Quantile-based Permutation Thresholds for QTL Hotspots

Elias Chaibub Neto[1]

Mark P Keller[2]

Andrew F Broman[2]

Alan D Attie[2]

Ritsert C Jansen[3]

Karl W Broman[4]

Brian S Yandell[5,6]

[1] Department of Computational Biology, Sage Bionetworks, Seattle, Washington, 98109.

[2] Department of Biochemistry, University of Wisconsin - Madison, Madison, Wisconsin, 53706.

[3] Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, Groningen, The Netherlands.

[4] Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison, Madison, Wisconsin, 53706.

[5] Department of Statistics, University of Wisconsin - Madison, Madison, Wisconsin, 53706.

[6] Department of Horticulture, University of Wisconsin - Madison, Madison, Wisconsin, 53706.

Running title: QTL Hotspot Quantile Thresholds

Key words:

1. Hotspots.
2. Permutation tests.
3. Multiple traits.
4. LOD scores.
5. Quantitative trait loci.

Corresponding author:

Brian S. Yandell

Department of Statistics

University of Wisconsin - Madison

1300 University Avenue

Madison, Wisconsin 53706

byandell@wisc.edu

Abstract

QTL hotspots (genomic locations affecting many traits) are a common feature in genetical genomics studies, and are biologically interesting since they may harbor critical regulators. Therefore, statistical procedures to assess the significance of hotspots are of key importance. One approach, randomly allocating observed QTLs across the genomic locations separately by trait, implicitly assumes all traits are uncorrelated. Recently, an empirical test for QTL hotspots was proposed based on the number of traits that exceed a predetermined LOD value, such as the standard permutation LOD threshold. The permutation null distribution of the maximum number of traits across all genomic locations preserves the correlation structure among the phenotypes, avoiding the detection of spurious hotspots due to non-genetic correlation induced by uncontrolled environmental factors and unmeasured variables. However, by only considering the number of traits above a threshold, without accounting for the magnitude of the LOD scores, relevant information is lost. In particular, biologically interesting hotspots composed of a moderate to small number of traits with strong LOD scores may be neglected as non-significant. In this paper we propose a quantile-based permutation approach that simultaneously accounts for the number and the LOD scores of traits within the hotspots. By considering a sliding scale of mapping thresholds, our method can assess the statistical significance of both small and large hotspots. Although the proposed approach can be applied to any type of heritable high volume 'omic' data set, we restrict our attention to eQTL analysis. We assess and compare the performances of these three methods in simulations and we illustrate how our approach can effectively assess the significance of moderate and small hotspots with strong LOD scores in a yeast expression data set.

Introduction

QTL hotspots, groups of traits co-mapping to the same genomic location, are a common feature of genetical genomics studies. Genomic locations associated with many traits are biologically interesting since they may harbor influential regulators. Therefore, statistical procedures aiming to assess the significance of such hotspots are of key importance.

Brem et al. (2002) and Schadt et al. (2003) detected hotspots by dividing the genome of an organism into equally spaced bins, counting the number of expression traits with a QTL in each bin. A hotspot was considered significant if it had more traits with quantitative trait loci (QTL) than expected if the expression QTLs were randomly distributed across the genome. Darvasi (2003) and Perez-Enciso (2004) pointed out that these hotspots may arise as an artifact of high correlation among expression traits, rather than by the action of a common master regulator. Non-genetic mechanisms, uncontrolled environmental factors and unmeasured variables are capable of inducing strong correlations among clusters of traits. Hence, whenever a trait shows a spurious linkage, many correlated traits will likely map to the same locus, creating a spurious eQTL hotspot. Furthermore, multiple testing and relaxed mapping thresholds may inflate the hotspots (Darvasi 2003).

West et al. (2007) and Wu et al. (2008) proposed a permutation test where the positions of the eQTLs detected in the original data set are permuted across the genome separately by trait, using the distribution of the maximal number of expression traits across the genome to assess hotspot significance. This Q-method permutes QTL positions, not phenotype or genotype data and improves upon the permutation approaches of Brem et al. (2002) and Schadt et al. (2003) since it accounts for multiple testing across the genome. However, the Q-method implicitly assumes traits are uncorrelated and hence underestimates the clustered pattern of spurious eQTLs for correlated traits.

Breitling et al. (2008) proposed a permutation test that randomized rows in the

marker data relative to rows in the trait data, preserving the correlation structure among phenotypes. The null distribution for this N-method of hotspot sizes depends on the number N of traits with LOD score exceeding a predetermined LOD threshold at each locus. The choice of LOD threshold is important: higher LOD thresholds yield smaller sized spurious hotspots by chance under the null hypothesis of no hotspots. Two natural LOD threshold choices are the Churchill-Doerge (1994) single-trait LOD threshold, controlling genome-wide error rate (GWER) for one trait, and a conservative permutation threshold for the maximum LOD score across all traits and all genomic locations. The former allows large hotspots by chance under the null distribution (Breitling et al. 2008). The latter favors small hotspots composed of traits with high LOD scores under the null. Which threshold is more appropriate?

We propose a quantile-based permutation approach, the NL-method, with a sliding scale of thresholds ranging from the conservative to the single-trait threshold, jointly considering hotspot size and the distribution of LOD scores among correlated traits. Hence, even a small hotspot, with a modest number of correlated traits all having high LOD scores at a location, can be significant. The NL-method controls the genome wide error rate across a range of possible hotspot sizes. Explicitly, we examine spurious hotspot size n, ranging from 1 to N, with N the hotspot size threshold delivered by the N-method using the single-trait LOD threshold. While the N-method yields the minimum significant hotspot size for a fixed LOD threshold, we turn the problem around and determine an empirical LOD threshold given a spurious hotspot size.

We assessed and compared the performances of the NL-, N- and Q-methods using: (i) simulated examples, where we generated hotspots with varying LOD score distributions for data sets with correlated and uncorrelated traits; and (ii) simulation studies, where we generated null data sets, i.e., data sets where none of the phenotypes had any QTLs, and

assessed the error rates of the three procedures under different levels of correlation among the traits. Application of the NL-method to a yeast data set detected additional moderate and small hotspots considered non-significant by the N-method and avoided spurious hotspots detected by the Q-method. This ability to assess the statistical significance of hotspots with varying sizes and LOD score distributions has the potential to yield important additional biological discoveries.

Methods

The Q- and N methods

The now standard permutation threshold method for QTL mapping (Churchill and Doerge 1994) estimates the null distribution of the genome-wide maximum LOD score by shuffling the phenotypes relative to the genotype data, breaking the association between the phenotype and the genotypes. Our interest, though, is to assess the significance of QTL hotspots. This section presents two different permutation schemes that have been used in hotspot analysis. Supplementary Figure S1 shows a schematic of the genotype data, phenotype data and the output of hotspot analysis in genetical genomics experiments.

The first permutation scheme for the Q-method (Supplementary Figure S2) derives the null distribution of hotspot sizes by permuting the cells of the observed QTL matrix along its rows, independently for each column; that is, the QTLs are permuted across genomic locations separately by trait. The Q-method does not account for the correlation structure among the phenotypes and, contrary to the Churchill-Doerge and Breitling's permutation tests, does not break the connection between phenotypes and genotypes. The Q-method permutation null distribution is generated under the assumption that phenotypes are uncorrelated. Violation of this assumption leads to a severe underestimation of the null

distribution of hotspot sizes and to detection of an many spurious hotspots, as shown by Breitling et al. (2008) in the re-analysis of the Wu et al. (2008) data and illustrated in the simulation study and examples below.

