Biometrics 62, 19–27
March 2006
DOI: 10.1111/j.1541-0420.2005.00437.x

Statistical Methods for Expression Quantitative Trait Loci (eQTL) Mapping

C. M. Kendziorski,[1,*] M. Chen,[2] M. Yuan,[1] H. Lan,[3] and A. D. Attie[3]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53703, U.S.A.
[2]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53703, U.S.A.
[3]Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53703, U.S.A.
[*]*email:* kendzior@biostat.wisc.edu

Summary. Traditional genetic mapping has largely focused on the identification of loci affecting one, or at most a few, complex traits. Microarrays allow for measurement of thousands of gene expression abundances, themselves complex traits, and a number of recent investigations have considered these measurements as phenotypes in mapping studies. Combining traditional quantitative trait loci (QTL) mapping methods with microarray data is a powerful approach with demonstrated utility in a number of recent biological investigations. These expression quantitative trait loci (eQTL) studies are similar to traditional QTL studies, as a main goal is to identify the genomic locations to which the expression traits are linked. However, eQTL studies probe thousands of expression transcripts; and as a result, standard multi-trait QTL mapping methods, designed to handle at most tens of traits, do not directly apply. One possible approach is to use single-trait QTL mapping methods to analyze each transcript separately. This leads to an increased number of false discoveries, as corrections for multiple tests across transcripts are not made. Similarly, the repeated application, at each marker, of methods for identifying differentially expressed transcripts suffers from multiple tests across markers. Here, we demonstrate the deficiencies of these approaches and propose a mixture over markers (MOM) model that shares information across both markers and transcripts. The utility of all methods is evaluated using simulated data as well as data from an F_2 mouse cross in a study of diabetes. Results from simulation studies indicate that the MOM model is best at controlling false discoveries, without sacrificing power. The MOM model is also the only one capable of finding two genome regions previously shown to be involved in diabetes.

Key words: Bayesian hierarchical mixture model; Expression trait loci (eQTL) mapping; Gene expression; Microarray; Quantitative trait loci (QTL) mapping.

1. Introduction

Traditional genetic mapping has largely focused on the identification of loci affecting one, or at most a few, complex traits. Microarrays allow for measurement of thousands of gene expression abundances, themselves complex traits, and a number of recent investigations have considered these measurements as phenotypes in mapping studies. This type of approach has the potential to impact a broad range of biological endeavors (Cox, 2004). Utility has been demonstrated in identifying candidate genes (Schadt et al., 2003), in inferring not only correlative but also causal relationships between modulator and modulated genes (Brem et al., 2002; Schadt et al., 2003; Yvert et al., 2003), and in elucidating subclasses of clinical phenotypes (Schadt et al., 2003). As a result of these early successes, a number of efforts are now underway to localize the genetic basis of gene expression.

As part of one such effort, an experiment was designed to identify the genetic basis for differences between two inbred mouse populations (B6 and BTBR) that show diverse responses to a mutation in the leptin gene. Leptin is a protein hormone with important effects in regulating body weight, metabolism, and reproductive function (Zhang et al., 1994). A mutation in the leptin gene causes only mild and transient type II diabetes in B6 mice, but severe diabetes in BTBR mice. Microarray experiments have led to the identification of previously unappreciated genes that are differentially expressed between the populations (Lan et al., 2003a). To identify genetic modifiers and novel regulatory pathways, we have collected second-generation offspring from these populations. Each offspring has been genotyped at 145 markers across the genome and 45,265 expression traits have been obtained for each using Affymetrix chips.

It is clear that the experimental set up in an expression quantitative trait loci (eQTL) mapping study is similar in structure to a traditional quantitative trait loci (QTL) mapping study, but with thousands of phenotypes. The simplicity with which this difference can be stated obscures the resulting challenges posed for the statistical analysis of eQTL data. The statistical methods available for multi-trait QTL

mapping consider relatively few traits and are not easily extended to the eQTL setting as they require estimation of a phenotype covariance matrix, which is not feasible for hundreds or thousands of traits (for a review of multiple-trait QTL methods, see Lund et al., 2003 and references therein).

