Biometrics 69, 1–7
March 2013

DOI: 10.1111/j.1541-0420.2012.01825.x

Estimation of False Discovery Rate Using Sequential Permutation *p*-Values

Tim Bancroft,[1,*] Chuanlong Du,[2,] and Dan Nettleton[2,***]**

[1]Health Economics and Outcomes Research, OptumInsight, Eden Prairie, Minnesota 55344, U.S.A.
[2]Department of Statistics, Iowa State University, Ames, Iowa 50011-1210, U.S.A.
[*]*email:* Timothy.Bancroft@optum.com
[**]*email:* dclong@iastate.edu
[***]*email:* dnett@iastate.edu

Summary. 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 nonuniform 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 significance thresholds. We use real data analyses and simulation studies to evaluate and illustrate the performance of our proposed approach relative to standard, more computationally intensive strategies. We find that our sequential approach produces similar results with far less computational expense in a variety of scenarios.

Key words: Expression quantitative trait locus; Gene expression; Monte Carlo testing; Multiple testing; Sequential analysis.

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 $\binom{20}{10}/2 = 92,378$. An exact permutation (EP) 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 50th 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 (SP) p-value resulting from this procedure has a nonuniform 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 SP 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 separate permutation test, 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 problem of testing m null hypotheses with m separate SP tests is the main focus of this article.

Our work is primarily motivated by applications in gene expression 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, e.g., Benjamini and Hochberg, 2000; Storey, 2000; Mosig et al., 2001; Storey, 2002; Storey and Tibshirani, 2003, hereafter ST; Langaas, Lindqvist, and Ferkingstad, 2005; Nettleton et al., 2006, hereafter NHCW; Ruppert, Nettleton, and Hwang, 2007; Liang and Nettleton, 2012). 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 SP p-values do not have this uniformity property, a new approach for estimating m_0 and FDR from SP p-values is needed. To address this problem, we extend the histogram-based estimator of NHCW to obtain an estimator of m_0 and use a variation of Benjamini and Hochberg's (1995) procedure to estimate FDR from discrete p-values whose distribution is nonuniform under the null hypothesis.

Section 2 provides a formal definition of SP p-values. Sections 3 and 4 discuss estimation of FDR and m_0 using these p-values. Section 5 illustrates application of the proposed methods to two real datasets and compares the performance of the proposed approach with standard alternative methods. Section 6 provides an evaluation of the proposed methods through simulation. The article concludes with a brief summary of the results in Section 7.

2. Sequential Permutation *p*-Values

Let ψ denote the cumulative distribution function of a permutation distribution that places equal probability on every element in a discrete, finite support. Without loss of generality, suppose that small 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 EP p-value $\psi(z)$ is given by $n^{-1}\{\sum_{i=1}^{n-1} I(Z_i \leq z) + 1\}$, where $Z_1, \dots, Z_{n-1}$ is a simple random sample from the support of the permutation distribution 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 the permutation distribution is large. For establishing useful properties of Monte Carlo p-values, it is convenient to assume that sampling is without replacement and that the observed statistic z is treated as the first draw so that $z, Z_1, \dots, Z_{n-1}$ are guaranteed to be distinct values from the support of the permutation distribution. Note, however, that the distinction between sampling with or without replacement is practically unimportant when the cardinality of the permutation distribution's support is large and that computational algorithms for computing Monte Carlo p-values use with-replacement sampling.

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 smaller 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

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

where L denotes the number of values sampled at termination (not counting the observed test statistic z) and G denotes the number of sampled values that are smaller than z. When the permutation distribution defined by ψ serves as the null distribution in Besag and Clifford's sequential procedure, we call p an SP p-value. Pounds et al. (2011) have defined a very similar p-value as part of what they refer to as an adaptive permutation test.