The second permutation scheme (Breitling et al. 2008; Supplementary Figure S3), used for the N- and NL-methods breaks the connection between genotypes and phenotypes while preserving the correlation structure separately within each type of data by permuting the rows of the phenotype data matrix relative to the rows of the genotype data. Mapping analysis is redone for all traits with the permuted data. This scheme, a direct extension of Churchill and Doerge (1994), preserves the correlation structure among the phenotypes, accounting for spurious hotspots due to non-genetic correlation.

The NL-method

In linkage analysis of a single phenotype we are usually interested in controlling the genome wide error rate (GWER) of falsely detecting a QTL. For a given error rate α, we determine a single trait mapping LOD threshold λ such that

$$\Pr\Big(\max_k \{LOD_k\} \geq \lambda \mid \text{there is no QTL anywhere in the genome}\Big) = \alpha\ , \tag{1}$$

where $k = 1, \ldots, K$ represents the genomic locations being tested for linkage. In the presence of multiple phenotypes we need to account for multiple testing across traits, and seek to control the genome wide error rate of falsely detecting at least one QTL associated with any of the T traits at level α. We determine a LOD threshold λ_c such that

$$\Pr\Big(\max_{t,k} \{LOD_{t,k}\} \geq \lambda_c \mid \text{none of the traits have a QTL anywhere in the genome}\Big) = \alpha \tag{2}$$

where $t = 1, \ldots, T$, $k = 1, \ldots, K$ and λ_c represents a more conservative LOD threshold that controls the probability that any of the traits have one or more false linkages anywhere in the genome.

Which threshold is more adequate depends on the underlying situation. We discard small hotspots with strong LOD scores when we adopt λ, but we miss hotspots composed of many traits with linkages barely reaching the single trait mapping threshold using λ_c. Therefore, we propose a sliding scale of thresholds $\lambda_{n,\alpha}$ for n varying from 1 to N, where N represents the hotspot size (i. e., the number of traits with significant LOD at a given genomic location) expected by chance under the N-method's permutation scheme, computed using the LOD score threshold λ.

Given a fixed mapping threshold, the N-method determines the hotspot size expected by chance. We turn the problem around: given a fixed spurious hotspot size n, the NL-method determines the associated mapping threshold $\lambda_{n,\alpha}$. We adopt $\max_k \{qLOD_k(n)\}$ as a test statistic where $qLOD_k(n)$ corresponds to the nth LOD value of an ordered sample of T LOD scores, ordered from highest to lowest. Note that by taking the maximum of $qLOD_k(n)$ across the genome we are able to control the genome wide error rate associated with the $qLOD_k(n)$ statistic. Explicitly, we control

$$\Pr\Big(\max_k \{qLOD_k(n)\} \geq \lambda_{n,\alpha} \mid \text{none of the traits have a QTL anywhere in the genome} \Big) = \alpha. \tag{3}$$

In other words, by adopting a QTL mapping threshold $\lambda_{n,\alpha}$ we control GWER at level α of detecting at least one spurious hotspot of size n or higher somewhere in the genome, given that none of the traits have a QTL anywhere in the genome.

Observe that when $n = N$ the LOD threshold λ (that controls the detection of a false QTL at a GWER α), matches $\lambda_{N,\alpha}$ (that controls the detection of a false hotspot of size N or higher), and $qLOD_k(N)$ corresponds to LOD_k. Therefore, when $n = N$, the quantity

in (3) reduces to (1). Similarly, when $n = 1$ we have that $\lambda_c = \lambda_{1,\alpha}$ and $qLOD_k(1) = \max_t \{LOD_{t,k}\}$, so that $\max_k \{qLOD_k(1)\} = \max_{t,k} \{LOD_{t,k}\}$ and (3) reduces to (2).

Finally, the quantity in (3) is the probability of detecting at least one spurious hotspot of *size exactly* n somewhere in the genome given that none of the traits have a QTL anywhere in the genome. However, under the null hypothesis of no QTLs, detecting a hotspot of size $n^* > n$ is less likely than a hotspot of size n, therefore, if a threshold $\lambda_{n,\alpha}$ controls the full-null GWER for a hotspot of size n, it will also control the full-null GWER for a hotspot of size larger than n. Below, we detail the permutation algorithm for LOD quantiles.

NL-method algorithm: For a fixed hotspot size, $n = 1, \ldots, N$, we obtain the permutation LOD threshold that controls the genome-wide error rate of detecting at least one hotspot of size n or higher, at a fixed α level as follows:

1. Permute the data according to the N-method to break the associations among genotypes and phenotypes, while keeping the correlation structure among phenotypes intact. Compute the LOD scores for all phenotypes across all genomic locations.

2. Process the LOD profile of each trait as follows: (1) determine the LOD peak for each chromosome; (2) compute the LOD support interval around the peak (Lander and Botstein 1989, Dupois and Siegmund 1999, Manichaikul et al. 2006); and (3) set to zero the LOD scores outside the LOD support interval (and below the single trait mapping threshold).

3. For a fixed hotspot size n, compute $qLOD_k(n)$ for genomic positions $k = 1, \ldots, K$, and store its maximum.

4. Repeat steps (1) to (3), B times. The histogram of the B permutation samples of

$\max_k \{qLOD_k(n)\}$ is an estimate of the null distribution of the test statistic for at least one spurious hotspot of size n or higher anywhere in the genome, given that none of the traits have a QTL anywhere in the genome.

5. The upper $(1-\alpha)$-quantile of the permutation sample generated in step (4) is our threshold (denoted by $\lambda_{n,\alpha}$).

This algorithm is analogous to the traditional permutation test, replacing LOD scores by LOD quantiles. Chen and Storey (2006) perform permutation tests for distinct quantile-based statistics, in a different context, where they consider a set of relaxed significance thresholds to detect multiple QTLs for a single trait.

The LOD score processing step described in Step 2 of the NL-method algorithm confines the hotspot location on the chromosome. LOD support intervals are the most commonly used interval estimates for the location of a QTL. Following Manichaikul et al. (2006) we adopt 1.5-LOD support intervals for a backcross, and 1.8-LOD support intervals for an intercross, decreasing the spread of the hotspot as illustrated in Supplementary Figure S4.

Finally, note that instead of running a genuine, but infeasible, multiple trait joint analysis to account for the correlation structure among the traits, our strategy is to perform multiple single trait mapping analyses with an appropriate multiple trait permutation threshold. Justification for permutation tests in the context of QTL mapping of a single phenotype is given by Churchill and Doerge (1994). A sufficient condition for a permutation test to have type I error rate held at the nominal level is that the observations are exchangeable (Good 1994, p.203, Lehmann 1986, p.231). Violation of the exchangeability assumption can lead to an inflation of type I error rates (Churchill and Doerge 2008). Observations are exchangeable if the joint probability of any outcome is the same

irrespective to the order in which the observations are considered (Lehmann 1986, p. 231). Permutation tests remain valid for a multivariate response that can be reduced to a single-valued test statistic (Good 1994, chapter 5). Exchangeability of subjects under the null distribution follows by the construction of an experimental cross. At a fixed genomic location our test statistic corresponds to $qLOD_k(n)$. Across the whole genome, we adopt $\max_k\{qLOD_k(n)\}$ as our genome-wide and single-valued test statistic.