To circumvent this, one could apply single-trait QTL mapping methods to reduced summaries of expression obtained, for example, via principal components analysis (Lan et al., 2003b). Doing so has proven useful; however, transcript-specific information is oftentimes of primary interest. When this is the case, simple tests (such as the Wilcoxon–Mann–Whitney) for linkage between each marker and transcript can be carried out with combinations identified as important if the resulting p-value is sufficiently small (Brem et al., 2002). Alternatively, interval mapping methods (see Broman, 2001 for a review) can be used to obtain transcript-specific significance profiles that are then calibrated via a common critical value intended to account for the potential increase in type I error induced by testing at multiple markers (Schadt et al., 2003).

As we show here, the repeated application of a transcript-specific linkage analysis has a number of serious flaws. Most notably, although adjustments are made for multiple tests across markers, few if any adjustments are made for multiple tests across transcripts. Furthermore, information common across transcripts is not utilized, which can lead to a loss in power. The use of a single, approximate, critical value for all transcripts is also problematic as the exact critical value for a given transcript depends not only on the number of transcripts and genomic locations tested (fixed for every data set), but also on the expression levels of that transcript. Using a common critical value further reduces power for some transcripts while increasing type I error for others. To address some of these issues, a marker-based approach can be used.

As a main goal of eQTL mapping is to identify transcripts and genomic locations that are significantly linked, instead of testing each transcript for significant linkage across the genome as described above, one could test each genome location for linked transcripts. At a given marker, this consists of identifying all transcripts with significant differences among phenotype groups where groups are determined by the marker's genotype. In this context, any method for identifying differentially expressed (DE) genes could be applied (for a review of methods, see Parmigiani et al., 2003). An advantage of this marker-based approach is that most methods to identify DE transcripts adjust for the multiple tests across transcripts. However, none of the methods currently used to identify DE genes would be applicable to the eQTL setting between markers where genotypes are unknown; and furthermore, although multiple tests across transcripts would be accounted for, multiple tests across markers would not be.

We have developed an approach that combines advantages from both the transcript- and marker-based methods. Our method maps eQTL by combining information across transcripts while controlling for the multiplicities induced by tests at transcripts and markers. The advantages are demonstrated and validated using simulated data as well as data from the diabetes study described above.

Section 2 describes in detail a transcript-based and two marker-based approaches. As discussed, none of the approaches properly accounts for multiplicities. An empirical Bayes hierarchical mixture over markers (MOM) model, which adjusts for relevant multiplicities, is introduced in Section 3. A simulation study is presented in Section 4, demonstrating that the MOM model controls the false discovery rate (FDR), without a substantial loss in power. The data set of interest is discussed in detail in Section 5 and is analyzed using all methods considered. Section 6 gives a discussion and outlines open questions in the analysis of eQTL data.

2. eQTL Mapping Methods

Consider for simplicity a backcross population from two inbred parental populations, P1 and P2, genotyped as 0 or 1 at M markers (this simplification to a backcross is not required and is relaxed in our simulations and analyses). For the kth animal, let $y_{t,k}$ denote the expression level for transcript t and $g_{m,k}$ denote the genotype at marker m; $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots, n$. Of interest is the identification of significant linkages between transcripts and markers. To be precise, a transcript t is linked to marker m if $\mu_{t,0} \neq \mu_{t,1}$, where $\mu_{t,0(1)}$ denotes the latent mean level of expression of transcript t for the population of animals with genotype 0(1) at marker m. Suppose observations $y_{t,k}$ have density $f_{\text{obs}}(y_{t,k} \mid \mu_{t,g_{m,k}}, \theta)$ where θ denotes any remaining unknown parameters. Under the null hypothesis of no linkage, the data are governed by $\prod_{k=1}^{n} f_{\text{obs}}(y_{t,k} \mid \mu_{t,0} = \mu_{t,1}, \theta)$; and under the alternative, $\prod_{k=1}^{n} \{f_{\text{obs}}(y_{t,k} \mid \mu_{t,0}, \theta)\}^{1-g_{m,k}} \{f_{\text{obs}}(y_{t,k} \mid \mu_{t,1}, \theta)\}^{g_{m,k}}$. As discussed below, a main difference between the transcript-based (TB) and marker-based (MB) approaches arises from different assumptions regarding the latent means.

