Multiple Interval Mapping for Gene Expression QTL Analysis

Wei Zou* and Zhao-Bang Zeng*,†

*Bioinformatics Research Center & Department of Statistics, †Department of Genetics, North Carolina State University, Raleigh, NC 27695,

June 4, 2007

Running head: MIM for eQTL analysis

Corresponding address: Zhao-Bang Zeng
Bioinformatics Research Center
Department of Statistics
North Carolina State University
Raleigh, NC 27695-7566, U.S.A.

Telephone number: 919-515-1942
Fax number: 919-515-7315
E-mail: zeng@stat.ncsu.edu.

Key words: eQTL, MIM, FDR

ABSTRACT

To find the correlations between genome-wide gene expression variations and sequence polymorphisms in inbred populations, we developed a statistical method to claim expression quantitative trait loci (eQTL) from the genome. It is based upon multiple interval mapping (MIM), a model selection procedure, and uses false discovery rate (FDR) to measure the statistical significance of the large number of eQTL. We organized the computational procedures in an R package (`http://statgen.ncsu.edu/~wzou/MIM.eQTL.html`), which can estimate the FDR for positive findings from analogous model selection procedures. We compared our method with a similar procedure recently proposed by Storey *et. al.* and concluded that our method was more powerful: applied onto the same yeast data, our method claimed eQTL for far more expression traits with a lower estimated FDR. After examining the two procedures thoroughly, we identified two features in the statistical analysis that resulted in these differences. A simulation study confirmed the remarkable influence from these features on the power to detect eQTL in the yeast data. We also applied bioinformatics analysis on the 5182 eQTL we declared from the yeast data. The mapping results are available at `http://statgen.ncsu.edu/eQTLViewer/` through our eQTL Viewer. The eQTL distribution patterns seem compatible with the general rules of transcriptional regulations and cellular networks.

Introduction

Transcriptional control is one of the most important steps for an organism to express the genetic information stored in its sequence as well as to respond to environmental changes (Ihmels *et al.*, 2004). Recent advance of genomic technology has made it possible to quantify transcript abundance systematically, as well as to genotype genetic markers covering the whole genome in a segregating population. Provided with these tools, expression quantitative trait locus (eQTL) analysis has been applied to study inheritance of thousands of similar traits in the hope to find general rules of genetic control of transcriptional regulation (Brem *et al.*, 2002; Schadt *et al.*, 2003; Bystrykh *et al.*, 2005; Chesler *et al.*, 2005).

To find the correlation between expression profiles and genotypic variation, an obvious way is to compare trait means among groups with different allelic types in a mapping population at each genetic locus (Sax, 1923). This approach was implemented as Wilcoxon-Mann-Whitney Test (Brem *et al.*, 2002; Yvert *et al.*, 2003; Bing and Hoeschele, 2005) and applied to the yeast data published by Brem *et al.* (2002); Yvert *et al.* (2003). However, the analysis of the genetic architecture of these expression traits revealed that the majority of heritable traits are controlled by multiple genetic factors (Brem and Kruglyak, 2005). Single marker analysis has some difficulty in finding multiple QTL for a quantitative trait: genotypic correlation will 'inflate QTL effects and complicate the interpretation of QTL mapping results' (Valdar *et al.*, 2006).

Another challenge in eQTL analysis is the multiple testing issue due to multiple tests across the genome in search of QTL for one trait and these many tests for each of thousands of expression traits included in a study. To address this problem, Storey *et al.* (2005) presented a sequential search algorithm which uses false discovery rate (FDR) to assess the joint significance of multiple QTL for each expression trait. He declared 170 two QTL models from the yeast data (Brem and Kruglyak, 2005) while controlling FDR at 10%.

In this paper, we develop a procedure for eQTL mapping that combines features of multiple interval mapping (MIM) (Kao *et al.*, 1999) and FDR computation from Storey *et al.* (2005). We demonstrate the power of this procedure using the same yeast data (Brem and Kruglyak, 2005) for comparison. The aim of MIM is to fit multiple main and epistatic genetic effects in a statistical model to integrate the search for QTL with the inference of genetic architecture of quantitative traits (Zeng *et al.*, 1999). This approach is naturally more powerful than the single marker analysis. Combining the FDR estimation mechanism from Storey *et al.* (2005) with

MIM, we are able to declare 2 or more QTL on a trait for 1242 traits while keeping the overall FDR at 8%. We study the differences of these two search procedures and identify the features in the sequential genome scans that would increase the statistical power of QTL identification. Bioinformatics analysis of these declared QTL suggests that the significant correlation patterns that we have declared are compatible with the general rules of transcriptional regulation and cellular networks.

The Yeast Mapping Data

The data (Brem and Kruglyak, 2005) contain 112 haploid segregants obtained by crossing a standard laboratory (BY) and a wild (RM) strain of budding yeast (Brem *et al.*, 2002; Yvert *et al.*, 2003; Brem and Kruglyak, 2005). 2956 SNPs markers were genotyped. The average length of the 16 chromosomes is about 405 cM as estimated by Mapmaker/QTL (Whitehead Inst., Cambridge, MA, USA). There are 6195 unique expression traits in the data. This number differs from the originally reported one (Brem and Kruglyak, 2005) since we found 21 pairs of gene names were aliases of each other. For each pair of these probes, we averaged their hybridization signals to form a single expression profile. We used log ratios of signal intensities versus a common reference as trait values. They were normalized according to the two step procedure by Wolfinger *et al.* (2001) and standardized before QTL analysis.

Multiple Interval Mapping

We used MIM to fit a multiple QTL model for each expression trait in a sequential way. 'Sequential' means that multiple QTL for a trait were added into the model in a series of genome scans sequentially. Genome scans were performed by testing the existence of a QTL on and between markers across the genome using the likelihood ratio test (LRT) (Lander and Botstein, 1989) as implemented in QTL Cartographer (Basten *et al.*, 2002).

In the initial cycle, we performed a genome scan for each trait to identify the most likely position of the first QTL. We only kept the maximal LRT statistic across the genome for a trait. If this maximum was above a threshold, the genomic position of the peak was retained as the first QTL of the trait. We will discuss the way to determine thresholds later. Then, in the second cycle, we restricted our genome scans for those traits with a QTL found in the first cycle (3367 in total, see Table 1). The genome scans for the second QTL were conditional upon the first QTL, i.e., the model contained the effect of the first QTL as a fixed effect while searching for

the second QTL to maximize the likelihood. In this sense, the genome scans in the search for the first QTL were unconditional, and in all the rest cycles, the LRTs were conditional. We also only considered the genome-wide maximal LRT statistic for each trait in the second scans and compared them with the threshold. The sequential genome scans continued until no traits were found with the maximal LRT statistics above the threshold.