Like any p-value, p defined in (1) is a random variable that depends on the data through the test statistic Z. Like any Monte Carlo p-value, there is an additional source of randomness in p due to sampling from the permutation distribution. It is reasonable to think of p as an estimator, given the test statistic $Z = z$, of the permutation p-value $\psi(z)$ that could be obtained by examining the entire permutation distribution. However, it is not necessary to think of p as an estimator of $\psi(z)$ or a predictor of $\psi(Z)$. As we shall subsequently demonstrate, p can be used like any other p-value to conduct a valid test at any desired significance level $\alpha \in (0, 1]$. For some values of α, the resulting test will be exact while for other values of α the size of the test will be less than α. In particular, the unconditional null distribution of the SP p-value is presented in the following theorem. Note that this unconditional null distribution captures both variation in the original data and variation introduced by sampling from ψ using the sequential scheme.

THEOREM 1. *Suppose p is an SP p-value computed for given values of h and n. Let $S(h,n) = \{1/n, \dots, h/n, h/(n-1), \dots, h/(h+1), 1\}$. Under H_0,*

$$Pr(p \leq \alpha) \leq \alpha \text{ for all } \alpha \in (0, 1] \text{ and } Pr\,(p \leq \alpha) = \alpha \text{ for all } \alpha \in S(h, n). \tag{2}$$

Theorem 1 was first stated by Besag and Clifford (1991), and a proof is given in a recent paper by Silva, Assunção, and Costa (2009). Our proof of the result is provided in Web Appendix A. Theorem 1 implies that the null distribution of the SP p-value is $\Pr(p = S_j(h,n)) = S_j(h,n) - S_{j-1}(h,n)$, where $S_j(h,n)$ denotes the jth 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 article is to use SP 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 SP p-values.

3. Estimation of False Discovery Rate with Sequential Permutation *p*-Values

Suppose $p_{(1)} \leq \cdots \leq p_{(m)}$ are m ordered p-values used to test the corresponding m null hypotheses $H_{(01)}, \dots, H_{(0m)}$. 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 $m_0\gamma/m$ for any $\gamma \in (0,1)$ if $H_{(01)}, \ldots, H_{(0k)}$ are rejected, where

$$k = \max\{k^* : p_{(k^*)} m/k^* \leq \gamma\}. \tag{3}$$

When control of FDR at level γ 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 (3) to obtain control at level γ rather than $m_0\gamma/m$. Because m_0 is unknown, a natural strategy is to replace m in (3) with an estimate of m_0. This strategy was first proposed by Benjamini and Hochberg (2000) and by Storey (2002, 2003) and Storey, Taylor, and Siegmund (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 SP procedure introduced in Section 2. The corresponding m p-values each have a discrete nonuniform null distribution as described in (2). The next section focuses on estimating m_0 from a collection of such p-values.

Given an estimate of m_0 denoted as $\hat{m}_0$, we define, for any $\alpha \in S(h,n)$, an estimate of the FDR associated with rejecting all null hypotheses whose corresponding p-values are no larger than α by

$$\widehat{\mathrm{FDR}}(\alpha) = \min\{p_{(k)}\hat{m}_0/k : \; p_{(k)} \geq \alpha\}. \tag{4}$$

If (4) is calculated for each value of $\alpha \in \{p_{(1)}, \ldots, p_{(m)}\}$, the resulting values of $\widehat{\mathrm{FDR}}(\alpha)$ are essentially the q-values of Storey (2003). As noted by Storey (2002), the decision rule that rejects $H_{(0k)}$ if and only if $\widehat{\mathrm{FDR}}(p_{(k)}) \leq \gamma$ is equivalent to the decision rule obtained by replacing m with $\hat{m}_0$ in (3).

4. A Histogram-Based Estimator of the Number of True Null Hypotheses

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 number of equal-width bins) of p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, 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 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, nonuniformly distributed SP p-values, let o_j denote the number of p-values that equal $S_j(h,n)$, the jth 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), s_j is the probability under the null hypothesis that an SP p-value equals $S_j(h,n)$. Let $\hat{m}_0(i)$ denote the estimate of m_0 at iteration i. Let $e_{ij} = \hat{m}_0(i)s_j$, which is the expected number of null p-values equal to $S_j(h,n)$ based on the estimate of m_0 at iteration i. Initialize $\hat{m}_0(0) = \sum_{j=1}^{n} o_j$ and, for all $i = 0, 1, 2, \ldots,$ define

$$\hat{m}_0(i+1) = m - \sum_{j=1}^{j_i - 1}(o_j - e_{ij}), \tag{5}$$

where

$$j_i \equiv \min\{j = 1, \ldots, n : o_j \leq e_{ij}\}. \tag{6}$$

A simple example of this procedure is provided in Web Appendix B.