2.1 *Transcript-Based Approach*

A TB approach refers generally to the repeated application of any single-phenotype mapping method to each mRNA transcript, with locations identified as important if the test statistic of interest exceeds some critical value. The LOD score

$$\log_{10}\left\{\frac{\prod_{k=1}^{n} f_{\text{obs}}(y_{t,k} \mid \hat{\mu}_{t,0}, \hat{\mu}_{t,1}, \hat{\theta})}{\prod_{k=1}^{n} f_{\text{obs}}(y_{t,k} \mid \hat{\mu}, \hat{\theta})}\right\}$$

is often used as the statistic measuring evidence in favor of linkage, where $(\hat{\cdot})$ denotes the maximum likelihood estimate (MLE) of the associated parameter(s) and μ denotes the mean common across genotype groups (Lander and Botstein, 1989). Critical values can be obtained theoretically (Dupuis and Siegmund, 1999) or via permutations (Churchill and Doerge, 1994).

The specific TB approach that will be evaluated here assumes a Gaussian density for f_{obs} with tests performed at every marker and critical values determined theoretically by the formulas given in Dupuis and Siegmund (1999). This marker regression approach, referred to as TB-MR, is identical (at each marker) to that used by Schadt et al. (2003) to identify significantly linked expression traits in an F_2 mouse cross.

2.2 *Two Marker-Based Approaches*

To identify transcripts significantly linked to genomic locations, instead of testing each transcript for significant linkage

across markers, one could test at each marker for significant linkage across transcripts. This amounts to identifying DE transcripts at each marker, with groups determined by marker genotypes. The MB approach refers generally to the repeated application, at each marker, of any method for identifying DE transcripts. In this setting, a number of approaches could be used (for a review, see Parmigiani et al., 2003). We consider two: an empirical Bayes approach, *EBarrays*, described in detail in Kendziorski et al. (2003) and an approach based on the Student's *t*-test followed by p-value adjustment, similar to that proposed by Dudoit et al. (2002).

EBarrays assumes measurements $y_{t,k}$ arise as conditionally independent random deviations from an observation distribution $f_{\text{obs}}(\cdot \mid \mu_{t,\cdot}, \theta)$. Instead of treating the $\mu_{t,\cdot}$'s as fixed effects as in TB-MR, the underlying means are described by a distribution $\pi(\mu)$. In this case, an equivalently expressed (EE) transcript t presents data $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \ldots, y_{t,n})$ according to the distribution

$$f_0(\mathbf{y}_t) = \int \left\{ \prod_{k=1}^{n} f_{\text{obs}}(y_{t,k} \mid \mu) \right\} \pi(\mu)\, d\mu, \tag{1}$$

where $\mu = \mu_{t,0} = \mu_{t,1}$.

For a DE transcript, let $\mathbf{y}_t^l$ denote the set of observations for animals with genotype $l = 0, 1$. The data $\mathbf{y}_t = (\mathbf{y}_t^0, \mathbf{y}_t^1)$ are governed by the distribution

$$f_1(\mathbf{y}_t) = f_0\big(\mathbf{y}_t^0\big)\, f_0\big(\mathbf{y}_t^1\big), \tag{2}$$

owing to the fact that different mean values, $\mu_{t,0}$ and $\mu_{t,1}$, govern the different subsets $\mathbf{y}_t^0$ and $\mathbf{y}_t^1$ of samples and are considered independent draws from $\pi(\mu)$. As a transcript's expression state is never known a priori, the marginal distribution of the data is given by $pf_1(\mathbf{y}_t) + (1-p)f_0(\mathbf{y}_t)$ where p denotes the proportion of DE transcripts. With estimates of p, f_0, and thus f_1 obtained via the EM algorithm, the posterior probability of DE is calculated by Bayes' rule.