We only considered QTL main effects in the above sequential genome scans. As the next step, we tried to detect significant pairwise interaction effects between multiple QTL for each trait. Then we applied a backward elimination process to insure all genetic effects were statistically significant based on the same threshold.

We obtained the threshold by using a permutation test (Churchill and Doerge, 1994). We first randomly shuffled the pairing between gene expression phenotypes of yeast samples and their genotypes 20 times. Thus, we had 20 permutation samples for each expression trait. We performed a unconditional genome scan in each permuted sample and collected the maximal LRT statistic for each trait. These statistics from all traits were pooled to form the null distribution for the initial cycle of genome scans, assuming a common null distribution for these traits (Storey *et al.*, 2005). The 90% quantile of this distribution was used as the threshold in all the steps above.

Controlling type I error at 10% for a genome scan is quite liberal. We used this threshold to favor possible interaction effects with minor main effects (Marchini *et al.*, 2005). For the second and later cycles of genome scans, we could get the corresponding null distributions conditional upon previous QTL using CET method (Doerge and Churchill, 1996). But instead, we used the same 90% quantile of the unconditional null distribution because of the computational burden for generating those null statistics. According to the CET method, markers around previously declared QTL are excluded in the conditional genome scans. Such a reduction in the number of LRTs would lower the maximal statistic under the null hypothesis as well as the threshold for the conditional LRTs.

We ran regression analysis between trait values and marker genotypes on permuted data (5 permutations for each of the 6195 traits) to validate our reasoning. The F test statistics associated with markers are asymptotically equivalent to LRT statistics. Given the dense marker set, the marker based analysis can approximate the interval mapping procedures well, but is much faster, thus makes CET feasible. The null distribution of the maximal F statistics from unconditional genome scans has mean 11.16 and standard deviation 2.86. We used the CET procedure (Doerge and Churchill, 1996) with minor modification to obtain the null F statistic

distribution conditional upon the first 'QTL', i.e., the markers with the maximal F statistics in the original data. We permuted data within each genotype group of the first QTL before performing the conditional scans. In the scans, we excluded the two 50 cM flanking regions surrounding the first QTL. The distribution looks very similar to the unconditional one (Figure 1), with mean 11.14 and standard deviation 2.86. Thus, compared with the unconditional null distribution, the conditional one is skewed a little bit towards small values of test statistics. Using the unconditional threshold for conditional LRTs will lead to very slightly more conservative tests.

After the above model selection process, we ran a conditional genome scan for each QTL main effect again to obtain a LRT statistic profile surrounding each QTL. The LRTs in this scan were conditional upon all the other significant genetic effects in the model. For those QTL with LRT peak above 15.4 (95% quantile of the null distribution), a 1.5 LOD support interval around a peak was declared as the QTL region. Thus, the genome-wide type I error rate is kept at 5% for each QTL.

We declared 5182 QTL for 3367 traits finally. Table 1 shows the detailed process of the sequential search. Very few interaction cases were retained: 49 interacting QTL pairs for 38 expression traits. The overall FDR was about 0.08 (see the next section for details). The average 1.5 LOD support interval of QTL is about 50kb in the physical map, containing 27.7 ORFs in average. We visualized the mapping results as a myriad of relationships between the expression trait genes and the candidate genes in the QTL regions through our eQTL Viewer at `http://statgen.ncsu.edu/eQTLViewer/` (Zou *et al.*, 2007).

False Discovery Rate (FDR) in Sequential Genome Scans

Empirical thresholds discussed above are supposed to control genome-wide type I error for a QTL in each genome scan. To find the overall errors for all these QTL, we followed the FDR definition from Storey *et al.* (2005):

$$
\begin{aligned}
FDR &= \frac{1}{N}\sum_i q_i && (1)\\
q_i &= 1-\prod_{j=1}^{J_i}\Pr(Q_{ij}=1|Data) && (2)\\
&= 1-\Pr(Q_{i1}=1|Data)\Pr(Q_{i2}=1|\ Q_{i1}=1,Data)\cdots\\
&\quad \Pr(Q_{iJ_i}=1|\ Q_{ij}=1, j=1,\cdots,J_i-1,Data)
\end{aligned}
$$

where q_i is the probability that the multiple QTL model for trait i is false; J_i means trait i has J_i

QTL declared; Q_{ij} is an indicator variable for the j-th QTL of trait i: $Q_{ij} = 1$ denotes the QTL is true, 0 for otherwise. $\Pr(Q_{ij} = 1|Data)$ takes different forms for different j: $\Pr(Q_{i1} = 1|Data)$ is the posterior probability that the first QTL for trait i is true given the data (or given the summary statistics of the data, e.g., LRT statistics); $\Pr(Q_{iJ_i} = 1|Q_{ij} = 1, j = 1, \cdots, J_i - 1, Data)$ is the probability that the J_i-th QTL is true given all the previous QTL are true and the data. Thus, q_i summarizes the errors that might happen in any step of sequential genome scans for trait i. N is the number of traits with at least one QTL declared through MIM. It was fixed when we calculated FDR. In the original work by Storey *et al.* (2005), they actually adjusted N to control FDR at a certain level.

From Equation 1, we can see that the FDR is defined as the average rate of declaring a false genetic model, rather than a false QTL. In sequential genome scans like MIM, hypotheses regarding multiple QTL of an expression trait are logically dependent: when we are searching for the second QTL of a trait, we assume there is the first QTL. If that assumption turns out to be invalid, the whole model should be regarded as false. Such a dependency structure among hypotheses differs apparently from the usual setting of multiple testing where hypotheses are statistically independent, or positive regression dependent on a subset (PRDS) (Benjamini and Yekutieli, 2001), or weakly correlated in detecting differentially expressed genes on microarrays (Storey and Tibshirani, 2003). The definition of FDR in Equation 1 specially fits a model selection procedure like MIM.