4.2 *Convergence Extension and the Proposed Estimator of m_0*

NHCW proved the existence of and characterized the limit of the iterative algorithm introduced in Subsection 4.1 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). The result is illustrated via application to a simple example in Web appendix B. A proof is provided in Web Appendix C.

Theorem 2. *Let* $J = \min\{j = 1, \ldots, n : \frac{o_j}{o_j + \cdots + o_n} \leq \frac{s_j}{s_j + \cdots + s_n}\}$. *Then,*

$$\hat{m}_0 \equiv \lim_{i\to\infty} \hat{m}_0(i) = \sum_{j=J}^{n} o_j \Big/ \sum_{j=J}^{n} s_j. \tag{7}$$

Given the result in Theorem 2, a natural estimator of m_0 is $\hat{m}_0 = \sum_{j=J}^{n} o_j / \sum_{j=J}^{n} s_j$. However, rather than working directly with individual p-values in the support set $S(h,n)$, we group the SP p-values in $S(h,n)$ into bins such that the probability under the null hypothesis associated with each bin is approximately 0.05. This is analogous to the recommendation of NHCW to use 20 equal-width bins when using the histogram-based approach to estimate m_0 with continuous p-values. NHCW also proposed a bootstrap approach for selecting the number of histogram bins, but ultimately recommended the simpler 20-bin estimator due to its ability to effectively manage variance and bias across a wide range of simulation settings.

Because the SP p-values have a discrete nonuniform support, it is not possible to construct 20 equally probable bins for practical choices of h and n. Thus, we use a simple algorithm described in Web Appendix D to obtain as many bins as possible that each have probability at least 0.05 under the null. We then use the estimator of m_0 given in (7) of Theorem 2, except that o_j, s_j, and n are replaced with observed number of SP p-values in the jth bin, the sum of the null probabilities associated with the values of $S(h,n)$ assigned to the jth bin, and the number of bins, respectively. We use this estimator of m_0 in (4) for FDR estimation when working with SP p-values. In the next section, we illustrate the proposed procedure in the analysis of two datasets.

5. Data-Based Comparison of the Proposed Approach with Standard Alternatives

In Subsection 5.1, we consider an example where each test involves a straightforward two-sample comparison using a computationally inexpensive test statistic. This allows us to compare the proposed SP approach with more computationally intensive strategies. In particular, we also carry out a standard Monte Carlo permutation (MP) approach that uses a total of n draws (counting the observed statistic) from the permutation distribution for every test. This is equivalent to the SP approach with $h = n$. We also conduct EP tests, where the entire permutation distribution is computed separately for each test and used as the reference distribution to obtain a p-value. In Subsection 5.2, we consider a specific example of a more general problem where the cost of computing each test statistic is less trivial and the number of distinct values of the test statistic in the permutation distribution is extremely large. In such situations, only the SP and MP approaches are computationally tractable.

5.1 *Acute Lymphoma Leukemia Data Analysis*

In this subsection, 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 levels—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 SP procedure with $h = 10$ and $n = 1,000$, the MP procedure with $n = 1000$, and the EP procedure were employed to find genes that are differentially expressed.

The solid lines in Figure 1 depict the discrete null distribution of the SP p-values. The dashed lines illustrate the observed distribution of the SP p-values computed from the ALL data. The top panel of the figures illustrates the entire distribution while the bottom panel focuses on the region involving the p-values no larger than $h/n = 0.01$. The region involving the p-values from 10/999 to 10/980 is depicted in Web Figure 1 of Web Appendix D. The overabundance of small p-values and deficit of large p-values in the observed distribution relative to the null suggest that some genes are differentially expressed. Note that the discreteness of the null distribution and its asymmetric V-shape depicted in the top panel of Figure 1 make existing m_0 estimation methods that assume a decreasing p-value density with a continuous uniform$(0, 1)$ component inappropriate for these SP p-values.

