

Statistical Applications in Genetics and Molecular Biology

Volume 3, Issue 1

2004

Article 1

Using Alpha Wisely: Improving Power to Detect Multiple QTL

Katy L. Simonsen*

Lauren M. McIntyre†

*Purdue University, simonsen@stat.purdue.edu

†Purdue University, lmcintyre@purdue.edu

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/sagmb>

Using Alpha Wisely: Improving Power to Detect Multiple QTL*

Katy L. Simonsen and Lauren M. McIntyre

Abstract

The increase in the number of available markers for many experimental populations has led to QTL studies with ever increasing marker numbers and densities. The resulting conundrum is that as marker density increases, so does the multiple testing problem. It is important to re-examine the detection of multiple QTL in light of increasing marker density. We explore through simulation whether existing methods have achieved the maximum possible power for detecting multiple QTL and whether increasing the marker density is an effective strategy for locating multiple QTL. In addition to existing methods, such as the maximum, the CET, and the Benjamini-Hochberg and Benjamini-Yekutieli procedures, we propose and evaluate the complete set of order statistics with their corresponding empirical joint distribution. We examine these statistics in conjunction with a novel application of the alpha-spending approach, providing a less conservative solution to the problem of controlling the false discovery rate (FDR) in multiple tests. We conducted a simulation study to assess the relative power of these approaches as well as their ability to control FDR. We find that several of the new approaches have a reasonable FDR, and can substantially improve the experimenter's ability to detect multiple QTL compared to existing approaches in many cases; however, the Benjamini-Hochberg procedure remains a very reasonable choice. The methods are applied to a nine-trait Oat vernalization dataset.

KEYWORDS: order statistics, power, alpha-spending, FDR, QTL

*ACKNOWLEDGEMENTS: The authors are very grateful to Jim Holland for generously sharing his Oat data. This work was supported in part by NSF DBI-9904704. We thank R. W. Doerge for helpful discussions, and the anonymous reviewers whose suggestions improved the manuscript. ADDRESSES: Katy L. Simonsen, Department of Statistics, Purdue University, 150 N. University Ave, West Lafayette IN 47907-2068, Email: simonsen@stat.purdue.edu, Phone: 765-494-6036. Lauren M. McIntyre, Department of Agronomy and Computational Genomics, Purdue University, 915 West State Street, West Lafayette IN 47907, Email: lmcintyre@purdue.edu, Phone: 765-496-3662. SOFTWARE: The software used for the data analysis is available at <http://www.stat.purdue.edu/~simonsen/AlphaWisely/>.

1 INTRODUCTION

A pressing problem in the detection of QTL is that when more information is available in the form of more markers, the power of an individual test for association between a marker and a trait can decrease when false positives are tightly controlled. In a genome scan, where marker loci are spaced throughout the genome, it is common to test for association between each marker and the phenotype or trait of interest. When a test for association is performed at each locus, multiple testing occurs and some correction for the statistical significance must be applied. Failure to control for multiplicity can overwhelm the investigator with false positive findings. However, if the false positive rate is conservative, well below the nominal level, an unacceptable decrease in power results. Most procedures which aim to control false positive rate become more conservative as more tests are performed.

As technology advances, more markers are available for genotyping, thus increasing marker density and exacerbating the problem. Since association between markers increases with marker density, the issue of how to appropriately control false positives while maximizing detection becomes increasingly acute. A similar problem occurs after a region of interest has been identified and that region is saturated with markers. In this case, the problem is more pronounced, as markers in a single region are more correlated than markers across the genome, making classical corrections such as the Bonferroni exceedingly conservative. As technology advances, we can imagine experiments where maps are genome wide saturation maps, and the best way to deal with this level of correlation between marker loci has not yet been addressed.

Using the maximum test statistic (statistic with the minimum individual p-value), in combination with permutation methods has been shown to be effective in circumventing multiple testing issues for procedures to detect a single quantitative trait locus (QTL) in experimental populations (Churchill & Doerge, 1994) and a binary trait locus in human populations with the Transmission Disequilibrium Test (TDT) (McIntyre *et al.*, 2000) and in general statistical theory (Westfall & Young, 1993). However, the maximum statistic is a single statistic, and is designed to detect single QTL, as has been pointed out (Churchill & Doerge, 1994; Doerge & Churchill, 1996). Nonetheless, the use of the maximum threshold on more than one statistic is common practice. If the maximum threshold is to be used in this manner, it is necessary to establish the effect of this threshold on testing for association in subsequent marker loci.

For the case of an infinitely dense map and large sample size, a threshold for LOD scores based on extreme-value properties has been proposed (Lander & Botstein, 1989). This theory is based on the completely null hypothesis of no QTL. When multiple QTL are present, thresholds based upon theoretical distributions become more complicated because of the mixture distributions involved. If only one major QTL is present, one could accommodate the underlying bimodal distribution, but with variable numbers of QTL present, this correction becomes increasingly difficult.

A conditional elimination test (CET) was proposed as a mechanism for detecting multiple QTL (Doerge & Churchill, 1996), and was shown to control the family-wise type I error in the completely null case. In this procedure, the maximum statistic is tested for significance. If the test is rejected, the data are partitioned according to the genotype at the significant marker, then the significant marker and all markers linked to that marker are removed from

the analysis. The second maximum is found, and tested for significance after re-permutation. The key to the control of the overall family-wise type I error rate is the re-permutation and thus re-estimation of the null distributions after the marker elimination. A natural companion of this procedure is the examination of the joint distribution of order statistics. We propose such a procedure, and examine its performance, both with and without the elimination of linked markers.

Many different types of corrections for multiple testing exist (Hsu, 1996; Hochberg & Tamhane, 1987). The Bonferroni correction is the most well known and the most straightforward procedure to implement. It implicitly assumes that all tests are equally important, since the risk of type I error is shared equally across all tests. A more general allocation of type I error allows any distribution of the overall type I error such that the sum across all tests does not exceed the nominal level (Demets & Lan, 1994). In a case where the tests can be ordered in terms of priority, the type I error can be allocated unequally, with the largest portion going to tests with higher priority. Considering the set of test statistics as order statistics provides a natural prioritization of the type I error allocation.

The false discovery rate (FDR) (Benjamini & Hochberg, 1995) has been proposed as an alternative criterion to the more conventional family wise type I error rate (FWER) in controlling false positives for multiple hypotheses and has been used in genetic studies (Lee *et al.*, 2002). To paraphrase Storey & Tibshirani (2003), the FDR is the expected proportion of rejections which are false, whereas the FWER is the expected proportion of true nulls that are rejected. The positive false discovery rate (pFDR) is the expected proportion of rejections which are false given that at least one rejection has occurred (Storey & Tibshirani, 2003). In the completely null case the FDR and FWER are the same; with either criterion the probability of at least one false rejection is controlled at level α . In this case the pFDR and the FDR differ. In the case where some of the null hypotheses are false (and true rejections are expected), the FDR and pFDR are essentially equivalent and consider the proportion of false rejections rather than the absolute number of them. That is, supposing 20 rejections occur, FDR control at level 0.05 specifies that on average one of these should be false, but if 100 rejections occur, on average 5 should be false. In contrast, FWER control specifies that the probability of at least one false rejection is less than 0.05 regardless of the total number of rejections.

We feel that the FDR is in many cases the more appropriate criterion to use in QTL mapping studies. Often, QTL mapping is a first step in a longer-term project, where follow-up steps may include fine mapping (Wayne *et al.*, 2004). In this case, the cost of pursuing a few false leads is not large compared to the cost of pursuing the true ones, and the researcher is willing to accept that cost for the tradeoff of finding the true leads. Controlling the FDR allows researchers to set their tolerance for false positives depending upon their individual resources for follow-up experiments.

Procedures for controlling the FDR have been proposed for independent tests (Benjamini & Hochberg, 1995) and dependent tests (Benjamini & Yekutieli, 2001), and these procedures are examined in this paper. These approaches use a linear step-up procedure to allocate type I error. While these procedures guarantee a FDR below the nominal level, it is unknown how their detection rates compare to existing procedures for detecting multiple QTL.

In this paper we focus on experimental populations to address the following four issues: 1) to determine the performance of existing methods; 2) to determine whether using the joint

distribution of order statistics, and their corresponding distributions, improves power to detect association between markers and traits compared to existing approaches; 3) whether an unequal allocation of type I error combined with different statistics has the potential to increase detection rates; and 4) whether increasing marker density improves power to detect multiple QTL.

The overall goal is to determine how best to maximize power for QTL detection while controlling false positive rates appropriately. Approaches which focus only on type I error control can lead to conservative solutions and a subsequent loss of overall power for detection of secondary QTL, or QTL of small effect. What we seek to do with the proposed approach is to develop concrete recommendations for managing false positives while maximizing QTL detection.

2 METHODS

The basic structure of a hypothesis test is as follows. First, a test statistic is calculated. Second, that statistic is compared to a particular distribution to calculate a p-value. Thirdly, that p-value is compared to a significance level (α value) and a decision to reject the null hypothesis is made if $P < \alpha$. Finally, for simulated data that decision can be declared appropriate, or in error (true or false). We use this basic structure in outlining our methods.

2.1 Test Statistics

2.1.1 Locuswise tests

When m marker loci are each tested separately for association with a trait, m null hypotheses are used. For locus i , the null hypothesis is H_{0i} : marker i is not linked to the trait. A set of statistics S_i measuring the extent of association at each locus is calculated. The form of the test statistic used can be completely general, such as t- or F-statistics, or likelihood ratios, but we assume here that it is continuous. For descriptive purposes, it is assumed here that a larger value of S indicates more association. To test for significance, the observed value s_i is compared to the distribution of S_i under H_{0i} . It is often the case that the distributions of the test statistics are different at different loci. This could arise, for example, if different amounts of data were missing at different loci, or if a different number of alleles are present, resulting in different degrees of freedom. In that case it is difficult to compare the raw test statistic values across different loci. To avoid this complication, the p-value associated with the individual test statistic and its marginal distribution can be used in place of the raw value of the test statistic. Under the null hypothesis H_{0i} , the definition of the marginal p-value for s_i is $X_i = \text{Prob}_{H_{0i}}(S_i > s_i)$. The distribution of S_i under H_{0i} is used to calculate X_i (the probability of observing a value at least that extreme). This is a simple transformation of each test statistic S_i into an equivalent test statistic X_i , with the property that each X_i has the same marginal distribution, Uniform(0,1), under H_{0i} (assuming S is continuous). There are two principal advantages to using X_i rather than S_i . First, it is easy to compare values of X_i for different loci, since they all have the same null distribution. Second, the X_i have a clear interpretation as the p-values of a locus-by-locus test. To avoid confusion with the p-values for order statistics discussed later, the X_i will be referred to as test statistics.