We estimated $\Pr(Q_{i1} = 1|Data)$ as a 'local' FDR (Efron *et al.*, 2001; Efron and Tibshirani, 2002; Storey *et al.*, 2005). That is, we assumed those 6195 LRT statistics obtained in the initial cycle of genome scans for all traits (called 'test set' later) were sampled independently from a mixture of 2 distributions. Let f_1 be the probability density of LRT statistics of true QTL, while f_0 be the one for false QTL. Denoting π_0 as the probability that the maximal LRT statistic for a trait was collected from a false QTL, the density for the mixture distribution, f_m, could be expressed as $f_m = \pi_0 f_0 + (1 - \pi_0) f_1$. Since $\pi_0 = 1$ for all the maximal LRT statistics from the permuted data (called 'null set' later), we could estimate f_0 from the null set. The above argument is valid for other statistics (for example, F statistics) from genome scans. To reflect such generality, We will use $f_0(.), f_m(.)$ instead of f_0, f_m later on. Thus, from Efron *et al.* (2001),

$$\Pr(Q_{i1} = 1|Data) = 1 - \pi_0 \frac{f_0(.)}{f_m(.)} \tag{3}$$

Methods to estimate π_0 generally use a tuning parameter λ, below which (for F, χ^2 statistics,etc) or above which (for p values corresponding to these statistics) all statistics in the test set are

assumed to be purely from the null hypothesis, or false QTL here. We set λ as the 25% quartile of the test set and estimated π_0 according to Equation 6.7 from EFRON *et al.* (2001). The 6 estimated π_0 for the 6 cycles of genome scans ranged between 0.33 and 0.53. It is a conservative way of estimating π_0, since there could be true QTL with relative small LRT statistics. We did not use the more powerful smoothing method (STOREY and TIBSHIRANI, 2003) because its estimates for π_0 were quite small (ranged between 0.03 and 0.17) especially for those later cycles of genome scans where the numbers of tests decreased greatly (Table 1). Inflated estimates that we have picked would lead to a deflated estimate of $\Pr(Q_{i1} = 1|Data)$ (Equation 3), and hence conservative estimate of FDR (Equation 2). $f_0(.)$ and $f_m(.)$ could be directly estimated from the null set and the test set respectively, using a general smoothing spline. $f_0(.)/f_m(.)$ could also be estimated from these two sets combined using a non-parametric logistic regression and a natural cubic spline (EFRON *et al.*, 2001; STOREY *et al.*, 2005). We used the later procedure to calculate $\Pr(Q_{ij} = 1|Data)$ in this paper. Thus, by estimating $\pi_0, f_0(.), f_m(.)$ from the LRT statistics, which summarized the correlation patterns in the data, we effectively estimated the probability of $Q_{i1} = 1$ given the data. The actual $\Pr(Q_{i1} = 1|Data)$ for each declared QTL was calculated by plugging in its LRT statistic into the estimated $f_0(.)/f_m(.)$.

We used the same Equation 3 to estimate $\Pr(Q_{i2} = 1|Q_{i1} = 1, Data)$, where $\pi_0, f_0(.), f_m(.)$ were calculated based on the LRT statistics for the second QTL (STOREY *et al.*, 2005). Such a statistic indicated the strongest correlation between one trait and a sequence polymorphism selected from the genome, given the position and the effect of the first QTL for the trait. We collected a new test set from the conditional genome scans for the 3367 traits with one QTL retained in the first cycle (Table 1), and the corresponding null set from the unconditional genome scans for the same 3367 traits using permuted data. $f_0(.)/f_m(.)$ was then estimated as before. Note that we used the unconditional null distribution in place of the conditional one again to reduce the computational burden. Such probabilities for QTL declared in the later cycles were obtained similarly.

We developed an R package called 'MIM.eQTL' to do the computation described in this section. This package can also be used to declare a set of genetic models from an expression QTL study with FDR controlled at a desired level. We will discuss this feature later.

In our sequential genome scans, we used the genome-wide threshold to control the type I error rate of the statistical tests in each genome scan for each trait. We did not directly tackle the issue of multiple testing arisen from multiple genome scans for all the expression traits, but used an estimated FDR to assess the reliability of the detected QTL. Here we provide some

intuitive explanation for why more than 10000 tests (sum of '#Scanned' column in Table 1) with 5% error rate each would end up with FDR at 8%. For example, in the initial cycle of MIM, the up-bound of expected number of false QTL is roughly $6195 \times 5\% = 310$. Such a number could be devastating for certain inference. But with 3354 declared QTL (Table 1), the percentage of false QTL is only about $310/3354 = 9\%$. Similar arguments can be applied to the other cycles of genome scans.

Choice of Sequential Genome Scan Procedures

Storey *et al.* (2005) presented an elegant way to do sequential genome scans while declaring QTL with FDR controlled at a certain level. In this paper, we actually followed his way of defining FDR but used MIM to do sequential genome scans. However, using the same yeast data, Storey *et al.* (2005) claimed only 170 traits with 2 QTL while controlling FDR at 10%. Using MIM, we are able to claim 1242 traits with 2 or more QTL with FDR 8%. Then why is there such a large difference?

It is unlikely due to the differences in pre-processing expression trait data, handling missing markers, testing hypotheses (marker based F tests versus interval mapping and LRTs) or estimating π_0 in Equation 3. They may have some effects on the results, but they do not seem sufficient to explain the large discrepancy.

A careful examination of the algorithm by Storey *et al.* (2005) reveals two differences in their way of performing sequential genome scans compared with MIM:

- They selected the second QTL for each trait from the genome because it has either a strong main effect, or a strong interaction effect with the first QTL, or both. This F test has two degrees of freedom in the numerator.

- They attempted to find the second QTL for each trait no matter how significant the first QTL is.

We replicated their genome scan procedures and then, modified it by incorporating features of MIM while keeping the FDR control mechanism. We collected three sets of F test statistics using three searching schemes:

- Original method: Storey's original sequential genome scan method.

- Main effect method: We modified the original method by removing the interaction effect from Model 2 (Equation 2 of Storey *et al.* (2005)).

- Restricted method: We restricted the search for the second QTL within the 4231 traits with the F statistics larger than 13.94 in the initial scans. Since F tests and LRTs are asymptotically equivalent, we used the same threshold value as the one in MIM. The intention is just to show what we could gain by restricting the search for the second QTL to only those traits that show a relatively strong phenotype-genotype association in the first scan. We kept the interaction term when searching for the second QTL.

Then, following Storey *et al.* (2005), we used Equation 2 and 3 to find q_i for each trait. In each cycle, test statistics were pooled to estimate $f_m(.)$. To estimate $f_0(.)$, we permuted the data 5 times and applied the same tests to generate the test statistics. We used CET method (Doerge and Churchill, 1996) to obtain the conditional null statistics for the second QTL as described above. Null statistics from all the traits in a cycle of sequential genome scan were pooled together to estimate the corresponding $f_0(.)$, except in the restricted method. In this method, only those 4231 traits, for which we did the second cycle of genome scans, contributed to the null distributions of the second F statistics. Traits were then sorted according to their q_i values. A threshold would then be applied to q_i to find the rejection region where the average of q_i, i.e. FDR, was 10%. We implemented this FDR control procedure in our R package 'MIM.eQTL'.

