypotheses ($R(c)$), estimated FDR ($\widehat{\text{FDR}}\,(c)$), and false positive fraction ($V(c)/\max\{1, R(c)\}$). The number of tests was $m = 10,000$ for each simulation run. The number of true null hypotheses (m_0) and the effect sizes for differentially expressed genes (λ) varied across simulation settings. Data were generated independently from actual microarray data as described in Subsection 6.3.1.

2.7 Conclusion

This paper proposes using the sequential analysis of Besag and Clifford (1991) when testing multiple hypotheses via permutation testing. The procedure is particularly applicable to Monte Carlo approximations when drawing a value from the permutation distribution is computationally expensive. When using the proposed procedure, the number of required draws from the permutation distribution is greatly reduced while sustaining little loss of information. This paper also describes how to use these discrete non-uniformly distributed null p-values to estimate m_0 and FDR. Simulations have shown that the proposed procedure provides estimates of FDR that should prove quite useful in practice.

2.8 References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

[2] Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

[3] Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*, **Vol. 78, No. 2**, 301-304.

[4] Brem. R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science*, **296**, 752-755.

[5] Churchill, G.A., and Doerge, R.W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, **138**, 963-971.

[6] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[7] Jansen, R.C. and Nap, J.P. (2001). Genetical Genomics: The Added Value From Segregation. *Trends in Genetics*, **17**, 388-391.

[8] Lander, E.S., and Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, **121**, 185-199.

[9] Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data. *Journal of the Royal Statistical Society, Series B*, **67**, 555–572.

[10] Mosig, M.O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Per-

centage in Israeli-Holstein Cattle by Means of Selective Milk DNA Pooling in a Daughter Design Using an Adjusted False Discovery Rate Criterion. *Genetics*, **151**, 1683-1698.

[11] Nettleton, D., Hwang, J.T.G., Caldo, R.A., and Wise, R.P. (2006). Estimating the Number of True Null Hypotheses From a Histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, **Vol. 11, No. 3**, 337-356.

[12] Ruppert, D., Nettleton, D., Hwang, J.T.G. (2007). Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics*, **63**, 483–495.

[13] Schadt, E. E., Monks, S.A., and Drake, T.A. (2003). Genetics of Gene Expression Survived in Maize, Mouse, and Human. *Nature*, **422**, 297-302.

[14] Storey, J.D. (2000). False Discovery Rates: Theory and Applications to DNA Microarrays. Unpublished Ph.D. thesis, Department of Statistics, Stanford University.

[15] Storey, J.D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*, Series B, **64**, 479-498.

[16] Storey, J.D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics*, **Vol. 31, No. 6**, 2013-2035.

[17] Storey, J.D., Taylor, J.E., Siegmund, D. (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society*, **Vol. 66, No. 1**, 187-205.

[18] Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T., and Lander, E.S. (1992). Identification of Genetic Factors Contributing to Heterosis in a Hybrid from Two Elite Maize Inbred Lines using Molecular Markers. *Genetics*, **132**, 823-839.

[19] Weller, J.I., Soller, M., and Brody, T. (1988). Linkage Analysis of Quantitative Traits in an Interspecific Cross of Tomato by means of Genetic Markers. *Genetics*, **118**, 329-339.

Appendix

This appendix contains the proof of the Convergence Result given in Section 4.3 . First, four facts used to facilitate the proof are given. Then, the proof is shown followed by the proofs of the four facts.

FACTS

A) $\tilde{n}_{1:n} > \tilde{n}_{2:n} > \ldots > \tilde{n}_{J:n}$.

B) If $j^{(k)} < J$, then (i) $j^{(k)} \leq j^{(k+1)}$, (ii) $j^{(k+1)} \leq J$, and (iii) $j^{(k)} < j^{(k^*)}$ for some $k^* > k$.

C) If there exists k^* such that $j^{(k^*)} = J$, then $j^{(k)} = J$ for all $k \geq k^*$.

D) Suppose $\{a_k\}_{k\geq 0}$ is an infinite sequence of real numbers. If there exists $\lambda \in [0, 1)$, an integer k^*, and a real number a such that $a_k = \lambda a_{k-1} + (1 - \lambda)a$ whenever $k \geq k^*$, then $\lim_{k\to\infty} a_k = a$.

Proof of Convergence Result : We first show that $j^{(k)}$ converges to J in a finite number of iterations.

CASE I: ($J = 1$) By the definition of J, $n_1 \leq s_1 \cdot \tilde{n}_{1:n} = s_1 \cdot m_0^{(0)}$. Since $j^{(k)} \equiv \min\left\{j : n_j \leq s_j \cdot m_0^{(k-1)}\right\}$, then $j^{(1)} = \min\left\{j : n_j \leq s_j \cdot m_0^{(0)}\right\}$ and so $j^{(1)} = 1 = J$. Fact C then implies $j^{(k)} = J = 1$ for all $k \geq 1$, i.e. $\left\{j^{(k)}\right\}$ converges to 1 at iteration 1.

CASE II: ($J > 1$) The definition of J and fact A imply $\frac{n_J}{s_J} \leq \frac{\sum_{i=J}^{n} n_i}{\sum_{i=J}^{n} s_i} < \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} s_i} = m_0^{(0)}$, i.e. $n_J \leq s_J \cdot m_0^{(0)}$. Then, since $j^{(1)} = \min\left\{j : n_j \leq s_j \cdot m_0^{(0)}\right\}$, $j^{(1)} \leq J$.

SUBCASE I: ($j^{(1)} = J$) By fact C, if $j^{(1)} = J$, then $j^{(k)} = J$ for all $k \geq 1$, i.e. $\left\{j^{(k)}\right\}$ converges to J at iteration 1.

SUBCASE II: ($j^{(1)} < J$) Fact B implies i) $j^{(1)} \leq j^{(2)}$, ii) $j^{(2)} \leq J$, and iii) $j^{(1)} < j^{(k^*)}$ for some $k^* > 1$. Therefore, $j^{(k^*)}$ has to equal J for some $k^* > 1$. To see this, note that we know $j^{(2)} \leq J$ from fact B ii). If $j^{(2)} = J$, use argument of SUBCASE I. If $j^{(2)} < J$, then use fact B again, etc. Thus, for some $k^* > 1$, $j^{(k^*)}$ must equal J. Then, by fact C, $j^{(k)} = J$ for all $k \geq k^*$, i.e. $\left\{j^{(k)}\right\}$ converges to J at iteration k^*.

Whatever the case, we have shown there exists a k^* such that $j^{(k)} = J$ for all $k \geq k^*$. Therefore, for all $k \geq k^*$,

$$m_0^{(k)} = \left(\textstyle\sum_{i=1}^{J-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \textstyle\sum_{i=1}^{J-1} s_i\right) \tilde{n}_{J:n}.$$

Last, by fact D, $\lim_{k\to\infty} m_0^{(k)} = \tilde{n}_{J:n} = \frac{\sum_{i=J}^{n} n_i}{\sum_{i=J}^{n} s_i}$.

□

Proof of Fact A : Since $J \equiv \min\{j : n_j \leq s_j \cdot \tilde{n}_{j:n}\}$, then for all $j < J$, $n_j > s_j \cdot \tilde{n}_{j:n}$. Therefore, for all $j < J$, we have

$$\begin{aligned}
n_j &> s_j \cdot \tilde{n}_{j:n} \\
&= s_j \cdot \frac{\sum_{i=j}^{n} n_i}{\sum_{i=j}^{n} s_i} \\
&= \frac{s_j}{\sum_{i=j}^{n} s_i} n_j + \frac{s_j}{\sum_{i=j}^{n} s_i} \sum_{i=j+1}^{n} n_i \\
&= \frac{s_j}{\sum_{i=j}^{n} s_i} n_j + \left(\frac{\sum_{i=j+1}^{n} s_i}{\sum_{i=j+1}^{n} s_i}\right) \frac{s_j}{\sum_{i=j}^{n} s_i} \sum_{i=j+1}^{n} n_i \\
&= \frac{s_j}{\sum_{i=j}^{n} s_i} n_j + s_j \left(\frac{\sum_{i=j+1}^{n} s_i}{\sum_{i=j}^{n} s_i}\right) \left(\frac{\sum_{i=j+1}^{n} n_i}{\sum_{i=j+1}^{n} s_i}\right) \\
&= \frac{s_j}{\sum_{i=j}^{n} s_i} n_j + s_j \left(1 - \frac{s_j}{\sum_{i=j}^{n} s_i}\right) \left(\frac{\sum_{i=j+1}^{n} n_i}{\sum_{i=j+1}^{n} s_i}\right) \\
&= \frac{s_j}{\sum_{i=j}^{n} s_i} n_j + s_j \left(1 - \frac{s_j}{\sum_{i=j}^{n} s_i}\right) \tilde{n}_{j+1:n},
\end{aligned}$$

and so we have $n_j > s_j \cdot \tilde{n}_{j+1:n}$. Next,

$$\begin{aligned}
\tilde{n}_{j:n} &= \frac{1}{\sum_{i=j}^{n} s_i} n_j + \left(1 - \frac{s_j}{\sum_{i=j}^{n} s_i}\right) \tilde{n}_{j+1:n} \\
&> \frac{s_j}{\sum_{i=j}^{n} s_i} \tilde{n}_{j+1:n} + \left(1 - \frac{s_j}{\sum_{i=j}^{n} s_i}\right) \tilde{n}_{j+1:n} \\
&= \tilde{n}_{j+1:n},
\end{aligned}$$

which implies $\tilde{n}_{j:n} > \tilde{n}_{j+1:n}$. Thus, $\tilde{n}_{1:n} > \tilde{n}_{2:n} > \ldots > \tilde{n}_{J:n}$.

□

Proof of Fact B(i) : The definition of J implies $n_j > s_j \cdot \tilde{n}_{j:n}$ for all $j < J$. Since $j^{(k)} < J$, then $n_{j^{(k)}} > s_{j^{(k)}} \cdot \tilde{n}_{j^{(k)}:n}$. Also, the definition of $j^{(k)}$ implies $n_{j^{(k)}} \leq s_{j^{(k)}} \cdot m_0^{(k-1)}$. Together, $s_{j^{(k)}} \cdot \tilde{n}_{j^{(k)}:n} < n_{j^{(k)}} \leq s_{j^{(k)}} \cdot m_0^{(k-1)}$. Then,

$$\begin{aligned} m_0^{(k)} &= \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \tilde{n}_{j^{(k)}:n} \\ &< \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(k-1)} \\ &= m_0^{(k-1)}. \end{aligned}$$

Hence, $s_{j^{(k)}} \cdot m_0^{(k)} < s_{j^{(k)}} \cdot m_0^{(k-1)}$. This implies, if $n_j \leq s_j \cdot m_0^{(k)}$, then $n_j < s_j \cdot m_0^{(k-1)}$. Therefore, $\left\{j : n_j \leq s_j \cdot m_0^{(k)}\right\} \subseteq \left\{j : n_j < s_j \cdot m_0^{(k-1)}\right\} \subseteq \left\{j : n_j \leq s_j \cdot m_0^{(k-1)}\right\}$, which implies $j^{(k)} = \min\left\{j : n_j \leq s_j \cdot m_0^{(k-1)}\right\} \leq \min\left\{j : n_j \leq s_j \cdot m_0^{(k)}\right\} = j^{(k+1)}$.

□

Proof of Fact B(ii) : Again, the definition of $j^{(k)}$ implies $n_{j^{(k)}} \leq s_{j^{(k)}} \cdot m_0^{(k-1)}$ and the definition of J implies both $n_J \leq s_J \cdot \tilde{n}_{J:n}$ and $n_j > s_j \cdot \tilde{n}_{j:n}$ for all $j < J$. Using these and fact A, we have for $j^{(k)} < J$, $\frac{n_J}{s_J} \leq \tilde{n}_{J:n} < \tilde{n}_{j^{(k)}:n} < \frac{n_{j^{(k)}}}{s_{j^{(k)}}} \leq \frac{s_{j^{(k)}} \cdot m_0^{(k-1)}}{s_{j^{(k)}}} = m_0^{(k-1)}$, i.e. $\frac{n_J}{s_J} < m_0^{(k-1)}$. Next,

$$\begin{aligned} m_0^{(k)} &= \left(\sum_{i=1}^{J-1} s_i\right) m_0^{(k-1)} + \left(1 - \sum_{i=1}^{J-1} s_i\right) \tilde{n}_{J:n} \\ &> \left(\sum_{i=1}^{J-1} s_i\right) \frac{n_J}{s_J} + \left(1 - \sum_{i=1}^{J-1} s_i\right) \frac{n_J}{s_J} \\ &= \frac{n_J}{s_J}, \end{aligned}$$

i.e. $n_J < s_J \cdot m_0^{(k)} = s_J \cdot m_0^{(k+1)-1}$. Hence, $j^{(k+1)} \leq J$.

□

Proof of Fact B(iii) : Suppose $j^{(k)} < J$ and $j^{(l)} \leq j^{(k)}$ for all $l > k$. Fact B i) states that if $j^{(k)} < J$, then $j^{(k)} \leq j^{(k+1)}$. Therefore, $j^{(l)} = j^{(k)}$ for all $l > k$. Then, for all $l > k$,

$$\begin{aligned} m_0^{(l)} &= \left(\sum_{i=1}^{j^{(l)}-1} s_i\right) \cdot m_0^{(l-1)} + \left(1 - \sum_{i=1}^{j^{(l)}-1} s_i\right) \tilde{n}_{j^{(l)}:n} \\ &= \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(l-1)} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \tilde{n}_{j^{(k)}:n}. \end{aligned}$$

Then, $\lim_{l\to\infty} m_0^{(l)} = \tilde{n}_{j^{(k)}:n}$ by fact D. But since $j^{(k)} < J$, that implies, by the definition of J, that $n_{j^{(k)}} > s_{j^{(k)}} \cdot \tilde{n}_{j^{(k)}:n}$. Together, $\lim_{l\to\infty} m_0^{(l)} = \tilde{n}_{j^{(k)}:n} < \frac{n_{j^{(k)}}}{s_{j^{(k)}}}$. By the definition of limit, there exists a $k^* > k$ such that for all $l \geq k^*$, $n_{j^{(k)}} > s_{j^{(l-1)}} \cdot m_0^{(l-1)} = s_{j^{(l)}} \cdot m_0^{(l-1)}$. Since $j^{(l)} \equiv \min\left\{j : n_j \leq s_j \cdot m_0^{(l-1)}\right\}$, then $n_{j^{(l)}} \leq s_{j^{(l)}} \cdot m_0^{(l-1)}$. But, $j^{(k)} = j^{(l)}$ for all $l \geq k$ and so $n_{j^{(k)}} = n_{j^{(l)}} \leq s_{j^{(l)}} \cdot m_0^{(l-1)}$. Thus, a contradiction, and so we have, for $j^{(k)} < J, j^{(k)} < j^{(k^*)}$ for some $k^* > k$.

□

<u>*Proof of Fact C*</u> : Suppose $j^{(k)} = J$. We want to show that $j^{(k+1)} = J$. First, since $j^{(k)} = J$, we have by the definitions of $j^{(k)}$ and J, $n_J \leq s_J \cdot \tilde{n}_{J:n}$ and $n_J \leq s_J \cdot m_0^{(k-1)}$. Then,

$$\begin{aligned} m_0^{(k)} &= \left(\sum_{i=1}^{J-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \sum_{i=1}^{J-1} s_i\right) \tilde{n}_{J:n} \\ &\geq \left(\sum_{i=1}^{J-1} s_i\right) \frac{n_J}{s_J} + \left(1 - \sum_{i=1}^{J-1} s_i\right) \frac{n_J}{s_J} \\ &= \frac{n_J}{s_J}. \end{aligned}$$

That implies $n_J \leq s_J \cdot m_0^{(k)} = s_J \cdot m_0^{(k+1)-1}$ and so we have, by the definition of $j^{(k+1)}$, $j^{(k+1)} \leq J$. Next, again using the definitions of $j^{(k)}$ and J and fact A, we have for all $j < J, n_j > s_j \cdot m_0^{(k-1)}$ and $\frac{n_j}{s_j} > \tilde{n}_{j:n} > \tilde{n}_{J:n} = \tilde{n}_{j^{(k)}:n}$. Then,

$$\begin{aligned} m_0^{(k)} &= \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot m_0^{(k-1)} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \tilde{n}_{j^{(k)}:n} \\ &< \left(\sum_{i=1}^{j^{(k)}-1} s_i\right) \cdot \frac{n_j}{s_j} + \left(1 - \sum_{i=1}^{j^{(k)}-1} s_i\right) \frac{n_j}{s_j} \\ &= \frac{n_j}{s_j} \end{aligned}$$

That implies $n_j > s_j \cdot m_0^{(k)} = s_j \cdot m_0^{(k+1)-1}$ for all $j < J$, which, by the definition of $j^{(k+1)}$, implies $j^{(k+1)} \geq J$. Since $j^{(k+1)} \leq J$ and $j^{(k+1)} \geq J$, we have $j^{(k+1)} = J$.