Results

Simulated examples

In this section we illustrate the application of the Q-, N- and NL-methods to two simulated data sets: one with highly correlated traits, and the other with uncorrelated traits. We generated data from backcrosses composed of 112 individuals with 16 chromosomes of length 400cM containing 185 equally spaced markers each, and phenotype data on 6,000 traits. The phenotype data was generated according to the following models,

$$Y_k = \beta\, M + \theta\, L + \epsilon_k, \quad \text{if } Y_k \text{ belongs to a hotspot,}$$
$$Y_k = \theta\, L + \epsilon_k, \qquad\quad \text{if } Y_k \text{ does not belong to a hotspot,}$$

where $L \sim N(0, \sigma^2)$ represents a latent variable that affects all $k = 1, \ldots, 6000$ traits; θ represents the latent variable effect on the phenotype and works as a tuning parameter to control the strength of the correlation among the traits; $M = \gamma\, Q + \epsilon_M$ represents a master regulator trait that affects the phenotypes in the hotspot; β represents the master regulator effect on the phenotype; Q represents the QTL giving rise to the hotspot. Note that traits composing the hotspot are directly affected by the master regulator M and map to Q indirectly; γ represents the QTL effect on the master regulator; and ϵ_k and ϵ_M represent independent and identically distributed error terms following a $N(0\,,\,\sigma^2)$ distribution.

In both examples we simulated 3 hotspots: (i) a small hotspot located at 200cM on chromosome 5 showing high LOD scores, see panels (a) and (d) on Supplementary Figure S5; (ii) a big hotspot located at 200cM on chromosome 7 showing LOD scores ranging from small to high, see panels (b) and (e) on Figure S5; and (iii) a big hotspot located at 200cM on chromosome 15 showing LOD scores ranging from small to moderate, see panels (c) and (f) on Figure S5.

In both simulations we set $\sigma^2 = 1$ and $\gamma = 2$. QTL analysis was performed using Haley-Knott regression (Haley and Knott, 1992) with the R/qtl software (Broman et al. 2003). We adopted Haldane's map function and genotype error rate of 0.0001. Because we adopted a dense genetic map, our markers are approximately 2.16cM apart, we did not consider putative QTL positions between markers.

In the first example, denoted simulated example 1, we adopted latent effect equal to 1.5. In the second example, denoted simulated example 2, we adopted latent effect equal to 0 and simulated uncorrelated traits. Panels (a) and (b) of Supplementary Figure S6 shows the distribution of all pairwise correlations among the 6,000 traits for both simulated examples. These extreme examples illustrate the effect of phenotype correlation on QTL hotspot sizes. The correlation of the real data is actually intermediate (see panel c).

Figure 1 shows the results for the Q- and N-methods for simulated example 1 using $\alpha = 0.05$. Panel (a) shows the hotspot architecture computed using a single trait LOD threshold of 3.65, i.e., at each genomic location the plot shows the number of traits with LOD score above 3.65. In addition to the simulated hotspots on chromosomes 5, 7 and 15, panel (a) shows a few spurious hotspots, including a big hotspot on chromosome 8. The blue and red lines show the N- and Q-method's thresholds, 560 and 7, respectively. In this example the N-method was unable to detect any hotspots, whereas the Q-method detected false hotspots on chromosomes 3, 6, 8, 9, 12 and 16. Panels (b) and (c) show

the hotspot size null distributions, and the 5% significance thresholds for the N- and Q-methods, respectively.

Figures 2 and 3 show the NL-method analysis results for simulated example 1 using $\alpha = 0.05$. Panels (a)-(d) on Figure 2 present the hotspot architecture inferred using 4 different quantile-based permutation thresholds. Panel (a) presents the hotspot architecture inferred using a LOD threshold of 7.07. Only the true hotspots (on chromosomes 5, 7 and 15) were significant by this conservative threshold. Panel (b) presents the hotspot architecture computed using a LOD threshold of 4.93, that aims to control GWER ≤ 0.05 for spurious hotspots of size 50. The hotspots on chromosomes 5, 7 and 15 were detected by this threshold. Panels (c) and (d) show the hotspot architectures using LOD thresholds of 4.21 and 3.72, respectively. Only the hotspot on chromosome 7 was detected as significant for these thresholds. Note that neither the big spurious hotspot on chromosome 8, or any of the other spurious hotspots we see in Figure 1(a), were picked up by the quantile-based thresholds.

Figure 3 connects hotspot size to quantile-based threshold. This hotspot size significance profile depicts a sliding window of hotspot size thresholds ranging from $n = 1, \ldots, N$, where $N = 560$ corresponds to the hotspot size threshold derived from the N-method. For each genomic location, the hotspot size (left axis) is significant for the LOD threshold (right axis). For example, the chromosome 5 hotspot was significant up to size 49, meaning that more than 1 trait mapped to the hotspot locus with LOD higher than 7.07, more than 2 traits mapped to the hotspot locus with LOD higher than 6.46, and so on up to hotspot size 49 where more than 49 traits mapped to the hotspot locus with LOD higher than 4.93. The hotspot on chromosome 7 was significant up to size 499, and the hotspot on chromosome 15 (higher peak) was significant for hotspot sizes 2 to 129 and 132 to 143.

The NL-method only detected the real hotspots on chromosomes 5, 7 and 15, whereas the N-method did not detect any hotspots and the Q-method detected 6 spurious hotspots, in addition to the real hotspots. The sliding window of quantile-based thresholds detected the small hotspot composed of traits with high LOD scores on chromosome 5 as well the big hotspots on chromosomes 7 and 15. Equally important, the NL-method dismissed spurious hotspots, such aschromosome 8, composed of numerous traits with LOD scores smaller than 5.57.

Figure 4 shows the results for the Q- and N-methods for simulated example 2 using $\alpha = 0.05$. Panel (a) shows the hotspot architecture. The blue and red lines show the N- and Q-method's thresholds, 19 and 8, respectively. In this example, both the N- and the Q-methods were able to correctly pick up the hotspots on chromosomes 5, 7 and 15.

Comparison of panels (a) on Figures 1 and 4 shows that the spurious hotspots tend to be much smaller when the traits are uncorrelated (compare chromosome 8 on both plots) leading to much smaller N-method thresholds (compare the blue lines). The Q-method thresholds, on the other hand, are quite close. This is expected since the Q-method threshold depends on the number of significant QTLs (we observed 3,162 significant linkages in simulated example 1, against 3,586 significant linkages in example 2) and not on the correlation among the traits.

Supplementary Figure S7 displays the hotspot size significance profile for simulated example 2. The NL-method also detected the hotspots on chromosomes 5, 7 and 15.

Simulation study

In this simulation study we assess and compare the error rates of the Q-, N- and NL-methods under three different levels of correlation among the traits. In order to determine whether the methods are capable of controlling the GWER at the target levels,

we conduct separate simulation experiments as follows:

1. We generate a "null genetical genomics data set" from a backcross composed of (i) 6,000 traits, none of which is affected by a QTL, but that are nevertheless affected by a common latent variable in order to generate a correlation structure among the traits; and (ii) genotype data on 2960 equally spaced markers across 16 chromosomes of length 400cM (185 markers per chromosome). Any detected QTL hotspot is spurious, arising from correlation among the traits.