2.1.2 Order statistics

A set of order statistics is constructed by sorting the statistics X_1, \dots, X_m in ascending order, and labelling these $X_{(1)}, \dots, X_{(m)}$. The first order statistic $X_{(1)}$ corresponds to the smallest p-value or the “largest” of the original test statistics S_i . The locus affiliated with $X_{(j)}$ will be denoted ℓ_j , so that $X_{(j)} = X_{\ell_j}$. Thus the locus ℓ_1 is the locus with the greatest amount of evidence for association with the trait (as measured by S). The distribution under H_0 of the m order statistics $X_{(j)}$ differs from the distribution under H_0 of the m unranked statistics X_i . The X_i all have a marginal Uniform(0,1) distribution, but since $X_{(1)}$ is the *smallest* of a set of m uniform random variables, it is more likely to be in the lower end of the $[0,1]$ interval. Thus the simple uniform distribution cannot be used to assess significance for the set of order statistics; rather, the distributions of the set of order statistics under H_0 must be ascertained.

2.1.3 Interval mapping

While all the calculations in this paper are based upon tests that occur at the marker loci, the methodology is easily extended into interval mapping. In the interval mapping framework test statistics such as the LOD score are calculated at many points along the interval. QTL are detected if at any point on the interval the LOD score exceeds the threshold set for detection. In other words, QTL are detected for that interval if the maximum LOD score on that interval crosses the relevant threshold. Thus, the maximum LOD score on the interval would be the test statistic to employ when evaluating the performance of an interval mapping procedure. If the maximum LOD score does exceed the threshold, that interval is said to contain a QTL. The maximum LOD score for an interval can therefore be viewed as the statistic upon which the above order statistics can be based. Therefore, as with the tests at marker loci, the number of tests evaluated is limited by the number of marker loci examined.

2.2 Null Distributions and P-Values

2.2.1 P-values for locuswise statistics

P-values arise in several contexts in this paper. These are all “uncorrected” p-values, in the sense that no correction for multiple testing is employed in calculating p-values. Corrections for multiple testing will be made by adjusting the significance level of individual tests (discussed later). The first context is the conversion of a raw test statistic S_i into its corresponding p-value X_i , according to the marginal distribution of S_i under H_{0i} . The testing procedure that uses these p-values is designated “Loc” (for locuswise). In many cases the marginal distributions of test statistics S_i at each locus will be known explicitly (for example, a t, F, or χ^2 -distribution with known degrees of freedom). In these cases the X_i are simple to calculate from the known distributions using standard numerical lookup methods. If the distributions of the test statistics S_i are not known explicitly, then standard permutation methods (Good, 1994; Churchill & Doerge, 1994) can be used to obtain the statistics X_i . Briefly, the permutation procedure used is as follows. Let s_i represent the observed test statistic at locus i from the original dataset. The permutation procedure randomly assigns the n trait values from the original dataset to the n genotypes N times, resulting in N new

permuted datasets for which H_{01}, \dots, H_{0m} are all true. New test statistics S_i are calculated for each dataset. At each locus i , X_i is estimated by counting the proportion of the N datasets for which S_i exceeds s_i .

2.2.2 P-values for order statistics

P-values for order statistics, p_j , can be estimated empirically using a permutation procedure to estimate their distributions. Permutation procedures were initially introduced by Fisher (1935). The purpose of a permutation procedure is to empirically determine the distribution of a test statistic under the null hypothesis. Permutation procedures have gained in popularity (Good, 1994; Edgington, 1995), have been applied to testing for marker-trait association (Churchill & Doerge, 1994), and are routinely used to empirically determine the distribution of the maximum order statistic (Churchill & Doerge, 1994; Doerge & Churchill, 1996; McIntyre *et al.*, 2000; Westfall & Young, 1993). It is relatively straightforward to extend this procedure to calculate the distributions for the entire set of order statistics. The null hypothesis H_0 is that the trait is not linked to *any* of the markers, while the existing linkage (and correlation) between markers is maintained. Let $x_{(1)}, \dots, x_{(m)}$ represent the observed order statistics from the original dataset. The original dataset consists of n genotypes and their corresponding n trait values. Let N be a large integer. N datasets for which H_0 is true are randomly generated by assigning the n trait values to the n genotypes N times while the genotypes remain intact, thus preserving the relationships among markers. In each permuted dataset, order statistics $X_{(j)}$ are calculated as described earlier. Then p_j is estimated by counting the proportion of the N datasets in which $X_{(j)}$ is less than $x_{(j)}$. Tests using the set of order statistics are referred to as “Ord” for the remainder of the paper.

When H_{0i} is true, the statistic X_i has a uniform distribution on the interval $[0,1]$. These statistics are ordered to obtain the set $X_{(1)}, \dots, X_{(m)}$, where $X_{(j)}$, the j^{th} smallest of the statistics, is the value corresponding to the locus ℓ_j . The distribution of $X_{(j)}$ under H_0 is needed to calculate a p-value. Let F_j denote the cumulative distribution function of $X_{(j)}$ under H_0 . When a particular value $x_{(j)}$ is observed, its p-value is defined as $p_j = F_j(x_{(j)}) = \text{Prob}_{H_0}(X_{(j)} < x_{(j)})$.

It should be carefully noted that the distribution of the first order statistic is different from the distributions of the remaining order statistics. The only exception would occur in the case where all the loci are completely linked ($r = 0$), where the m statistics would all be identical for every locus, and so the $X_{(j)}$ would jointly have a uniform distribution. If the m loci were all unlinked (recombination rate $r = \frac{1}{2}$) and independent, then the null distribution of $X_{(j)}$ would be that of the j^{th} smallest of a set of m independent uniform random variables, which is a Beta($j, m - j + 1$) distribution. (See, e.g., (Casella & Berger, 1990).) In all realistic situations, ($0 < r < \frac{1}{2}$), the distributions of the $X_{(j)}$ are different for each j and will depend upon the degree of linkage, which varies from case to case, and so cannot practically be derived for each case in order to determine appropriate critical values. Instead, the empirical distributions of the $X_{(j)}$ under incomplete linkage will be examined, and their behavior is expected to vary between the two extremes of the beta and uniform distributions as the amount of linkage varies (Figure 1).

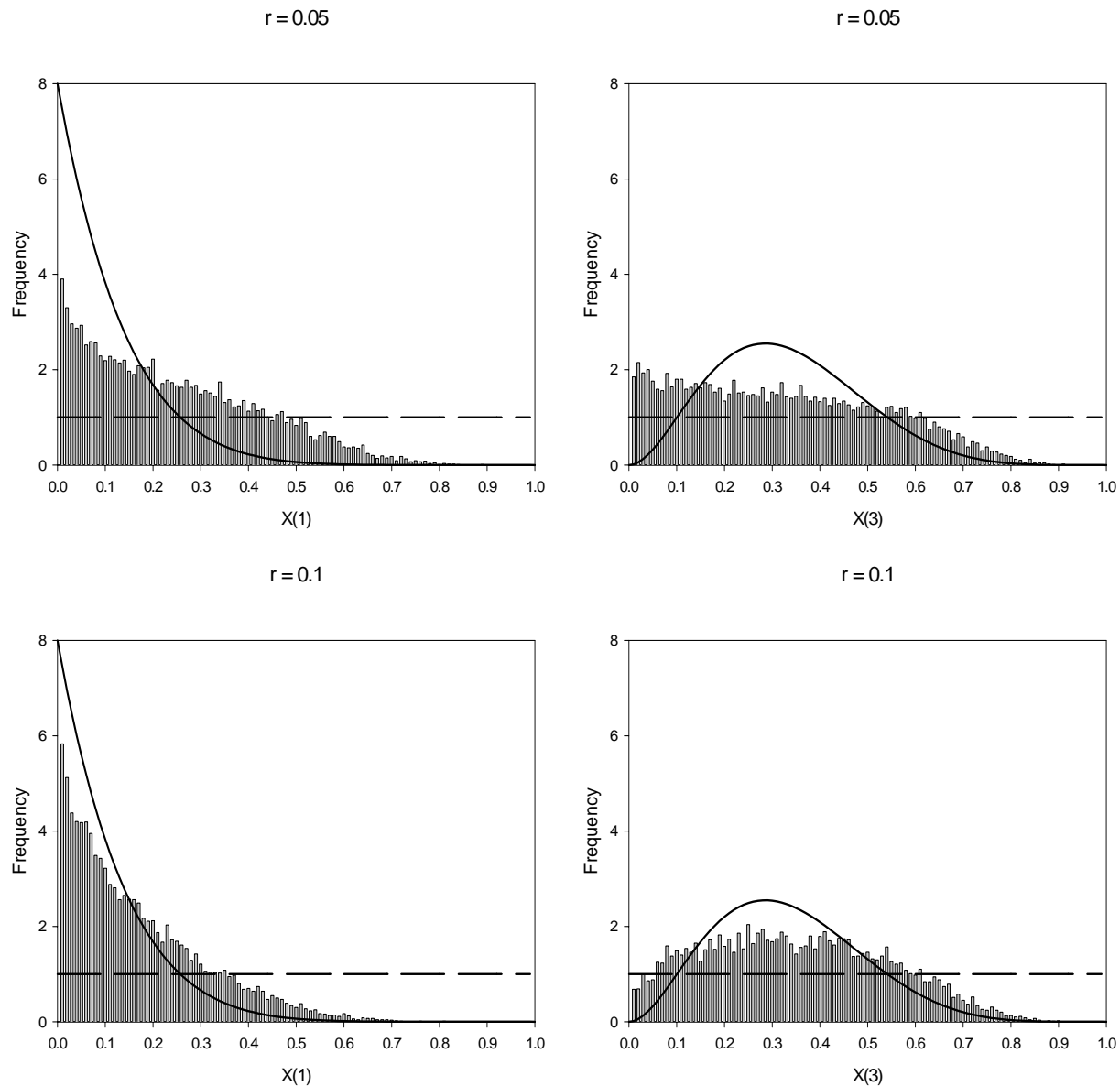


Figure 1: The empirical distributions of $X_{(j)}$ for $j = 1$ (left) and 3 (right) under H_0 are shown as a histogram. Histograms are based on $N = 9999$ null permutations, $m = 8$ marker loci, and marker spacings $r = 0.05$ (top) and 0.1 (bottom). For comparison, the Uniform(0, 1) distribution is shown with a dashed line and the Beta($j, m - j + 1$) distribution is shown with a solid line.