Table 1 provides results for ALL data analyses using p-value significance threshold 0.001 applied to EP, MP, and SP p-values. When using the EP and MP p-values, two different methods for estimating m_0 were considered: the method of ST, which is perhaps the most commonly used approach, and the method of NHCW, which is the method most closely related to the approach we have proposed for use with SP p-values.

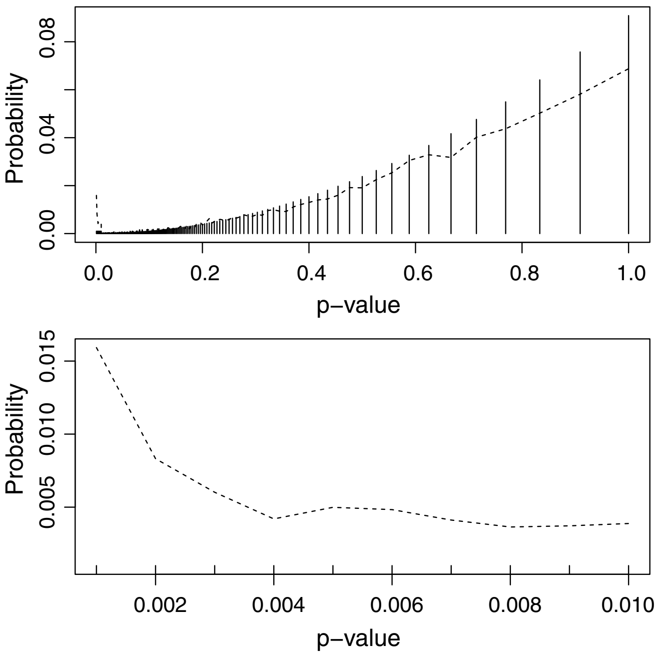

Figure 1. The null (solid) and observed (dashed) SP p-value distributions for the ALL analysis with $h = 10$ and $n = 1000$ for all p-values (top) and p-values no larger than 0.01 (bottom).

Table 1
The estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected hypotheses at p-value threshold 0.001 ($R(0.001)$), and the estimated FDR at p-value threshold 0.001 ($\widehat{\text{FDR}}(0.001)$) for EP, MP, and SP analyses of the ALL dataset

	m_0			$\widehat{\text{FDR}}$
p-value method	estimator	$\hat{m}_0$	$R(0.001)$	(0.001)
EP	ST	9031	229	0.0394
EP	NHCW	9060	229	0.0395
MP ($n = 1,000$)	ST	9166	201	0.0456
MP ($n = 1,000$)	NHCW	9040	201	0.0450
SP ($h = 10$, $n = 1,000$)	Proposed	9548	201	0.0475

The results from all five methods of analysis are similar. The most noticeable differences are between the analyses based on exact p-values (EP) and those based on Monte Carlo p-values (MP and SP). The exact analysis identifies more significant genes at the 0.001 p-value threshold and estimates a lower FDR. This is not surprising because the exact analysis is more powerful, as discussed in Web Appendix E and illustrated in the simulation study of Section 6.

Although the EP analyses are more powerful, their computation cost is extravagant compared to the Monte Carlo

Table 2
The estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected hypotheses at p-value threshold 0.01 ($R(0.01)$), and the estimated FDR at p-value threshold 0.01 ($\widehat{\text{FDR}}(0.01)$) for MP and SP analyses of the barley eQTL dataset

Method	n	h	$\hat{m}_0$	$R(0.01)$	$\widehat{\text{FDR}}(0.01)$
MP	1,000	1,000	16,240	3838	0.0423
SP	1,000	10	16,340	3857	0.0424
MP	10,000	10,000	16,285	3865	0.0421
SP	10,000	100	16,304	3851	0.0423

approaches. The EP tests required the computation of $\binom{26}{5} = 65,780$ test statistics for each of 12,625 genes for a total of $12,625 \times 65,780 = 830,472,500$ test statistics. Alternatively, the MC p-values required the computation of $12,625 \times 1,000 = 12,625,000$ statistics. In contrast, the total number of test statistics computed to obtain the SP p-values was just $1,626,171$. Thus, in this case, the EP analysis was approximately 500 times more computationally intensive than the SP analysis. Furthermore, the MP analysis was approximately eight times more computationally intensive than the SP analysis, even though the two approaches produced nearly identical results.