Before submitting these statistics to the package, we took a look at the empirical P value distributions for the second QTL obtained by comparing their F statistics with the corresponding null distributions. Figure 2 shows the distributions. Table 2 gives more information as well as π_0, the prior probability that a second QTL is false, estimated using the smoothing method (Storey and Tibshirani, 2003).

Comparing the P values from the original searching method with those from the two modified methods, we found a consistent pattern. After the modification, P values systemically shifted towards smaller values, which means the null distribution shifted away from the test statistics. Thus, $f_0(.)/f_m(.)$ is expected to decrease for those large F statistics. The estimated π_0 decreased as well. Based on the equations to calculate the FDR, both changes are expected to increase the estimated probability that a second QTL is true (see Equation 3), decrease q_i (Equation 2) and hence allow more two QTL genetic models into the rejection region (Table 3). This also explains the different results between the 'Original method' and the 'Restricted method': although both results shared many F statistics in common, the same statistics were mapped to different q_i in different methods.

Thus, each feature of sequential genome scans from MIM can declare more QTL with FDR

controlled at the same level. We think that Table 3 explains why there is such a big discrepancy between Storey's results and our findings.

Though we could declare more QTL pairs using the 'Main effect' method, interaction effects between these main effect pairs are much weaker than those in the two QTL models declared using Storey's original method (Figure 3). But apparently we have traded them for more main effects in the rejection region.

We also investigated whether genome scans with only main effects tended to declare linked QTL for a trait. For each trait, both the 'Original method' and the 'Main effect method' found a 2 QTL model respectively, though only part of the models were statistically significant. Each model was based upon a pair of markers. From the 'Original method', for 535 traits, the pair was on the same chromosome, including 22 traits with a 2 QTL model declared, or 12.6% of all models passing the 10% FDR threshold (Table 3). From the 'Main effect method', there were 473 such traits, including 77 traits with a 2 QTL model, or 10.3% of all genetic models declared using the 'Main effect method' (Table 3). The histograms for the genetic distances between linked QTL can be found in Figure 4. Thus, including interaction terms in searching for additional genetic factors slightly increased the chance to declare more closely spaced QTL.

Figure 5 suggests one reason for the improved power from the 'Restricted method'. There is quite strong positive correlation between F statistics of the first QTL and the second QTL no matter whether we included the interaction effect or not (Figure 5). Especially, if a trait's first F statistic is less than the cut-off value in the 'Restricted method', i.e., 13.94, in most cases its second F statistic is also less than 13.94. Thus, if a trait has a small F statistic for its first QTL, it will very likely have a small F statistic for its second QTL. Thus, by utilizing the correlation of test statistics between cycles of sequential search, and making decision at each step, MIM effectively excludes those traits with high chance of generating insignificant results. In Bayesian terms of the FDR estimation (Efron *et al.*, 2001), such an exclusion reduces π_0, the prior probability that a test statistic belongs to a false QTL. In terms of exploring the parameter space, it discards certain parts of the space where the models tend to fit the data poorly. Given the huge number of statistical tests in QTL analysis, the restricted sequential test in MIM is a very effective way to relieve the multiple testing burden and to increase the power to declare significant association patterns.

The 'Restricted method' is quite heuristic in deciding for which traits we would not search for the second QTL. Though we used a F statistic larger than 13.94 as a cutoff value here, there was no strict statistical justification for choosing the specific value. We simply did not estimate

any two QTL models for certain traits following this heuristic procedure. We could miss some true 2 QTL models; but as a compensation, we had larger $\Pr(Q_{i2} = 1 |\ Q_{i1} = 1, Data)$ for the remaining traits, and an increased number of significant QTL models than otherwise. Actually, only an exhaustive 2-D search can guarantee to find the best two QTL model for each trait given a sufficiently large sample. In this sense, the original method by Storey is also a heuristic one with better statistical power to achieve more positive findings. Though about 2000 traits were excluded in the second cycle, their associated test statistics and null statistics for their first QTL contributed to the estimated $f_0(.), f_m(.)$ and π_0 (Equation 3), and hence $\Pr(Q_{i1} = 1|Data)$ and q_i for each of the remaining traits. We could also declare single QTL models for them (see Table 3 legend). Another interesting thing is that the number of strong interaction effects detected using the restricted search was even larger than that from the original method (Figure 3).

A Simulation Study

We used simulations to further compare the statistical powers of the 'Original method', the 'Main effect method' and the 'Restricted method'. To save computational time while still addressing this issue, in each simulation, we generated 620 traits for 112 haploid subjects. We filtered out those markers with no recombination with their neighbors or with a high percentage of missing genotypes. All the analysis in this section used 868 markers across the genome.

For each of the 6195 traits, we recorded the 2 QTL models obtained from the real data using the 'Original method' and the 'Main effect method' respectively. We formed a repository of genetic models by randomly combining the following components:

- the marker with the maximal F statistic in the first cycle of genome scan, from which the first QTL will be chosen;
- the marker with the maximal F statistic in the second cycle of genome scan (with or without an interaction effect), from which the second QTL will be chosen;
- additional QTL (discussed later);
- regression coefficients for each factor in the models, which were used as genetic effects in the simulation;
- an independent error item from a standard normal distribution.

We excluded 6 markers with an allele frequency larger than 0.65 or smaller than 0.35 as potential QTL in the simulation. We also excluded those combinations where the 2 QTL were on the same chromosome or in perfect correlation ($r^2 = 1$).

When generating trait values, we sampled from the repository a portion of genetic models derived from the 'Original method' and the rest from the 'Main effect method'. We tried three different mixing proportions: 5%,50% and 95%.

Focusing on comparing the powers to detect 2 QTL genetic models, all trait values were generated from a randomly selected model with at least 2 QTL, though a variable portion of these models included an interaction effect. About half traits were given a third QTL from a binomial sampling procedure with the probability of success equal to 0.5. This probability was to approximate the π_0 estimated for the third cycle of genome scans. This QTL was randomly picked from the genome excluding the chromosomes where the first or the second QTL was located. The effect of this QTL was decided by multiplying the coefficient of the previous QTL (or the second QTL here) with a decaying factor (which was 0.9 in the simulation) and with a random sign. Among those traits with a third QTL, about half traits were given a fourth QTL in the same way. This procedure continued to allow a trait to have as many as 6 QTL.