2. We perform QTL mapping analysis, and 1.5-LOD support interval processing, of the 6,000 traits. For each one of the the following single trait QTL mapping permutation thresholds (that control GWER at the $\alpha = 0.01, 0.02, \ldots, 0.10$ levels, respectively):

   (a) We compute the observed QTL matrix and generate the Q-method hotspot size threshold based on 1,000 permutations of the observed QTL matrix. We record whether or not we see at least one spurious hotspot of size greater than the Q-method threshold anywhere in the genome.

   (b) For each genomic location we count the number of traits above the single trait LOD threshold. We compute the N-method hotspot size threshold based on 1,000 permutations of the null data set. We record whether at least one spurious hotspot of size greater than the N-method threshold anywhere in the genome.

   (c) We compute the NL-method LOD thresholds for spurious hotspots size thresholds ranging from 1 to the N-method threshold. For each NL-method LOD threshold, $\lambda_{n,\alpha}$ where $n = 1, \ldots, N$, we count, at each genomic location, how many traits mapped to that genomic location with a LOD greater than $\lambda_{n,\alpha}$,

and record whether there is at least one spurious hotspot of size greater than n anywhere in the genome.

3. We repeated the first two steps 1,000 times. For each one of the three methods, the proportion of times we recorded spurious hotspots, out of the 1,000 simulations, gives us an estimate of the empirical GWER associated with the method.

QTL analysis was performed as described above. Figure 5 shows the simulation results for null data sets generated using latent variable effects of 0.0, 0.25 and 1.0. The Q- and N-methods, with observed GWER (red), and target error rate (black), have two α levels, α_1 for QTL mapping, and α_2 for the tail area of the hotspot size permutation null distribution; panels display the results when $\alpha_1 = \alpha_2 = 0.01, 0.02, \ldots, 0.10$. The NL-method has a single α level; the red curves are the observed GWERs for spurious hotspot sizes $n = 1, \ldots, N$, where N represents the N-method's permutation threshold.

Panels (a-c) on Figure 5 show that for uncorrelated traits the Q- and N-methods were conservative, below target levels, whereas the NL-method shows error rates about the right target levels for most of hotspot sizes. Panels (d) and (g) show that error rates for the Q-method are higher than target levels when the traits are correlated, and increase with correlation strength among the phenotypes. These results are expected since the Q-method thresholds depend on the number of QTLs detected in the un-permuted data and tend to increase with the number of phenotypes. Because we generated the same number of phenotypes on the three simulation studies, the Q-method's thresholds were similar. Therefore, the number and the size of the spurious QTLs tend to be proportional to the correlation strength of the phenotypes. The N- and NL-methods on the other hand, are designed to cope with the correlation structure among the phenotypes and show error rates close to the target levels as shown in panels (e), (f), (h) and (i).

Yeast data set example

In this section we illustrate and compare the Q-, N- and NL-methods using data generated from a cross between two parent strains of yeast: a laboratory strain, and a wild isolate from a California vineyard (Brem and Kruglyak, 2005). The data consists of expression measurements on 5,740 transcripts measured on 112 segregant strains, with dense genotype data on 2,956 markers. Processing of the expression measurements raw data was done as described in Brem and Kruglyak (2005), with an additional step of converting the processed measurements to normal quantiles by the transformation $\Phi^{-1}[(r_i - 0.5)/112]$ where Φ is the standard normal cumulative density function, and the r_i are the ranks. We performed QTL analysis using Haley-Knott regression (Haley and Knott, 1992) with the R/qtl software (Broman et al. 2003). We adopted Haldane's map function, genotype error rate of 0.0001, and set the maximum distance between positions at which genotype probabilities were calculated to 2cM.

Hotspot analysis of the yeast data, based on the N-method (Figure 6a), detected significant eQTL hotpots on chromosomes 2 (second peak), 3, 12 (first peak), 14 and 15 (first peak), at a GWER of 5% according to null distribution of hotspot sizes shown in Figure 6b. The blue line represents the N-method's significance threshold of N=96. The maximum hotspot size on chromosome 8 was 95 and almost reached significance. Nonetheless, Figure 6a also shows suggestive (although substantially smaller) peaks on chromosomes 1, 4, 5, 7, 9, 12 (second peak), 13, 15 (second peak) and 16 that did not reach significance according to the N-method's significance threshold.

The red line on (Figure 6a) represents the Q-method's significance threshold of 28, derived from the null distribution of hotspot sizes shown in Figure 6c. The Q-method detected significant hotspots on chromosomes 2 (both peaks), 3, 4, 5 (both peaks), 7, 8, 12 (both peaks), 13, 14, and 15 (both peaks).

Figure 7 shows the hotspot significance profile for the NL-method. The major hotspots on chromosomes 2, 3, 12 (first peak), 14 and 15 (first peak) were significant across all thresholds tested up, and the hotspot on chromosome 8 was significant up to size 93. Furthermore, the NL-method showed that the small hotspots detected by the Q-method on chromosomes 5, 12 (second peak), 13 and 15 (second peak) might indeed be real. Nonetheless, the small hotspots on chromosomes 4 and 7, detected by the Q-method, are less interesting than the small hotspot on chromosome 1, that was actually missed by the Q-method.

Discussion

A common feature in genetical genomics studies of expression traits is the presence of eQTL hotspots where a single polymorphism leads to widespread downstream changes in the expression of distant genes. These genomic loci associated with many distant genes are biologically interesting since they may harbor important regulators. Statistical procedures aiming to assess the significance of such hotspots are of key importance.

Breitling et al. (2008) were the first to propose a permutation test (the N-method) for eQTL hotspots that accounts for the correlation structure among phenotypes due to the effect of confounders. However, the authors restricted their attention to the single trait empirical threshold only, and may have overlooked interesting hotspots composed of moderate to small numbers of traits with strong LOD scores.

In this paper, we adopt the Breitling et al. permutation scheme and propose a method to determine a range of quantile-based permutation thresholds (the NL-method) that allows us to assess the significance of hotspots based on the number and on the linkage strength of the traits composing those hotspots. For a fixed error rate α, our approach investigates the significance of a hotspot using a range of N distinct mapping thresholds,

where N is the smallest hotspot size that is significant by the N-method. For each $n = 1, \ldots, N$ we determine the LOD threshold that controls the genome wide error rate of detecting at least one spurious hotspot of size n or higher somewhere in the genome, at an error rate less or equal than α.

Our simulated examples and simulation studies show that Q-method performs well when the traits are uncorrelated, but detects spurious hotspots at high rates when the traits are correlated. This result is not surprising since the Q-method implicitly assumes that the traits are uncorrelated. Molecular traits such as mRNA expression levels, metabolite concentrations and protein levels are often highly correlated and the Q-method is not adequate in these situations. On the other hand, our simulations suggest that the N- and NL-methods perform adequately for correlated or uncorrelated traits, showing genome wide error rates close the target levels.

The advantage of the NL-method over the N-method is that it can assess the significance of hotspots with any type of LOD score distribution. For instance: i) a hotspot composed of many traits with moderate LOD scores will be found with thresholds close to the single trait threshold; ii) a hotspot consisting of a few traits with strong LOD scores will be detected with thresholds close to the conservative threshold; iii) a large hotspot with a range of moderate to large LOD scores will be significant at all thresholds in our sliding scale. The ability to assess the significance of these different types of hotspots can lead to important additional biological findings that might be overlooked by previous approaches, while still avoiding the detection of spurious hotspots. In the analysis of the yeast data, the hotspots on chromosomes 5, 8, 12 (second peak), 13 and 15 (second peak) have a LOD distribution of type ii. The hotspots on chromosomes 2, 3, 12 (first peak), 14 and 15 (first peak) have LOD distributions of type iii. No hotspot with LOD distribution of type i is present in the yeast data set. Note that hotspots composed of moderate to

small number of traits with moderate LOD scores will be missed by all thresholds in our sliding scale, and will be discarded as non-significant by our analysis. Application of the N-method detected only the 5 big hotspots on chromosomes 2, 3, 12, 14 and 15. Additionally, the simulated example 1 shows an example were the NL-method was able to pick up the 3 simulated hotspots missed by the N-method.