2.2.3 P-values using the distribution of the maximum

The distributions of the order statistics vary with j . That is, the distribution of the first order statistic is different from that of the second, third, and so on. Nonetheless, as a simple extension of the method of Churchill and Doerge (1994), people may use the distribution of the first order statistic in an attempt to be conservative in the evaluations of the subsequent order statistics. We additionally determined a “p-value” for the subsequent order statistics

using the (inappropriate) distribution of the first order statistic and we refer to that as pm_j .

2.2.4 P-values using the CET

Doerge and Churchill (1996) proposed the Conditional Elimination Test (CET) for detecting multiple QTL. The CET is a sequential procedure for detecting QTL conditional on previously detected ones.

The CET starts with the first order statistic, and if that rejects according to the null distribution of the first order statistic, the corresponding marker (ℓ_1) is declared significant and all markers linked to that locus are eliminated from further consideration. The individuals are then partitioned into two (or more) classes according to the genotype at the significant marker. Within each class, test statistics are recalculated at all remaining loci, resulting in two test statistics for each remaining locus. Then, the individuals are re-permuted within their own marker classes. Test statistics are compared to the distribution within their own marker class. Significance at a locus is declared if a test statistic for a locus in either class was found to be significant. Linked loci are eliminated, the individuals within each marker class are again partitioned according to their genotype. The procedure repeats until either no more markers are significant, or insufficient data remain. In our implementation of the CET we chose to stop the process if no partition contained more than 5 individuals.

Because of the repeated sub-partitioning employed, the CET is self-limiting in the number of loci it can detect. For example, if four loci have been detected, there are $2^4 = 16$ partitions into which the sample must be divided. For a sample of size 100, this would result in an average of about 6 individuals per partition, too few to perform the necessary tests for the detection of a potential fifth locus. We therefore considered a simplified version of the CET, in which the individuals are not partitioned before re-permuting, and in which all tests are conducted with the full sample. We call this method “ERP” for “eliminate and re-permute”. Like the CET, the ERP eliminates markers linked to previous detections, and sets a threshold based on the re-permuted values. A comparison of the CET and ERP approaches would show whether the power of the CET comes primarily from the elimination of linked markers (in which case the ERP should be similar) or from the stratification given by the partitions (in which case the CET should be more powerful).

2.3 Correction for Multiple Testing

2.3.1 Bonferroni correction and alpha-spending

Multiple tests require careful consideration of type I error. When m tests are performed, each test has some probability of type I error. The total probability of falsely rejecting H_0 on at least one test, known as the family-wise error rate (FWER), should be less than or equal to some fixed α . Let the probability of type I error on the i^{th} test be α_i . Then Bonferroni’s inequality states that

$$\text{total probability of type I error} = \text{Prob}(\text{reject on at least one test}) \leq \sum_{i=1}^m \alpha_i,$$

with equality only if the rejections are mutually exclusive. If the α_i are chosen so that $\sum_{i=1}^m \alpha_i = \alpha$, the total probability of type I error is guaranteed not to exceed α .

An alpha spending function assigns a particular weight α_i to the i^{th} test, where $\sum_{i=1}^k \alpha_i = \alpha$. The simplest of these is the standard Bonferroni procedure, which assigns $\alpha_i = \frac{\alpha}{m}$, where m is the number of tests performed. This weights all m tests equally, and is reasonable when all tests are considered equally important. We will refer to this weight function as B. So the method using order statistics with a Bonferroni correction for multiple testing will be denoted Ord/B. There are many possible alpha spending functions with unequal weights that can be used when tests are not considered equally important. It seems reasonable to suppose that tests performed with the smallest order statistics should have higher priority than those with the larger order statistics, since a smaller X indicates more association. We examine two such functions. One of these is a geometric series (G), where $\alpha_i = \begin{cases} \frac{\alpha}{2^i} & i = 1 \dots m-1 \\ \frac{\alpha}{2^{m-1}} & i = m \end{cases}$. For two markers the Bonferroni and geometric alpha spending functions are equivalent. The choice of the geometric series is arbitrary; thus, a “spend-as-you-go” function (S), where the amount of the α spent on each test is equal to the observed p-value is intrinsically appealing. We consider such a function, namely $\alpha_i = \begin{cases} p_i & \text{when } \sum_{j=1}^i p_j \leq \alpha \\ 0 & \text{otherwise} \end{cases}$. Because this is a post-hoc comparison it will be necessary to carefully examine the impact of this approach upon both false positive and true positive rates.

The unequal allocation of the alpha discussed previously is applied to the set of order statistics without regard to marker linkage. Alternatively, known information about the map can be exploited to increase the available alpha for detection of unique QTL. In this approach, alpha is allocated to a particular test only if the marker locus being examined is not linked to any previously detected loci. This results in the most frugal allocation of alpha possible. We denote this as ELM for Eliminating Linked Markers. It is similar in spirit to the CET, in that markers are sequentially eliminated from consideration. However, whereas the CET requires partitioning and re-permuting to calculate new p-values after each rejection, the ELM does not. The p-values used with the ELM are identical to those used with the order statistic approach; only the allocation of alpha is changed so that markers linked to previously rejected markers are given an alpha allocation of 0. The geometric or spend-as-you-go functions can be used to allocate alpha to the remaining marker loci, but the Bonferroni is not appropriate since the number of tests being performed is not known in advance. We denote ELM used in conjunction with the spend-as-you-go alpha as ELM/S, and with the geometric spending function as ELM/G.

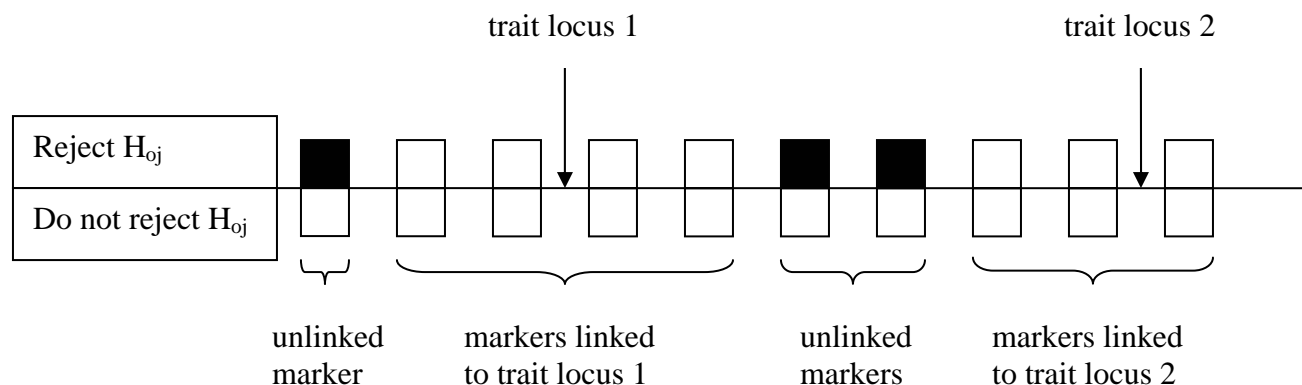
2.3.2 FDR-controlling procedures

Benjamini and Hochberg (1995) proposed the following method (which we denote H) to control FDR. Let $P_{(i)}$ denote the sorted set of p-values (for any given approach). Let $k = \max \{i : P_{(i)} \leq i \frac{\alpha}{m}\}$; declare the k p-values $P_{(1)} \dots P_{(k)}$ significant. When m_0 of the m null hypotheses are true, this method was shown to control FDR at level $\alpha \frac{m_0}{m} \leq \alpha$ when test statistics are independent (Benjamini & Hochberg, 1995) and when they have positive regression dependency (Benjamini & Yekutieli, 2001). Benjamini and Yekutieli (2001) additionally proposed the following adjustment (which we denote Y) to the Benjamini-Hochberg method for dependent tests. Let $\gamma = \sum_{j=1}^m \frac{1}{j}$ and let $k = \max \{i : P_{(i)} \leq i \frac{\alpha}{m\gamma}\}$; declare the k p-values $P_{(1)} \dots P_{(k)}$ significant. This procedure was shown to control FDR at level $\alpha \frac{m_0}{m} \leq \alpha$

under a completely general dependence structure. We apply the two procedures (H) and (Y) to two sets of p-values: the locuswise p-values X_i (Loc) and the order statistic p-values p_i (Ord). These procedures cannot be applied to the CET or ELM p-values because the number m of hypotheses to be tested is not known in advance.

2.4 Evaluating the Decision

Every test has one of four possible outcomes: true positive, false positive, true negative or false negative. In the case of linked genetic markers, rejecting the null hypothesis at a marker linked to the trait locus is a true positive. Failure to reject the null hypothesis for a marker linked to a trait locus is a false negative. Rejecting the null hypothesis at a marker not linked to a trait locus is a false positive, while failure to reject when the marker is not linked to a trait locus is a true negative (Figure 2).