□

Proof of Fact D : Let $b_n = a_{k^*+n-1}$ for all $n \geq 0$. Then $b_1 = \lambda b_0 + (1-\lambda)a$, and an induction argument shows that $b_n = \lambda^n b_0 + a(1-\lambda)\sum_{i=0}^{n-1} \lambda^i$ for all $n \geq 1$. Now $\lambda \in [0,1)$ implies that

$$\lim_{n\to\infty} \lambda^n = 0 \text{ and } \sum_{i=0}^{\infty} \lambda^i = \tfrac{1}{1-\lambda}.$$

Hence, $\lim_{n\to\infty} b_n = a$, and $\lim_{k\to\infty} a_k = a$ since $\{a_k\}_{k\geq k^*} = \{b_n\}_{n\geq 1}$

□

CHAPTER 3. Estimation of False Discovery Rate using *P*-Values with Different Discrete Null Distributions

Abstract

The false discovery rate (FDR) is a multiple testing error rate which describes the expected proportion of type I errors among the total number of rejected hypotheses. Benjamini and Hochberg introduced this quantity and provided an estimator that is conservative when the number of true null hypotheses, m_0, is smaller than the number of tests, m. Replacing m with m_0 in Benjamini and Hochberg's procedure reduces the conservative bias but requires estimation as m_0 is unknown. Methods exist to estimate m_0 when each null p-value is distributed as a continuous uniform (0,1) random variable. This paper discusses how to estimate m_0 and therefore FDR when the m_0 null p-values are from a mixture of different discrete distributions. The method will be demonstrated through a permutation analysis of data with many ties and by conducting multiple Fisher's exact tests.

Key Words: Permutation testing; False discovery rate; Fisher's exact test; Gene set enrichment; Multiple testing

3.1 Introduction

Multiple testing problems are common in many modern applications due to technological advances that permit the simultaneous measurement of many response variables. In some cases, the number of hypotheses tested can be in the thousands so that conducting individual tests at traditional type I error rates may lead to many type I errors. One multiple testing adjustment is the family wise error rate (FWER). Controlling FWER at level α amounts to choosing a p-value cutoff such that the probability of at least one false positive, i.e. rejecting a hypothesis which is true, is less than or equal to α. Because researchers are often willing to accept some type I errors when conducting many tests as long as the number of type I errors is a small proportion of the total number of null hypotheses rejected, focus has shifted away from FWER control to the pioneering work on the false discovery rate (FDR) of Benjamini and Hochberg (1995).

FDR is formally defined as $\mathrm{E}(V/\max\{1, R\})$, where V is the number of type I errors and R is the number of rejected hypotheses among m hypotheses tested. Benjamini and Hochberg (1995) described a procedure to control FDR at a prespecified level α, which is conservative when the number of true null hypotheses, m_0, is smaller than the number of tests, m. Ideally, the number of true null hypotheses, m_0, would be used in place of m in their procedure to reduce the conservative bias. However, m_0 is unknown, and thus estimation of m_0 is of interest. Many methods exist to estimate m_0 including Benjamini and Hochberg (2000), Storey (2000, 2002), Storey and Tibshirani (2003), Langaas et al. (2005), Nettleton et al. (2006) among others. Most of the existing methods assume the p-values corresponding to true null hypotheses have a continuous uniform distribution on the interval (0,1). This paper describes how to estimate m_0, and therefore FDR, when the null p-values are from a mixture of discrete non-uniform distributions.

Section 2 describes example applications that can give rise to p-values that have different discrete non-uniform null distributions. In Section 3, the proposed procedure is introduced along with some review of existing techniques. The proposed procedure is applied to a proteomics data set in Section 4. A data-based simulation study is presented in Section 5 which compares the average estimated FDR to the average false positive fraction ($V/\max\{1,R\}$) and also evaluates how well the proposed procedure estimates m_0.

3.2 *P*-values with Different Discrete Supports

This section describes multiple testing scenarios where the analysis yields a collection of p-values from a mixture of different discrete null distributions, each with support elements that are not equally likely. The examples can be viewed as special cases of permutation testing using data with many tied observations.

3.2.1 Proteomics Data Analysis

Proteomics technologies allow researchers to simultaneously measure the relative amounts of hundreds of proteins in each of several biological samples. Often, the goal of subsequent data analysis is to identify which proteins differ in relative abundance across samples of different types.

Multidimensional Protein Identification Technology (MudPIT, Whitelegge 2002) is one technique used in proteomics. MudPIT generates peptide (small pieces of a protein) counts for each of hundreds of proteins in a biological sample. For a given protein and a given sample, the sum of counts for peptides matching the protein divided by the sum of all peptide counts provides a measure that is correlated with the abundance of the protein in the sample. A matrix of such proportions can be computed with one row for each protein and one column for each sample. This data matrix may contain many zeros because no peptides matching a protein may be found for some protein/sample combinations. In Section 4, we will analyze such a dataset involving over 1,000 proteins and 15 samples. The 15 samples were collected in three blocks containing five samples each. One sample from each of five soybean genotypes was processed in each block. The goal is to identify proteins whose abundance differs significantly across genotypes. A conceptual presentation of the data for one protein is provided in Table 3.1.

The example depicted in Table 3.1 is quite extreme as only two of 15 observations are

	Genotype				
Repetition	G_1	G_2	G_3	G_4	G_5
1	0	0	0	0	0
2	0	0	0	0	y_1
3	0	0	0	0	y_2

Table 3.1 Example data from a block design with many zero observations. The data points y_1 and y_2 represent arbitrary positive distinct values.

non-zero. Unfortunately, this situation is not uncommon in our dataset; more than 10% of the proteins in the dataset have only two nonzero data points. Because of the challenges of parametrically modeling data with many zeros, we consider conducting a permutation test for each protein, using the traditional ANOVA treatment sum of squares as our test statistic. It is straightforward to show that the permutation test based on this statistic is identical to the permutation test based on the ANOVA F-statistic for testing treatment main effects in a randomized complete block design.

Under the null hypothesis of identical distributions across genotypes and conditioning on the observed data values, each permutation that involves permuting observations within blocks is equally likely. There are $(5!)^3$ such permutations, but these permutations give rise to only two distinct values of the test statistic for the example in Table 3.1 because of the many tied observations. In particular, it is easy to see that the value of the test statistic will be maximized over permutations when the nonzero observations y_1 and y_2 are assigned to the same genotype. The number of within-block permutations that yield this maximum value is $5! \cdot 5 \cdot (4!)^2$, and thus, the permutation distribution assigns probability $5! \cdot 5 \cdot (4!)^2/(5!)^3 = 1/5$ to the maximum value of the test statistic. Likewise, the test statistic is minimized when y_1 and y_2 are assigned to different genotypes, which occurs with probability $5! \cdot 5 \cdot 4 \cdot (4!)^2/(5!)^3 = 4/5$ based on the permutation distribution. Because the permutation test rejects for large values of the test statistic, the only possible permutation p-values are $1/5$ and 1, and these values occur with probability $1/5$ and $4/5$, respectively, under the null hypothesis of identical distributions across genotypes. Thus, in this case, the null distribution of the permutation p-value has only two values in its support, and these values are not equally likely.

Now consider a protein with one nonzero value in each block. Suppose the nonzero values are $y_1 < y_2 < y_3$ and that $y_2 \neq (y_1 + y_3)/2$. In this case, the support of the permutation distribution has five distinct values ($F_1 < F_2 < F_3 < F_4 < F_5$) as discussed below.

[1] The test statistic takes value F_1 when y_1, y_2, and y_3 are assigned to three different genotypes, which occurs with probability $5 \cdot 4 \cdot 3 \cdot (4!)^3/(5!)^3 = 12/25$.

[2] The test statistic takes value F_2 when y_1 and y_3 are assigned to one genotype and y_2 is assigned to another, which occurs with probability $5 \cdot 4 \cdot (4!)^3/(5!)^3 = 4/25$.

[3] The test statistic takes value F_3 when y_2 and the observation that it is farthest from are assigned to one genotype and the remaining observation is assigned to another, which occurs with probability $5 \cdot 4 \cdot (4!)^3/(5!)^3 = 4/25$.

[4] The test statistic takes value F_4 when y_2 and the observation that it is closest to are assigned to one genotype and the remaining observation is assigned to another, which occurs with probability $5 \cdot 4 \cdot (4!)^3/(5!)^3 = 4/25$.

[5] The test statistic takes value F_5 when y_1, y_2, and y_3 are assigned to the same genotype, which occurs with probability $5 \cdot (4!)^3/(5!)^3 = 1/25$.

Thus, the potential permutation p-values are $1/25$, $1/25 + 4/25 = 1/5$, $1/25 + 4/25 + 4/25 = 9/25$, $1/25 + 4/25 + 4/25 + 4/25 = 13/25$, and 1; with probabilities 1/25, 4/25, 4/25, 4/25, and 12/25, respectively, under the null hypothesis. This null p-value distribution shares two support elements with the null p-value distribution derived for the protein characterized in Table 3.1, but the probabilities assigned to these common support elements differ between the proteins. Thus, a single p-value should not necessarily be interpreted the same way across different proteins. In Section 4, we use such permutation p-values to identify proteins differentially expressed (DE) across genotypes.

3.2.2 Testing for Differential Expression using Next Generation Sequencing of RNA Samples

Microarray experiments have become a typical way to identify genes whose expression is different between conditions. The results of microarray analysis provide a list of genes whose function may be related to the treatment conditions. Advances in technology provide a more accurate measure of the expression of genes. Next generation sequencing is the newest technology for sequencing genomes of organisms. Ribonucleic acid (RNA) is extracted from a sample of interest, fragmented, converted to complementary DNA (cDNA), and sequenced. Once sequenced, each fragmented piece becomes a string of base pairs and compared to a database of base pair sequences (genes) that are a subset of the entire genome for the organism of interest. If a fragmented sequenced piece matches to the sequence for gene A, say, the piece is counted as a gene A 'read.' RNA samples from two treatments, conditions, populations, genotypes etc., can be separately sequenced and compared to identify genes with relatively more reads in one sample that the other. The dataset resulting from these next generation sequencing applications is a breakdown of the number of reads from an RNA sample that matched to each of thousands of genes in each of two or more samples. In the case with just one biological sample for each of two sample types, the entire dataset can be viewed as a collection of many 2-by-2 tables, one for each gene. Each table breaks down the number of reads that matched a particular gene and the number of reads that matched all other genes for each of the two samples. A hypothetical example of one gene in a comparison of samples of two genotypes is given in Table 3.2.

	Genotype 1	Genotype 2	Total
Gene A Reads	$n_{11} = 1$	$n_{12} = 3$	$n_{1+} = 4$
Reads Matching Other Genes	$n_{21} = 3453080$	$n_{22} = 3977732$	$n_{2+} = 7431812$
Total	$n_{+1} = 3453081$	$n_{+2} = 3977735$	$n_{..} = 7431816$

Table 3.2 Example table dreived from next generation sequencing technology for comparing the relative abundance of gene A RNA transcripts across samples from two different genotypes.

Note that there is only one biological replication for each genotype, and therefore, differences between these two samples cannot be generalized to the genotype populations. Ideally, multiple independent biological replicates of each genotype would be separately sequenced. However, due to the high cost of sequencing, researchers often begin with a pilot study that involves measurement of only two samples (with perhaps each sample consisting of a pool of RNA from independent biological replications) before investing in replicated studies necessary for distinguishing treatment effects or population differences from sample to sample variation.

We assume that the observed number of reads for each gene in a sample is a draw from a sample-specific multinomial distribution with cell probabilities given by the proportional abundance of each gene-specific RNA sample. We wish to test, separately for each gene, the null hypothessis of equal cell probabilities across samples using the two-by-two table from a given gene. This null hypothesis can be tested with a Pearson X^2 test when the estimated cell frequencies are large, e.g. greater than 5 or 10. When this is not case, as it is with the data in Table 5, the chi-square approximation to the distribution of the Pearson X^2 test statistic is not reliable. An alternative test is Fisher's exact test which relies on exact small-sample distributions. Under the null hypothesis, the probability of observing a specific arrangement of a two-by-two table given the marginal counts, can be modeled using the hypergeometric distribution. Conditioning on n_{1+}, n_{+1}, and n, extreme values of n_{11} will provide evidence against the null hypothesis. Define $n_{-} = \max\{0, n_{1+} + n_{+1} - n\}$, $n_{+} = \min\{n_{1+}, n_{+1}\}$, and $S = \{n_{-}, \dots, n_{+}\}$. Then, Fisher's exact test defines the two-sided p-value as $\sum_{t \in S} \mathrm{I}\left[p(t) \leq p(n_{11})\right] \cdot p(t)$, where $p(t)$ is the hypergeometric null probability of observing $n_{11} = t$.

For the data in Table 3.2, $S = \{0, 1, 2, 3, 4\}$ and the corresponding hypergeometric probabilities are $\left\{\binom{3453081}{i}\binom{3977735}{4-i}/\binom{7431816}{4} : i \in S\right\} = \{0.285, 0.371, 0.215, 0.047\}$. Since the observed value of n_{11} is 1 and $\mathrm{P}(n_{11} = 1) = 0.285$, then the two sided p-value equals $0.082 + 0.285 + 0.215 + 0.047 = 0.629$. The potential p-values for a table with the marginal counts equal to those in Table 3.2 are 0.047, 0.129, 0.344, 0.629, and 1 with probabilities 0.047, 0.082, 0.215, 0.285, and 0.371, respectively, under the null hypothesis.

Since the total number of reads across all genes for each genotype is fixed, the p-value support is simply a function of the total number of reads for a gene summed over both genotypes (n_{1+}). There are many genes where the number of reads to that share the same value of n_{1+}, and hence, will share the same null distribution. Once all genes are tested, the result will be a collection of m p-values whose null distributions are from a mixture of different discrete distributions.

3.2.3 Gene Set Testing

Another multiple testing application that results in a collection of m p-values whose null distributions are different discrete distributions is testing for "overrepresentation" or "enrichment" of DE genes in a given gene set. Gene Sets can be formed based on past biological research. A prime example are the gene sets corresponding to Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000). Genes belonging to a gene set that is considered to contain an unusually large number of DE genes are deemed to jointly play a role in how a treatment affects the organism of interest.

One popular way to test for overrepresentation or enrichment of DE genes in a given set is to use Fisher's exact test to identify sets where the proportion of DE genes is more than would be expected if the DE genes were randomly sampled without replacement from all genes of interest. Criticisms of this and similar tests are that genes are not independent, a clear violation of Fisher's exact test assumptions. Also, the outcome of whether or not a category is overrepresented depends on the threshold used to declare a gene DE. Other drawbacks can be found in Allison et al. (2006), Barry et al. (2006) and Subramanian et al. (2005).

Even though some assumptions of these statistical analyses are in question, it is common to use procedures such as Fisher's exact test to for overrepresentation of gene sets. Since Fisher's exact test relies on small-sample distributions, the collection of p-values resulting from testing numerous gene sets are discrete and follow different null distributions depending on the size of the gene set.

The next section describes how to estimate m_0 for each unique p-value null distribution and how to estimate FDR across all unique p-value null distributions.

3.3 Proposed Procedure

3.3.1 Estimating m_0

As mentioned in the Introduction, many methods exist to estimate m_0, the number of true null hypotheses out of m total hypotheses tested, but most are in the framework of parametric testing which results in p-values that are continuous uniform (0,1) under the null hypothesis. As discussed in the previous subsections, it is possible to obtain a collection of m p-values that have different discrete null distributions.

Suppose we tested m null hypotheses and obtained $p_1, \ldots, p_m$ each of which has one of n unique discrete null distributions, namely $F_1, \ldots, F_n$, with respective support sets $S_1, \ldots, S_n$. Let m_i denote the number of p-values that have null distribution F_i, and let m_{0i} denote the number of p-values that correspond to a true null hypothesis among those with null distribution F_i. If we can obtain an estimate $\widehat{m}_{0i}$ of m_{0i} for each $i = 1, \ldots, n$; we can estimate m_0 by $\widehat{m}_0 \equiv \sum_{i=1}^{m} \widehat{m}_{0i}$. Therefore, we now focus how to estimate m_{0i} for an arbitrary discrete null distribution, F_i.

3.3.1.1 Estimating m_{0i} using an Iterative Algorithm

Mosig et al. (2001) proposed a histogram based iterative algorithm that estimates the number of true null hypotheses given a collection of p-values all with a continuous uniform (0,1) null distribution. The basic idea of the iterative algorithm is the following. Given a histogram (with any bin size) of p-values that have a null distribution that is continuous uniform on (0,1), start by assuming that each p-value is truly distributed continuous uniform on (0,1), i.e. all the hypotheses are in fact true null hypotheses. Then, the expectation of the number of p-values in each bin can be calculated. Find the first bin where the observed frequency fails to exceed expectation and for each bin to the left of this bin, count the excess of observed frequencies to expected frequencies. The algorithm declares this total excess as an estimate of $m - m_0$, and therefore yields an estimate of m_0. Now, recalculate the expected number of p-values that should fall in each bin using the updated estimate of m_0 as the number of true null hypotheses. Again, find the first bin where the observed frequency fails to exceed the updated expectation and for each bin to the left of this bin, find the excess of observed frequencies to the updated expected frequencies. This total excess replaces the estimate of $m - m_0$, and therefore replaces the estimate of m_0. The algorithm continues in this fashion until convergence.