Although a number of parametric assumptions are available in *EBarrays*, for comparison with the TB-MR approach, here we also consider a Gaussian model on the log observations for f_{obs} and a Gaussian model for π. Specifically, for a log-transformed expression measurement $y_{t,k}$,

$$y_{t,k} \sim N\big(\mu_{t,g_{m,k}}, \sigma^2\big) \quad \text{and} \quad \mu_{t,\cdot} \sim N\big(\mu_0, {\tau_0}^2\big). \tag{3}$$

At a particular marker, a transcript is identified as significantly linked if the posterior probability of differential expression exceeds some threshold. These posterior probabilities have been referred to as "local FDRs" (Efron et al., 2001; Efron and Tibshirani, 2002; Efron, 2004) and it has been shown that to control the posterior expected FDR at $\alpha \cdot 100\%$, the appropriate threshold is the smallest posterior probability such that the average posterior probability of all transcripts exceeding the threshold is larger than $1-\alpha$ (Efron, 2004; Newton et al., 2004). This marker-based empirical Bayes approach will be referred to as MB-EB.

The second MB approach consists of calculating Student's *t*-statistics at a marker and obtaining adjusted p-values. Dudoit et al. (2002) propose methods that control the family-wise error rate across transcripts. Here, we control the FDR using q-values (Storey and Tibshirani, 2003). In particular, to control the FDR at α, transcripts with q-values $\leq \alpha$ are considered significant; MB-Q will denote this MB approach.

2.3 *TB and MB Combined*

To test transcript and marker combinations simultaneously, one could consider the p-value matrix obtained from calculating Student's *t*-statistics for every transcript at every marker, and calculate q-values for the entire matrix at once. The FDR can be controlled using q-values as described above. We refer to this approach as Q-ALL. This approach is justified provided the p-values are weakly dependent (Storey, Taylor, and Siegmund, 2004). Storey and Tibshirani (2003) hypothesize that weak dependence is the most likely form of dependence in genomewide studies such as the eQTL study of Brem et al. (2002).

3. Mixture over Markers Model

Although the TB and MB approaches described above are in many ways fundamentally different, they share an important flaw. Separate tests are conducted for each transcript-marker pair, and each measures evidence that the transcript maps to that marker relative to evidence that it maps nowhere. Since a transcript can map to any of many marker locations, the evidence that a transcript maps to a particular marker should not be judged relative only to the possibility that it maps nowhere, but rather relative to the possibility that it maps nowhere *or* to some other marker. This idea motivates the MOM model.

Suppose a transcript t maps nowhere with probability p_0 or to any marker m with probability p_m where $\sum_{i=0}^{M} p_i = 1$ and M denotes the total number of markers. (In fact, this is only an approximation as the transcript could map in between markers. This possibility is discussed in Section 6.) The marginal distribution of the data $\mathbf{y}_t$ is then given by

$$p_0 f_0(\mathbf{y}_t) + \sum_{m=1}^{M} p_m f_m(\mathbf{y}_t), \tag{4}$$

where f_m describes the distribution of data if transcript t maps to marker m (f_0 describes the data for nonmapping transcripts). A density of the form given by equation (1) ((2)) describes the marginal distribution of data for nonmapping (mapping) transcripts. In the degenerate case of a single marker, equation (4) reduces to the mixture model given below equation (2) that forms the basis for MB-EB. For most eQTL mapping data sets, including the one discussed in Section 5, M is large (>100).