The simulated data captured many characteristics of the original data. Majority of the sequence polymorphisms remained in the simulation to replicate the burden of multiple testing during genome scans. QTL estimated from the yeast data are not distributed evenly along the chromosomes (Zou *et al.*, 2007), which accounts for in the genetics level the correlated expression variation observed in the mapping population. We utilized the empirical QTL location distributions to perverse the correlation structure among traits. Additional QTL were introduced to simulate expression variations with various numbers of underlying genetic factors (Table 1). Using the estimated regression coefficients as the genetic effects in the simulation, we reproduced the relative contributions from each genetic factor, including the interaction effects, towards each trait value. Using a standard normal error term did change the effect sizes of genetic factors, and hence, the heritability for individual trait. Though the statistical powers to detect 2 QTL models were different between the real data and the simulated data, the changes affected the three genome scan methods similarly.

We applied the three methods on each set of 620 simulated traits and compared the counts of 2 QTL genetic models passing the 10% FDR threshold. We generated 100 such sets for each mixing proportion. Results were summarized in Table 4. It suggests that: 1) the 'Main effect method' is more powerful than the 'Original method' as long as statistical interaction effects

do not dominate the data, which seems to be the case for the yeast data; 2) the 'Restricted method' is consistently more powerful than the 'Original method'. Thus, this simulation study confirms that the MIM procedure comprises the features which can improve the power to extract expression regulatory patterns from the data.

Bioinformatics Analysis

***Cis*-acting and *trans*-acting QTL:** The large number of QTL declared through MIM gives a rich source of information for bioinformatics analysis. Given the 1.5 LOD support interval around a QTL LRT peak that defines the region for an QTL, we first identified genes located in each QTL region. Comparing a trait gene with genes in its QTL region, we found 737 QTL overlap the physical location of their trait genes, which suggest potential *cis*-acting regulations for the QTL. This number of '*cis*-acting' QTL is much larger than what would be obtained when QTL were distributed in the genome randomly ($P = 0$) (See Appendix).

In yeast, 3417 genes are known to contain regulatory targets for certain transcriptional factors (TF) (Luscombe *et al.*, 2004). In this mapping population, expression profiles of 1886 target genes (TG) are mapped to at least one QTL. However, there are only 49 cases, where a TG's QTL overlaps with the genomic location of its corresponding TF. These 49 TF-TG pair-wise relationships include the 3 TGs mapped to their common TF YJL206C, which is the only TF regulation case reported by Yvert *et al.* (2003). If QTL were distributed randomly across the genome, the probability of observing 49 or less TGs mapped onto their TF is close to zero. Thus, contrary to *cis*-acting QTL, there is a clear statistical tendency for TG to avoid being mapped onto their TF.

We offer two explanations for the observed deficiency of TF-TG pairing and the abundance of *cis*-acting QTL in the statistical association patterns. First, the mapping population is composed of yeast cells at various stages of their cell cycles. Most TFs are only active during a certain phase of cell cycle (Luscombe *et al.*, 2004). There could be internal inconsistency of the active regulatory network topology among yeast segregants. Such inconsistency would lower the statistical power to detect co-segregating patterns between expression traits and polymorphic sites near their regulator genes. In this sense, a mapping population under certain external stimuli is expected to represent more faithfully the underlying regulation by TFs in response to stimuli. On the other hand, sequence variability in *cis*-acting elements can affect the transcriptional level of nearby genes in a consistent way across cell cycles. Thus, it is much easier to catch *cis*-acting mechanisms in the population. Second, TF might be too critical to accumulate sequence varia-

tion in normal laboratory strains or wild strains. They are more likely to exert large phenotypic effect and push the organism out of the phenotypic threshold of being normal. On the other hand, the direct influence from a *cis*-acting genetic polymorphism is always limited in its neighborhood. Purifying selection pressure is expected to be much stronger on transcriptional factors than on their targets.

Scale-free network: Many cellular networks, including both protein-protein interaction (Han *et al.*, 2004) and regulatory networks by transcriptional factors (Luscombe *et al.*, 2004), belong to a scale-free network (Barabasi and Oltvai, 2004). Let K be the number of links from one node (gene) to the rest of the genome; $p = \Pr(K = k)$ be the probability density of K. The numerical characteristic of a scale free network is that, K has a power-law distribution, i.e., $p \propto k^{-r}$, where r is a positive constant. On the contrary, $p \propto e^{-k}$ is the feature of a random network.

In expression QTL analysis, K can be interpreted as the number of expression traits affected by a genomic region, and $p = \Pr(K = k)$ can be estimated as the frequency that a gene sequence is included in k QTL intervals. The mapping results from MIM analysis on the yeast data indicate that most genes were in one or two QTL intervals, while a few genomic segments affect as many as 600 expression traits. See Figure 6 for the change of the estimated p as K. p could be well fitted with a function proportional to $K^{-1.66}$. The index was obtained from a linear regression of $log(p)$ onto $log(K)$ (Figure 6).

Compared with a random network, such a scale free network has a much shorter mean path length to connect gene pairs (Barabasi and Oltvai, 2004), allowing a local fluctuation in biochemical pathways easily amplified to hundreds of its near or far neighbors. Thus, a single sequence variation can have pleiotropic effects on the expressions of hundreds of genes. Such a wide spread change in the transcriptome has certain potential to buffer the adverse effect associated with the sequence variation. In this way, a high heritable expression variation can be maintained in the population because of its minor effect on fitness (Barton and Keightley, 2002).

CONCLUSION

The LRT based interval mapping strategy has advantages when compared with a direct correlation study between traits and marker genotypes, even in this mapping population with dense markers. EM algorithm, an integrated part of the LRT in MIM, can handle the occasional situations like missing markers and large marker intervals systemically and efficiently. LOD

support intervals generated by LRTs provide a straightforward and robust way to find an interval estimate of QTL location.

In this paper, we presented MIM as a heuristic way of finding multiple QTL for expression traits. We focused on QTL main effects in model selection. Interaction effects were essentially introduced into the models to refine the results of model selection. We collected LRT statistics in the search for main effects in real and permuted data. Based on these, we estimated the FDR for the subset of the main effect LRTs which were declared as statistically significant with a selected genome-wide type I error rate.

Interaction effects were added into a genetic model after QTL were detected based on their main effects. An alternative approach is to search for these two categories of effects together, but in this way we might sacrifice our ability to detect many more QTL main effects as we have shown in this paper. On the other hand, the complex biochemical interactions in transcriptional regulation networks or other biochemical pathways would not guarantee complex statistical interactions (Barton and Keightley, 2002). All gene products interact biochemically with other gene products in a cell, directly or indirectly; however, most of them have main effects that can be attributed to their own variation, and a lot of them have no detectable interaction effects statistically.