Note that the SP p-values were computed from a subset of the permutations used to compute the MP p-values. Thus, the same 201 genes were identified as differentially expressed by both MP and SP methods. (175 of these 201 genes were also identified by the EP analyses.) The exact agreement between the MP and SP rejection sets will hold for any p-value significance threshold $\alpha \leq h/n$. To see this, note that rejection of the null at level $\alpha = k/n$ for $k \leq h$ will occur for the SP test if and only if the SP p-value is less than or equal to k/n, i.e., if and only if G in (1) is less than $k \leq h$. When $G < h$, the SP p-value takes the same value as the MP p-value, namely $(G+1)/n$.

In Web Appendix E, we discuss how this agreement between the SP and MP rejection sets can be used to form a strategy for selecting h and n. In particular, we recommend choosing h and n such that $h/n = \alpha$, where α is an approximate upper bound on the p-value threshold for significance expected to yield acceptable FDR levels. Furthermore, we provide a formula for choosing n, given the maximum number of test statistics that an investigator is willing to compute, the number of tests m, and hypothesized values for α and m_0.

To study the variability of the Monte Carlo approaches, we repeated the MP and SP analyses of the ALL data a total of 1000 times. The results presented previously in this subsection are for only the first of these 1000 analyses. On average, the Monte Carlo approaches identified 212 (standard deviation of 7, 0.05, and 0.95 quantiles of 200 and 223, respectively) genes as significant at the p-value threshold 0.001. The average number of genes that were identified as significant by both the Monte Carlo and EP analyses was 179 (standard deviation of 5, 0.05, and 0.95 quantiles of 171 and 188, respectively). Web Tables 2 and 3 in Web Appendix F summarize the m_0 and FDR estimates from all 1000 analyses.

5.2 *Identification of Expression Quantitative Trait Loci in Barley*

Lander and Botstein (1989) wrote the seminal paper on interval mapping of quantitative trait loci (QTL), 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 are associated with quantitative characteristics. Hundreds or thousands of genetic positions (loci) are examined for association with a quantitative trait. A test statistic is computed at each locus. This analysis is repeated for thousands of data permutations, and the most extreme test statistic across all loci is computed for each permutation. To control the familywise error rate when testing hundreds or thousands of loci, the permutation distribution of the most extreme test statistic across all loci is used as the reference distribution when testing each locus for significant association with the trait. Hence, obtaining each single test statistic value in the permutation distribution requires computing hundreds or thousands of test statistics. Due to these computational challenges, Nettleton and Doerge (2000) proposed a permutation-based sequential procedure for estimating a QTL significance threshold for a single trait.

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 gene expression traits (see, e.g., Jansen and Nap, 2001; Brem et al., 2002; or Schadt, Monks, and Drake, 2003). The goal of eQTL analysis is to identify the loci that are associated with the expression of each gene. Employing SP p-values for such applications can substantially reduce the number of test statistics that must be computed and lead to considerable computational savings.

To illustrate the performance of our proposed approach in eQTL analysis, we consider a study of gene expression regulation in barley during stem rust infection (Moscou et al., 2011). The dataset has genotypes at 378 loci and expression measurements for 22,840 genes for each of 75 doubled haploid barley lines. For each gene, we wish to determine if expression is significantly associated with any of the genetic loci and if so, which locus is most strongly associated with expression. For this dataset, there are two genotypes observed at each locus. We consider there to be association between a locus and gene expression if the gene expression distribution for one genotype at that locus is different from the gene expression distribution for the other genotype. One way to measure evidence of such an association is to compute a two-sample t-statistic for each combination of gene and locus. For gene i, let t_{ik} denote the t-statistic for locus k, and let $M_i = \max\{|t_{ik}| : k = 1, \ldots, 378\}$. We use the test statistic M_i to test the null hypothesis of no association between the expression of gene i and any of the 378 genetic loci.