The NL-method is in a certain sense analogous to the approach proposed by Chen and Storey (2006). In the same way that Chen and Storey relax the single trait mapping threshold by controlling the probability that a trait falsely maps to k or more genomic locations, we relax the conservative threshold by controlling the probability that n or more traits falsely map to a common genomic location.

Even though the sliding window of thresholds delivered by the NL-method is more informative than the single hotspot size threshold of the N-method, these approaches have the same computational complexity. They use exactly the same permutations but summarize the results differently. Both methods are computationally intensive: reliable results require 1,000 or more permutations, and for each permuted data set we perform mapping analysis of several thousand traits. Thus, in general, parallel computation on a cluster are required. In order to reduce the computational burden, we adopted Haley-Knott regression and mapped traits with common missing phenotype data patterns as blocks. An R package called qtlhot is being submitted to CRAN.

The approach in this paper relied on single-QTL mapping methods. To examine whether an apparent hotspot could be an artifact, such as a ghost QTL (Haley and Knott 1992), we used multiple QTL methods (Manichaukal et al. 2008) for some smaller hotspots (data not shown). Most traits from these hotspots continued to map to the same location detected by single trait analysis when we allowed for other possible QTL on the same or other chromosomes. It would be possible to extend our quantile-based permutation

approach to multiple QTL mapping (Jansen 1993, Jansen and Stam 1994, Manichaukal et al. 2008) by considering the LOD profile for each QTL adjusted for all other QTLs (e.g. using the addqtl or the multiple QTL mapping functions in R/qtl, Broman and Sen 2009, Arends et al. 2010). However, this would require considerably more computation and is left for future research.

The analysis of data sets containing groups of repetitive traits (that is, distinct traits representing slightly different measurements of a same "baseline" phenotypic trait) must be conducted with care. Repetitive traits are artifacts of the experimental design rather than indications of underlying biological processes. For instance, traits derived from oligos of same gene are often highly correlated simply because they arise from the same gene, and might be picked up as a hotspot. Thus, repetitive traits can introduce artefactual hotspots that are indistinguishable statistically from biologically-driven hotspots, unless this is addressed by attention to the design. Other examples of repetitive traits include: (i) protein traits where one protein can exist in many variants due to post-translational modifications and the abundance of each variant is measured and used as a separate trait; and (ii) classical phenotypic traits such as flowering in Arabidopsis, where a major QTL has been investigated in a number of independent studies, under different environmental conditions, leading to a group of repetitive traits strongly mapping to the same QTL (see supplement for Fu et al. 2009). If repetitive traits are known ahead of time, they should be removed or otherwise accounted for in the analysis. For example, Fu et al (2009) proposed organizing repetitive classical traits into disjunct phenotypic groups based on trait annotations and performed hotspot analysis on the average trait per category.

Fu et al. (2009) point out that large eQTL hotspots may or may not persist when examining proteomic (pQTL), metabolic (mQTL) and phenotypic (phQTL) gene mapping. Now that we can infer smaller hotspots composed by any of these QTL types, it may be

possible to find more connections. A small hotspot could in fact be quite important to reveal genetic effects on whole-body phenotypes.

Acknowledgments

This work was supported by CNPq Brazil (ECN); NCI ICBP grant U54-CA149237 and NIH grant R01MH090948 (ECN); NIDDK grants DK66369, DK58037 and DK06639 (ADA, MPK, AB, BSY, ECN); NIGMS grants PA02110 and GM069430-01A2 (BSY); NBIC (Netherlands Bioinformatics Centre), Distinguished Scientist Traveling Stipend (RCJ); and by NIH grants R01GM074244 (KWB). We would like to thank Rachel Brem for sharing the yeast data set, and Bill Taylor from the Center for High Throughput Computing of UW-Madison for his assistance with cluster computation.

References

1. Arends D., P. Prins, R. C. Jansen, K. W. Broman, 2010 R/qtl: high throughput multiple QTL mapping Bioinformatics **26**: 2990-2992.

2. Breitling R., Y. Li, B. M. Tesson, J. Fu, C. Wu, et al., 2008 Genetical genomics: spotlight on QTL hotspots. PLoS Genetics **4:** e1000232.

3. Brem R. B., G. Yvert, R. Clinton, L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

4. Brem R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS **102:** 1572-1577.

5. Broman K. W., W. Wu, S. Sen, and G. A. Churchill GA, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889-890.

6. Broman K. W., and S. Sen, 2009 *A Guide to QTL Mapping with R/qtl.* Springer, New York.

7. Chen L., and J. D. Storey, 2006 Relaxed significance criteria for linkage analysis. Genetics **173:** 2371-2381.

8. Churchill G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963-971.

9. Churchill G. A., and R. W. Doerge, 2008 Naive application of permutation testing leads to inflated type I error rates. Genetics **178:** 609-610.

10. Darvasi A., 2003 Gene expression meets genetics. Nature **422:** 269-270.

11. Dupuis, J., and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics **151:** 373386.

12. Fu J., J. J. Keurentjes, H. Bouwmeester, T. America, F. W. Verstappen, et al., 2009 System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics **41:** 166-167.

13. Good P., 1994 *Permutation tests: a practical guide to resampling for testing hypothesis.* Springer-Verlag, New York.

14. Haley C., and S. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity *69:* 315-324.

15. Jansen R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205-211.

16. Jansen R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447-1455.

17. Lander E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185199.

18. Lehmann E. C., 1986 *Testing statistical hypothesis*, Ed.2. John Wiley and Sons, New York.

19. Manichaikul A., J. Dupuis, S. Sen, and K. W. Broman, 2006 Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics **174:** 481-489.

20. Manichaikul A., J. Y. Moon, S. Sen, B. S. Yandell, and K. W. Broman, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics **181:** 1077-1086.

21. Perez-Enciso M., 2004 In silico study of transcriptome genetic variation in outbred populations. Genetics **166:** 547-554.

22. Schadt E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, et al., 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

23. West M. A. L., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore, R. W. Doerge, D. A. St. Clair 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics **175:** 1441-1450.

24. Wu C., D. L. Delano, N. Mitro, S. V. Su, J. Janes, et al. 2008 Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genetics **4:** e1000070.