Similar to MB-EB, a Gaussian model is assumed for $f_{\text{obs}}(\cdot)$ and for $\pi(\cdot)$. However, here we allow for the possibility that clusters of transcripts present data with different variances. Thus, σ^2 as in equation (3) is no longer constant, but is cluster dependent. Cluster membership is determined by K means prior to model fitting. The total number of clusters is chosen by the Bayes Information Criterion (BIC). Model fit proceeds via EM (see details in Kendziorski et al., 2003). Multiple initial value configurations are used to check convergence. Diagnostics such as those described in Newton and Kendziorski (2003) should always be checked. For the moderately sized data set described in Section 5, parameter estimates were obtained via the EM algorithm implemented in `R` 1.9.1 (R Development Core Team, 2004). This took under 9 hours

on a Dell Precision 650 (Xeon, 3 GHz) with 4 GB of memory. We found that 20 iterations were sufficient to reach convergence. We also found that results were robust to different initial cluster centers, but dependent on the lower bound for the number of clusters chosen via BIC. If too few clusters were chosen (fewer than the optimal predicted by BIC), model diagnostics were poor.

Once parameter estimates are obtained, posterior probabilities of mapping nowhere or to any of the M locations are calculated via Bayes' rule. A transcript is identified as DE using the MOM approach if the posterior probability of EE is smaller than some threshold, where thresholds are chosen to bound the posterior expected FDR at 5% as described in Section 2.2. Expression QTL for identified transcripts are those contained in the 90% highest posterior density (HPD) region (Carlin and Louis, 1998). With thousands of transcripts, posterior uncertainty regarding θ is generally very small (Kendziorski et al., 2003) and so the anti-conservative nature of the HPD intervals should be minimal.

4. Simulation Studies

To assess the performance of these approaches, we performed a small set of simulation studies. The simulations are in no way designed to capture the many complexities of eQTL data, but rather to provide some preliminary information on operating characteristics of the approaches in simple settings. Marker genotype data were obtained from chromosomes 2 and 3 of the F_2 data described in Section 5. Chromosome 2 (3) contains 17 (6) markers with an average intermarker distance of 7.6 (17.7) cM. An eQTL at marker 5 on chromosome 2 was simulated; no eQTL is simulated on chromosome 3. Each transcript is simulated as either EE or in any one of four DE patterns ($aa \,|\, Aa, AA$; $aa, Aa \,|\, AA$; $aa, AA \,|\, Aa$; $aa \,|\, Aa \,|\, AA$) where "|" denotes inequality among the latent genotype group means. Pattern membership is determined by a multinomial where the expected proportion of transcripts in each pattern is specified at 3%, 3%, 1%, and 3%, respectively.

Conditional of the mean pattern, simulated log intensities follow a Gaussian distribution. Since both the TB and MB approaches assume a log-normal distribution (for TB, the intensities are logged before analysis), this assumption does not bias the simulation in favor of any method. Rather than specify arbitrary means and variances for the simulation, we use values derived from the F_2 data. Consider a single transcript t. Latent means for each genotype group are obtained by calculating sample averages within the groups. As the genotype groupings change at each marker, so too will these averages. To remedy this, the median value across markers within each genotype group specifies $\mu_{t,aa}$, $\mu_{t,Aa}$, and $\mu_{t,AA}$. This is done separately for each transcript. The differences between the aa and AA genotype groups are also considered. A length T vector δ is defined as the maximum of the differences across markers.

For one transcript t, the aa group mean is sampled from the vector $\mu_{\cdot,aa}$. If t is EE, the means in the heterozygous and homozygous AA group are set to the sampled value, $\mu_{s*,aa}$. If t is in any DE pattern, a random sample, δ_{s*}, is taken from the upper quartile of the vector δ. If $aa \,|\, Aa$ for t, the heterozygous mean is defined to be $\mu_{s*,aa} + \delta_{s*}$. If t is in pattern $aa \,|\, Aa \,|\, AA$, the homozygous AA mean is $\mu_{s*,aa} + 2 \times \delta_{s*}$.