The objective is to use this algorithm to estimate m_{0i} for each unique p-value distribution $F_i,\ i = 1, \ldots, n$. Even though this algorithm is setup to handle p-values that are distributed uniformly on (0,1) under the null hypothesis, all that the algorithm requires is the null expected frequency for each bin. For arbitrary i, the support S_i is a finite discrete set, and therefore we can allocate a bin for each element of S_i. Hence, to estimate m_{0i} for arbitrary i, gather all p-values with null distribution F_i, and construct a histogram allocating a bin for each element of S_i. Then use the histogram based estimator to calculate $\widehat{m}_{0i}$. Repeat for all i to obtain $\{\widehat{m}_{0i}\}_{i=1}^{n}$ and calculate $\widehat{m}_0 = \sum_{i=1}^{n} \widehat{m}_{0i}$. More formally, the iterative algorithm can be described as follows.

Let n_i be the number of elements in S_i and let S_{ij} be the j^{th} smallest element of S_i for $j = 1, \ldots, n_i$. Also, let n_{ij} be the number of observed p-values that have null distribution F_i and are equal to S_{ij}. Let $s_{ij} = S_{ij} - S_{i,j-1}$ for $j = 2, \ldots, n_i$ and $s_{i1} = S_{i1}$. Then s_{ij} is the null probability that a p-value with distribution F_i equals S_{ij}. Define

$$\tilde{n}_{i,j:n_i} = \frac{\sum_{l=j}^{n_i} n_{il}}{\sum_{l=j}^{n_i} s_{il}}, \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n_i. \tag{3.1}$$

Let $m_{0i}^{(0)} = \sum_{j=1}^{n_i} n_{ij} = m_i$ and define for any $j \in \{1, \ldots, n_i\}$,

$$m_{0i}^{(k)} = \left(\sum_{l=1}^{j_i^{(k)}-1} s_{il} \right) \cdot m_{0i}^{(k-1)} + \left(1 - \sum_{l=1}^{j_i^{(k)}-1} s_{il} \right) \tilde{n}_{i,j_i^{(k)}:n_i} \text{ for all } k \geq 1, \tag{3.2}$$

where

$$j_i^{(k)} \equiv \min \left\{ j : n_{ij} \leq s_{ij} \cdot m_{0i}^{(k-1)} \right\}. \tag{3.3}$$

One can think of $j_i^{(k)}$ as the index of the element of S_i where the frequency of observed p-values that equal S_{ij} does not exceed the null expectation given the estimated number of p-values with distribution F_i that correspond to true null hypotheses at iteration k. Then, $m_{0i}^{(k)}$ is the estimated number of p-values with distribution F_i that correspond to null hypotheses at iteration k.

3.3.1.2 A Simple Example

Suppose we have, for some arbitrary i, $m_i = 30$ and $S_i = \{0.04, 0.20, 0.36, 0.52, 1\}$ with corresponding null probabilities $s_i = \{0.04, 0.16, 0.16, 0.16, 0.48\}$. Table 3.3 displays the observed p-value frequencies and the expected p-value frequencies given $m_{0i}^{(k)}$ for the first $k = 2$ iterations of the algorithm.

Iteration k	$m_{0i}^{(k)}$	p-value	.04	.20	.36	.52	1
0	30	Observed Frequency	4	8	6	5	7
1	22.6	Expected Frequency	1.2	4.8	4.8	4.8	14.4
2	18.752	Expected Frequency	0.904	3.616	3.616	3.616	10.848

Table 3.3 Observed frequencies and expected frequencies for two iterations of the histogram based estimator for estimating the number of true null hypotheses for a simple example.

Start by constructing a histogram (or table) allocating a bin for each element in S_i and assuming all $m_i = 30$ hypotheses are null, i.e. $m_{0i}^{(0)} = 30$. Given $m_{0i}^{(0)} = 30$ and s_i, find the expected frequency for each bin. Then, find the first bin where the observed frequency does not exceed the expected frequency, which is the bin corresponding to a p-value equal to 1. Next, for all bins to the left of this bin, add up the difference between the observed and expected frequencies. This excess is $(4-1.2)+(8-4.8)+(6-4.8)+(5-4.8) = 2.8+3.2+1.2+0.2 = 7.4$ which implies $m_{0i}^{(1)} = 30 - 7.4 = 22.6$. Executing the same steps, we estimate $m_{0i}^{(2)} = 30 - 11.248 = 18.752$. Continuing to iterate, we have $\widehat{m}_{0i} = \lim_{k\to\infty} m_{0i}^{(k)} = 14.583$.

In terms of the formal definitions for $k = 1$, $j_{i1} = 5$ since $7 = n_{i5} \leq s_{i5} \cdot m_{0i}^{(0)} = 0.48 \cdot 30 =$

14.4. Then, (3.2) yields

$$
\begin{aligned}
m_{0i}^{(1)} &= \left(\sum_{l=1}^{5-1} s_{il}\right) m_{0i}^{(1-1)} + \left(1 - \sum_{l=1}^{5-1} s_{il}\right) \tilde{n}_{i,4:5} \\
&= \left(\sum_{l=1}^{4} s_{il}\right) m_{0i}^{(0)} + \left(1 - \sum_{l=1}^{4} s_{il}\right) \frac{\sum_{l=5}^{5} n_{il}}{\sum_{l=5}^{5} s_{il}} \\
&= (0.04 + 0.16 + 0.16 + 0.16) \cdot 30 + (1 - (0.04 + 0.16 + 0.16 + 0.16)) \frac{7}{0.48} \\
&= (0.52) \cdot 30 + (0.48) \frac{7}{0.48} \\
&= 15.6 + 7 \\
&= 22.6.
\end{aligned}
$$

Continuing with the formal definitions yields $m_{0i}^{(2)} = 18.752$, $m_{0i}^{(3)} = 16.75104$, $m_{0i}^{(4)} = 15.71054$, $m_{0i}^{(5)} = 15.16948$, $m_{0i}^{(6)} = 14.88813$, $m_{0i}^{(7)} = 14.74183$, $m_{0i}^{(8)} = 14.66575$, $m_{0i}^{(9)} = 14.62619$, $m_{0i}^{(10)} = 14.60562$, etc.

Although the algorithm usually converges quickly in pratice, a noniterative procedure can be used to compute the limiting value directly. Nettleton et al. (2006) characterize the limit for p-values that are continuous uniform (0,1) under the null hypothesis. The next section extends the limit characterization to handle p-values that have a discrete distribution where the elements in the support have unequal null probabilities.

3.3.1.3 Limit Characterization

Nettleton et al. (2006) showed the existence of and characterized the limit of the iterative algorithm when the p-values are continuous uniform on (0,1) under the null hypothesis. There is an analogous result to the iterative algorithm when the p-values have a discrete distribution where the elements in the support have unequal null probabilities. Consider an arbitrary discrete distribution F_i with support S_i.

Convergence Result: Let $J_i = \min\{j : n_{ij} \leq s_{ij} \cdot \tilde{n}_{i,j:n_i}\}$. Then,

$$
\widehat{m}_{0i} = \lim_{k\to\infty} m_{0i}^{(k)} = \frac{\sum_{l=J_i}^{n_i} n_{il}}{\sum_{l=J_i}^{n_i} s_{il}}.
$$

The proof of this convergence result for continuous uniform (0,1) p-values under the null hypothesis can be found in Nettleton et al. (2006) and the proof of this convergence result for discrete non-uniform p-values under the null hypothesis can be found in Bancroft and Nettleton (2009).

The limit characterization says to find the leftmost bin where the observed frequency n_{ij} does not exceed $s_{ij} \cdot \tilde{n}_{i,j:n_i}$. For the observed frequencies in Table 3.3, the values of $s_{ij} \cdot \tilde{n}_{i,j:n_i}$ are 1.2, 4.33, 3.6, 3, and 7. Thus, $J_i = 5$ and

$$\begin{aligned} \widehat{m}_{0i} &= \frac{\sum_{l=5}^{5} n_{il}}{\sum_{l=5}^{5} s_{il}} \\ &= \frac{7}{0.48} \\ &= 14.58333. \end{aligned}$$

Using the limit characterization, it is straightforward to obtain $\widehat{m}_{0i}$ for $i = 1, \ldots, n$ and, thus, $\widehat{m}_0 = \sum_{i=1}^{n} \widehat{m}_{0i}$. This estimate plays an important role in estimating FDR as described in the next section.

3.3.2 Estimating FDR

Suppose the m ordered p-values $p_{(1)} \leq \ldots \leq p_{(m)}$ are to be used to test the corresponding m hypotheses $H_{(01)}, \ldots, H_{(0m)}$. To control the number of false discoveries made in a collection of discoveries at level α, Benjamini and Hochberg (1995) proposed finding the largest integer k such that $\frac{p_{(k)}m}{k} \leq \alpha$ and rejecting $H_{(01)}, \ldots, H_{(0k)}$. The numerator $p_{(k)} \cdot m$ provides an estimate of the expected number of type I errors in m tests if $p_{(k)}$ is used as a significance threshold. This is clearly an overestimate of the expected number of type I errors when m_0 is less than m. Benjamini and Hochberg (1995) show that their procedure controls FDR at level $\frac{m_0}{m}\alpha$, and thus is conservative when $m_0 < m$. Replacing m with m_0 will reduce the conservative bias yielding control at the same level α while obtaining a larger list of discoveries.

Instead of specifying a level at which FDR is to be controlled and finding the corresponding p-value cutoff that gives the specified control, one could specify a p-value cutoff c, and find the corresponding FDR for that p-value cutoff. More explicitly, $\widehat{\mathrm{FDR}}(c)$, the estimated false discovery rate for a p-value cutoff c, is

$$\widehat{\mathrm{FDR}}(c) = \min\left\{\frac{p_{(k)}\widehat{m}_0}{k} \;:\; \text{for all } p_{(k)} \geq c\right\}. \tag{3.4}$$

If (3.4) is calculated for each value of $c \in \{p_{(1)}, \ldots, p_{(m)}\}$, the resulting values of $\widehat{\mathrm{FDR}}(c)$ are analogous to the q-values of Storey (2002).

As mentioned, the numerator in (3.4) estimates the expected number of type I errors. If a collection of p-values all have the same support and possess the property $\mathrm{P}(p\text{-value} \leq p^*) = p^*$ for all p^* in the support under the null hypothesis, then the definition in (3.4) provides an effective estimate of FDR (Bancroft and Nettleton, 2009). If these conditions do not hold, then $\widehat{\mathrm{FDR}}(c)$ as defined in (3.4) may be overly conservative. To see this, consider the following example.

Suppose we have $m_1 = 50$ p-values where $S_1 = \{0.20, 1\}$ with respective observed frequencies $\{n_{11}, n_{12}\} = \{15, 35\}$ and we have $m_2 = 50$ p-values where support $S_2 = \{0.04, 0.20, 0.36, 0.52, 1\}$ with respective observed frequencies $\{n_{21}, n_{22}, n_{23}, n_{24}, n_{25}\} = \{4, 9, 9, 8, 20\}$.

First, using the limit characterization of the histogram based estimator, we have $\widehat{m}_{01} = 43.75$ and $\widehat{m}_{02} = 41.\bar{6}$ and therefore, out of the 100 hypotheses tested, we have an experiment-wise estimate of m_0 as $\widehat{m}_0 = 43.75 + 41.\bar{6} = 85.41\bar{6}$. Next, if we ignored that these 100 p-values have two different supports, we would estimate, for a p-value cutoff of 0.04, the expected number of type I errors to be $0.04 \cdot 85.41\bar{6} = 3.41\bar{6}$. In reality, the estimated number of expected type I errors made from p-values with support $S_1 = \{0.20, 1\}$ should be $0 \cdot 43.75 = 0$, since 0.20 is the minimum value in the support, and the estimated number of type I errors made from p-values with support $S_2 = \{0.04, 0.20, 0.36, 0.52, 1\}$ should be $0.04 \cdot 41.\bar{6} = 1.\bar{6}$. Hence, the estimate of the total number of type I errors should be $0 + 1.\bar{6} = 1.\bar{6}$, not $3.41\bar{6}$. For this hypothetical example, the estimated FDR would be unnecessarily conservative as we have over estimated the expected number of type I errors by not recognizing the distinction between the two distributions. This illustrates that the number of expected type I errors should be individually estimated for each unique distribution when a collection of m discrete p-values do not all have the same null distribution. Then, retaining the spirit of the definition of FDR, we can estimate $\mathrm{FDR}(c)$ across the unique null distributions by adding up the estimated number of type I errors across the unique distributions and dividing by the total number of rejections. A formal description of our proposed procedure for estimating FDR is as follows.

For each $i = 1, \ldots, m$, and each p-value threshold for significance c, calculate $\widehat{V}_i(c) = S_{ij_i^*} \cdot m_i \cdot \widehat{\pi}_0$, where $j_i^* = \max\{j : S_{ij} \leq c\}$ and $\widehat{\pi}_0 = \frac{\widehat{m}_0}{m}$. This estimates the expected number of type I errors among tests with the i^{th} null p-value distribution when c is used as a threshold for significance. Note that if c is not an element of S_i, we use the next smallest element of S_i which will eliminate the conservative bias discussed previously. Also, calculate $R(c) = \#\{p\text{-values} \leq c\}$. Then, the estimated false discovery rate associated with a p-value cutoff c, is defined as

$$\widehat{\mathrm{FDR}}(c) = \min\left\{\frac{\sum_{i=1}^{n} \widehat{V}_i(c^*)}{R(c^*)} : c^* \in \left(\bigcup_i S_i\right) \cap [c, 1]\right\}. \tag{3.5}$$

The next section applies the proposed procedure to a dataset from a proteomics experiment.

3.4 Application to a Proteomics Experiment

Counts of peptide matches to proteins in a database of thousands of proteins were recorded on samples from 5 different genotypes over 3 blocks. Matches were made to 1,176 proteins yielding 1,176 3x5 matrices of peptide counts. Each of the 15 samples contains a different number of peptides and therefore, all matches emanating from the same sample were normalized by the total number of matches for that sample as described previously. Table 3.4 displays the data for a few proteins.

Note that with just three blocks and along with the sparse data, these data do not lend themselves well to typical parametric analysis. Therefore, permutation testing is employed to test, for each protein, whether all genotypes have the same abundance distribution of each protein. The null hypothesis implies that all permutations of data within blocks are equally likely. As described in Section 2, there are $(5!)^3$ ways to randomly assign the observations to the experimental units within blocks. For protein 18, only four samples had a peptide that matched to this protein. Because of the many ties in the data, the $(5!)^3$ assignments gives only 10 unique values of the F-statistic leading to a discrete p-value support of $S_{18} =$

		Genotype				
Protein ID	Repetition	A	B	C	D	E
18	1	0	0	0.0002	0	0.0003
	2	0	0	0	0	0.0003
	3	0	0	0	0	0.0002
39	1	0	0	0	0	0
	2	0.0003	0	0.0005	0	0
	3	0.0003	0.0009	0.0010	0	0
4	1	0.0034	0.0048	0.0042	0.0025	0.0056
	2	0.0049	0.0043	0.0037	0.0047	0.0040
	3	0.0020	0.0032	0.0030	0.0038	0.0028

Table 3.4 Subset of data (rounded) from a randomized complete block design measuring peptide counts on 1,176 different proteins on five genotypes over 3 blocks. Data is normalized by dividing the counts for each sample by the sum of the all sample specific counts across proteins.

$\{0.04, 0.08, 0.12, 0.16, 0.28, 0.40, 0.52, 0.64, 0.76, 1.00\}$. For example, a p-value of 1 corresponds to the smallest possible test statistic. Here, that occurs when the variance of the genotype means is smallest and is when there is only one non-zero observations per column. The number of ways to obtain only one non-zero observation per genotype is $5 \cdot 4 \cdot 3 \cdot 2 \cdot 3! \cdot (4!)^2 = 414,720$. Hence a p-value of 1 occurs with null probability $414720/(5!)^3 = 0.24$

For the actual arrangement for protein 18, the test statistic is the largest possible, and therefore, the p-value is equal to $\left(5 \cdot 4 \cdot 3! \cdot (4!)^2\right)/(5!)^3 = 0.04$. For protein ID 4, there are no tied observations, and the data are such that there are $5! \cdot 5! = 14,400$ unique test statistics, each occurring with equal probability of $\frac{1}{14,400}$. There were 140 total proteins with a p-value support of 14,400 elements. On the other extreme, there were 94 proteins with a p-value support of just two elements, namely $\{0.20, 1\}$. Also of note, there were 586 proteins with a p-value support of just one element, namely 1. These proteins were ignored in the computation of $\widehat{m}_0$ and estimating FDR as they contained no information. Table 3.5 displays the breakdown of number of proteins versus number of normalized counts that are 0.

Table 3.5 shows that there were 536 proteins with 14 of the 15 normalized counts equal to zero, suggesting that each of these proteins will have a p-value singleton support set of $\{1\}$. There are 50 other proteins that have a singleton support set $\{1\}$. This singleton support set

Number of 0 normalized counts	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# proteins	76	45	21	21	14	23	18	37	32	37	38	53	86	139	536	0

Table 3.5 Breakdown of number of proteins versus number of normalized counts that are 0.

can be obtained if there is only one non-zero normalized count or if all the non-zero normalized counts are observed in the same block. Hence, Table 3.5 does not summarize the number of unique p-value null distributions, but rather shows a general structure of the data at hand.