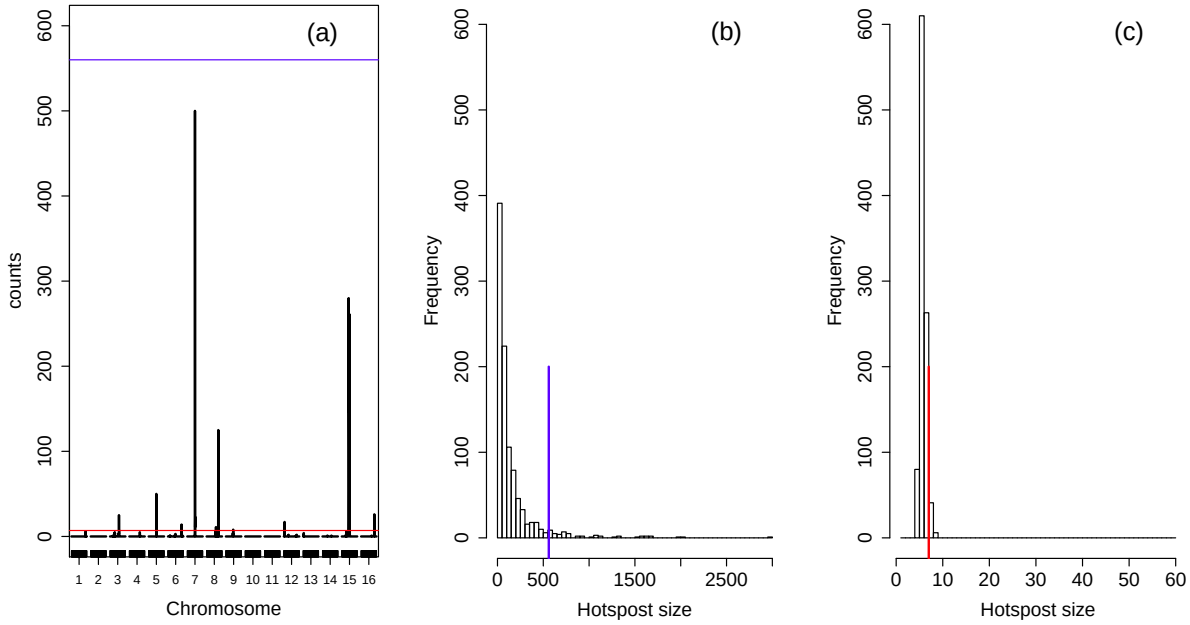

Figure 1: N- and Q-method analyzes for simulated example 1. Panel (a) depicts the inferred hotspot architecture using a single trait permutation threshold of 3.65 corresponding to a GWER of 5% of falsely detecting at least one QTL somewhere in the genome. The blue line at count 560 corresponds to the hotspot size expected by chance at a GWER of 5% according to the N-method permutation test. The red line at count 7 corresponds to the Q-method's 5% significance threshold. The hotspots on chromosomes 5, 7, 8 and 15 have sizes 50, 500, 125 and 280, respectively. Panel (b) shows the N-methods permutation null distribution of the maximum genomewide hotspot size. The blue line corresponds to the hotspot size 560 expected by chance at a GWER of 5%. Panel (c) shows the Q-methods permutation null distribution of the maximum genomewide hotspot size. The red line at 7 shows the 5% threshold. Results based on 1,000 permutations.

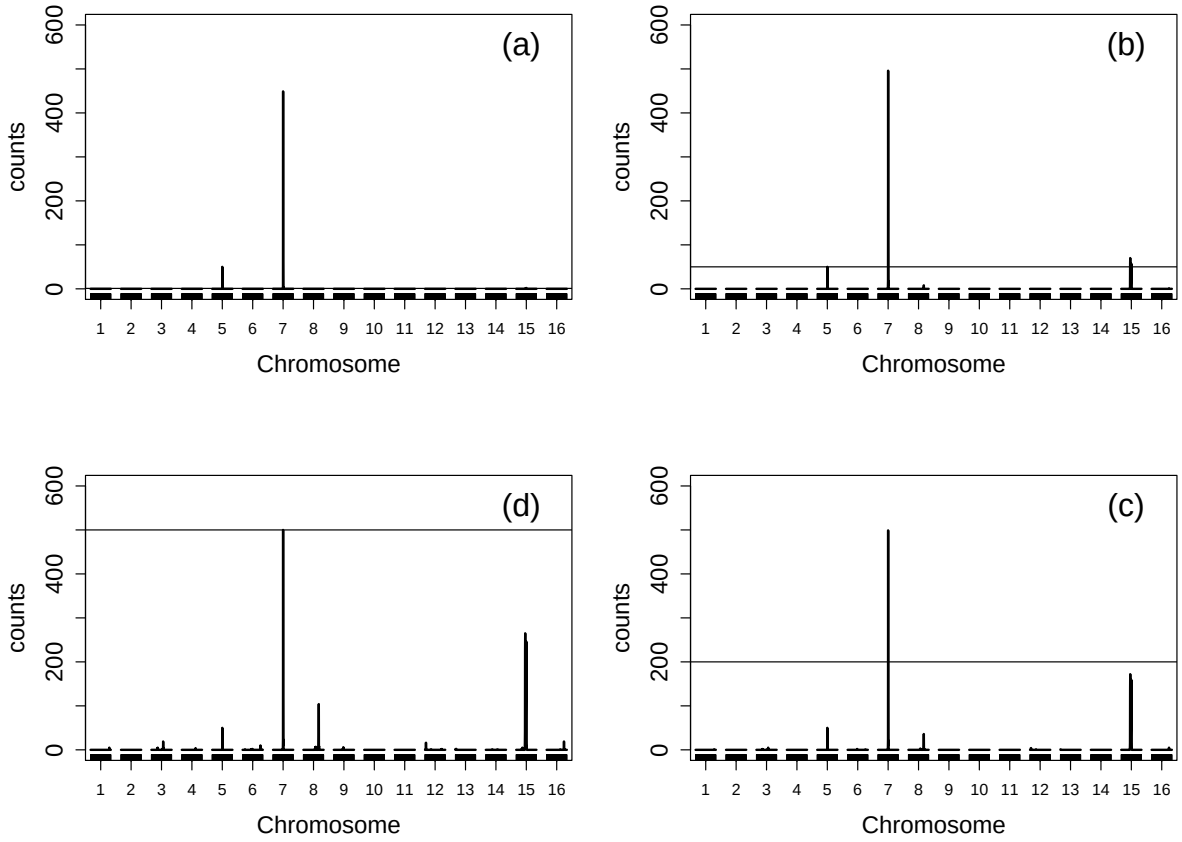

Figure 2: NL-method analysis for simulated example 1. Panels (a)-(d) present the hotspot architecture inferred using different quantile-based permutation thresholds, i.e., for each genomic location it shows the number of traits that mapped there with a LOD threshold higher than the quantile-based permutation threshold. Panel (a) presents the hotspot architecture inferred using a permutation LOD threshold of 7.07 corresponding to the LOD threshold that controls the probability of falsely detecting at least a single linkage for any of the traits somewhere in the genome under the null hypothesis that none of the traits have a QTL anywhere in the genome, at an error rate of 5%. Panels (b), (c) and (d) present the hotspot architectures computed using QTL mapping LOD thresholds of 4.93, 4.21 and 3.72 that aim to control GWER at a 5% error rate for spurious eQTL hotspots of size 50, 200 and 500, respectively.