Legend	Decision	
	Reject H_{0j}	Do not reject H_{0j}
Linked	True positive <input type="checkbox"/>	False negative <input type="checkbox"/>
Unlinked	False positive <input checked="" type="checkbox"/>	True negative <input type="checkbox"/>

Figure 2: A schematic of true positive, false positive, true negative and false negative decisions for the tests performed on a set of linked loci.

As a technical note, because the order statistics consider the joint set of marker loci, the null hypothesis for the joint set of statistics is simply H_0 : no locus is linked to the trait, and the alternative is H_1 : at least one locus is linked to the trait. If we used this definition in determining the false positive, false negative, true positive and true negative rates, we would be unable to interpret rejections of the null hypothesis as detection of a physically linked trait locus. Since this is contrary to the biological interpretation, we have defined our calculation of the true positive, false positive, true negative, and false negative rates according to the biological interpretation of the rejection of a specific marker locus. That is, if we reject with p-value p_j , it is considered a true positive only if marker ℓ_j is linked to the

trait (Figure 2). Using this definition, when QTL are present, false positives can exceed their nominal levels as a result of QTL previously located. We refer to this effect as shadowing.

The FDR is the expected proportion of rejections that are false. The FDR for a single dataset with multiple hypotheses tested is defined as $FDR = (\# \text{ false positives}) / (\text{total } \# \text{ rejections})$, or 0 if no rejections, and we estimate the expected FDR by averaging this value over all simulated datasets. In the completely null case, all positives are false, so the FDR is identical to the type I error rate, whereas if trait loci are present, some positives will be true (linked) and others false (unlinked). We calculated the FDR for all procedures.

In the case of a single trait locus the definition of power is straightforward: the probability of at least one true positive. However, when more than one trait locus is present, it is necessary to consider the set of contributing trait loci. For example, if three loci contribute to a quantitative trait, we would wish to consider the chance of 0, 1, 2, or 3 true positive loci detected as well as the chance of at least one true positive locus ($1 - \text{Prob}(0 \text{ loci are detected})$) and the average number of loci detected.

2.5 Case Study: Vernalization and Photoperiod Responses in Oat

In a recently published QTL study, Holland et. al (2002) look for QTL associated with nine different phenotypes. In this study of Oat vernalization, 136 F_6 derived lines with 283 markers in 34 linkage groups were used. There was a small amount of missing marker data in this dataset. We applied the full suite of methods described above to these data.

3 GENOTYPE AND TRAIT SIMULATIONS

3.1 Data Generation

To address our questions we simulated data according to a backcross design with trait values following a multiple gene additive model. No interactions were simulated. The contribution of each locus to the trait value was derived from one of two normal distributions, with homozygotes at a trait locus i having mean $\mu_{i1} = 0$ and heterozygotes having mean μ_{i2} , and common standard deviation $\sigma = 1$. Different values of μ_{i2} were used to create different effect sizes $E_i = (\mu_{i2} - \mu_{i1})/\sigma = \mu_{i2}$, which we label “small” ($E_i = 0.25$), “moderate” ($E_i = 0.5$), “large” ($E_i = 1.0$), and “huge” ($E_i = 1.5$). One hundred markers on five chromosomes were used throughout. Three trait loci were positioned between markers 10, 11, markers 50, 51 and markers 75, 76, contributing to the phenotype additively. The null case of no trait locus was also simulated. We examined both equally and randomly spaced marker maps, each with varying degrees of linkage. Equally spaced markers had recombination probabilities $r = 0.05, 0.10, \text{ or } 0.20$. Random spacing assigned distances between adjacent markers from an exponential distribution (truncated at 0.5) with the same three means. For each combination data were simulated 10,000 times. (See Table 1.)

3.2 Hypothesis Tests and Multiplicity Corrections

In the backcross design there are two marker classes, and thus a t-statistic was the raw test statistic S used at each locus. Five sets of p-values were calculated for each simulated

Table 1: Simulation Conditions

Recombination Rates	Marker Spacing	Effect			Number of Combinations
		μ_{12}	μ_{22}	μ_{32}	
0.05, 0.10, 0.20	random	0	0	0	3
	equal	0	0	0	3
	equal	0.25	0.25	0.25	3
	equal	0.5	0.5	0.5	3
	equal	1.0	1.0	1.0	3
	equal	1.5	1.5	1.5	3
	equal	0.5	1.5	1.0	3
	equal	0.5	2.0	0.5	3

24

dataset. For each marker, the set of locuswise test statistics S_j were calculated. Then the locuswise p-values X_j , representing the locuswise tests, were calculated (Loc). These were then sorted to obtain the set of ordered statistics $X_{(j)}$. P-values were calculated for these statistics in three ways. Using the correct distributions for $X_{(j)}$ gave the set of p-values p_j (Ord and ELM). Using the distribution for $X_{(1)}$ for all statistics gave the set of p-values pm_j (Max). P-values using the CET procedure pc_j were also calculated (CET).

Multiplicity was addressed in six ways. We considered a Bonferroni correction (B), a geometric alpha-spending function (G), a spend-as-you-go function (S), no correction (N), Benjamini-Hochberg (H) and Benjamini-Yekutieli (Y). Not every type of correction was applicable for every test. All six types of correction were applied to the locuswise (X_j) and order statistic (p_j) p-values. Using the distribution of the maximum statistic automatically corrects for multiplicity; therefore only (N) was performed for that test. For the CET, ERP, and ELM the number of tests to be done is not known in advance; therefore B, H, and Y cannot be applied. For the CET and ERP, we examined only (N), while for ELM, we examined G, S, and N. In total, eighteen combinations of p-values and multiplicity corrections were examined.

4 SIMULATION RESULTS

Results for the case of small effect size ($E_i = 0.25$ for each trait locus) were largely uninformative because detection rates were so low that none of the approaches had acceptable power. Similarly, when effect sizes were huge ($E_i = 1.5$ for each trait locus) the detection rates were all close to one, making it difficult to distinguish among the methods. We present detailed results on the subset of simulations that are informative in separating the behaviour of the different methods: large ($E_i = 1$) and moderate ($E_i = 0.5$) effects. The results for randomly spaced markers are similar to those for equally spaced markers, so we present the results for random spacing in the null cases only. The results for unequal effect sizes are qualitatively similar to those for equal effect size, so only the equal effect size results are shown.

4.1 No Trait Locus: False Discovery Rates

When H_0 is true, the goal is to maintain the false positive rate as close as possible to α . Table 2 gives the estimated false discovery rates (equal to the family-wise type I error rate in this null case) for each test procedure/multiplicity correction combination that was examined, and for equal or random marker spacing of $r = 0.20$ (“linkage map”), $r = 0.10$ (“dense map”), and $r = 0.05$ (“saturated map”). FDR values were the average fraction of 10,000 null datasets for which at least one (false) rejection occurred, and are accurate to within approximately 0.005.

Table 2: Estimated False Positive Rates in the Null Case: $r = 0.05, 0.10, 0.20$; equal/random spacing. Values accurate to within approximately ± 0.005 based on 10,000 simulations. All but Loc/S, Loc/G, Loc/N, Ord/N, and ELM/N control type I error at the nominal level of $\alpha = 0.05$.

Test	Multiplicity Correction					
	B	G	S	N	H	Y
Linkage Map ($r = 0.20$)						
Loc	0.0503/0.0400	0.8669/0.8230	0.9775/0.9668	0.9775/0.9668	0.0523/0.0442	0.0128/0.0101
Ord	0.0096/0.0089	0.0396/0.0319	0.0544/0.0495	0.3849/0.3639	0.0006/0.0003	0.0001/0.0000
Max	n/a	n/a	n/a	0.0544/0.0495	n/a	n/a
CET	n/a	n/a	n/a	0.0544/0.0495	n/a	n/a
ERP	n/a	n/a	n/a	0.0544/0.0495	n/a	n/a
ELM	n/a	0.0396/0.0319	0.0544/0.0495	0.3849/0.3639	n/a	n/a
Dense Map ($r = 0.10$)						
Loc	0.0379/0.0348	0.7694/0.7407	0.9351/0.9284	0.9351/0.9284	0.0430/0.0410	0.0080/0.0087
Ord	0.0069/0.0067	0.0290/0.0309	0.0519/0.0497	0.3123/0.3267	0.0003/0.0004	0.0000/0.0001
Max	n/a	n/a	n/a	0.0519/0.0497	n/a	n/a
CET	n/a	n/a	n/a	0.0519/0.0497	n/a	n/a
ERP	n/a	n/a	n/a	0.0519/0.0497	n/a	n/a
ELM	n/a	0.0290/0.0309	0.0519/0.0497	0.3123/0.3267	n/a	n/a
Saturated Map ($r = 0.05$)						
Loc	0.0280/0.0257	0.6270/0.6135	0.8395/0.8387	0.8395/0.8387	0.0359/0.0352	0.0078/0.0066
Ord	0.0059/0.0062	0.0267/0.0266	0.0504/0.0502	0.2589/0.2883	0.0009/0.0004	0.0003/0.0001
Max	n/a	n/a	n/a	0.0504/0.0502	n/a	n/a
CET	n/a	n/a	n/a	0.0504/0.0502	n/a	n/a
ERP	n/a	n/a	n/a	0.0504/0.0502	n/a	n/a
ELM	n/a	0.0267/0.0266	0.0504/0.0502	0.2589/0.2883	n/a	n/a

Examining Table 2, we see that results are similar for all three maps. For the Loc tests, procedures B, H, and Y controlled FDR while N, G, and S did not. For the Max, CET, and ERP methods, no multiplicity correction was necessary since N controlled FDR adequately. For Ord and ELM, all procedures except N gave an acceptable FDR. Procedures in general became more conservative as map density increased.

It is noteworthy that the geometric (G) and spend-as-you-go (S) procedures control FDR for tests based on order statistics (Ord and ELM) but not for locuswise tests (Loc). Clearly, an improper prioritization of the locuswise tests inflates the FDR. It is particularly interesting that S did not inflate the false positive rate for Ord and ELM, while giving a false positive rate near one for Loc. This is an exciting result, since it is a natural inclination to be frugal with alpha, and not overspend for any individual test. At first glance, this result

may seem incredible, in that no inflation of false positive rate is seen, even with a post-hoc definition of the alpha spending function. It should be realized that the sequence of tests in Ord and ELM is not post-hoc, and the order of tests is already taken into account in the null distributions used. In the truly post-hoc situation (Loc/S), for which no correction for the ranking procedure is employed, the expected hyper-inflation of false positive rate is seen.