There were 126 unique p-value null distributions (excluding the p-value null distribution with the singleton support $\{1\}$) and the estimated number of true null hypotheses, $\widehat{m}_0$, for each unique p-value distribution was found. For the subset of the 590 informative proteins, $\widehat{\pi}_0 = \frac{\sum_{i=1}^{n} \widehat{m}_{0i}}{m} = \frac{528.6259}{590} = 0.896$ and the estimated FDR for significance cutoffs of 0.001, 0.005, 0.01, and 0.05 are given in Table 3.6.

c	0.001	0.005	0.01	0.05
$\sum_{i=1}^{n} \widehat{V}_i(c)$	0.1911	1.1792	2.7994	18.5099
$\sum_{i=1}^{n} R_i(c)$	2	5	12	67
$\widehat{\text{FDR}}(c)$	0.0956	0.2358	0.2333	0.2763

Table 3.6 Estimated number of type I errors and number of rejections across the subset of 590 proteins and the estimated false discovery rate (FDR) for four different significance thresholds.

3.5 Simulation Study

This section investigates the performance of the proposed procedure via a simulation study that attempts to produce data similar to the data described in the previous section. In words, data for m proteins were simulated, of which, m_0 were simulated under the null hypothesis of equal protein abundance across genotypes and $m - m_0$ were generated under the alternative hypothesis of unequal protein abundance across genotypes. Once the data were simulated, permutation testing was employed to test the m null hypotheses which resulted in a collection of m permutation p-values that do not all have the same distribution under the null hypothesis. The proposed procedure was employed to estimate the number of true null hypotheses and to estimate the false discovery rate for prespecified significance cutoffs. The false positive fraction defined as $V/\max\{1, R\}$, where V is the true number of type I errors and R is the number of rejections, was also recorded. After repeating this simulation scheme N times, the average $\widehat{m}_0$ was compared to m_0 and the average estimated FDR was compared to the average $V/\max\{1, R\}$.

The follow procedure was used to simulated data under the null hypothesis for each simulation run. We can alternatively think of the data analyzed in Section 4 as a 1176 by 15 matrix, where each row corresponds to a protein and each column corresponds to a sample. Each column contains counts for the number of peptides in the sample that matched to each of the 1176 proteins. One of these 15 columns was randomly selected, and $m = 100$ counts were randomly drawn from the 1176 counts denoted $\lambda_1, ..., \lambda_m$. These m counts served as means for Poisson distributions from which data was generated. To simulate 15 counts (three for each of five groups) for the i^{th} protein, $i = 1, \ldots, m_0$, 15 observations were drawn from Poi(λ_i). These 15 draws were then placed in a 3 by 5 matrix where the rows serve as blocks and the columns serve as different genotypes. Note that in contrast to the real data and data analysis, these counts were not normalized and no block effects were added to the observations. Here, we act as if there are blocks so that data permutations must be done within block as in the actual analysis. Also, some of the λ_i were equal to zero. If zero was used as a mean for the Poisson distribution, all 15 simulated observations would be zero, and the p-value for that hypothetical protein would equal its only possible value of 1. These p-values with a singleton support were ignored in Section 4 since they contained no information. Here, if they were used they would only inflate the estimated FDR while contributing no information to the overall analysis. However, there were many zero counts in the actual data, which we would like to recreate since tied observations lead to non-uniform p-value supports. Hence, if λ_i was 0, each of the 15 observations were either set to 0 with probability 1/2 or simulated from Poi(1) with probability 1/2.

To simulate data under the alternative hypothesis of unequal protein abundance distributions across genotypes, the following was done for each simulation run. For each $i = m_0 + 1, \ldots, m$, three observations were simulated from each of Poi(λ_{i1}),...,Poi(λ_{i5}), where $\lambda_{ij} = \lambda_i + \gamma_{ij}$ for $\gamma_{ij} \sim \Gamma(1, \delta),\ j = 1, \ldots, 5$, with E[$\gamma_{ij}$]= $1/\delta$. Here, for a given i, we simply created five values that are not equal which we used as Poisson means to simulate three Poisson counts for each of the $j = 1, \ldots, 5$ groups.

The results of $N = 100$ simulation runs using $m = 100$ and $m_0 = 80$ are given in Table 3.7. FDR(c) is estimated very conservatively. This conservative bias is directly tied to the fact that the m_0 estimates are also conservative since an estimate of m_0 is used to estimate FDR(c) given by (3.5). The conservative bias in the m_0 estimates, especially for $m_0 = 75$, could be due to the fact that there are only 3 observations simulated per treatment condition. When there is low power, the distribution of the alternative p-values is similar to the distribution of the null p-values. If more repetitions were used, one would expect to see $\widehat{m}_0$ decrease closer to the true m_0. Still, as the power increases (E[γ_{ij}] increases), the conservative bias of $\widehat{\text{FDR}}(c)$ decreases and is larger than the false positive fraction in all cases. Hence, the procedure controls FDR at the advertised levels.

3.6 Summary

This paper proposes how to estimate the number of true null hypothesis, m_0, and the false discovery rate given a collection of m p-values from a mixture of discrete distributions. Further, each discrete null distribution does not have an evenly spaced support resulting in an unequal null probability of observing each element in the support. A simulation study has shown that the proposed procedure does control the false discovery rate and estimates m_0, both with a conservative bias.

		$\widehat{m}_0$			$\widehat{\text{FDR}}(c)$		$V(c)/\max\{1,R(c)\}$	
m_0	$\text{E}[\delta]$	mean	se	c	mean	std. err.	mean	std. err.
75	1	92	0.276	0.005	0.1165	0.0076	0.0517	0.0206
				0.010	0.1473	0.0079	0.1087	0.0250
				0.050	0.2915	0.0077	0.2034	0.0121
75	2	93	0.254	0.005	0.0254	0.0017	0.0152	0.0084
				0.010	0.0482	0.0031	0.0203	0.0084
				0.050	0.1790	0.0048	0.0934	0.0147
75	4	93	0.254	0.005	0.0238	0.0007	0.0054	0.0028
				0.010	0.0400	0.0010	0.0200	0.0038
				0.050	0.1520	0.0026	0.0972	0.0058
90	1	93	0.298	0.005	0.2360	0.0214	0.0733	0.0237
				0.010	0.2832	0.0200	0.1650	0.0311
				0.050	0.4993	0.0194	0.4831	0.0220
90	2	93	0.257	0.005	0.0933	0.0080	0.0570	0.0193
				0.010	0.1351	0.0078	0.1007	0.0200
				0.050	0.3835	0.0263	0.3195	0.0153
90	4	94	0.285	0.005	0.0443	0.0026	0.0165	0.0057
				0.010	0.0776	0.0032	0.0450	0.0083
				0.050	0.0273	0.0060	0.2207	0.0132

Table 3.7 Average $\widehat{\text{FDR}}(c)$ and average $V(c)/\max\{1,R(c)\}$ over $N = 100$ simulation runs where each run consisted of simulating data for $m = 100$ proteins, of which, 80 proteins were simulated under the null hypothesis of equal protein abundance across genotypes.

3.7 References

[1] Allison, D.B., (2004). Fatigo: A Web Tool for Finding Significant Associations of Gene Ontology Terms with Groups of Genes. *Nature*, **7**, 55-56.

[2] Barry, W. et al. (2005). Significance Analysis of Functional Categories in Gene Expression Studies: A Structured Permutation Approach. *Bioinformatics*, **21**, 1943-1949.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

[4] Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, **25**, 6083.

[5] The Gene Ontology Consortium (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, **25**, 25-29.

[6] Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data. *Journal of the Royal Statistical Society, Series B*, **67**, 555–572.

[7] Mosig, M.O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Percentage in Israeli-Holstein Cattle by Means of Selective Milk DNA Pooling in a Daughter Design Using an Adjusted False Discovery Rate Criterion. *Genetics*, **151**, 1683-1698.

[8] Nettleton, D., Hwang, J.T.G., Caldo, R.A., and Wise, R.P. (2006). Estimating the Number of True Null Hypotheses From a Histogram of *p*-values. *Journal of Agricultural, Biological, and Environmental Statistics*, **Vol. 11, No. 3**, 337-356.

[9] Storey, J.D. (2000). False Discovery Rates: Theory and Applications to DNA Microarrays. Unpublished Ph.D. thesis, Department of Statistics, Stanford University.

[10] Storey, J.D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*, Series B, **64**, 479-498.

[11] Subramanian, A. et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profile. *Proceedings of the National Academy of Sciences, USA*, **102**, 15545-15550.

[12] Whitelegge, J.P. (2002). Plant Proteomics: Blasting out of a MudPIT. *Proceedings of the National Academy of Sciences, USA*, **Vol. 99, No. 18**, 11564-11566.

CHAPTER 4. Expression Quantitative Trait Loci Mapping Studies

Abstract

Quantitative Trait Loci (QTL) are regions of the genome that affect quantitative traits. Computationally intensive QTL mapping studies consist of testing loci closely spaced throughout the genome to find the locus whose genotypic information is most associated with the trait in question. The number of testing positions can be in the thousands, and computation of each locus specific test statistic requires numerical methods such as the EM algorithm. Further, the theoretical reference distribution for the test statistic corresponding to the locus most associated with the trait value is not easily identified because of the correlation among test statistics due to the dependence along the genome. Hence, the reference distribution used is the empirical distribution of the test statistic generated through permutation, increasing the computational burden. For a given trait, the total number of EM iterations can easily be in the millions. Multiple trait cases, such as expression QTL (eQTL) mapping studies where the expression of one gene is one trait, make the computational expense in this multiple testing situation very substantial. This paper discusses how to reduce computation by generating gene specific reference distributions where the number of permutations is relatively small when there is little evidence against the null hypothesis. Estimating the false discovery rate from the resulting *p*-values is also discussed.

Key Words: EM algorithm; Microarray; Mixture Model; Monte Carlo testing; Sequential analysis

4.1 Introduction

Quantitative trait loci (QTL) analysis is an important tool for dissecting the genetic influences on biological traits. Researchers have attempted to map (locate) QTL for a wide variety of traits and organisms from fruit weight in tomatoes (Paterson et al., 1988) to cognitive ability in children (Chorney et al., 1998). The goal is to determine associations between genetic variability at known locations on chromosomes with variability in the observed traits, also known as phenotypes, to pinpoint the locations of genes controlling the phenotypes. Typically, genotypes are obtained at a large number of locations (referred to as markers) throughout the genome and separate analyses may be done at each location. To do this, an appropriate experimental population must be available. Typically, divergent inbred lines are systematically mated to obtain progeny for evaluating the genotype-phenotype associations. Initial breeding between parents differing substantially in both the quantitative trait of interest and in genetic makeup will facilitate mapping of QTLs. These designs are staples in agriculture and animal breeding, where the objective is to manipulate profitable traits. They are also used to investigate human diseases (e.g., diabetes and cancer) and can assist in the identification of causative genes.

The analysis of experimental populations was first considered by Sax (1923), who proposed t-tests for phenotypic means of different genotypic groups at a known marker. Lander and Botstein (1989) generalized the idea of Sax (1923) to interval mapping at loci between known markers where simple t-tests may not be appropriate since genotypes at such loci are unobserved. Fortunately, flanking markers contain information about the genotypes of the positions they surround because of the dependence in genotypes at nearby positions. Given a genetic map for the organism under study, the conditional probabilities of the two (or more) genotypes at hand, given the flanking marker genotypes, can be computed for any nonmarker position. The conditional probabilities depend on the location of the position relative to the flanking markers and vary from individual to individual with marker genotype. This leads to one standard model for interval mapping of the QTL where the phenotype distribution is modeled as a mixture of two (or more) components corresponding to two (or more) different genotypes at the putative (assumed) QTL (Lander and Botstein, 1989). The distributions of the mixture components are typically assumed to be normal with the same variance but different means. The mixing proportions in interval mapping are fixed known functions of the flanking marker positions as described above. A usual way to assess the genotype-phenotype association at each testing position is with the LOD score which is equivalent to a likelihood ratio statistic. The testing position with the largest LOD score is taken as the candidate QTL. Once the genome is scanned (i.e. loci have been tested at closely spaced uniform increments along the genome) and a candidate QTL is found (i.e. chosen to be the locus with the most association between the phenotypic trait values and the genotypic information), it then indicates a candidate region in the genome that may carry one or more genes controlling the trait. However, finding this candidate QTL comes with much computational expense as computing each LOD score requires a numerical procedure such as the Expectation-Maximization (EM) algorithm. Further, the appropriate reference distribution to assess the significance of the maximum LOD score across markers is not analytically tractable due to the dependence of the loci along the genome and thus, is generated through permutation.

The computational challenge doesn't end there. Recent advances in data collection technology have make it possible to consider simultaneous mapping of thousands of traits. The prime example involves the mapping of expression quantitative trait loci (eQTL). An eQTL is a locus or region on the genome associated with the expression of a gene. A gene is sometimes said to "map to" a locus if it is associated with that locus. The typical experiment involves gene expression measurements from microarrays with a genome-wide set of markers. Microarrays are assays that simultaneously measure the expression (i.e., mRNA abundance) of thousands of genes. The motivation for eQTL analysis is to discover how gene expression might be genetically controlled. To accomplish this, a single trait analysis is done for each of thousands of genes, markedly increasing the necessary computation.

Section 2 provides a brief background of the genetics and the terminology necessary to fully grasp the ideas and concepts of eQTL mapping studies. A statistical model for the single trait case is presented in Section 3 including the response distribution, estimation, and assessing significance of the maximum LOD score across testing positions. Section 4 discusses the multiple trait case and how to use the false discovery rate (FDR) to correct for multiple testing. The proposed methodologies are applied to an actual data set in Section 5. A simulation study is described in Section 6 where trait values are simulated from a distribution where the mean may depend on the genotypic information at a random QTL location. Comparing the mapped QTL locations with their true positions and comparing the estimated FDR to the true false discovery rate is the focus of the simulation study. Section 7 summarizes the findings of this paper.

4.2 Genetics Review

DNA (deoxyribonucleic acid) is arranged in pairs of chromosomes. Different organisms have different numbers of chromosomes. The human has 23, a mouse has 20, a fruit fly has 4, and barley has 7 chromosomes. A position on a chromosome corresponding to a segment of DNA is referred to as a locus. A particular DNA variant at a locus is known as an allele. Because chromosomes are arranged in pairs, there are two alleles at each locus. For example, a locus may have two alleles A and a. If an organism has two copies of allele A, then the genotype of the organism is AA, and it is said to be homozygous at the locus. If the organism has genotype Aa, it is said to be heterozygous at the locus. The other homozygote is genotype aa. Physical characteristics or measurements that are governed by a specific loci or multiple loci are called phenotypes. If genotypes AA, Aa, and aa at a genetic position show three distinct values for a phenotype, then the alleles are called codominant. If AA and Aa show the same phenotype, the A is defined as the dominant allele and a the recessive allele. These and other genetic concepts were derived from experimental populations resulting from controlled crossing (mating) of organisms with distinct phenotypes. Controlled crossing can be dated back to Mendel's garden peas in 1865 and is still the most common way to obtain experimental populations used in genomic research for animal and plant species.

Mating, in general, begins with a cross between two parents. The offspring, or progeny, are each comprised of a mixture of genetic material from each parent. At each locus along the genome, each parent passes on the genotype from only one of the two chromosomes. Typically, for a parent with two adjacent loci A and B with alleles Aa and Bb, if they pass on genotype A (a) from locus A, then it is more likely that they will also pass on genotype B (b) from locus B as there is dependence along the genome. This dependence is referred to as genetic linkage. If a parent with genotype $AaBbCc$ at loci A, B, and C passes on genotype ABc, then a 'crossover' event has occurred between loci B and C. Crossover events are rare and happen roughly once every Morgan (M), the unit of measure for chromosomes. Chromosomes are typically between 50 centiMorgans (cM) and 200 cMs depending on the organism. During mating, the collection of alleles passed on from each parent is called a haplotype. Haplotypes across all chromosomes along with other genetic material are carried by a gamete. In human males, sperm are gametes just as eggs are gametes in human females. The progeny of a cross between two parents is referred to as an F1 population. When an organism from an F1 population is crossed with one of its parents, the cross is called a backcross. F2 progeny results from F1 crossed with itself. Repeated self crossing creates inbred lines that have the same copy of the allele at each locus, i.e. eventually, the haplotypes from each parent will be identical. To obtain experimental populations, two differing homozygous parents, say $AABBCC$ and aabbcc in a three locus model, are mated resulting in an F1 population with genotype $AaBbCc$. A backcross could then be $AABBCC$ x $AaBbCc$. Clearly, the haplotype passed on by the original parent in this backcross (and all backcrosses of this form) is ABC. But the haplotype passed on by the F1 population parent in this backcross could vary from offspring to offspring. The haplotype could be, e.g., ABc. If so, there has been a crossover as the alleles A and B were inherited from loci A and B on one chromosome and allele c was inherited from locus C on the other chromosome. The concept of crossover events and recombination is vital to genetics, especially in QTL mapping studies. Figure 4.1 summarizes this paragraph for a three locus scenario.