To set the variance for a transcript t, we use the posterior mean of σ_t^2, given by $\sum_{k=1}^{n}(y_{t,k} - \bar{y}_{t,\cdot})^2 + \nu_0\sigma_0^2/\nu_0 + n - 2$ (derived assuming the variance is distributed as scaled inverse chi square $\sigma_t^2 \sim \text{Inv}\,\chi^2(\nu_0, \sigma_0^2)$). Note that as $\nu_0 \to 0$, the posterior mean approaches $(n-1)s^2/(n-2) \approx s^2$, the transcript-specific sample variance, which is the naive estimate of any EE transcript variance under TB-MR assumptions. Data simulated with small ν_0 are therefore consistent with assumptions made in TB-MR. As $\nu_0 \to \infty$, the posterior mean approaches a constant variance σ_0^2, which is assumed in MB-EB (note that this assumption implies a constant coefficient of variation on the raw gene expression scale). By varying ν_0, operating characteristics can be evaluated without biasing the results in favor of one method. Data simulated by this empirical method have marginal distributions that are virtually indistinguishable from the observed data.

Seven sets of simulations were obtained for ν_0 between 5^{-5} and 5^5. At each fixed ν_0, the profile marginal MLE is obtained for σ_0^2. For each simulated data set, thresholds are chosen as described in Section 2 to control the type I error rate across the two simulated chromosomes at 5% for TB-MR (by the formulas in Dupuis and Siegmund, 1999, the critical value for the simulations is 2.57) and to control the FDR at 5% for MB-EB, MB-Q, and Q-ALL. The location of the maximum LOD (TB-MR), maximum posterior probability of DE (MB-EB), or minimum q-value (MB-Q and Q-ALL) for each transcript was recorded. Mapping transcripts are defined as those for which the evidence in favor of linkage at the location of the maximum (minimum) exceeds the threshold (or is smaller than the threshold in the case of MB-Q and Q-ALL). With multiple transcripts and putative linkage locations, the definition of power and FDR in an eQTL study is not obvious. By only considering the single, most likely location of mapping for each transcript as we have done here (given by maximum LOD, minimum q-value, etc.), the definitions are simplified.

Power measures the ability to identify the DE transcripts exactly at marker 5 or either of the flanking markers that are 16.5 cM and 5.8 cM away, respectively (this definition is motivated by that used in Broman and Speed, 2002, where an identification is deemed correct if it is made within a 20 cM window containing the true QTL—in that work, unlike here, the QTL was located in the center of the window). As shown in Figure 1 (left panel), there is little variation in power across ν_0. MB-Q is the most powerful method, followed by TB-MR, Q-ALL, MOM, and MB-EB. Power-b only considers calls exactly at marker 5. Table 1 shows that there is only a slight decrease in power when the flanking markers are not considered. Although power is significantly different among some of the approaches at $\alpha = 5\%$, the magnitude of the differences is quite small. This is not the case for FDR.

FDR gives the proportion of transcripts, out of all that mapped to chromosome 2 or 3 that were not truly DE or that were DE but mapped to a region outside the flanking marker region. Figure 1 (right panel) shows that the MOM model is the only approach with well-controlled FDR over a variety of simulations (indexed by ν_0). For the TB and MB methods, FDR is well over the target level of 0.05 for virtually