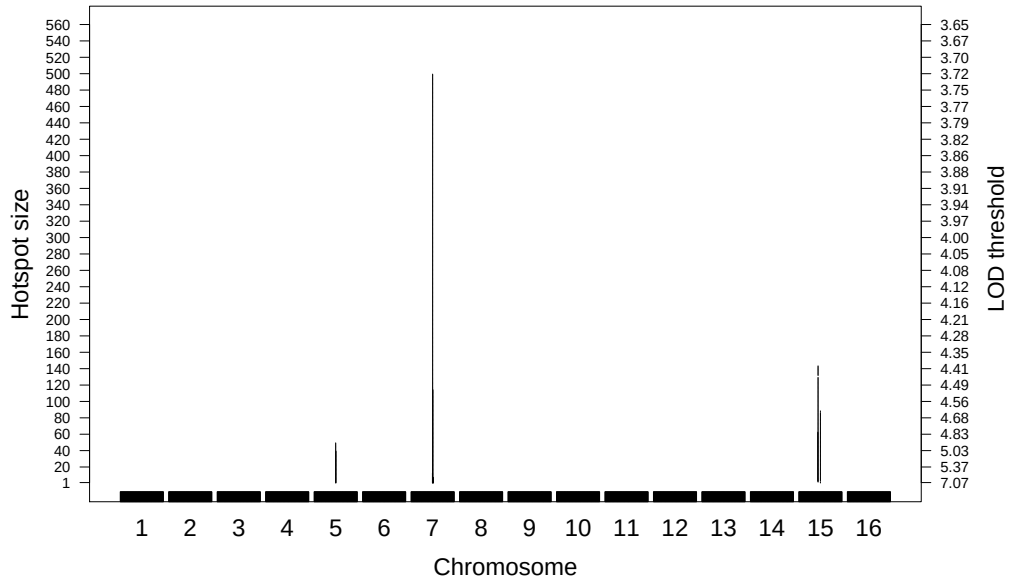

Figure 3: Hotspot size significance profile derived with the NL-method for simulated example 1. For each genomic location (i.e., x-axis position) this figure shows the hotspot sizes at which the hotspot was significant, that is, at which the hotspot locus had more traits mapping to it with a LOD score higher than the threshold on the right, than expected by chance. The scale in the left shows the range of spurious hotspot sizes investigated by our approach. The scale in the right shows the respective LOD thresholds associated with the spurious hotspot sizes in the left. The range is from 7.07, the conservative empirical LOD threshold associated with a spurious "hotspot of size 1", to 3.65, the single trait empirical threshold, associated with a spurious hotspot of size 560. All permutation thresholds were computed targeting GWER ≤ 0.05, for $n = 1, \ldots, 560$.

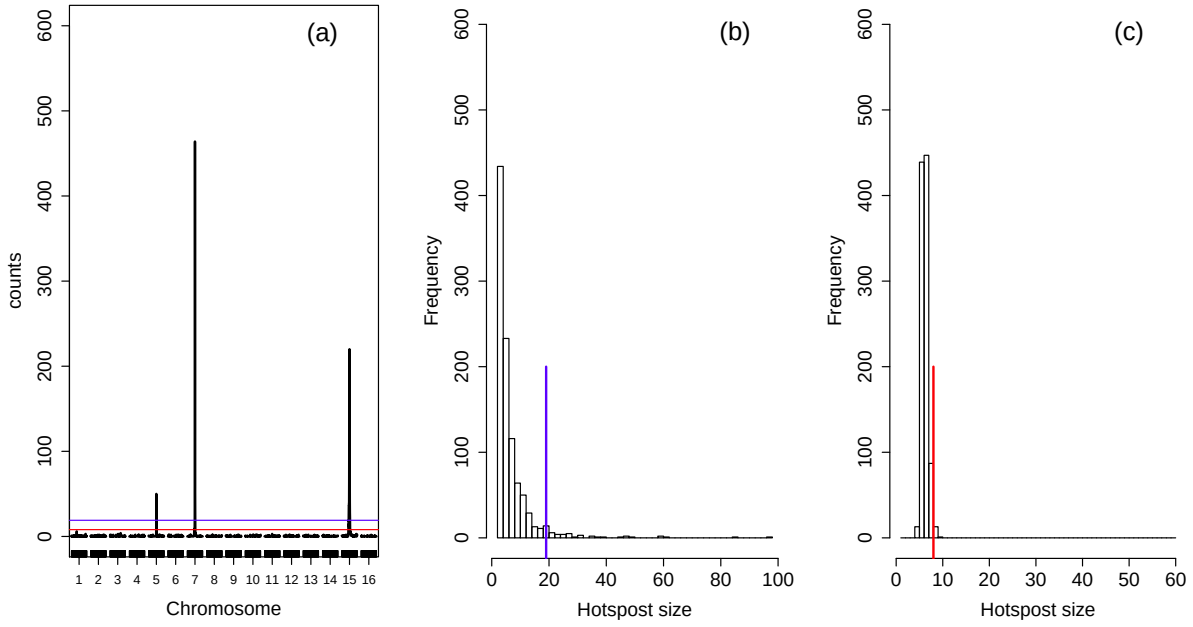

Figure 4: N- and Q-method analyzes for simulated example 2. Panel (a) depicts the inferred hotspot architecture using a single trait permutation threshold of 3.65 corresponding to a GWER of 5% of falsely detecting at least one QTL somewhere in the genome. The blue line at count 19 corresponds to the hotspot size expected by chance at a GWER of 5% according to the N-method permutation test. The red line at count 8 corresponds to the Q-method's 5% significance threshold. The hotspots on chromosomes 5, 7 and 15 have size 50, 464 and 220, respectively. Panel (b) shows the N-method's permutation null distribution of the maximum genomewide hotspot size. The blue line at 19 corresponds to the hotspot size expected by chance at a GWER of 5%. Panel (c) shows the Q-methods permutation null distribution of the maximum genomewide hotspot size. The red line at 8 shows the 5% threshold. Results based on 1,000 permutations.