Recombination is defined as an odd number of crossover events. Suppose, for a five locus model, at some stage of some mating scheme a parent with genotype $AaBbCcDdEe$ passes on haplotype $A\_\_\_e$. Then a recombination has occurred as the forms of the passed alleles are different (capital and lower case). Note that only one crossover may have occurred, and

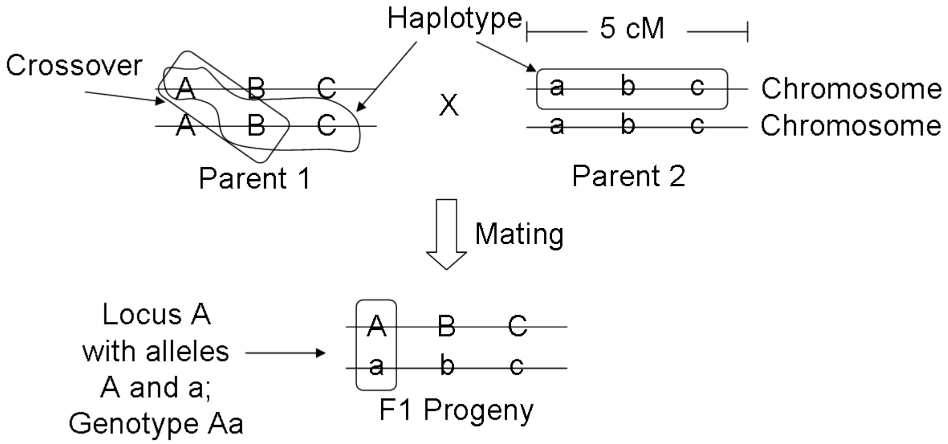


Figure 4.1 Simple three locus mating example. Both parents are homozygous at all three loci. During mating, a crossover has occurred in Parent 1 between locus A and locus B. The offspring, or progeny, of parent 1 and parent 2 is called the F1 population. The F1 population is comprised of one haplotype from Parent 1 and one haplotype from Parent 2. The F1 progeny shown here is heterozygous at all three loci.

the haplotype could have looked like *Abcde*, *ABcde*, *ABCde*, or *ABCDe*. However, three crossovers (albeit with a very low probability) may have occurred and the haplotype could have looked like *AbCde*, *AbcDe*, *AbCDe*, or *ABcDe*. Consider a chromosome pair and any two loci, say 1 and 2, along the chromosome pair. If the allele passed on from locus 1 is from one chromosome and if the allele passed on from locus 2 is from the other chromosome, then an odd number of crossovers has taken place and is defined as a recombination. As the distance between loci 1 and 2 increase, the chance of recombination increases towards $\frac{1}{2}$. But note that the recombination fraction (roughly speaking, the chance of a recombination between two loci along the same segment of the genome) is not additive along the genome. If A, B, and C represent three loci along the genome and if we assume crossovers occur at random and the event of a recombination between loci A and B is independent of the event of a recombination between loci B and C, then the relationship between the recombination fractions is

$$r_{AC} = r_{AB}(1 - r_{BC}) + (1 - r_{AB})r_{BC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC} \tag{4.1}$$

where $r_{AB}r_{BC}$ is the expected double recombination frequency, i.e. recombination between loci

A and B and between B and C, which by definition, is not a recombination between loci A and C since there would have been an even number of total crossover events. In words, (4.1) is the probability of a recombination between loci A and B multiplied by the probability of no recombination between loci B and C plus the probability of no recombination between loci A and B multiplied by the probability of a recombination between loci B and C.

When there are more than three loci, the recombination relationships become more complex. Mapping functions are designed to solve this problem. Haldane's mapping function is one such mapping function and models crossover events as a Poisson Process with a rate of one crossover per Morgan. Then the recombination fraction between any two loci, X and Y, located at a distance d Morgan's apart is defined by Haldane as

$$\begin{aligned}
r_{XY} &= \mathrm{P}(\text{recombination}) \\
&= \mathrm{P}(\text{odd number of crossovers}) \\
&= \mathrm{P}(\text{one crossover}) + \mathrm{P}(\text{three crossovers}) + \mathrm{P}(\text{five crossovers}) + \ldots \\
&= \sum_{i=0}^{\infty} \frac{\exp(-d)d^{2i+1}}{(2i+1)!} \\
&= d\exp(-d) + \frac{d^3\exp(-d)}{3!} + \frac{d^5\exp(-d)}{5!} + \ldots \\
&= \frac{1}{2}\left[2d\exp(-d) + \frac{2d^3\exp(-d)}{3!} + \frac{2d^5\exp(-d)}{5!} + \ldots\right] \\
&= \frac{1}{2}\left[1 - \left\{1 - 2d\exp(-d) - \frac{2d^3\exp(-d)}{3!} - \frac{2d^5\exp(-d)}{5!} - \ldots\right\}\right] \\
&= \frac{1}{2}\left[1 - \exp(-d)\left\{\exp(d) - 2d - \frac{2d^3}{3!} - \frac{2d^5}{5!} - \ldots\right\}\right] \\
&= \frac{1}{2}\left[1 - \exp(-d)\left\{\left(\sum_{i=0}^{\infty}\frac{d^i}{i!}\right) - 2d - \frac{2d^3}{3!} - \frac{2d^5}{5!} - \ldots\right\}\right] \\
&= \frac{1}{2}\left[1 - \exp(-d)\left\{1 + (d - 2d) + \frac{d^2}{2!} + \left(\frac{d^3}{3!} - \frac{2d^3}{3!}\right) + \frac{d^4}{4!} + \left(\frac{d^5}{5!} - \frac{2d^5}{5!}\right) + \ldots\right\}\right] \\
&= \frac{1}{2}\left[1 - \exp(-d)\left\{1 - d + \frac{d^2}{2!} - \frac{d^3}{3!} + \frac{d^4}{4!} - \frac{d^5}{5!} + \ldots\right\}\right] \\
&= \frac{1}{2}\left[1 - \exp(-d)\exp(-d)\right] \\
&= \frac{1}{2}\left[1 - \exp(-2d)\right].
\end{aligned}$$

Hence, given a distance in Morgan's, d, between two loci, this mapping function can be used to find the probability of a recombination, r, between the two loci. Haldane's mapping function is most useful in conjunction with a genetic map, details of which will not be discussed in this paper except to say that a genetic map of an organism is an abstract model of the linear arrangement of a group of loci. Essentially, it gives the location of markers (loci whose genotypes are observable) along the genome of an organism. The genetic map, the genotypes at the marker positions, and Haldane's mapping function give enough information to test any locus on the genome for association with the trait in question. Reiterating the following four points is necessary in understanding the subsequent parts of this paper.

- Only the genotypes of marker loci are observed.
- There is a spatial dependence in genotype across any chromosome.
- This dependence satisfies the Markov property that says the conditional probability of each parental genotype at any particular position depends only on the given genotypes of the markers flanking that position.
- Genetic distances are often reported in centiMorgans. The dependence of genotypes at genetic positions increases as the distance in cM between the positions decreases.

The next section provides a few brief details of the experimental population and breeding scheme particular to this paper followed by the statistical model.

4.3 Single Trait Statistical Modelfor Data from Doubled Haploid Lines

In this subsection, we describe a statistical model from mapping QTL with data from doubled haploid lines. Doubled haploid lines are generally created as follows. Two parents differing substantially in phenotype are crossed to produce an F1 population. Then, one offspring is taken from the F1 population, and the chromosomes are each copied to produce chromosome pairs that are homozygous at every locus. Because every locus is homozygous, it is reasonable to suppress the second allele and write A for genotype AA and write for a

for genotype aa in this doubled haploid scenario. Figure 4.2 shows the production of doubled haploid lines with genotypes $ABQC$, $abQC, \ldots, ABqc$ in the abbreviated notation.

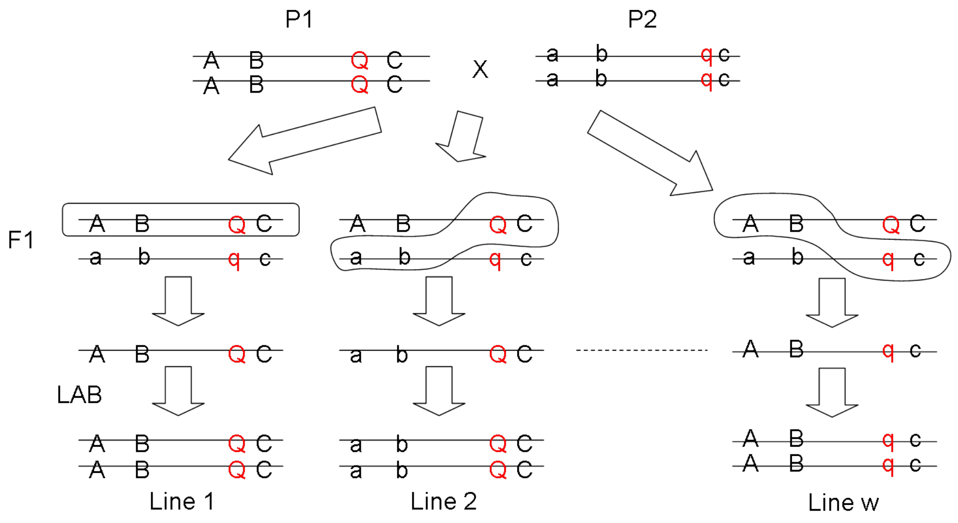
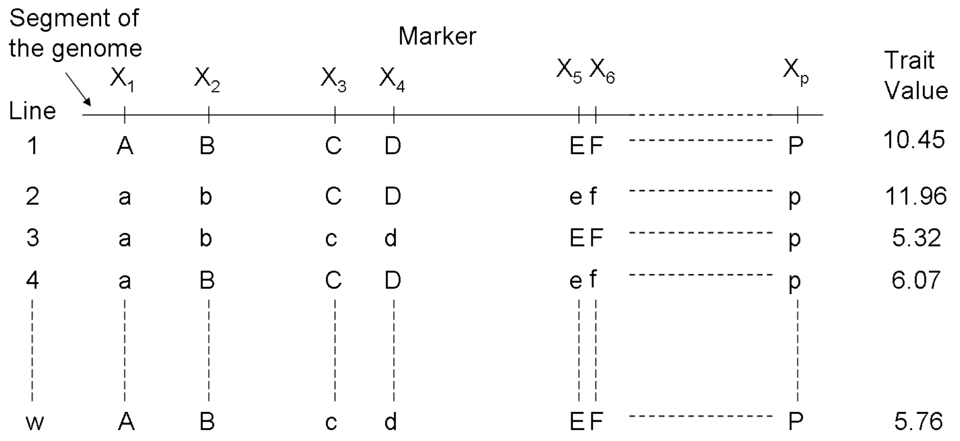


Figure 4.2 Cartoon illustration of double haploid creation.

The statistical model to be described next is designed for w doubled haploid lines each produced from copying a chromosome from a different F1 population organism. Information is available on the genotypes at p markers along the genome on w doubled haploid lines as well as a trait value (phenotype) for each of the w lines. Figure 4.3 shows a portion of a hypothetical dataset.

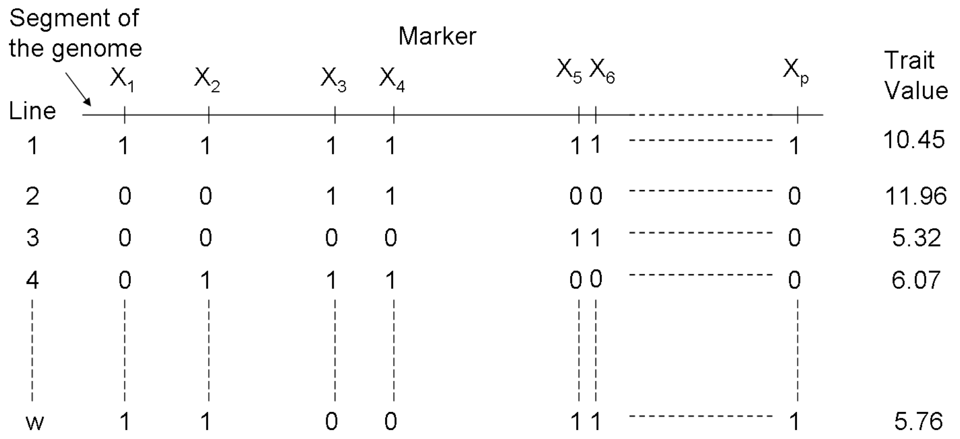
The goal is to find a locus (not necessarily a marker) along the genome where the genotypes of the w lines exhibit the most association with the w phenotype values. Since we are only concerned with analyzing positions on the genome, we can further simplify Figure 4.3 by replacing all capital genotypes with a 1 and all lowercase genotypes with a 0 as illustrated in Figure 4.4.

A model will now be discussed where the QTL position and the genotypes of the w lines at this position are known. Then, a model is discussed where only the QTL position is known, but the genotypes of the w lines at this position are not known. Last, a model is discussed that reflects the true situation. The QTL position is unknown and the genotypes of the w lines at



Line	X_1	X_2	X_3	X_4	X_5	X_6	X_p	Trait Value
1	A	B	C	D	E	F	P	10.45
2	a	b	C	D	e	f	p	11.96
3	a	b	c	d	E	F	p	5.32
4	a	B	C	D	e	f	p	6.07
w	A	B	c	d	E	F	P	5.76

Figure 4.3 Genotypes at p markers along a segment of the genome for w doubled haploid lines.



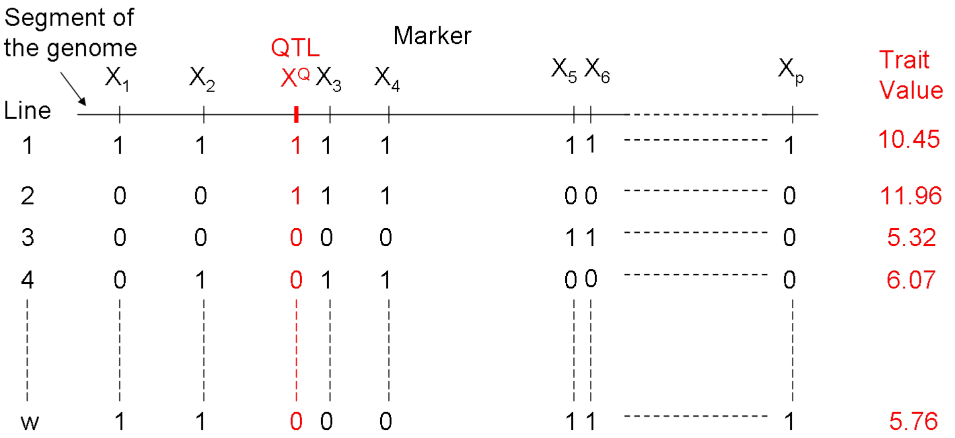
Line	X_1	X_2	X_3	X_4	X_5	X_6	X_p	Trait Value
1	1	1	1	1	1	1	1	10.45
2	0	0	1	1	0	0	0	11.96
3	0	0	0	0	1	1	0	5.32
4	0	1	1	1	0	0	0	6.07
w	1	1	0	0	1	1	1	5.76

Figure 4.4 Genotypes at p markers along a segment of the genome for w doubled haploid lines. For example, at marker X_1, genotype AA (aa) was replaced by A (a) in Figure 4.3 and is now replaced by a 1 (0).

the unknown QTL position are not known.

4.3.1 QTL Location Known and QTL Genotypes Known

To motivate the true QTL mapping situation, we start by assuming we know the QTL position and the genotypes of the w lines at the QTL position. Figure 4.5 displays this fictional situation.



Line	X_1	X_2	X^Q	X_3	X_4	X_5	X_6	X_p	Trait Value
1	1	1	1	1	1	1	1	1	10.45
2	0	0	1	1	1	0	0	0	11.96
3	0	0	0	0	0	1	1	0	5.32
4	0	1	0	1	1	0	0	0	6.07
w	1	1	0	0	0	1	1	1	5.76

Figure 4.5 Fictional situation where the QTL is known and the genotypes of the w lines at the QTL are known.

Before the statistical model is defined, it is worthwhile to note the association between the genotypes at the QTL and the trait values. Also, note that the association (theoretically) diminishes as the distance between a marker and QTL increases. Marker X_p is essentially unassociated with the trait value since marker X_p is located farthest from the QTL. Now, to the statistical model.