The threshold in MIM plays an important role to detect QTL:

- In the later cycles of genome scans, the search is restricted within the parameter space where the chance to detect strong association is high when we focus on those traits that have shown significant QTL in the previous cycles based on the threshold;
- It serves as the stopping rule to decide how many QTL we want to find for each trait;
- It can be relaxed in the search for QTL main effects to favor those QTL that are involved in strong interaction effects but are below a more stringent threshold in terms of their main effects.

However, when specifying such a threshold for the genome scans as in MIM, it becomes hard to control the final FDR at a pre-defined level. A solution is to run MIM and the FDR estimation procedure iteratively and adjust the MIM threshold according to the estimated FDR. In this way, it is possible to control the FDR in the QTL model selection by MIM. As compared to Storey *et al.* (2005), our strategy for expression QTL mapping analysis has more statistical power to find significant QTL.

ACKNOWLEDGMENT

This work was partially supported by the USDA Cooperative State Research, Education and Extension Service, grant number 2005-00754. We sincerely thank Dr. Leonid Kruglyak for the access of the wonderful yeast data set.

References

Barabasi, A. L., and Z. N. Oltvai, 2004 NETWORK BIOLOGY: Understanding The Cell'S Functional Organization. *Nat Rev Genet* **5**: 101–113.

Barton, N. H., and P. D. Keightley, 2002 Understanding Quantitative Genetic Variation. *Nat Rev Genet* **3**: 11–21.

Basten, C., B. Weir, and Z. Zeng, 2002 *QTL Cartographer, Version 1.17*. Department of Statistics, North Carolina State University, Raleigh, NC.

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**: 1165–1188.

Bing, N., and I. Hoeschele, 2005 Genetical Genomics Analysis of a Yeast Segregant Population for Transcription Network Inference. *Genetics* : genetics.105.041103.

Brem, R., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.

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

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, D. W. Threadgill, K. F. Manly, and R. W. Williams, 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37**: 233–242.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–71.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285–94.

Efron, B., and R. Tibshirani, 2002 Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol* **23**: 70–86.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher, 2001 Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**: 1151–1160.

Han, J.-D. J., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, 2004 Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**: 88–93.

Ihmels, J., R. Levy, and N. Barkai, 2004 Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat Biotech* **22**: 86–92.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* **152**: 1203–1216.

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

Luscombe, N. M., M. Madan Babu, H. Yu, M. e. Snyder, S. A. Teichmann, and M. Gerstein, 2004 Genomic analysis of regulatory network dynamics reveals large topological changes. *nature* **431**: 308–312.

Marchini, J., P. Donnelly, and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**: 413–417.

Sax, K., 1923 The Association Of Size Differences With Seed-Coat Pattern And Pigmentation In Phaseolus Vulgaris. *Genetics* **8**: 552–560.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend, 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.

Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440–5.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint, 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* **38**: 879–87.

Wolfinger, R., G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**: 625–637.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak, 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the rol e of transcription factors. *Nat Genet* **35**: 57–64.

Zeng, Z., C. Kao, and C. Basten, 1999 Estimating the genetic architecture of quantitative traits. *Genet Res.* **74**: 279–289.

Zou, W., D. Aylor, and Z.-B. Zeng, 2007 eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics* **8**: 7.

APPENDIX

We simulated random mapping results in the following way:

- Number of QTL for each trait was generated according to the empirical distribution of number of QTL per trait in the real mapping result;
- The chromosome where a QTL would be located was picked with the probability proportional to the chromosomal length; a QTL position was then picked from the chromosome randomly;
- Length of the QTL was generated according to a gamma distribution with shape parameter 1.96 and scale parameter 25659.8. These two values were obtained by fitting the versatile gamma density function to the histogram of observed QTL physical length.

After generating 100000 mapping results, we counted the number of *cis*-acting and *trans*-acting QTL in each result.

Table 1: Sequential genome scan using MIM

Cycle	#Scanned[1]	#Retained[2]	#Claimed[3]
1	6195	3367	3354
2	3367	1617	1242
3	1617	578	422
4	578	197	122
5	197	66	37
6	66	10	5

[1] Number of traits for which a genome scan was performed in the cycle

[2] Number of traits with one QTL retained in the cycle, controlling type I error rate at 10%.

[3] Number of traits with one QTL finally claimed from the cycle, controlling type I error rate at 5% to declare each QTL.

Table 2: Empirical P value distributions for the second QTL

	Original	Main effect	Restricted
75% quantile	0.62	0.59	0.53
50% quantile	0.33	0.30	0.25
25% quantile	0.12	0.09	0.09
π_0	0.58	0.49	0.31

Table 3: Number of two QTL genetic models claimed

Method	# traits[1]	Threshold on q_i
Original method	174	0.15
Main effect method	747	0.19
Restricted method	662[2]	0.15

[1] Number of traits with a two QTL genetic model declared while controlling FDR at 10%.

[2] Those traits which were excluded from the second genome scans were not included. We could also declare a single QTL model for them. If including their q_i in calculating the FDR, we could declare significant genetic models for 862 traits, including 526 traits with 2 QTL and 336 traits with 1 QTL. The corresponding threshold on q_i is 0.14.

Table 4: Power comparison through simulations

Proportion[1]	# Original	# Main effect	# Restricted[2]
5%	99(37)[3]	238(71)	168(72)
50%	161(40)	194(39)	236(67)
95%	226(62)	190(58)	306(64)

[1] The proportion of genetic models with an interaction effect in generating trait values.

[2] Number of 2 QTL models declared using the 'Original method', the 'Main effect method', and the 'Restricted method' while controlling FDR at 10%.

[3] Data are shown as mean(standard deviation) numbers of significant 2 QTL models, estimated from 100 iterations of simulations.

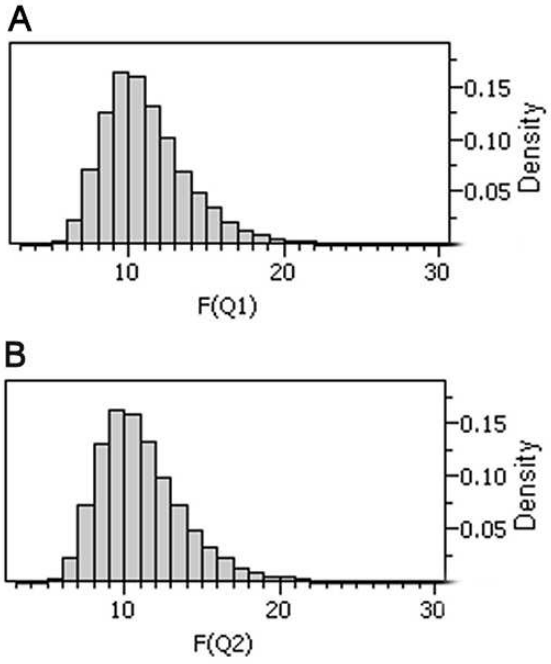

Figure 1: Null distributions of F statistics. A: The null distribution of genome-wide maximal F statistics for the first QTL. B: The conditional null distribution of genome-wide maximal F statistics for the second QTL, given the existence of the first QTL, obtained using CET.