It is apparent that the distribution for the order statistics partially corrects for the number of tests, and the additional correction using B, G, H, and Y are overly conservative (particularly Ord/Y) in the null case. For Loc, only Y is extremely conservative, as predicted by its authors (Benjamini & Yekutieli, 2001) in this situation. Interestingly, the uncorrected order statistic (Ord/N and ELM/N) has an estimated false positive rate only between 0.25 and 0.39. While still greatly inflated, this is much lower than that for Loc/N. This is due to the nature of the order statistic distributions, where by definition we are looking at a set of 100 tests, whereas in the locuswise tests no consideration is given to the number of tests performed.

Based on the results of the null simulations, we recommend against the combinations Loc/G, Loc/S, Loc/N, Ord/N, and ELM/N. These approaches are excluded from further consideration based on their inflated false positive rates in the null case. The remaining thirteen approaches control the false positive rate in the completely null case, and so will be examined further in non-null cases. The Max, CET, and ERP do not appear to require any multiplicity correction.

4.2 Non-null Cases: False Discovery Rates

In situations where some markers are linked to a QTL while others are unlinked, it is possible to get false positives at unlinked loci (See Figure 2). It is therefore important to remember to examine the FDR in such cases; control of the FDR in null cases does not automatically imply FDR control when QTL are present. Therefore, each approach is evaluated in terms of its ability to control the FDR. Refer to the column labelled FDR in Tables 3 (moderate effect) and 4 (large effect).

When the effect size was moderate ($E_i = 0.5$ for each of 3 trait loci), nine out of the thirteen approaches examined had FDR at or below the nominal level. ELM/S had a mildly inflated FDR for all maps (values 0.06 to 0.12), and ELM/G and Ord/B had slightly inflated FDR (0.08 and 0.06 respectively) for the saturated map ($r = 0.05$) only. When the effect size was large ($E_i = 1$ for each of 3 trait loci), the approaches found to be inflated in the moderate effect size were also inflated for the large effect. In addition, Ord/G and Ord/S had FDR values ranging from 0.07 – 0.09 and 0.15 – 0.19, respectively, tending to increase with marker density. This inflation is likely due to the shadowing effect of order statistics described earlier. The CET had one mild inflation to 0.07 in the dense map only. The remaining eight approaches controlled FDR in all cases, most quite conservatively. In particular, Ord/H and Ord/Y were extremely conservative (FDR 0 to 0.0002).

4.3 Non-null Cases: Power For Detection

The power to detect QTL can be examined in several ways when multiple QTL are involved. We focus on three quantities, the probability of detecting *at least one* QTL, the probability

Table 3: For Moderate Effect simulations, saturated, dense, and linkage maps: FDR, probability of detecting at least one, exactly one, two, and all three loci, and the average fraction (of three) loci found.

Map	FDR	Prob Find > 0	Prob Find 1	Prob Find 2	Prob Find 3	Fraction Found
saturated						
LocB	0.00840	0.4363	0.3553	0.0760	0.0050	0.1741
Loc/H	0.01692	0.5146	0.2270	0.2002	0.0874	0.2965
Loc/Y	0.00307	0.2493	0.1785	0.0618	0.0090	0.1097
Ord/B	0.06040	0.3164	0.0664	0.0687	0.1813	0.2492
Ord/G	0.03249	0.4563	0.2823	0.0781	0.0959	0.2421
Ord/S	0.03768	0.5515	0.2855	0.1198	0.1462	0.3212
Ord/H	0.00026	0.0629	0.0576	0.0052	0.0001	0.0228
Ord/Y	0	0.0212	0.0209	0.0003	0	0.0072
Max/N	0.01429	0.5502	0.4124	0.1264	0.0114	0.2331
CET/N	0.02824	0.5512	0.4065	0.1142	0.0305	0.2421
ERP/N	0.01622	0.5511	0.3859	0.1413	0.0239	0.2467
ELM/G	0.07962	0.4580	0.1377	0.1117	0.2086	0.3290
ELM/S	0.12428	0.5537	0.1578	0.0945	0.3014	0.4170
dense						
Loc/B	0.01195	0.3702	0.3169	0.0503	0.0030	0.1422
Loc/H	0.01706	0.4106	0.2473	0.1288	0.0345	0.2028
Loc/Y	0.00295	0.1847	0.1543	0.0275	0.0029	0.0727
Ord/B	0.04410	0.2135	0.0661	0.0494	0.0980	0.1530
Ord/G	0.02424	0.3572	0.2240	0.0806	0.0526	0.1810
Ord/S	0.03333	0.4240	0.2262	0.0966	0.1012	0.2410
Ord/H	0.00023	0.0326	0.0318	0.0008	0	0.0111
Ord/Y	0	0.0112	0.0111	0.0001	0	0.0038
Max/N	0.01469	0.4207	0.3485	0.0682	0.0040	0.1656
CET/N	0.02517	0.4213	0.3399	0.0644	0.0170	0.1732
ERP/N	0.01640	0.4216	0.3322	0.0814	0.0080	0.1730
ELM/G	0.05651	0.3583	0.1447	0.0942	0.1194	0.2304
ELM/S	0.08741	0.4252	0.1560	0.0866	0.1826	0.2923
linkage						
Loc/B	0.02236	0.2338	0.2125	0.0205	0.0008	0.0853
Loc/H	0.02794	0.2518	0.1951	0.0495	0.0072	0.1052
Loc/Y	0.00578	0.0948	0.0854	0.0089	0.0005	0.0349
Ord/B	0.03870	0.1109	0.0511	0.0279	0.0319	0.0675
Ord/G	0.03867	0.2437	0.1601	0.0617	0.0219	0.1164
Ord/S	0.03999	0.2500	0.1526	0.0585	0.0389	0.1288
Ord/H	0.00040	0.0134	0.0133	0.0001	0	0.0045
Ord/Y	0.00010	0.0043	0.0043	0	0	0.0014
Max/N	0.02518	0.2451	0.2215	0.0228	0.0008	0.0898
CET/N	0.03237	0.2440	0.2162	0.0256	0.0022	0.0913
ERP/N	0.02565	0.2454	0.2191	0.0249	0.0014	0.0910
ELM/G	0.05285	0.2447	0.1393	0.0697	0.0357	0.1286
ELM/S	0.06166	0.2513	0.1351	0.0601	0.0561	0.1412

Table 4: For Large Effect simulations, saturated, dense, and linkage maps: FDR, probability of detecting at least one, exactly one, two, and all three loci, and the average fraction (of three) loci found.

Map	Prob	Prob	Prob	Prob	Fraction	
saturated	FDR	Find > 0	Find 1	Find 2	Find 3	Found
Loc/B	0.00184	0.9835	0.1644	0.4546	0.3645	0.7224
Loc/H	0.01783	0.9935	0.0185	0.1347	0.8403	0.9363
Loc/Y	0.00288	0.9396	0.1088	0.3204	0.5104	0.7603
Ord/B	0.13525	0.9624	0.0202	0.0397	0.9025	0.9357
Ord/G	0.08688	0.9866	0.0698	0.1062	0.8106	0.9047
Ord/S	0.19380	0.9949	0.0222	0.0349	0.9378	0.9685
Ord/H	0.00016	0.6765	0.2857	0.2725	0.1183	0.3952
Ord/Y	0	0.4368	0.2811	0.1296	0.0261	0.2062
Max/N	0.00263	0.9949	0.0986	0.4186	0.4777	0.7896
CET/N	0.06735	0.9947	0.1610	0.3110	0.5227	0.7837
ERP/N	0.01137	0.9949	0.0801	0.3340	0.5808	0.8302
ELM/G	0.29400	0.9866	0.0110	0.0242	0.9514	0.9712
ELM/S	0.37491	0.9950	0.0087	0.0092	0.9771	0.9861
dense						
Loc/B	0.00355	0.9625	0.2458	0.4589	0.2578	0.6457
Loc/H	0.01827	0.9755	0.0670	0.2520	0.6565	0.8468
Loc/Y	0.00363	0.8701	0.2128	0.3600	0.2973	0.6082
Ord/B	0.12036	0.8963	0.0520	0.0958	0.7485	0.8297
Ord/G	0.06912	0.9648	0.0916	0.2036	0.6696	0.8359
Ord/S	0.16258	0.9731	0.0421	0.0602	0.8708	0.9250
Ord/H	0.00019	0.5023	0.3228	0.1498	0.0297	0.2372
Ord/Y	0	0.2823	0.2269	0.0513	0.0041	0.1139
Max/N	0.00421	0.9729	0.2095	0.4573	0.3061	0.6808
CET/N	0.05865	0.9727	0.3005	0.3458	0.3264	0.6571
ERP/N	0.01042	0.9730	0.1731	0.4003	0.3996	0.7242
ELM/G	0.25204	0.9649	0.0328	0.0634	0.8687	0.9219
ELM/S	0.36231	0.9731	0.0228	0.0266	0.9237	0.9490
linkage						
Loc/B	0.01247	0.8437	0.4173	0.3360	0.0904	0.4535
Loc/H	0.02725	0.8715	0.2234	0.3550	0.2931	0.6042
Loc/Y	0.00556	0.6462	0.3421	0.2312	0.0729	0.3411
Ord/B	0.12658	0.6938	0.1280	0.1684	0.3974	0.5523
Ord/G	0.07640	0.8770	0.1524	0.2727	0.4519	0.6845
Ord/S	0.15164	0.8547	0.1069	0.1377	0.6101	0.7375
Ord/H	0.00003	0.2359	0.1982	0.0353	0.0024	0.0920
Ord/Y	0	0.1054	0.0984	0.0067	0.0003	0.0376
Max/N	0.01292	0.8516	0.4087	0.3460	0.0969	0.4638
CET/N	0.04894	0.8501	0.4989	0.2616	0.0896	0.4303
ERP/N	0.01663	0.8518	0.3790	0.3442	0.1286	0.4844
ELM/G	0.19286	0.8772	0.1096	0.1974	0.5702	0.7383
ELM/S	0.32332	0.8549	0.0799	0.1067	0.6683	0.7661

of detecting *all three* QTL, and the average number of QTL detected. See Table 3 for the moderate effect results and Table 4 for the large effect results. These values are also presented as bar charts in Figure 3.