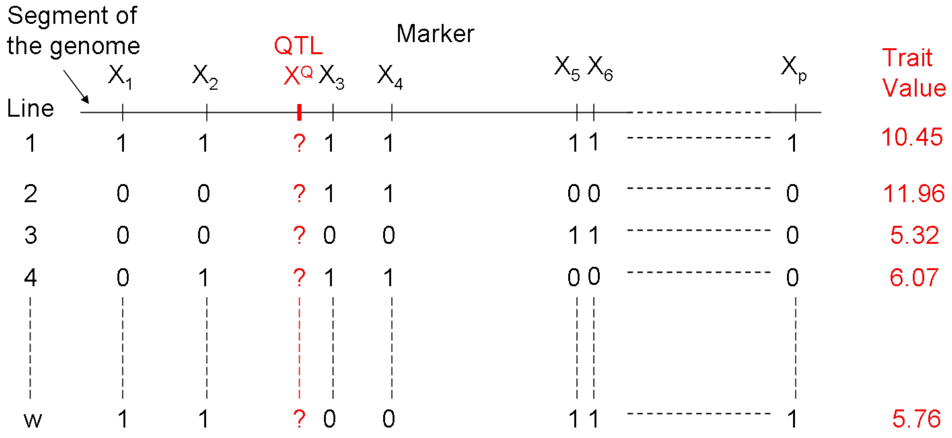
Define $Y_i,\ i = 1, \ldots, w$, as the phenotype of line i. Let $X_i^Q = 0$ if the genotype on line i at the QTL is 0 and let $X_i^Q = 1$ if the genotype on line i at the QTL is 1. Then, the density of Y_i is written as

$$f(Y_i|X_i^Q, \mu_0, \mu_1, \sigma) = (1 - X_i^Q) \cdot \phi(Y_i; \mu_0, \sigma) + (X_i^Q) \cdot \phi(Y_i; \mu_1, \sigma)$$

where $\phi(\cdot; \mu, \sigma)$ represents the normal density with mean μ and standard deviation σ. To determine whether X^Q is associated with the trait values, the hypothesis $H_0 : \mu_1 = \mu_2$ could be tested using a simple two-sample t-test. The null hypothesis declares that X^Q is not a QTL since the trait values emanate from the same distribution regardless of the genotypic information, i.e. there is no association between X^Q and the trait. The null hypothesis may be rejected for many loci near the QTL. Hence, we interpret the alternative hypothesis H_A : $\mu_1 \neq \mu_2$ by stating X^Q is associated with the trait and should not interpret H_A by stating X^Q is the QTL. Rejection of H_A implies that there is an association because the trait values emanate from distributions with means that depend on the genotypic information. Next, a statistical model is presented that addresses a midpoint between this fictional situation and the true situation.

4.3.2 QTL Location Known and QTL Genotypes Unknown

Now, the location of the QTL remains known, but we drop the assumption that the genotypes of the w lines at the location of the QTL are known. This scenario is shown in Figure 4.6.



Line	X_1	X_2	X^Q	X_3	X_4	X_5	X_6		X_p	Trait Value
1	1	1	?	1	1	1	1	---------------	1	10.45
2	0	0	?	1	1	0	0	---------------	0	11.96
3	0	0	?	0	0	1	1	---------------	0	5.32
4	0	1	?	1	1	0	0	---------------	0	6.07
w	1	1	?	0	0	1	1	---------------	1	5.76

Figure 4.6 Fictional situation where the QTL is known. The genotypes at the QTL are not known.

Here, the density of Y_i is written

$$f(Y_i|\pi_i, \mu_0, \mu_1, \sigma) = (1 - \pi_i) \cdot \phi(Y_i; \mu_0, \sigma) + (\pi_i) \cdot \phi(Y_i; \mu_1, \sigma) \tag{4.2}$$

where $\pi_i \in [0,1]$ is the probability of genotype 1 at the QTL on line i. The π_i's are computed by taking advantage of the dependence along the genome. Haldane's mapping function is one function which exploits this dependence to find conditional probabilities of genotypes at positions given the genotypes of surrounding markers. For example, consider line 2 in Figure 4.6. We know the genotype at X_2 is 0 and the genotype at X_3 is 1. Hence,

$$\pi_2 = \mathrm{P}(\_1\_|0\_1) = \frac{\mathrm{P}(011)}{\mathrm{P}(0\_1)} = \frac{\mathrm{P}(011)}{\mathrm{P}(001){+}\mathrm{P}(011)}, \tag{4.3}$$

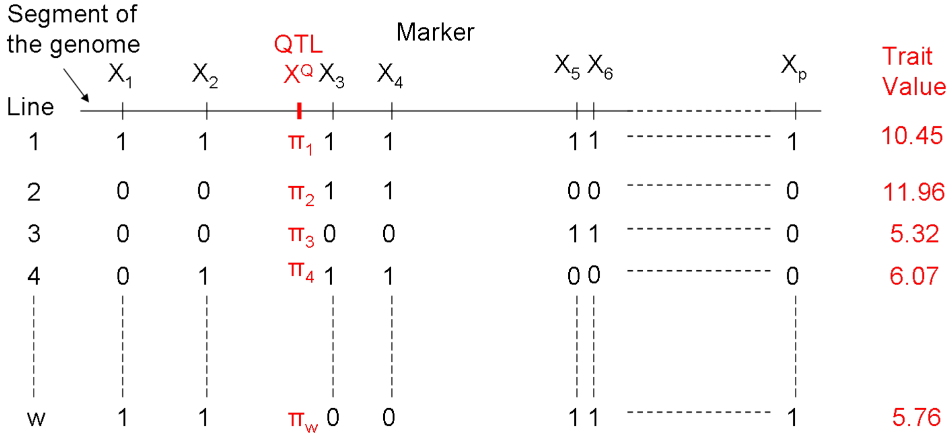
i.e. for line 2, π_2 is the conditional probability of genotype 1 at the QTL given that the genotype of the nearest marker to the left of the QTL is 0 and that the genotype of the nearest marker to the right of the QTL is 1.

To compute P(011) using Haldane's mapping function, suppose marker X_2 is d_1 M from the QTL, X^Q. Then, the probability of a recombination between X_2 and X^Q is $r_1 = \frac{1}{2}\left[1-\exp(-2d_1)\right]$, i.e. r_1 is the probability that the genotype at marker X_2 is 0 (1) and then an odd number of crossovers takes place between X_2 and X^Q resulting in genotype 1 (0) at X^Q. Similarly, suppose X^Q is d_2 M from X_3, then the probability of a recombination between X^Q and X_3 is $r_2 = \frac{1}{2}\left[1-\exp(-2d_2)\right]$. Thus,

$$\begin{aligned}
\mathrm{P}(011) &= \mathrm{P}(011|01\_)\mathrm{P}(01\_) \\
&= (1-r_2)\mathrm{P}(01\_) \\
&= (1-r_2)\mathrm{P}(01\_|0\_\_)\mathrm{P}(0\_\_) \\
&= (1-r_2)\cdot r_1 \cdot \frac{1}{2}
\end{aligned}$$

and similarly, P(001)$= (1-r_1)\cdot r_2 \cdot \frac{1}{2}$. Using equation (4.3), we have $\pi_2 = \frac{(1-r_2)\cdot r_1 \cdot \frac{1}{2}}{(1-r_2)\cdot r_1\cdot\frac{1}{2}+(1-r_1)\cdot r_2\cdot\frac{1}{2}} = \frac{(1-r_2)\cdot r_1}{(1-r_2)\cdot r_1+(1-r_1)\cdot r_2}$. In this example, generic distances d_1 and d_2 were used. Without these, the π's cannot be found. Fortunately, the genetic map gives the distances between the markers, and hence, π_i can be calculated for $i = 1, \ldots, w$. Figure 4.7 displays the updated situation.

Now that (4.2) is fully defined, the hypothesis $H_0 : \mu_1 = \mu_2$ could be tested to determine if there is association between the locus and trait. Again, the null hypothesis declares that X^Q is not a QTL while the alternative hypothesis $H_A : \mu_1 \neq \mu_2$ declares that X^Q is associated with



Line	X_1	X_2	X^Q	X_3	X_4	X_5	X_6		X_p	Trait Value
1	1	1	π_1	1	1	1	1	---------------	1	10.45
2	0	0	π_2	1	1	0	0	---------------	0	11.96
3	0	0	π_3	0	0	1	1	---------------	0	5.32
4	0	1	π_4	1	1	0	0	---------------	0	6.07
w	1	1	π_w	0	0	1	1	---------------	1	5.76

Figure 4.7 Fictional situation where the QTL is known. The genotypes at the QTL are not known, but estimated probabilities of genotype 1 (π) are known using a genetic map in conjunction with Haldane's mapping function.

the trait. A typical test statistic that is used to determine the amount of evidence against H_0 is the LOD score defined as

$$\text{LOD}= \log_{10} \frac{\sup_{H_A} L(\mu_0,\ \mu_1,\ \sigma|\mathbf{Y},\ \boldsymbol{\pi})}{\sup_{H_0} L(\mu_0,\ \mu_1,\ \sigma|\mathbf{Y},\ \boldsymbol{\pi})}$$

where $L\left(\mu_0,\ \mu_1,\ \sigma|\mathbf{Y},\ \boldsymbol{\pi}\right) = \prod_{i=1}^{w} f\left(Y_i|\pi_i,\ \mu_0,\ \mu_1,\ \sigma\right)$. Under H_0, data are a sample from a single normal population, so the maximum likelihood estimates of the mean and variance are $\hat{\mu} = \hat{\mu}_0 = \hat{\mu}_1 = \bar{Y} \equiv \frac{1}{w}\sum_{i=1}^{w} Y_i$ and $\hat{\sigma} \equiv \sqrt{\frac{1}{w}\sum_{i=1}^{w}\left(Y_i - \bar{Y}\right)^2}$. Under H_A, maximization of $L\left(\mu_0,\ \mu_1,\ \sigma|\mathbf{Y},\ \boldsymbol{\pi}\right)$ requires numerical methods. A common choice in mixture model settings is the EM algorithm.

The missing information in this case is the genotypic information at the QTL. Let $G_i = 0$ if the genotype of line i at the QTL is 0, i.e. Y_i originates from $\mathrm{N}(\mu_0, \sigma^2)$ and let $G_i = 1$ if the genotype of line i at the QTL is 1, i.e. Y_i originates from $\mathrm{N}(\mu_1, \sigma^2)$. Then, the complete data likelihood function is

$$L\left(\mu_0,\ \mu_1,\ \sigma|\mathbf{Y},\ \mathbf{G},\ \boldsymbol{\pi}\right) = \prod_{i=1}^{w} L_i\left(\mu_0,\ \mu_1,\ \sigma|Y_i,\ G_i,\ \pi_i\right) =$$
$$\prod_{i=1}^{w}\left[\left\{(1-\pi_i)\cdot\phi\left(Y_i;\mu_0,\sigma\right)\right\}^{1-G_i}\cdot\left\{\pi_i\cdot\phi\left(Y_i;\mu_1,\sigma\right)\right\}^{G_i}\right]$$

and the complete data log likelihood function is

$$\ell\left(\mu_0,\ \mu_1,\ \sigma \mid \mathbf{Y},\ \mathbf{G},\ \boldsymbol{\pi}\right) = \sum_{i=1}^{w} \ell_i\left(\mu_0,\ \mu_1,\ \sigma \mid Y_i,\ G_i, \pi_i\right) =$$
$$\sum_{i=1}^{w} \left[(1 - G_i)\left\{\log\left(1 - \pi_i\right) + \log\left(\phi(Y_i; \mu_0, \sigma)\right)\right\} + G_i\left\{\log \pi_i + \log\left(\phi\left(Y_i; \mu_1, \sigma\right)\right)\right\}\right].$$

Let $\boldsymbol{\theta} = (\mu_0, \mu_1, \sigma)$ and let $\boldsymbol{\theta}^{(t)}$ denote the result of the t^{th} iteration of the EM algorithm where $\boldsymbol{\theta}^{(0)}$ denotes the staring value. Then, determine $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$ by iterating the E-step and M-step described next. It can be shown that the E-step is

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \sum_{i=1}^{w} \mathrm{E}\left[\ell\left(\mu_0,\ \mu_1,\ \sigma \mid \mathbf{Y},\ \mathbf{G},\ \boldsymbol{\pi}\right)\right] \\ &= \sum_{i=1}^{w} \left[\left(1 - G_i^{(t+1)}\right)\left\{\log\left(1 - \pi_i\right) + \log\left(\phi\left(Y_i; \mu_0, \sigma\right)\right)\right\} + G_i^{(t+1)}\left\{\log \pi_i + \log\left(\phi\left(Y_i; \mu_1, \sigma\right)\right)\right\}\right] \end{aligned}$$

for

$$G_i^{(t+1)} = \mathrm{P}\left(G_i = 1|Y_i,\ \pi_i,\ \boldsymbol{\theta}^{(t)}\right) = \frac{\pi_i \cdot \phi\left(Y_i; \mu_1^{(t)}, \sigma^{(t)}\right)}{\pi_i \cdot \phi\left(Y_i; \mu_1^{(t)}, \sigma^{(t)}\right) + (1 - \pi_i) \cdot \phi\left(Y_i; \mu_0^{(t)}, \sigma^{(t)}\right)}$$

and simple calculus shows that the M-step yields

$$\begin{aligned} \mu_0^{(t+1)} &= \frac{\sum_{i=1}^{w}\left(1 - G_i^{(t+1)}\right)Y_i}{\sum_{i=1}^{w}\left(1 - G_i^{(t+1)}\right)} \\ \mu_1^{(t+1)} &= \frac{\sum_{i=1}^{w} G_i^{(t+1)} Y_i}{\sum_{i=1}^{w} G_i^{(t+1)}} \\ \sigma^{(t+1)} &= \sqrt{\frac{1}{w}\sum_{i=1}^{w}\left[\left(1 - G_i^{(t+1)}\right)\left(Y_i - \mu_0^{(t+1)}\right)^2 + G_i^{(t+1)}\left(Y_i - \mu_1^{(t+1)}\right)^2\right]}. \end{aligned}$$

Repeat execution of the E-step and M-step until a convergence criterion is satisfied can be used to obtain the maximum likelihood estimates of $\boldsymbol{\theta}$.

Many convergence criteria are possible, including what seems as a logical choice to stop the algorithm when the Euclidean distance between the parameter vector in successive steps or when a change in the likelihood is sufficiently small. However, when the likelihood surface is relatively flat, a small change in the parameters from one iteration to the next could easily mislead someone to thinking that these estimates correspond to the solution of the maximum of the likelihood. Also misleading, different values of the parameters could lead to the very similar values of the likelihood. To avoid these potential problems, Böhning, Dietz, Schaub, Schlattman, and Lindsay (1994) use the *Aitken Acceleration* to predict the value of the log likelihood at the maximum likelihood solution. This is particularly applicable when the algorithm is linearly convergent and has a slow rate of convergence, as is the case with the EM algorithm. To define the stopping rule, let ℓ_{i-2}, ℓ_{i-1}, and ℓ_i be log likelihood values for three consecutive steps of the algorithm. The final value of the log likelihood can be predicted with

$$\ell_\infty = \ell_{i-2} + \tfrac{1}{1-c_i}\left(\ell_{i-1} - \ell_{i-2}\right)$$

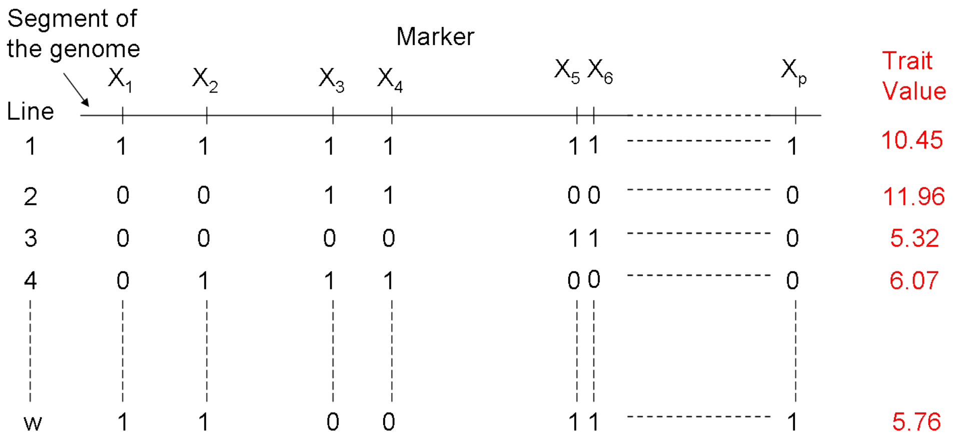
where $c_i = \frac{\ell_i - \ell_{i-1}}{\ell_{i-1} - \ell_{i-2}}$. Iterating stops if

$$\ell_\infty - \ell_i < \epsilon$$

for some small ϵ.

4.3.3 QTL Location Unknown and QTL Genotypes Unknown

In the true QTL mapping scenario, a QTL may or may not exist. If there is a QTL, its location is unknown, and the genotypes at the unknown QTL are unobservable unless the QTL happens to coincide with a marker. Figure 4.8 depicts the actual situation.



Line	X_1	X_2	X_3	X_4	X_5	X_6		X_p	Trait Value
1	1	1	1	1	1	1	----------------	1	10.45
2	0	0	1	1	0	0	----------------	0	11.96
3	0	0	0	0	1	1	----------------	0	5.32
4	0	1	1	1	0	0	----------------	0	6.07
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	
w	1	1	0	0	1	1	----------------	1	5.76

Figure 4.8 The true situation. The QTL is unknown.