Computing EP p-values in this case is not computationally feasible. However, we can sample from the permutation distribution of M_i by randomly permuting the 75 gene expression values for the ith gene, computing the absolute value of the t-statistic for each of the 378 loci, and finding the maximum of these absolute values. Thus, a traditional Monte Carlo p-value based on $n-1$ draws from the permutation distribution for each gene requires the computation

Table 3

Means and standard errors of the means (sem) computed from $N = 1,000$ *simulation runs for the estimated number of true null hypotheses* $(\hat{m}_0)$, *the number of rejected null hypotheses* (R), *estimated FDR* $(\widehat{FDR})$, *and false positive fraction* $(V/max\{1, R\})$ *for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were* $m = 10,000$ *and* $m_0 = 7,500$, *respectively. Data were generated as described in Section 6 with* $\lambda = 2$

	$\hat{m}_0$		R		$\widehat{\mathrm{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	7851	8.968	1526	0.8153	0.0518	0.0001	0.0490	0.0002
EP/NHCW	7908	3.417	1526	0.8153	0.0522	0.0000	0.0490	0.0002
MP/ST	7920	9.120	1516	0.8221	0.0523	0.0001	0.0492	0.0002
MP/NHCW	7907	3.389	1516	0.8221	0.0522	0.0000	0.0492	0.0002
SP	7906	3.761	1516	0.8221	0.0522	0.0000	0.0492	0.0002

of $n \times 378 \times 22{,}840 = n \times 8{,}633{,}520$ t-statistics to obtain $n \times 22{,}840$ maximum absolute t-statistics.

In four independent analyses, we computed MP p-values for $n = 1,000$ and $n = 10,000$ and SP p-values using $h = 10$, $n = 1000$ and $h = 100$, $n = 10,000$. (For a discussion of the choice of h and n see Web Appendix E.) To obtain estimates of m_0, the method of NHCW was applied to the 22840 MP p-values while the method proposed in Section 4 was applied to the 22,840 SP p-values. Table 2 displays results for p-value significance threshold 0.01. Results for additional thresholds and for estimation of m_0 using the method of ST are presented in Web Tables 4 through 6 of Web Appendix G. Although these results were obtained with independently selected permutations using different Monte Carlo strategies, the agreement among the results was quite good. For example, 3690 genes were declared to be significantly associated with one or more loci by all four methods. Pairwise intersections of significant results ranged in size from 3727 to 3813 genes.

For both $n = 1000$ and $n = 10,000$, the computing time for the tradition Monte Carlo analysis was approximately 3.4 times greater than the corresponding SP analysis with $h = n/100$. For datasets with a greater proportion of true null hypotheses, the time savings of the sequential approach would be even greater because the expected number of permutations sampled at termination $(E(L))$ is minimized when the null is true. In particular, Besag and Clifford (1991) showed that $E(L)$ under the null is approximately

$$\hat{L} = h + h \log\{(n - 1/2)/(h + 1/2)\}.$$

For either $n = 1,000$ or $n = 10,000$ and $h = n/100$, as in our example analyses, $n/(\hat{L} + 1) \approx 17.8$. Thus, the sequential approach would be expected to be more than 17 times faster than the traditional approach when almost all tested null hypotheses are true.

6. A Simulation Study

In this section, we conduct a simulation study to investigate the performance of the SP approach relative to the more computationally intensive MP and EP strategies. For each run of the simulation, each of m independent null hypotheses was tested for a difference in distribution 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.

For $j = 1, \ldots, m$, $H_{0j} : \delta_j = 0$ was tested using EP, MP $(n = 1000)$, and SP $(h = 10, n = 1000)$ tests based on a two-sample t-statistic. The SP p-values were computed from a subset of the permutations used by the MP approach. As in Section 5, the methods of ST and NHCW were used to estimate m_0 from EP and MP p-values, while the method proposed in Section 4 was used to estimate m_0 from SP p-values. For each simulation run, the number of rejected null hypotheses (R), the estimated FDR, and $V/\max\{1, R\}$ were recorded for the p-value significance threshold 0.01. 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 nonnull effects on performance. The results for each parameter combination were based on $N = 1000$ simulation runs with $m = 10,000$.