Nine approaches were roughly equivalent in the detection of at least one QTL (Ord/G, Loc/H, CET/N, Max/N, Loc/B, ELM/G, ELM/S, ERP/N and Ord/S). Four approaches (Loc/Y, Ord/Y, Ord/B, and Ord/H) tended to have lower power. The power tended to increase with marker density.

ELM/S was certainly the most powerful test/multiplicity combination in detecting all three QTL, with either Ord/S or ELM/G the next most powerful and Loc/H not far behind this set. In some cases ELM/S had a probability of detecting all three QTL several times that of the CET/N or Max/N. ERP/N, CET/N, Max/N behaved similarly with ERP/N performing the best of the three.

Similar patterns held for the average number detected, with ELM/S the highest and either Ord/S or ELM/G second best and Loc/H just behind those results. In the case with the largest difference (large effect, linkage map) the ELM/S had average power 0.766 and Loc/H had average power 0.604.

4.4 The Best Combinations Of Test And Multiplicity Correction

In order to decide which approach is “best” it is necessary to balance the FDR and the probability of detection. To assist in visualizing the trade-offs involved in this process, two figures are presented. We plot the probability of detecting all three QTL versus FDR (Figures 4 – 5) for the three maps and for two effect sizes. Interestingly, there was no overall ‘best’ (uniformly most powerful level alpha) procedure for all effect sizes and maps. If we consider only those procedures which strictly control FDR at level 0.05 or less, the best choice for the large effect is Loc/H, while the best choice for the moderate effect is Ord/S. Ord/S is more powerful than Loc/H for the large effect also, but at the cost of some FDR inflation.

5 CASE STUDY RESULTS

Table 5 summarizes the results of the case study. This study examined nine related phenotypic traits with 136 individuals and 283 markers on 34 linkage groups with some missing data (Holland *et al.*, 2002). For each test and trait combination, Table 5 gives a list of the linkage groups on which the test procedure detected a QTL. The results are quite consistent with the simulation results, though of course in a data analysis one cannot be sure whether detections are true or false positives.

ELM/S detects the most loci, even detecting all 34 linkage groups for some traits (EFLD, LFLD, and NSGC). From our simulations, we know that this procedure has the highest power, but also that FDR is inflated. Note, however, that for the traits VSGC and PHOTO, ELM/S is consistent with most other tests in detecting no loci. Ord/B, Ord/S, ELM/G and Ord/G follow in rough order of number detected (though the VSGC and PHOTO and traits are an exception). These first five procedures (ELM and Ord) are exactly the ones that showed high power but risked some FDR inflation in the simulations.

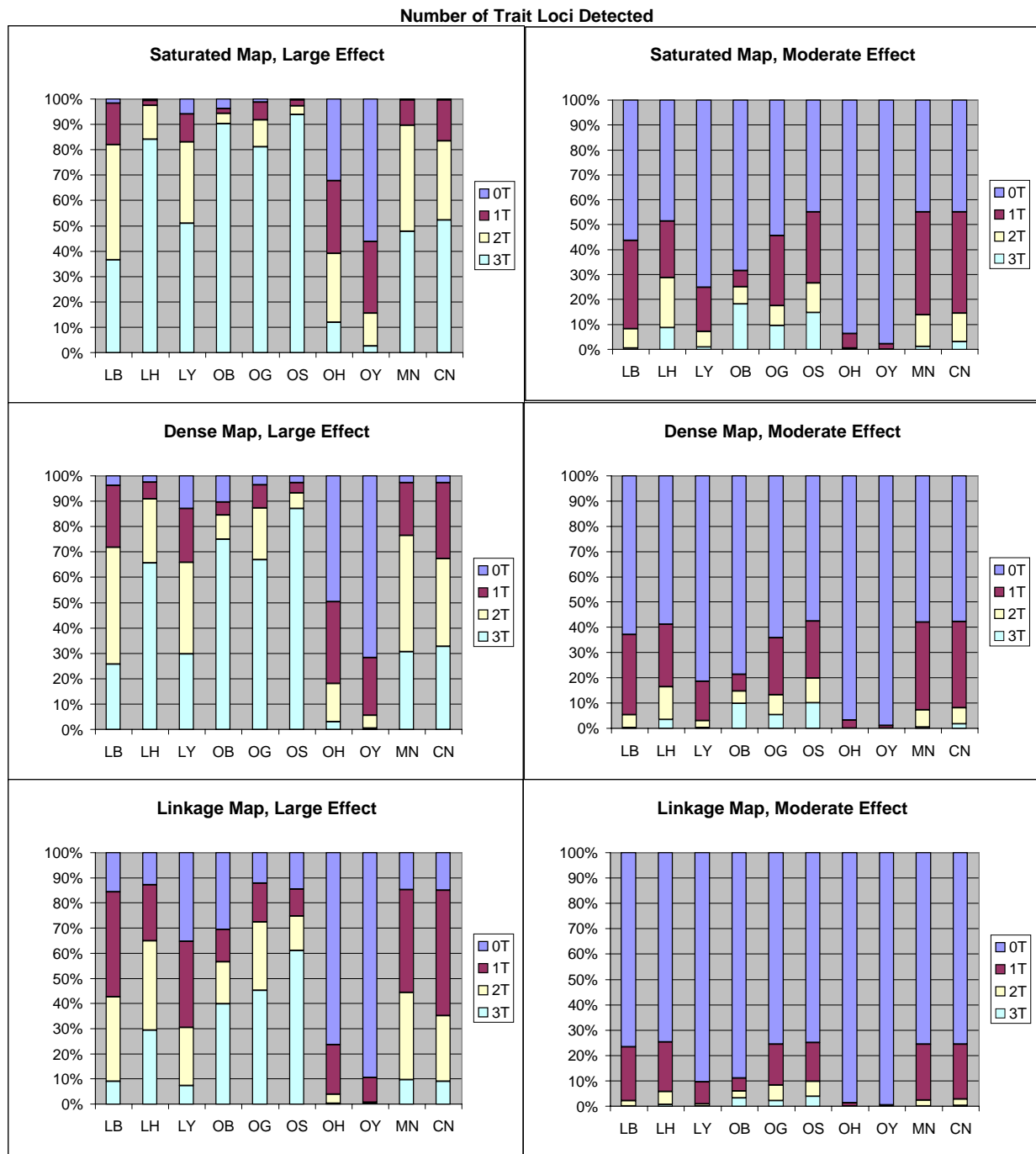


Figure 3: A bar chart comparing the ability of each procedure to correctly detect 0, 1, 2, or all 3 loci. The portion of the bar labelled 3T indicates the percentage of simulations in which all 3 loci were detected, and similarly 2T, 1T, and 0T indicate the percentage of simulations in which exactly 2, 1, or 0 loci were detected. The top of the 2T and 1T bars indicate the percentage of simulations in which at least two and at least one loci were detected, respectively.

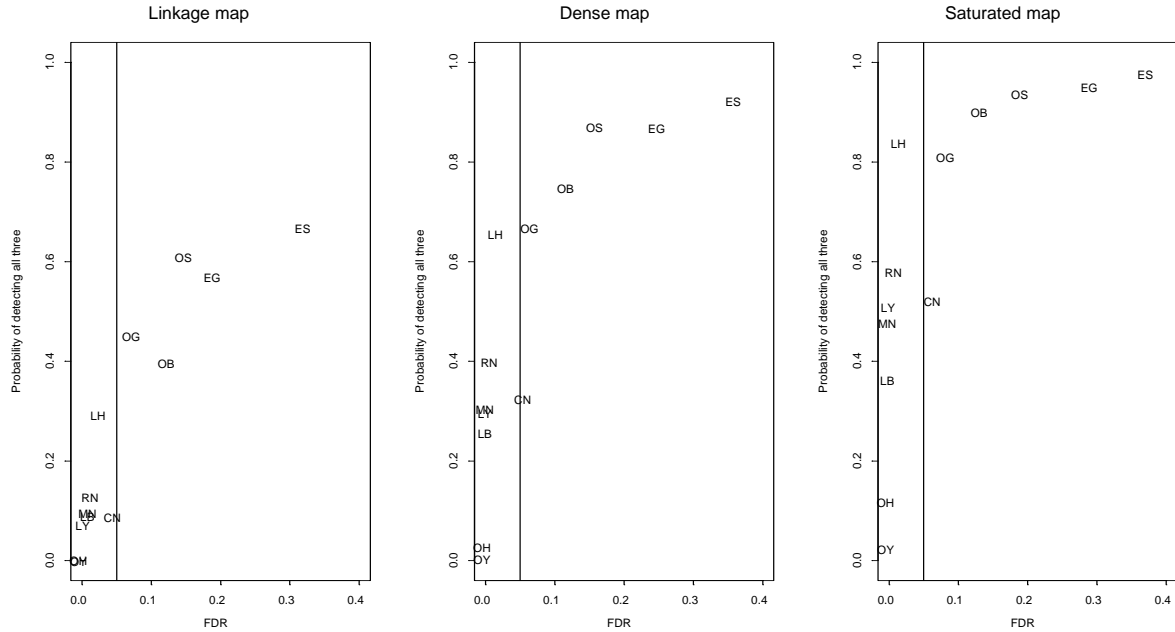


Figure 4: The probability of detecting all three QTL is plotted on the Y axis for effect sizes of 1.0 for each trait locus (“large effect”). On the X axis is the false discovery rate. The three panels are for the linkage map ($r=0.20$), the dense map ($r=0.10$), and the saturated map ($r=0.05$). Plotting symbols are as follows: Loc/B is LB, Loc/H is LH, Loc/Y is LY, CET is CN, Max/N is MN, ERP/N is RN, Ord/G is OG, Ord/S is OS, Ord/B is OB, Ord/H is OH, Ord/Y is OY, ELM/S is ES, and ELM/G is EG.