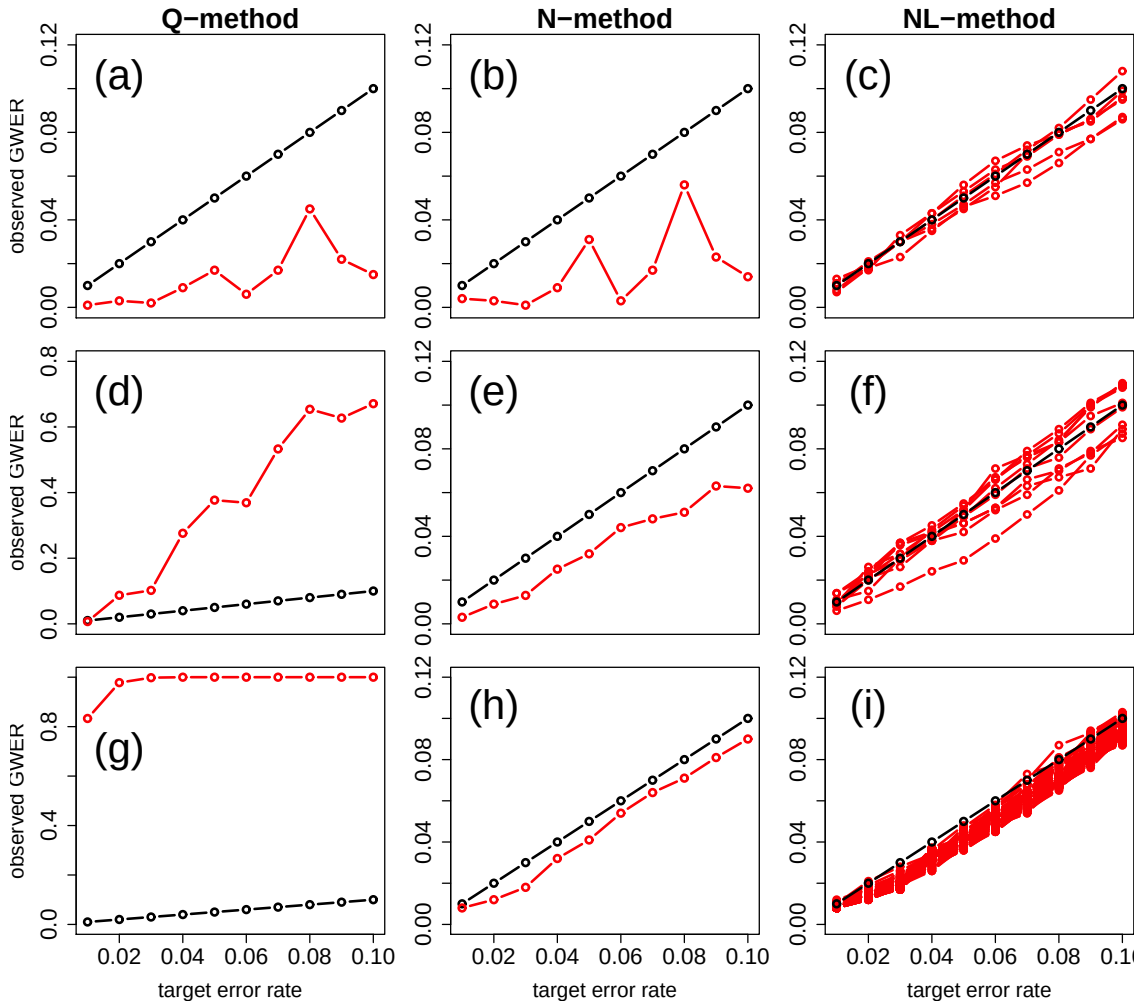

Figure 5: Observed GWER for the Q-, N- and NL-methods under varying strengths of phenotype correlation. Black lines show the targeted error rates. Red curves show the observed GWER. Panels (a), (b) and (c) show the results for uncorrelated phenotypes. Panels (d), (e) and (f) show the results for weakly correlated phenotypes generated using latent variable effect equal to 0.25. Panels (g), (h) and (i) show the simulation results for highly correlated phenotypes generated using latent effect set to 1. The left, middle and right panel columns show the results for the Q-, N- and NL-methods, respectively. Note the different y-axis scales for the Q-method panels. The red curves on the NL-method panels show the observed GWER for hotspot sizes ranging from 1 to N, where N is the median N-method threshold for $\alpha = 0.10$.

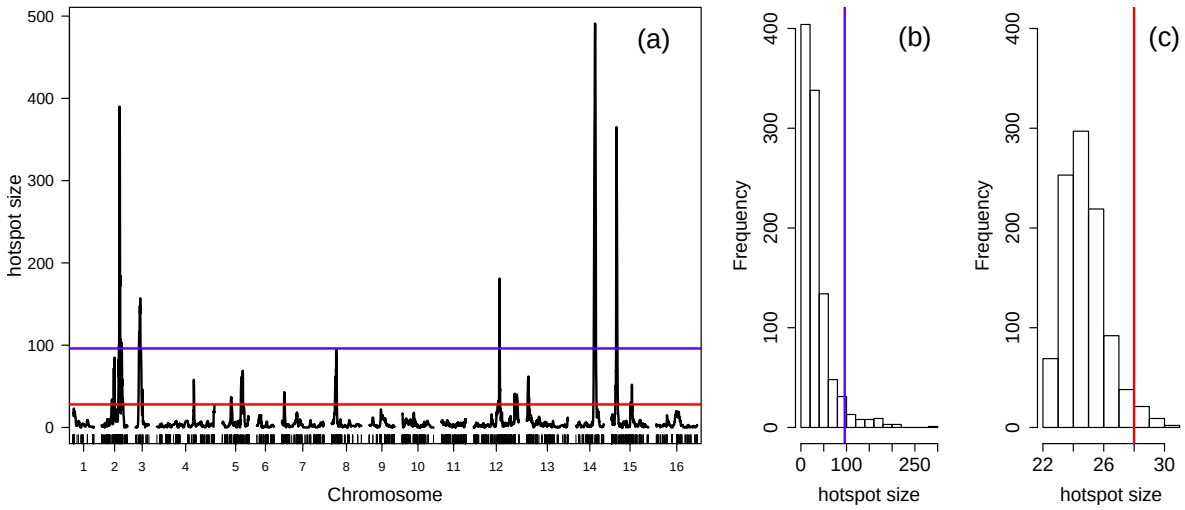

Figure 6: N- and Q-method analyzes for the yeast data. Panel (a) depicts the inferred hotspot architecture using a single trait permutation threshold of 3.44 corresponding to a GWER of 5% of falsely detecting at least one QTL somewhere in the genome. The blue and red lines at counts 96 and 28 correspond to the hotspot size expected by chance at a GWER of 5% according to the N- and the Q-method permutation tests, respectively. Panels (b) and (c) show, respectively, the permutation null distributions of the maximum genomewide hotspot size based on 1000 permutations. The blue and red lines at 96 and 28 correspond, respectively, to the hotspot size expected by chance at a GWER of 5% for the N- and Q-methods.

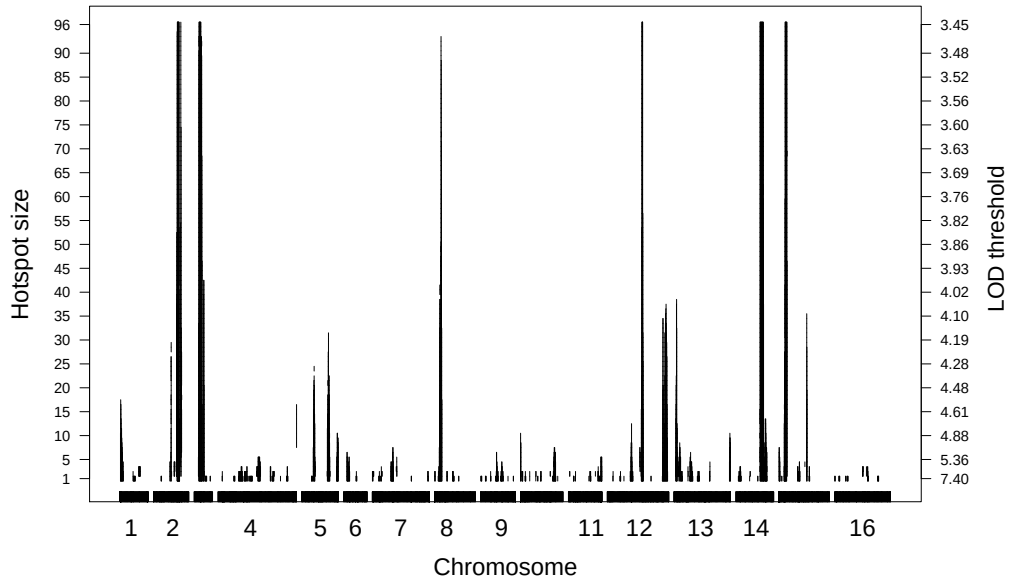

Figure 7: Hotspot size significance profile derived with the NL-method. The range is from 7.40, the conservative empirical LOD threshold associated with a spurious "hotspot of size 1", to 3.45, the single trait empirical threshold, associated with a spurious hotspot of size 96. All permutation thresholds were computed targeting GWER $\leq$ 0.05, for $n = 1, \ldots, 96$.