Table 3 displays summary statistics for the case of $m_0 = 7500$ and $\lambda = 2$. Results for other scenarios and more information about variability of the estimators are provided in Web Tables 7 through 18 of Web Appendix H. The results for all five methods are remarkably similar to each other within each scenario. In Table 3, we see that the EP approach identifies a few more significant results on average than the Monte Carlo approaches while incurring approximately the same true FDR (estimated to be about 0.049). Thus, the EP approach is slightly more powerful than the Monte Carlo approaches. The average of the estimated FDR levels is quite similar across all methods, in part due to the similar estimates of m_0.

When looking across all scenarios, the results show that the average estimated FDR compares very well to the estimate true FDR level and is slightly conservative due to conservative estimation of m_0. The degree of conservative bias is reduced as the power increases. The performance of the m_0 estimator seen here is consistent with past performance for continuous p-values from parametric t-tests as discussed by NHCW. 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 nonnull tests are likely to fall in the right tail of the p-value distribution, and estimates of m_0 are inflated as a result. Most existing estimators of m_0 exhibit a similar positive bias when effects are small. 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. Furthermore, the approaches studied here are far less conservative than the original strategy of Benjamini and Hochberg (1995), which is equivalent to always using m as an estimate of m_0 when estimating FDR through (4).

This simulation study has focused on the case where tests are conducted for m independent sets of observations. Web Appendix I presents a second simulation study that mimics the study described here except that subsampling from actual data is used to simulate data with a realistic dependence structure across multiple tests. The results of that study show that dependence induces greater variation across simulation replications than is seen with independent data. However, the overall conclusions remain the same.

7. Conclusion

This article proposes using sequential analysis when testing multiple hypotheses via permutation testing and describes how to use the resulting discrete nonuniformly distributed null p-values to estimate m_0 and FDR. The proposed procedure requires far fewer draws from the permutation distribution and sustains little loss of information relative to both exact and traditional Monte Carlo methods. In particular, the sequential approach obtains precisely the same set of rejected null hypotheses as the traditional Monte Carlo approach for p-value significance thresholds less than or equal to h/n. Furthermore, the estimated FDR for the set of rejected null hypothesis is quite similar for both the sequential and traditional Monte Carlo approaches and tends to be close to—but no smaller than—the actual FDR level. Thus, the proposed approach can be quite useful for conducting multiple permutation tests at reduced computational expense.

8. Supplementary Materials

Web Appendices A through I and an R package for computing SP p-values and using them to estimate m_0 and FDR are available with this article at the Biometrics website on Wiley Online Library.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grants 0500461, 0714978, and 0811804.

References

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.

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.

Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78,** 301–304.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.

Jansen, R. C. and Nap, J. P. (2001). Genetical genomics: The added value from segregation. *Trends in Genetics* **17,** 388–391.

Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–199.

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.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society, Series B* **74,** 163–182.

Moscou, M. J., Lauter, N., Steffenson, B., and Wise, R. P. (2011). Quantitative and qualitative stem rust resistance factors in barley are associated with transcriptional suppression of defense regulons. *PLoS Genetics* **7,** e1002208.

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.

Nettleton, D. and Doerge, R. W. (2000). Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* **56,** 52–58.

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* **11,** 337–356.

Pounds, S., Cao, X., Cheng, C., Yang, J., Campana, D., Evans, W. E., Pui, C.-H., and Relling, M. V. (2011). Integrated analysis of pharmacologic, clinical, and SNP microarray data using projection onto the most interesting statistical evidence with adaptive permutation testing. *International Journal of Data Mining and Bioinformatics* **5,** 143–157.

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.

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.

Silva, I., Assunção, R., and Costa, M. (2009). Power of the sequential Monte Carlo test. *Sequential Analysis* **28,** 163–174.

Storey, J. D. (2000). False discovery rates: Theory and applications to DNA microarrays. Unpublished Ph.D. thesis, Department of Statistics, Stanford University.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64,** 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31,** 2013–2035.

Storey, J. D., Taylor, J. E., and 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, Series B* **66,** 187–205.

Received May 2011. *Revised August* 2011.
Accepted August 2012.