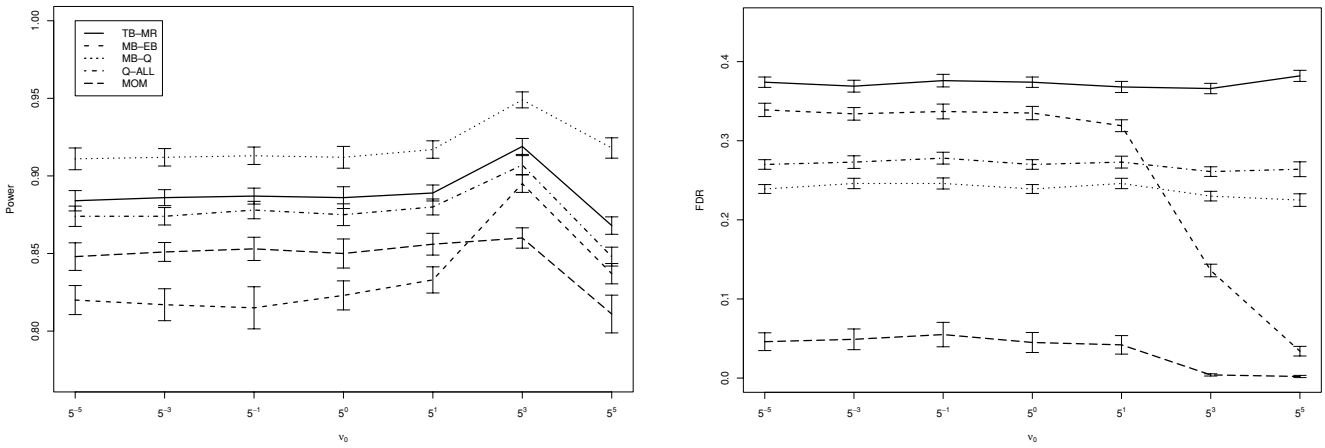

Figure 1. For each value of ν_0, 20 simulated data sets are generated (see Section 4). Operating characteristics are evaluated for each of the five methods on each data set. Table 1 reports the average performance at each value of ν_0. Shown here are two operating characteristics—power (left panel) and FDR (right panel)—along with the 95% pointwise confidence intervals.

all values of ν_0. FDR for MB-EB is controlled at the target level of 5% only when ν_0 is large. This is somewhat expected since as $\nu_0 \to \infty$, the simulation more closely approximates the assumptions made in MB-EB. Any increase in FDR due to repeated tests at markers in this case is minimal. Here, a false discovery can be made due to identification of EE genes or DE genes at nonflanking markers. Over two thirds of the false calls for each method are made from the former for every value of ν_0 (results not shown). The number of false calls made on chromosome 3 (N-chr3) alone is also considered. As shown in Table 1, TB-MR identifies the most transcripts on chromosome 3.

These results suggest that it is difficult in most cases to control FDR using a simple application of a TB or MB approach. This is because the TB-MR approach considers each transcript in isolation, controlling a type I error rate across markers, with no control for multiple tests across transcripts. MB-EB and MB-Q share information across transcripts to

Table 1
Average operating characteristics (OCs) for TB-MR, MB-EB, MB-Q, Q-ALL, and MOM

		ν_0						
OC	Method	5^{-5}	5^{-3}	5^{-1}	5^{0}	5^{1}	5^{3}	5^{5}
Power	TB-MR	0.884	0.886	0.887	0.886	0.889	0.919	0.868
	MB-EB	0.820	0.817	0.815	0.823	0.833	0.895	0.837
	MB-Q	0.911	0.912	0.913	0.912	0.917	0.949	0.918
	Q-ALL	0.874	0.874	0.878	0.875	0.880	0.907	0.848
	MOM	0.848	0.851	0.853	0.850	0.856	0.860	0.811
Power-b	TB-MR	0.852	0.85	0.856	0.854	0.853	0.878	0.816
	MB-EB	0.807	0.803	0.804	0.811	0.818	0.881	0.818
	MB-Q	0.893	0.893	0.896	0.895	0.898	0.928	0.887
	Q-ALL	0.844	0.841	0.848	0.846	0.846	0.868	0.799
	MOM	0.848	0.85	0.852	0.85	0.856	0.86	0.811
FDR	TB-MR	0.286	0.286	0.293	0.285	0.286	0.28	0.301
	MB-EB	0.282	0.281	0.285	0.279	0.269	0.117	0.034
	MB-Q	0.24	0.246	0.246	0.24	0.245	0.23	0.226
	Q-ALL	0.202	0.209	0.213	0.202	0.209	0.195	0.207
	MOM	0.038	0.041	0.046	0.037	0.036	0.005	0.002
N-chr3	TB-MR	86.5	82.2	82.7	86.45	82.1	85.95	82.2
	MB-EB	48.8	46	44.95	47.9	43.25	11.45	0.15
	MB-Q	0.55	0.65	0.25	0.55	0.65	0.55	0.55
	Q-ALL	51	49.55	50.2	50.95	49.6	49.65	42.9
	MOM	3.75	4.15	4.3	4.1	3.25	0	0