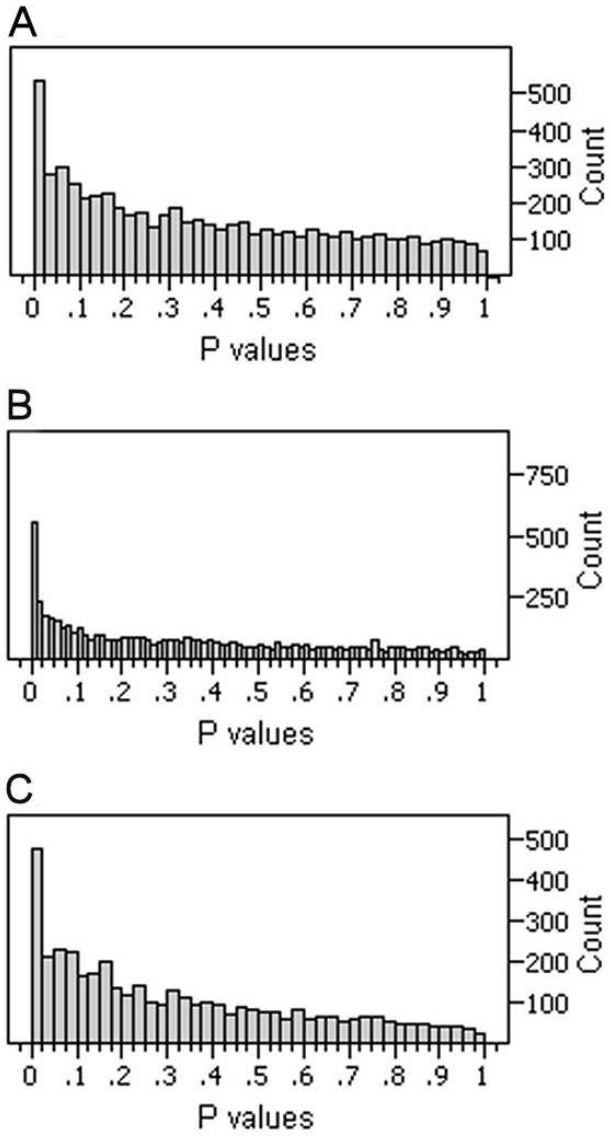

Figure 2: Empirical P value distributions for the second QTL. A: 6195 P values obtained using the 'Original method' from Storey *et al.* (2005) for all traits. B: 6195 P values obtained using the 'Main effect method' for all traits. C: P values obtained using the 'Restricted method'. The histogram contains P values from the second QTL of the 4231 traits. Quantiles of these distributions can be found in Table 2.

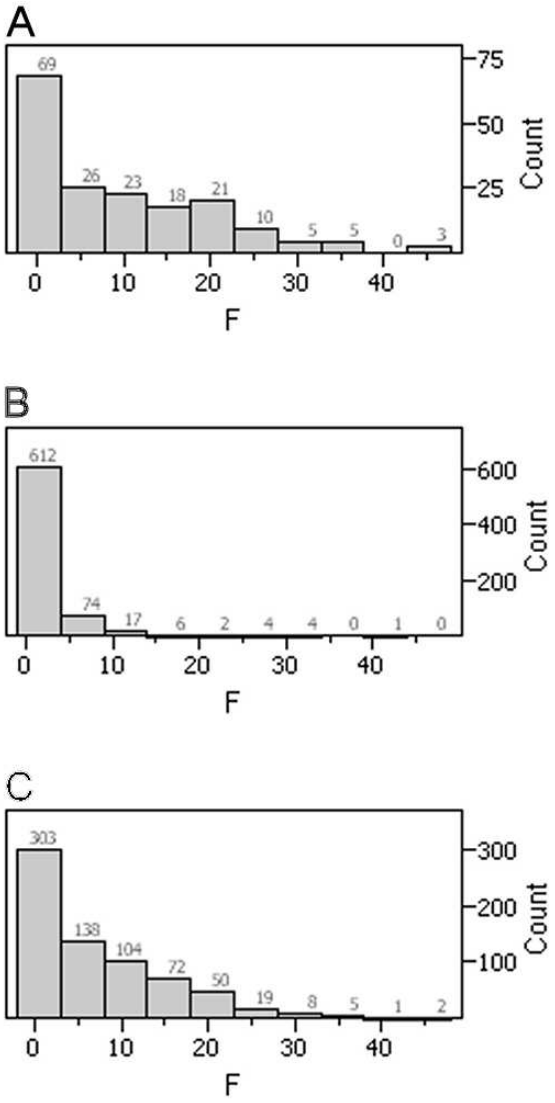

Figure 3: Histograms of F statistics of the interaction effects between two QTL detected through various methods. Numbers above each bar show the number of traits, or the number of two QTL interaction effects with type I F statistics in the bin. A: The significances of the interaction effects in the two QTL models claimed using the original method from Storey *et al.* (2005). B: We tested the significance of the interaction between each QTL pair for a trait obtained using the main effect method. C: F statistics of the interaction effects in two QTL models obtained using the restricted method.