To formalize the notion of QTL existence versus nonexistence, we introduce some additional notation. Let $\mathcal{G}$ denote the set all loci on the genome. For any $X \in \mathcal{G}$, define $\mu_0(X)$ and $\mu_1(X)$ to be the mean trait values for individuals with genotypes 0 and 1, respectively. At locus X, the null hypothesis of no QTL is then $H_0 : \mu_0(X) = \mu_1(X) \; \forall \; X \in \mathcal{G}$. The alternative corresponding to the existence of a QTL is $H_A : \mu_0(X) \neq \mu_1(X)$ for some $X \in \mathcal{G}$. To test H_0 and locate the QTL if it exists, LOD scores are computed at loci closely spaced throughout the genome. These LOD scores, denoted $\text{LOD}_1, \ldots, \text{LOD}_K$, correspond the testing positions $X_1^*, \ldots, X_K^*$. The candidate QTL is taken as the locus with the largest LOD score across testing positions. To determine if the candidate QTL is significant, the test statistic $\max_{1 \leq k \leq K} \text{LOD}_k$ can be used to assess the evidence of H_0. To analytically determine the distribution of $\max_{1 \leq k \leq K} \text{LOD}_k$ under H_0 is not an easy task since the LOD scores are dependent. Therefore, Churchill and Doerge (1994) proposed a shuffling approach where the reference distribution is generated through permutation.

To test H_0 at level γ, the idea is to compare the actual $\max_{1 \leq k \leq K} \text{LOD}_k$ score to the (1-γ) quantile of the $w!$ $\max_{1 \leq k \leq K} \text{LOD}_k$ scores computed for each of the $w!$ assignments of trait values $Y_1, \ldots, Y_w$ to the observed genotypic information of the w lines. As discussed in the previous section, the EM algorithm (or some alternative iterative numerical procedure) is required to compute each LOD score. Thus, generating all $w!$ $\max_{1 \leq k \leq K} \text{LOD}_k$ scores is computationally undesirable. Instead, a valid test can be based on only a random sample of the $w!$ possible arrangements of trait values to lines. The 1-γ quantile of the $\max_{1 \leq k \leq K} \text{LOD}_k$ values that arise from the analysis of the random sample of the $w!$ assignments can serve as an estimate of the 1-γ quantile of all $w!$ $\max_{1 \leq k \leq K} \text{LOD}_k$ values.

4.4 Multiple Trait Statistical Model

One drawback to using permutation to assess significance in QTL mapping studies is that it can be computationally intensive as the testing positions are typically located every 1 cM apart, resulting in hundreds or thousands of testing positions. Since each LOD score (computed at each testing position for each permutation) requires the EM algorithm, the total number of times the EM algorithm is employed can be $\sim 10^6$. For a single trait analysis, this is not necessarily attractive, but if is affordable with basic computing equipment and patience. On the other hand, if there are multiple traits to analyze, the amount of computation required can become very unattractive.

Expression QTL (eQTL) mapping studies is one such multiple trait case where the expression of a gene is treated as a single trait, and expression values are available on tens of thousands of genes. Hence, the number of times the EM algorithm (or similar iterative numerical procedure) is employed is $\sim 10^{10}$. Some of this computational expense is not worth the investment as there will not be significant evidence of a QTL for many of the thousands of genes. Hence, this would be an ideal situation to employ the sequential p-values of Besag and Clifford (1991) described next.

4.4.1 Sequential P-values

Suppose a test statistic Z has some distribution ψ under a null hypothesis H_0 and that large values of Z provide evidence against H_0. Given an observed test statistic z, suppose $1 - \psi(z)$ cannot be computed in a straightforward manner, but we are able to draw values from ψ, namely $z_1, \ldots, z_{n-1}$. These values can be used to approximate the true p-value $1-\psi(z)$ by $\frac{\sum_{i=1}^{n-1} I(z \leq z_i)+1}{n}$. Conducting a permutation test uses this Monte Carlo idea, especially when the computational burden is too great to produce the entire permutation distribution.

Suppose n is chosen to be 1,000. Regardless of the computational expense, the amount of evidence against H_0 maybe be substantially lacking after relatively few draws from ψ. For example, suppose after the first 49 draws from ψ, $\frac{\sum_{i=1}^{49} I(z \leq z_i)+1}{50} = 0.88$, or a similar value close to one. At this point, little information is gained by continuing to draw the remaining 950 values from ψ. Besag and Clifford (1991) recognized this, and proposed sampling (up to) $n-1$ values, but if at any point during sampling, h of the values are strictly larger than z, the sampling would terminate early as there is lack of evidence against H_0. The p-value under this scheme is defined as

$$p = \left\{ \begin{array}{ll} h/L & \text{if } G = h, \\ (G+1)/n & \text{if } G < h \end{array} \right\},$$

where L denotes the number of values sampled at termination and G denotes the number of sampled values that are strictly larger than the observed test statistic z.

Since this is a Monte Carlo procedure, the p-value can only take on a finite number of values which clearly depend upon the values of h and n and can described by the set $S(h,n) = \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{h-1}{n}, \frac{h}{n}, \frac{h}{n-1}, \ldots, \frac{h}{h+1}, 1\right\}$. These possible p-values possess the property $\mathrm{P}\left(p \leq p^*\right) = p^*$ for all $p^* \in S(h,n)$, but since the elements of $S(h,n)$ are not equally spaced, the null probabilities of the elements of $S(h,n)$ are not equal. The relevance of these characteristics will be evident in the next section.

4.4.2 False Discovery Rate

Suppose m hypotheses are tested, e.g. a QTL is attempted to be located for each of m traits of interest, based on the corresponding ordered continuous p-values $p_{(1)}, \ldots, p_{(m)}$ Table 4.1 summarizes the outcomes of this situation.

	Declare Non-Significant No Discovery Negative Result	Declare Significant Declare Discovery Positive Result	Total
True Nulls	U	V	m_0
False Nulls	T	S	$m - m_0$
Total	W	R	m

Table 4.1 Table of outcomes when testing m null hypotheses.

The quantities m and m_0 are constants with m known while m_0 is unknown. The quantities U, V, T, S, W, and R are all random variables with W and R observable, while U, V, T, and S are unobservable. Note that these random variables are all functions of a given significance threshold c, for $c \in [0,1]$. Since the unobservable random variable V is the number of false discoveries, most multiple testing error rates attempt to control the size of V, or some function thereof. Benjamini and Hochberg attempt to control the rate of false findings, not just V itself, by controlling the expectation of the random variable Q, where

$$Q = \left\{ \begin{array}{ll} V/R & \text{if } R > 0, \\ 0 & \text{if } R = 0 \end{array} \right\}.$$

This quantity describes the ratio of the number of falsely rejected hypotheses to the total number of rejections, and thus, Benjamini and Hochberg (1995) define the false discovery rate (FDR) as E$[Q]$. They show that rejecting the k hypotheses $H_{(01)}, \ldots, H_{(0k)}$, where

$$k = \max\left\{k^* : \frac{p_{(k^*)}m}{k^*} \leq \alpha;\ k^* = 1, \ldots, m\right\} \tag{4.4}$$

for $\alpha \in (0,1)$, will control FDR at level α. Alternatively, instead of declaring a level at which the FDR is to be controlled and finding the associated p-value cutoff using (4.4), one could declare a significance threshold c, $c \in (0,1)$, and find the associated FDR using c as the p-value cutoff. To do this, (4.4) can be alternatively expressed as

$$\widehat{\text{FDR}}(c) = \min\left\{\frac{m \cdot c^*}{\#\ p\text{-values}\ \leq\ c^*},\ \forall\ c^* \geq c\right\}. \tag{4.5}$$

The numerators in (4.4) and (4.5) are the expected number of type I errors if all null hypotheses are true, and the denominator is simply the total number of rejections. Hence, if $m_0 < m$, then the numerators in (4.4) and (4.5) overestimates the number of type I errors, which produces unnecessarily conservative control and estimation of FDR, respectively. In other words, if FDR is to be controlled at level α, then replacing the unknown quantity m_0 for m in (4.4) and (4.5) will result in a list of discoveries at least as large compared to a list of discoveries obtained when using m in (4.4) or (4.5). Given an estimate of m_0, approximate control of FDR at level α can be obtained. Similarly, for any given $c \in (0,1)$, the estimated FDR defined by (4.5) will be no larger than the estimated FDR if m had not be replaced with $\widehat{m}_0$.

Estimating m_0 has been given much attention in the past decade (see, for example, Benjamini and Hochberg, 2000; Storey, 2000a,b; Mosig et al., 2001; Storey and Tibshirani, 2003; Langaas, Lindqvist, and Ferkingstad, 2005; Nettleton et al., 2006; Ruppert, Nettleton, and Hwang, 2007) and most methods assume each p-value is distributed continuous uniform on (0,1) under its null hypothesis. If the m hypotheses were tested using the sequential analysis of Besag and Clifford (1991), then the existing methods to estimate m_0 cannot be directly applied as these sequential p-values have unequal null probabilities. Bancroft and Nettleton (2009) propose a histogram based estimator to estimate the number of true null hypothesis when the collection of p-values have any discrete distribution.

To account for the fact that the p-values have a discrete support, redefine the estimated FDR only for $c \in S(h,n)$ as

$$\widehat{\mathrm{FDR}}(c) = \min\left\{ \frac{\hat{m}_0 \cdot c^*}{\#\ p\text{-values} \ \leq \ c^*} : \ c^* \in S(h,n) \cap [c,1] \right\}. \tag{4.6}$$

In summary, in multiple trait QTL mapping studies, such as eQTL mapping studies, the expression of each gene is treated as a single trait and a single trait analysis is applied to each gene. The sequential analysis of Besag and Clifford (1991) is employed to substantially reduce the computation required. Since the number of genes investigated is typically in the tens of thousands, a multiple testing correction has to be made. A variation of the false discovery rate of Benjamini and Hochberg (1995) will summarize the rate of false findings by using an estimate of m_0 instead of m in (4.4). The histogram based approach of Bancroft and Nettleton (2009) will be employed to estimate m_0. This procedure is demonstrated in the following section.

4.5 Case Study

Two parent lines of barley, where one is resistant to the fungus *Puccinia graminis* f. sp. *tritici* (*Pgt*) and one is susceptible, were crossed producing $w = 75$ doubled haploid lines. These lines were inoculated (intentionally infected with the fungus) and mRNA abundance was measured on each of $m = 22,860$ genes for each of the w lines for both treatments. The genetic map consisted of genotype information at $p = 378$ markers over the 7 chromosomes. The testing positions were located every 1cM. For each gene, the proposed sequential permutation procedure with $h = 10$ and $n = 1,000$ was employed to determine the significance of the locus with the maximum LOD score across markers. Haldane's mapping function was used to compute conditional probabilities of the genotype at each testing position. These probabilities were used as initial probabilities of group indicators to generate starting values for the EM algorithm. The estimated m_0's along with the number of rejections, $R(c)$, and the estimated FDR, $\widehat{\mathrm{FDR}}(c)$, are given in Table 4.2 for four different significance thresholds.

Table 4.2 shows that we can obtain a list of 3841 genes by using a significance threshold of $c = 0.01$ if we are willing to tolerate an estimated FDR around 5%. To better understand

$\widehat{m}_0/m$	c	0.001	0.005	0.01	0.05
0.84	$\widehat{\mathrm{FDR}}(c)$	0.0069	0.0277	0.0500	0.1670
	$R(c)$	2780	3458	3841	5745

Table 4.2 The estimated proportion of true null hypotheses ($\widehat{m}_0/m$), the number of rejections ($R(c)$), and the estimated false discovery rate ($\widehat{\mathrm{FDR}}(c)$) for each of four significance thresholds (c) based on a sequential permutation analysis of data from an eQTL mapping study.

how a locus is declared to be an eQTL based on the data, Figure 4.9 displays plots for two genes. The top two plots correspond to a gene whose expression is significantly ($p = 0.001$) mapped to a locus, while the bottom two plots correspond to a gene whose expression is not significantly ($p = 1$) mapped to a locus. The left-hand column shows plots of LOD scores versus testing positions. The right-hand column shows plots of $\log_2$ expression values versus genotype (based on the marker closest to the candidate eQTL).

Figure 4.9 shows that the LOD score for a significant candidate locus is much larger compared to LOD scores computed at markers not close to the candidate locus whereas the LOD score for a non-significant locus is similar to LOD scores computed at other loci. We can also see that the distributions of expression values are well separated based on the genotypic information for genes with a significant candidate locus whereas the distributions of expression values for genes with a non-significant candidate locus are hard to distinguish from one another. The genotypic information used here is based on the marker closest to the candidate eQTL position. It turns out that some markers have also had expression values measured. Significant eQTL for these genes can be very revealing. One may expect that eQTL in this case would mapped to themselves. However, one gene may need to be expressed in order to 'turn on' an entirely different gene. To illustrate this phenomena, for significant genes ($p <= 0.05$) whose location is known and had expression values measured, the actual position versus the mapped position are plotted in Figure 4.10. For this subset of genes, Figure 4.10 shows how a genes expression may be controlled by a gene on a completely different chromosome.

The authors propose eQTL mapping studies as an ideal application to use the sequential

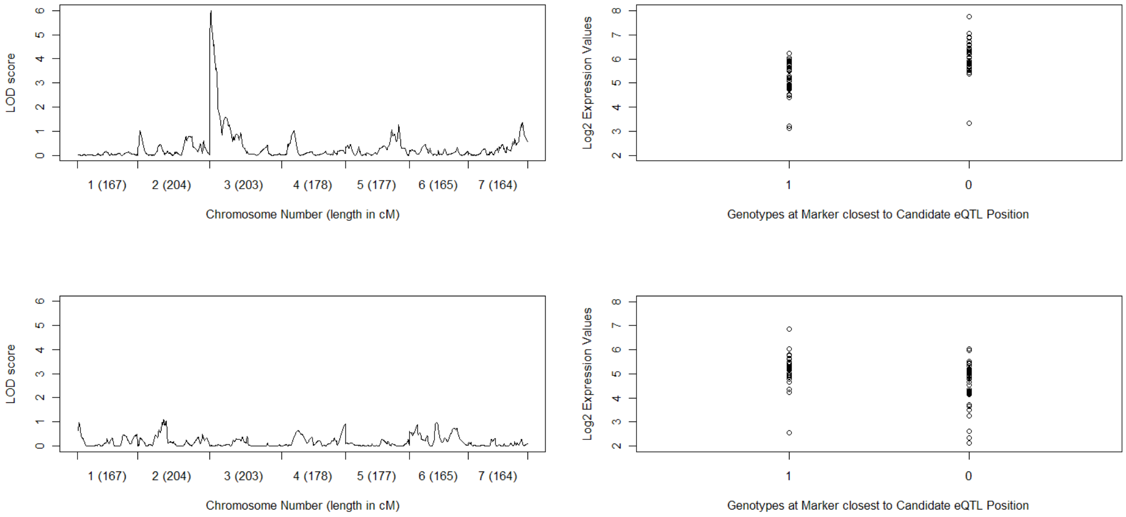


Figure 4.9 Plots of LOD scores vs testing position and expression values versus genotype. The top two plots correspond to a gene with a significantly ($p = 0.001$) mapped eQTL while the bottom two plots correspond to a gene with a non-significantly ($p = 1$) mapped eQTL. Significance was determined through a sequential permutation approach.

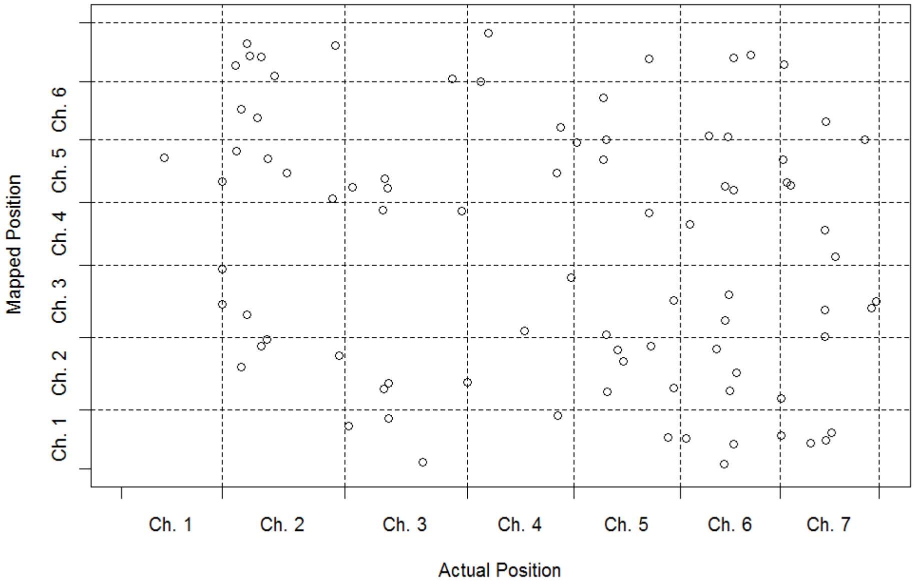


Figure 4.10 Plot of known gene position vs the position of its mapped eQTL for significant genes ($p <= 0.05$).

procedure described in section 4.1 as computing $\max_{1\leq k\leq K} \mathrm{LOD}_k$ is computationally intensive. If, for each gene, the $\max_{1\leq k\leq K} \mathrm{LOD}_k$ was compared to 1,000 draws from its permutation distribution (999 from the distribution plus the actual $\max_{1\leq k\leq K} \mathrm{LOD}_k$), then the total number of test statistics (each requiring K calls to the EM algorithm) computed would be $1000 \cdot 22840 = 22,840,000$. By employing the described sequential procedure, the total number of test statistics computed for was 5,331,831, a savings of around 75%.