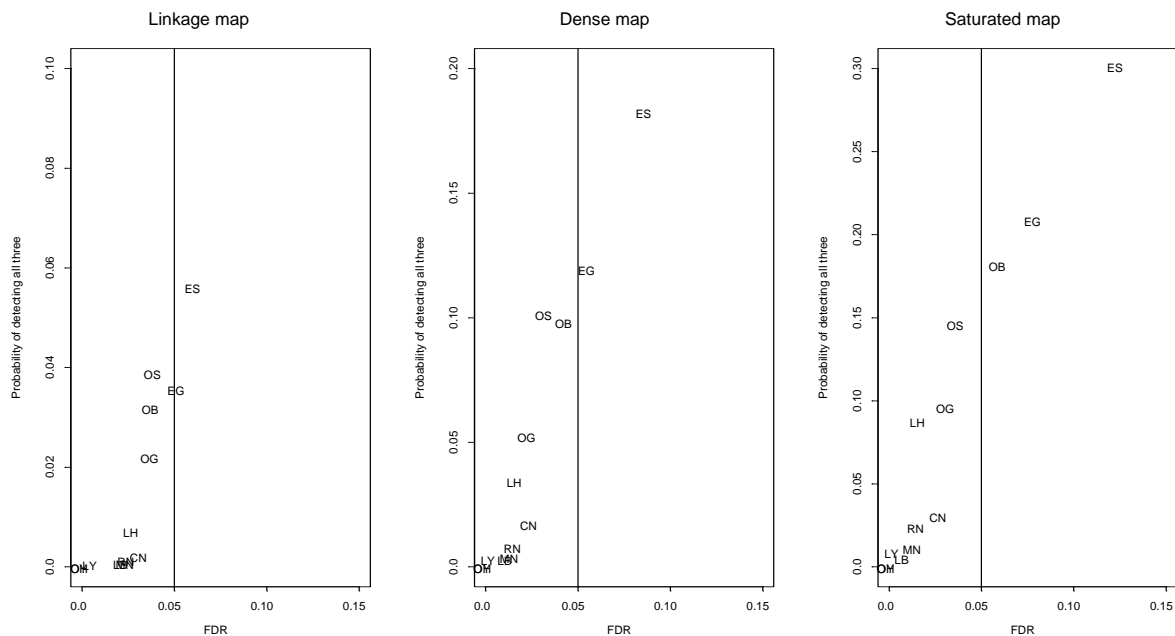


Figure 5: The probability of detecting all three QTL is plotted on the Y axis for effect sizes of 0.5 for each trait locus (“moderate effect”). On the X axis is the false discovery rate. The three panels are for the linkage map ($r=0.20$), the dense map ($r=0.10$), and the saturated map ($r=0.05$). Note the differing scales on the Y-axes. Plotting symbols are as in Figure 4.

Table 5: Performance of Test/Correction procedures in detecting linkage groups for nine traits with markers on 34 linkage groups (Holland, 2002). Number of linkage groups on which a test rejected is given (**bold**): followed by a list of those linkage groups.

	EFLD	LFLD	FLDATE	NLGC	NSGC	VLGC	VSGC	VERN	PHOTO
ELM/S	(34) : all	(34) : all	(17) : 2, 4-6, 9-12, 15-18, 20, 27, 31, 33, 34	(33) : 1-18, 20-34	(34) : all	(29) : 1-5, 7-16, 19-23, 25, 26, 28-34	(0) :	(12) : 1, 3, 6, 7, 9-12, 14, 27, 31, 33	(0) :
Ord/S	(34) : all	(34) : all	(11) : 2, 5, 9-11, 16, 20, 27, 31, 33, 34	(30) : 1-13, 15-17, 20-24, 26-34	(31) : 1-16, 19-24, 26-34	(27) : 1-5, 7-16, 20-23, 25, 26, 28, 29, 31-34	(0) :	(10) : 1, 6, 7, 9, 10, 11, 12, 14, 31, 33	(0) :
Ord/B	(32) : 1-16, 18-24, 26-34	(34) : all	(4) : 9, 31, 33, 34	(26) : 1-13, 15, 16, 20-22, 24, 26, 29-34	(26) : 1-14, 16, 20, 23, 24, 26, 28-34	(20) : 1, 3, 5, 7-10, 12-14, 16, 20-22, 26, 29, 31-34	(2) : 8, 12	(5) : 1, 9, 10, 31, 33	(14) : 1, 3, 4, 8, 10, 15, 16, 20, 22, 24, 31-34
ELM/G	(25) : 1, 2, 4-8, 10-13, 16, 18, 20-24, 26, 29-34	(24) : 1, 2, 4-8, 10, 12, 13, 16, 18, 20-24, 26, 29-34	(5) : 9, 11, 31, 33, 34	(21) : 1-3, 5-13, 16, 20-22, 29, 31-34	(9) : 1, 2, 7, 8, 10, 13, 20, 31, 33	(14) : 1, 5, 8, 10, 12-14, 16, 20, 29, 31-34	(0) :	(6) : 1, 6, 9, 10, 31, 33	(1) : 31
Ord/G	(25) : 1, 2, 4-8, 10-13, 16, 18, 20-24, 26, 29-34	(25) : 1, 2, 4-8, 10, 12, 13, 16-18, 20-24, 26, 29-34	(1) : 31	(21) : 1-3, 5-13, 16, 20-22, 29, 31-34	(2) : 7, 31	(14) : 1, 5, 8, 10, 12-14, 16, 20, 29, 31-34	(0) :	(2) : 10, 31	(1) : 31
Loc/H	(9) : 7, 8, 12, 16, 20, 29, 31, 32, 34	(7) : 2, 7, 8, 20, 29, 31, 32	(2) : 31, 33	(8) : 1, 8, 10, 29, 31, 22, 33, 34	(2) : 7, 31	(4) : 8, 13, 32, 34	(0) :	(2) : 10, 31	(1) : 31
CET/N	(3) : 8, 31, 32	(5) : 2, 8, 15, 31, 32	(1) : 31	(6) : 5, 8, 11, 31, 32, 33	(3) : 7, 20, 31	(5) : 4, 8, 11, 31, 32	(0) :	(1) : 31	(0) :
Max/N	(2) : 31, 32	(2) : 31, 32	(1) : 31	(2) : 31, 32	(1) : 31	(2) : 32, 34	(0) :	(1) : 31	(0) :
ERP/N	(2) : 31, 32	(2) : 31, 32	(1) : 31	(2) : 31, 32	(1) : 31	(2) : 32, 34	(0) :	(1) : 31	(0) :
Loc/B	(2) : 31, 32	(2) : 31, 32	(1) : 31	(2) : 31, 32	(1) : 31	(1) : 32	(0) :	(1) : 31	(0) :
Loc/N	(15) : 1, 2, 5, 7, 8, 12, 16, 20, 22, 23, 29, 31-34	(15) : 1, 2, 5, 7, 8, 12, 13, 16, 20, 23, 29, 31-34	(10) : 2, 9-11, 16, 20, 27, 31, 33, 34	(14) : 1, 5, 8, 10, 12-14, 16, 20, 29, 31-34	(10) : 1, 6, 7, 9-12, 14, 31, 33	(12) : 2, 4, 7, 8, 12, 13, 16, 20, 29, 31, 32, 34	(16) : 1, 5-8, 10, 12, 13, 16, 20, 22, 29, 31-34	(13) : 1, 2, 7, 8, 10, 11, 12, 13, 16, 20, 29, 31, 33	(14) : 1, 2, 4, 7, 8, 10, 12, -14, 16, 20, 28, 29, 31

Though statistical theory and our simulations clearly demonstrate that Loc/N is undesirable because of its enormous FDR inflation, it is a common first step when people are exploring model selection techniques. In addition to this FDR inflation, our results also indicate that it generally detects fewer loci than the ELM and Ord procedures, which have comparatively mild FDR inflation. In the few cases where Loc/N detects loci that other methods do not (VSGC, PHOTO, VERN), given the high power of the other methods we suspect many are false positives.

The Loc/H procedure detected fewer loci than the Ord and ELM procedures, which is expected, since as shown in the simulations Loc/H controls FDR below the nominal level, resulting in a conservative procedure. Loc/H in general had the most detections among procedures which controlled FDR, such as the CET/N.

The CET/N procedure detected more loci than ERP/N, Max/N, and Loc/B, which all behaved similarly. However, there were instances in which the CET detected loci which Loc/H did not. For example, for the VLGC trait, CET detected QTL on 4 and 11 which were detected by almost no other procedures, while ERP/N and Max/N detected QTL on linkage group 34 which CET did not detect. Similarly, for the NLGC trait, CET detected 11 which was not found by Loc/H. The CET is unique in that its test statistics are recalculated based on partitioning at previously detected loci. Thus it may have the ability to detect epistatic loci with weak main effects.

6 DISCUSSION

We proposed new QTL detection methods and multiplicity correction procedures and undertook a comprehensive comparison of existing approaches for their ability to detect multiple QTL and control FDR in the case of multiple additive QTL. The new approaches using order statistics improve upon existing methods in several situations. The choice of multiplicity correction is found to have a substantial impact on detection ability. With the right combination of approaches, increased marker density can improve power to detect multiple QTL.

Previous work has shown that when the total type I error (α) is allocated to the first order statistic (Max/N), the resulting test is always at least as powerful as the locuswise test with a Bonferroni correction (Loc/B) at detecting at least one QTL. We confirmed this result, and we have determined the properties of using the maximum as a mechanism for detecting multiple QTL.