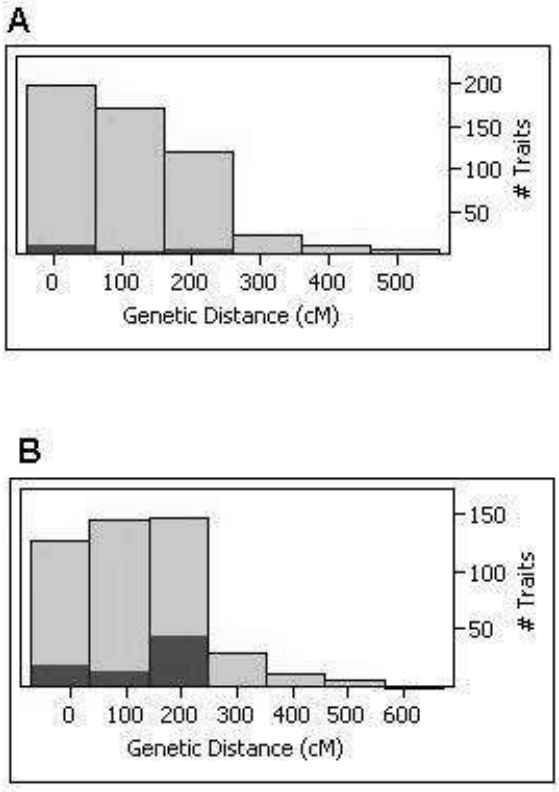

Figure 4: Genetic distances between 2 markers with the maximal F statistics in the 2 cycles of genome scans for each trait. Marker pairs on different chromosomes are not shown. Dark regions correspond to those pairs that were declared as 2 QTL genetic models in Table 3. A: from the 'Original method'. B: from the 'Main effect method'.

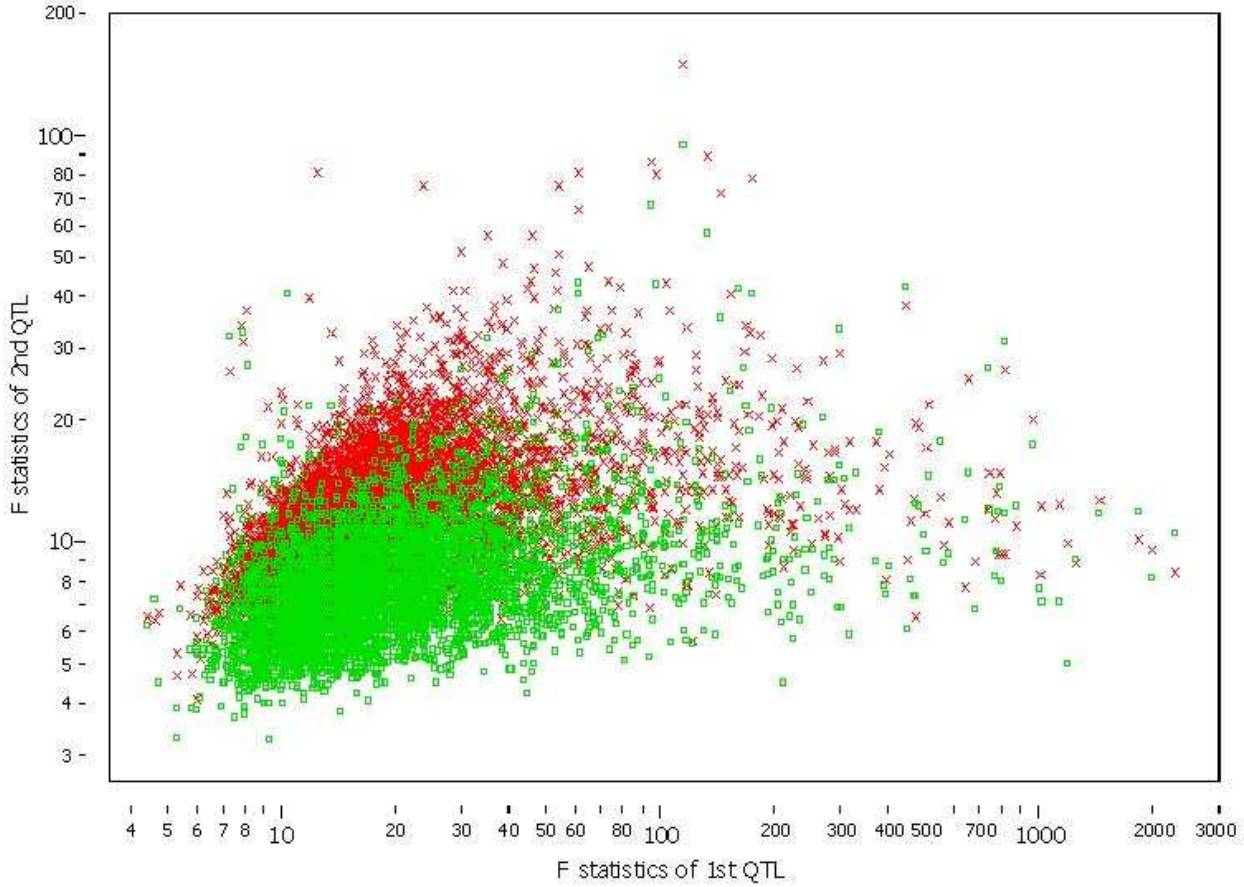

Figure 5: A scatter plot for F statistics of the first QTL and the second QTL of all traits. Green squares for F statistics obtained using Storey's original method. Red crosses for F statistics obtained using the main effect method.

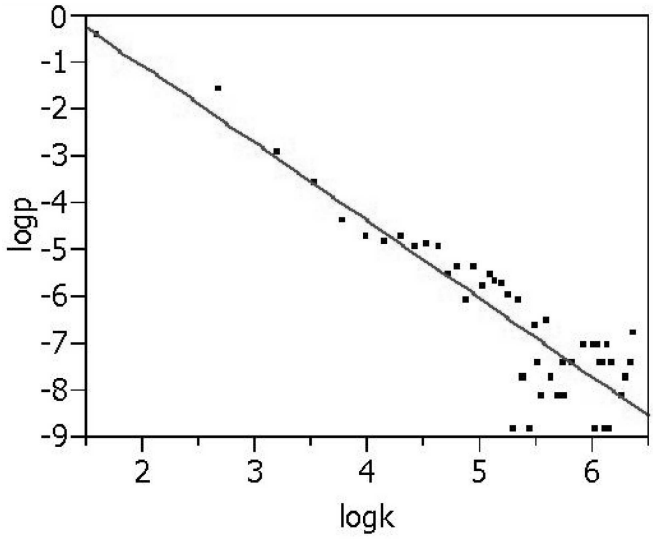

Figure 6: QTL mapping results suggest transcriptional regulation network has the scale free property. After collecting the numbers of expression traits mapped onto each gene, these numbers were put into a frequency table with bin size of 10. Let K be the mid-bin value (the mean of the numbers of expression traits affected by a gene, averaged within the bin), p be the frequency (the percentage of the genes that affect $[K-5, K+5]$ traits). Based on this frequency table, we estimated the empirical density function using parametric curve fitting. Log-transformed frequencies $log(p)$ were regressed onto log-transformed mid-bin values $log(K)$. The linear regression function is $logp = 2.24 - 1.66 \times logk$.