4.6 Simulation Study

4.6.1 Simulation Setup

The simulation study is setup into two parts. The first is simulating the genetic map and the genotypes for all w lines at the simulated marker positions. The second is to simulate gene expression data for m genes where there is not QTL for m_0 of the genes and there is a QTL for $m - m_0$ genes.

A genetic map is first simulated for a single chromosome of fixed length, P, i.e. $p-1$ marker positions are randomly simulated from a uniform distribution on the interval $[0,P]$ and position 0 is taken as the first marker for a total of p markers. Then, starting at position 0, one of two genotypes are randomly simulated with equal probability for each of w lines. Using Haldane's mapping function, genotypes are simulated at each of the remaining $p-1$ markers for each of the w lines, but based only on the previous marker. That is, suppose marker X_k has simulated genotype 0 for line 1, and the next simulated marker X_{k+1} is located 1.5cM from marker k. To simulate the genotype at marker $k+1$ for line 1, the probability of a recombination, i.e. conditional probability of genotype 1, is $\frac{1}{2}\left[1 - \exp(-2(0.015 - 0))\right] = 0.0148$. Hence, marker $k+1$ is simulated to be genotype 1 with probability 0.0148 and is simulated to be genotype 0 with probability of $1-0.0148 = 0.9852$. This process is repeated sequentially for $k = 1, \ldots, p-1$ for all w lines.

Simulated gene expression should reflect the fact that QTL may not exist for some genes in a mapping population. Therefore, m_0 of the genes were simulated with no QTL while there was a QTL for the other $m - m_0$ genes. For each of the m_0 genes with no QTL, a candidate QTL position was simulated on the chromosome, but the w expression values were all simulated from one single distribution. For each of the $m - m_0$ genes with a QTL, a candidate QTL position was simulated on the chromosome. Then, Haldane's mapping function was used to compute the probability of each genotype based on both flanking markers (if the QTL was simulated after the last simulated marker, then the single flanking marker was used) for each of the w lines. Next, the w genotypes at the QTL were simulated based on the previously calculated probabilities obtained using Haldane's mapping function. If the simulated genotype for line i was 0, the expression (trait) value was drawn from $\mathrm{N}\big(\mu_0, \sigma^2\big)$ and if the simulated genotype for line i was 1, simulate the expression (trait) value was from $\mathrm{N}\big(\mu_1, \sigma^2\big)$. The values of μ_0, μ_1, and σ^2 are gene specific and are chosen based on the Heretability index.

Heretability is defined as the ratio of genotypic variance to phenotypic variance, which in this case is equivalent to

$$H = \frac{\frac{1}{4}\left(\mu_0 - \mu_1\right)^2}{\sigma^2 + \frac{1}{4}\left(\mu_0 - \mu_1\right)^2}. \tag{4.7}$$

H is a function of μ_0 and μ_1 only through the difference $|\mu_0 - \mu_1|$. Thus, setting μ_0 equal to 0 and solving equation (4.7) for μ_1 as a function of σ^2 and H, we have

$$\mu_1\left(H, \sigma^2\right) = \frac{H\sigma^2}{\frac{1}{4}(1-H)}. \tag{4.8}$$

So, given a value for H and σ^2, we have a value for μ_1. To choose H and σ^2, note that for the data in the Case Study section, the maximum likelihood estimates of μ_0, μ_1, and σ^2, and therefore H, were recorded for each gene-by-marker combination. A histogram of the heretability values suggests that their distribution can be approximated by a gamma distribution. Let $\tilde{\alpha}_H$ and $\tilde{\beta}_H$ be method of moments estimates of the shape and scale parameters of a gamma distribution corresponding to the heretability values. A histogram of the σ^2 values suggests that their distribution can also be approximated by a gamma distribution. Let $\tilde{\alpha}_\sigma$ and $\tilde{\beta}_\sigma$ be method of moments estimates of the shape and scale parameters of a gamma distribution corresponding to the σ^2 values. Thus, for each gene that was simulated to have a QTL, a heretability value was drawn from $\Gamma(\tilde{\alpha}_H, \tilde{\beta}_H)$ and a variance was drawn from $\Gamma(\tilde{\alpha}_\sigma, \tilde{\beta}_\sigma)$. Then, (4.8) was used to obtain μ_1 recalling that μ_0 was set to 0.

Once the genetic map, genotypes for all w lines, and expression values for m genes were simulated, a single trait analysis was performed on the expression values for each of the m genes. The significance of the identified candidate QTL is assessed through permutation, using the sequential permutation p-value approach of B & N (2009) to reduce the number of permutations required in building the reference distribution. This yielded a collection of m p-values, some of which indicated a presence of a eQTL when no eQTL was present, i.e. false discovery. An estimate of m_0, obtained via the histogram based estimator, was used in the variation of Benjamini and Hochberg (1995) defined by (4.5) to estimate FDR. This entire simulation process was repeated N times, and the average FDR was compared to the true false positive fraction for different values of c.

The settings for the simulation are as follows. The simulation was executed for N=1,000 simulation runs, each of which simulated data for $m = 1,000$ genes, $p = 29$ markers along one chromosome of length $P = 1$M (to match the map density in the case study), and $w = 75$ lines. The case study data yielded $\tilde{\alpha}_H = 3.024$, $\tilde{\beta}_H = 0.116$, $\tilde{\alpha}_\sigma = 0.505$, and $\tilde{\beta}_\sigma = 0.437$. The number of genes whose expression values had a QTL was 200. The testing positions, the X_k^*'s, were located every 1cM along the chromosome. Hence, there were $K = 100$ testing positions. The convergence criterion for the EM algorithm was $\epsilon = 0.0001$. The starting values used for each employment of the EM algorithm were based on the given probabilities of genotype 1 calculated using the genetic map. Those gene-by-permutation-by-testing position specific probabilities were used as the $G_i^{(0)}$ values to find $\mu_0^{(0)}$, $\mu_0^{(0)}$, and $\sigma^{(0)}$, defined in section 3.2, which were used as gene-by-permutation-by-testing position specific starting values for the EM algorithm. The values used in the sequential procedure of Besag and Clifford (1991) were $h = 10$ and $n = 1,000$.

4.6.2 Simulation Results

Table 4.3 shows the estimated FDR and the true false positive fraction, $V(c)/\max\{1, R(c)\}$, each averaged over the $N = 1,000$ simulation runs. Also displayed are their standard errors and similar quantities for the number of rejections, $R(c)$, for $c = 0.001, 0.005, 0.01$, and 0.05.

	$R(c)$		$\widehat{\mathrm{FDR}}\,(c)$		$V(c)/\max\{1, R(c)\}$	
c	mean	se	mean	se	mean	se
0.001	96.44	0.8351	0.0089	0.0001	0.0082	0.0009
0.005	138.92	0.7874	0.0308	0.0002	0.0279	0.0014
0.010	156.88	0.7358	0.0544	0.0003	0.0493	0.0017
0.050	216.56	0.7193	0.1969	0.0008	0.1872	0.0023

Table 4.3 Average estimated FDR and average false positive fraction along with their respective standard errors over N=1,000 simulation runs where each run consisted of testing m=1,000 null hypotheses of no eQTL present.

4.7 Conclusion

This paper discusses how to dramatically reduces the computation required to analyze eQTL data using permutation methods. This paper employs the procedure of Bancroft and Nettleton (2009) to reduce the size of the permutation distribution used to assess the significance of the maximum LOD score across markers for each gene. Their procedure was shown here to control FDR through simulation in eQTL scenarios.

4.8 References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

[2] Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

[3] Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*, **Vol. 78, No. 2**, 301-304.

[4] Bhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B.G. 1994. The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-Parameter Exponential Family. *Annals of the Institute of Statistical Mathematics*, **Vol. 46, No. 2**, 373-388.

[5] Chorney, M.J., et al. (1998). A Quantitative Trait Locus Associated with Cognitive Ability in Children. *Psychological Science*, **Vol. 9, Iss. 3**, 159-166.

[6] Churchill, G. A., and Doerge, R. W. (1994). Empirical Threshold Values for Quantitative Train Mapping. *Genetics*, **138**, 963-971.

[7] Lander, E.S., and Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, **121**, 185-199.

[8] Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B* **67**, 555–572.

[9] Mosig, M.O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Per-

centage in Israeli-Holstein Cattle by Means of Selective Milk DNA Pooling in a Daughter Design Using an Adjusted False Discovery Rate Criterion. *Genetics*, **151**, 1683-1698.

[10] Nettleton, D., Hwang, J.T.G., Caldo, R.A., and Wise, R.P. (2006). Estimating the Number of True Null Hypotheses From a Histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, **Vol. 11, No. 3**, 337-356.

[11] Paterson, A., et al. (1988). Resolution of Quantitative Traits into Mendelian Factors by using a Complete RFLP Linkage Map. *Nature*, **335**, 721-726.

[12] Ruppert, D., Nettleton, D., Hwang, J.T.G. (2007). Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics*, **63**, 483–495.

[13] Sax, K. (1932). The Association of Size Differences with See-Coat Pattern and Pigmentation in $Phaseolus\ Vulgaris$. *Genetics*, **8**, 552-560.

[14] Storey, J.D. (2000a). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*, Series B, **64**, 479-498.

[15] Storey, J.D. (2000b). False Discovery Rates: Theory and Applications to DNA Microarrays. Unpublished Ph.D. thesis, Department of Statistics, Stanford University.

CHAPTER 5. Summary

5.1 Conclusion

This dissertation has introduced methodology to estimate the number of true null hypothesis in a multiple testing framework given a collection of p-values with a mixture of discrete non-uniform distributions under the null hypothesis. With this estimate, a less conservative estimate of the false discovery rate can also be obtained. There is more than one way in which a multiple testing scenario can bring about a collection of p-values with a mixture of discrete non-uniform distributions under the null hypothesis. One example is when permutation testing is applied to data with many tied observations. Because of the tied observations, some permutations of the data will produce the same value of the chosen test statistic leading to unequal null probabilities for the possible p-values. The number of tied observations play a role in determining the support of the p-value. This phenomena is also seen when the data set is comprised of many 2 by 2 contingency tables and Fisher's Exact test is used to assess dependency between the two categorical variables for each 2 by 2 table. Data that can be summarized in a 2 by2 table can be thought of, in a sense, as data with many ties since there are only two possible responses. Another scenario that has been discussed where p-values have the said properties is when the sequential analysis of Besag and Clifford (1991) is employed to reduce the size (and therefore computational expense) of the permutation distribution. This last case is particularly applicable when the test statistic is compuationally burdensome to compute, e.g. a maximum LOD score (each LOD score requires the EM algorithm) across hundreds of markers as seen in Chapter 4. Some ideas for potential future research are discussed next.

5.2 Future Work

5.2.1 Drawbacks and a Potential Improvement to the Histogram Based Estimator of the Number of True Null Hypotheses

One topic that is seen throughout this dissertation is estimating the number of true null hypotheses using the histogram based estimator given a collection of discrete p-values. One of the following examples shows a case where the estimator can be extremely conservative. The other example shows a possible way to reduce the conservative bias. First, suppose we have a collection of p-values, each of which belongs to one of n unique discrete non-uniform distributions under its null hypothesis. Say only one of these p-values belongs to the support $S_1 = \{0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.50, 1\}$. Note that if the lone p-value equals any value in S_1 other than 0.01, the histogram based estimator would estimate that the number of true null hypotheses is one. But, clearly, if the p-value is 0.02, there is some evidence that this p-value may correspond to a false null hypothesis.

Consider a more extreme example. Suppose again that we have a collection of p-values, each of which belongs to one of n unique discrete non-uniform distributions under its null hypothesis. Say one of those supports is S_2. The support, the observed frequencies for the subset of the 50 p-values that belong to S_2, and the 'tail expectations' are given below in Table 5.1.

p-value	.0001	.05	.36	.52	1
Observed Frequency	0	40	7	2	1
'Tail Expectation'	.005	2.495250	3.263158	0.75	1

Table 5.1

In this situation, the histogram based estimator yields $\widehat{m}_{02} = 50$. To see this, the first 'bin' whose observed frequency does not exceed its 'tail expectation' is the bin corresponding to a p-value of 0.0001, and hence, $\widehat{m}_{02} = \frac{50}{1} = 50$. But, note that even under the alternative hypothesis, the power may not be great enough to observe a p-value of 0.0001 in 50 tries. Clearly, however, there is evidence of some false null hypotheses as there are many p-values that equal the second smallest element in S_2. If all 50 p-values correspond to null hypotheses, then the expected number of p-values that equal 0.05 would be $(0.05 - 0.0001) \cdot 50 = 2.495$, a value much less than the observed frequency of 40, providing strong evidence that many of the 50 hypotheses tested are false null hypotheses. To circumvent the problem that none of the p-values equal the smallest element in S_2, we could combine the first two 'bins' and Table 5.1 would then look like

p-value	.0001 or .05	.36	.52	1
Observed Frequency	40	7	2	1
'Tail Expectation'	2.5	3.263158	0.75	1

Table 5.2

To estimate m_{02}, find the first bin whose observed frequency does not exceed its 'tail expectation.' Here, it is the 'bin' whose corrsponding p-value equals 1, and hence, $\widehat{m}_{02} = \frac{1}{0.48} = 2.08333$. Clearly, the estimates for m_{02} differ greatly depending on if the 'bins' are collapsed. And since the estimates of m_0 via the histogram based estimator are valid (i.e. have been shown to be slightly conservative in the simulations studies in Chapters 2, 3, and 4), the question remains whether or not to combine bins, and if so, which 'bins' should be combined and how do the estimates of m_0 change depending on which and how many 'bins' are collapsed.

Reducing the conservative bias of $\widehat{m}_0$ can be critical in the sense that it is used in the adaptive variation of Benjamini and Hochberg (1995), i.e. the list of discoveries obtained at the same level of FDR control grows larger as the conservative bias of $\widehat{m}_0$ decreases.

5.2.2 Assessing Significance for the Second most Associated Locus with the Trait in Question in eQTL Mapping Studies

It is well know that biological systems are very complex and are controlled by many different genes and proteins working simultaneously. In Chapter 4, we discuss how to identify and assess the significance of a single locus whose genotypic information is most associated with the trait in question. But since biological functions are usually controlled by more than one gene, it may also be of interest to identify and assess significance for a second locus whose genotypic information displays the second most associated with the trait in question. It should be noted that when we refer to the locus with the second most association, we really mean the locus with the second local maximum of the LOD score across testing positions. The locus with the second most association will typically be near the locus taken as the candidate QTL. We are interested in finding a second locus on the genome that is not necessarily close to the intially identified locus. Analytically determining the reference distribution in this case is not tractable (as it is also not tractable when determining the signficance of the locus that shows the most association with the trait in question) and to use permutation to generate the reference distribution does not provide a clear solution as it does when determining the significance for the locus with the most association with the trait in question.

In light of the point made above, consider assesing significance for the first and second largest LOD scores. Using the following general idea to assess significance for the locus with the second local maximum of LOD scores across markers would also encounter the following difficulty. Perhaps one may immediately think to record, for each permutation, the largest LOD score and the second largest LOD score. Then, one could compare the actual largest LOD score to the permutation distribution of the max LOD score and the actual second largest LOD score to the permutation distribution of the second largest LOD score. There is a problem with this however. Suppose we have two test statistics Z_1 and Z_2 whose null distributions are both N(0,1) and let $W_1 = \max\{|Z_1|, |Z_2|\}$ and $W_2 = \min\{|Z_1|, |Z_2|\}$. To generate our reference distributions, sample $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ iid from N(0,1) and take $U_i = \max\{|X_i|, |Y_i|\}$ and $V_i = \min\{|X_i|, |Y_i|\}\,,\ i = 1, \ldots, n$. Then the reference distribution for W_1 is the empirical distribution of $U_1, \ldots, U_n$ and the reference distribution for W_2 is the empirical distribution of $V_1, \ldots, V_n$. Now, if both hypotheses are null, these reference distributions will yield p-values with the appropriate properties, e.g. uniformity. But, if one null hypothesis is true and one null hypothesis is false, i.e. $Z_1 \sim N(0,1)$ and $Z_2 \sim N(10,1)$, then the empirical distribution of $V_1, \ldots, V_n$ is not the appropriate reference distribution for $W_2 = \min\{|Z_1|, |Z_2|\}$ as this procedure would not have the correct size. To see this, note that W_2 will take on the value of $|Z_1|$ with high probability and so $W_2 \overset{d}{\approx} |Z_1|$=N(0,1) which is stochastically larger than the distribution of the minimum of two indepedent N(0,1) absolute values.

Once the appropriate reference distribution is identified for the locus with the second local maximum of the LOD score across testing positions, then ideally that could be extended to obtain appropriate reference distributions for the 3rd largest LOD score, 4th largest LOD score, etc.