The CET, proposed by Doerge and Churchill (1996), is an elegant solution to the control of the false positive rate and in all the cases we examined, we found CET/N to have FDR at or less than the nominal type I error rate with no additional multiplicity correction. In addition, the CET can detect at least one QTL with very high power and has reasonable power to detect multiple QTL. The CET is the most computationally intensive of the methods explored. However, in the data analysis for the case study in this paper the total computing time for 10,000 permutations (all methods including the CET) was under 5 real time minutes for a 1Ghz PC. The final results, based upon a million permutations, took under 3 hours real time. Even the additional calculations required by interval mapping are unlikely to slow

down the computation enough to render it impractical. Though increasing marker density will also increase the time involved, it is unlikely to take more than an overnight run on a fast PC particularly as PC technology is improving at pace with our ability to type markers. Thus we do not see computational effort as a barrier to the use of any of the procedures examined in this paper.

The ERP, a simplified version of the CET, showed very similar results to the CET in the simulated moderate effect cases, and was slightly more powerful than the CET for the large effect cases. The ERP additionally had a lower FDR than the CET. These observations suggest that the major advantage of this approach comes from the elimination of linked markers rather than the stratification and partitioning on previously detected loci. However, it is noteworthy that in the data analysis, the CET detected some loci that the ERP did not. While the simulated effects were purely additive, the real data had several loci thought to be epistatic (Holland *et al.*, 2002). The partitioning may allow the CET to detect epistatic loci with weak main effects, an advantage the CET has over every other method examined. However, the CET does not identify every locus the MAX identified, i.e. in the case study for the trait VLGC the MAX identified linkage group 34, but the CET did not. This is perhaps due to the reduction in sample size from the partitioning in CET.

An existing procedure to control the FDR (Loc/H) is among the best of the procedures examined. Loc/H offers a balanced tradeoff between FDR control and power, and appears to be the best choice for large effects and dense maps. Though it was not the most powerful approach in all scenarios examined, it was not dramatically underpowered compared to approaches that fail to control the FDR.

We proposed alternative methods for detecting multiple QTL using the joint set of order statistics combined with an alpha spending function (Ord, ELM). The results presented here show that the combination of order statistics with alpha spending is a reasonable alternative to current practice in many situations. We discuss three possible spending functions: the Bonferroni, the geometric and the spend as you go. Each of these approaches is relying on the premise that the sum of the individual α_i should be less than or equal to the nominal α . However, these functions are not optimized, in that an optimal function would allocate the total type I error depending upon the number of true null and alternative hypotheses. This is an interesting area to explore and it is possible that these functions could be improved using the methodology discussed in (Storey, 2002).

The procedure proposed for estimating p-values for order statistics is completely general, allowing test statistics to have different distributions at different loci. We find that the Ord, ELM and ERP approaches have comparable power to CET/N in the detection of at least one locus. In detecting multiple loci an interesting trend appears. In the linkage map, Ord/G has substantially more power than ERP, CET/N, and Loc/H to detect all three loci, while in the dense map the powers are comparable, and in the saturated map Loc/H and CET/N have slightly higher power than Ord/G.

In all cases the FDR for Ord/G is controlled appropriately or has very mild inflation. Of the approaches using order statistics that include all loci (Ord), Ord/S tends to have the highest detection rates, and the FDR is controlled for Ord/S except when the effect size is large. The elimination of linked markers approaches (ELM) give the highest detection rates and ELM/S has consistently the highest detection rate. This is largely due to the parsimonious allocation of alpha in these cases, since once a QTL is detected, markers linked

to it are eliminated from further testing. In the null case, the FDR is less than or equal to the nominal level, even when the spend-as-you-go function is used with Ord or ELM. However, the FDR in the non-null cases is inflated and while the inflation is modest in moderate effect sizes, when effect sizes are large the inflation increases and can be as large as 0.40. This is due to the “shadow” cast by the true QTL in the joint set of order statistics. This effect is less using the geometric alpha spending function.

These results point to some clear recommendations depending upon the experimental circumstance. In cases where the effect size is moderate or small, Ord/S controls the FDR and has higher detection rates than all other procedures. However, the ELM/S has substantially higher detection rates, sometimes even doubling the detection probability. Since the FDR is controlled for this procedure in the completely null case, the experimenter may be willing to risk the inflation of the FDR found by using the ELM/S in order to increase the detection probability.

When the effect size is large, the choice of the best statistic will depend both upon the map density and the researcher’s willingness to accept FDR inflation to obtain the highest possible detection rate. In saturated maps, Loc/H has the highest power among tests that control the FDR. Conversely, in linkage maps, Ord/G has higher power than Loc/H and still controls the FDR. In dense maps, the two approaches are comparable both in detection rates and FDR. As with the moderate effects, the ELM/S has a higher detection rate than any of these methods, with the cost of an inflation of the FDR which increases with effect size.

Some clear results come from examining the behavior of Loc/N. While Loc/N is sometimes advocated as a way to be inclusive, and to avoid type II error, it is clear from both simulation and case study results that the other procedures evaluated identify more contributing loci. For example, in the trait EFLD, Loc/N identifies 15 linkage groups while Ord/G and ELM/G identify 25. For the trait VSGC, Loc/N identifies 16 linkage groups while these other approaches identify no linkage groups. The Loc/N approach inflates type I error to such an extent that reasonable interpretation of loci detected is difficult. The Loc/N procedure clearly is unacceptable because of its enormous FDR inflation. But these results also indicate that it is not even as powerful as the Ord and ELM procedures, which have relatively mild FDR inflation. Thus, even if one is tolerant of FDR inflation, there are smarter choices than Loc/N.

Among the methods we examined, higher detection power can be achieved by using methods that fail to control the FDR. Allowing FDR inflation must be carefully balanced either by the use of subsequent model selection techniques or by planned follow-up experiments. In the case where the QTL analysis is part of a series of experiments, the decision about which method to use may well depend upon the resources available for follow-up as well as the individual investigator’s concern over false negatives and false positives. If a whole genome microarray experiment is used as the next step (Wayne & McIntyre, 2002) then there is no incremental dollar cost to examining additional regions and it may be appropriate to consider minimizing the false negative rate in the QTL mapping. In this case the ELM/S or the Ord/S may be worth examining. However, if other techniques are used, such as fine mapping, the experimenter will need to weigh available resources against an increased possibility of a false positive and the decision about which region to follow up should be a consequence of whether false positives or false negatives are more costly in that particular

situation.

As QTL mapping matures, scientists are looking for multiple QTL and are not content with the detection of “at least one” QTL. Instead, QTL mapping has become the first step in unravelling the genetic contributions to complex traits. It is important to recognize the multistep process to uncovering underlying genetic contributions in designing the statistical methodology for each step in the analysis. Careful consideration of steps to come after an initial QTL experiment will lead to informed choices about the appropriate balance between false positive and false negative rates. In this paper we present new procedures that improve power to detect multiple QTL and in cases where effect sizes are moderate, dramatically improve the probability of detecting multiple QTL, without increasing false discovery rates. The examination of a case study clearly underlines the differences among the approaches. A C++ program which executes these different approaches for a particular data set has been written and is available by request (KLS). The methodology described here is easily extended to interval mapping, and can be used in conjunction with any continuous test statistic.

References

- BENJAMINI, YOAV, & HOCHBERG, YOSEF. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society, series B*, **51**(1), 289–300.
- BENJAMINI, YOAV, & YEKUTIELI, DANIEL. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, **29**(4), 1165–1188.
- CASELLA, GEORGE, & BERGER, ROGER L. 1990. *Statistical inference*. Belmont, CA: Duxbury. pp. 232–233.
- CHURCHILL, GARY A, & DOERGE, REBECCA W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138**(3), 963–971.
- DEMETS, D L, & LAN, K K G. 1994. Interim analysis – the alpha-spending function-approach. *Statistics in medicine*, **13**(13–14), 1341–1352.
- DOERGE, REBECCA W, & CHURCHILL, GARY A. 1996. Permutation tests for multiple loci affecting a quantitative trait. *Genetics*, **142**(1), 285–294.
- EDGINGTON, EUGENE S. 1995. *Randomization tests*. 3rd edn. New York: Marcel Dekker, Inc.
- FISHER, RONALD A. 1935. *The design of experiments*. 3rd edn. London: Oliver & Boyd Ltd.
- GOOD, PHILLIP. 1994. *Permutation tests*. New York: Springer.
- HOCHBERG, Y, & TAMHANE, A C. 1987. *Multiple comparison procedures*. New York: Wiley.

- HOLLAND, J. B., PORTYANKO, V. A., HOFFMAN, D. L., & LEE, M. 2002. Genomic regions controlling vernalization and photoperiod responses in oat. *Theoretical and applied genetics*, **105**, 113 – 126.
- HSU, JASON C. 1996. *Multiple comparisons. Theory and methods*. Chapman & Hall.
- LANDER, ERIC S., & BOTSTEIN, DAVID. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**(January), 185–199.
- LEE, H., DEKKERS, J. C. M., SOLLER, M., MALEK, M, FERNANDO, R. L., & ROTH-SCHILD, M. F. 2002. Application of the false discovery rate to quantitative trait loci: interval mapping with multiple traits. *Genetics*, **161**(2), 905–914.
- MCINTYRE, LAUREN M, MARTIN, EDEN R, SIMONSEN, KATY L, & KAPLAN, NORMAN L. 2000. Circumventing multiple testing: A multilocus Monte Carlo approach to testing for association. *Genetic epidemiology*, **19**(1), 18–29.
- STOREY, JOHN D. 2002. A direct approach to false discovery rates. *Journal of the Royal statistical society series B*, **64**(3), 479–498.
- STOREY, JOHN D., & TIBSHIRANI, ROBERT. 2003. Statistical significance for genomewide studies. *Proceedings of the national academy of sciences, USA*, **100**(16), 9440 – 9445.
- WAYNE, MARTA L., & MCINTYRE, LAUREN M. 2002. Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the national academy of sciences, USA*, **99**(23), 14903–14906.
- WAYNE, MARTA L., JACOBS, L., KUNTZ, A., & SHEN, L.-Y. 2004. Fine scale mapping of QTL for ovariole number in *drosophila melanogaster* using deficiency mapping. In review.
- WESTFALL, PETER H., & YOUNG, S. STANLEY. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, NY.