Note: Averages are calculated over 20 data sets; standard errors were less than 0.005 for power, power-b, and FDR and less than 2 for N-chr3. Power measures the ability to identify DE transcripts at marker 5 or either of the flanking markers; Power-b considers calls exactly at marker 5. Other OC definitions and details of the simulation are given in the text (see Section 4).

control an expected FDR at each marker, but do not account for tests at multiple markers. When model assumptions do not hold, MB-EB performs poorly. The MOM approach addresses these deficiencies. It allows for information sharing across transcripts while controlling for multiplicities across both transcripts and markers; and, as a result, much improved FDR control is observed.

5. eQTL Data Analysis

The *ob* mutation in the C57BL/6J mouse background (B6-*ob/ob*) causes obesity, but only mild and transient diabetes (Coleman and Hummel, 1973). In contrast, the same mutation in the BTBR genetic background (BTBR-*ob/ob*) causes severe type II diabetes (Stoehr et al., 2000). A (B6 × BTBR)F_2-cross was generated yielding 110 animals. Selective phenotyping (Jin et al., 2004) was employed to identify 60 F_2 *ob/ob* mice. For each of the 60 mice, pancreatic islets were isolated and 45,265 mRNA abundance traits were collected at 10 weeks of age using Affymetrix Gene Chips (MOE430A,B). The probe level data were processed using Robust Multi-array Average (RMA) to give a single, normalized, background-corrected summary score of expression for each transcript (Irizarry et al., 2003). Low abundance transcripts, defined as transcripts with average expression level below the tenth percentile, were removed leaving 40,738 traits. Genotypes for 145 markers were also obtained (over 90% of the animals provided genotype data at any given marker).

The TB-MR, MB-EB, MB-Q, Q-ALL, and MOM methods were each applied to the F_2 data. As in the simulation study, the location of the maximum LOD (TB-MR), maximum posterior probability of DE (MB-EB and MOM), or minimum q-value (MB-Q and Q-ALL) for each transcript was recorded. Mapping transcripts are defined as those for which the evidence in favor of linkage at the location of the maximum (minimum) exceeds the threshold (or is smaller than the threshold in the case of MB-Q and Q-ALL). For TB-MR, the threshold is 3.5 as determined by Dupuis and Siegmund (1999). To control the FDR at 5% with MB-Q or Q-ALL, q-values smaller than 0.05 are deemed significant (Storey and Tibshirani, 2003). For MB-EB and MOM, the threshold is chosen to control the FDR at 5% as described in Section 2.2.

The approaches named above identified 3689, 4083, 1913, 652, and 3039 transcripts, respectively, that map to at least one location. The most similarity was between MB-Q and Q-ALL with 92% of the Q-ALL transcripts also identified by MB-Q; MB-EB and MOM followed with 84% of the MOM transcripts also identified by MB-EB; the least similarity was between MB-EB and TB-MR with 23% of the TB-MR transcripts identified by MB-EB. A main reason for these differences is shown in Figure 2. The sample standard deviations of transcripts identified as DE by TB-MR and MB-Q are relatively small compared to those identified by the Bayes approaches, MB-EB and MOM. This is perhaps expected, considering the Bayes approaches share information across transcripts to estimate variance; Q-ALL, which uses transcript-specific p-values but considers the entire p-value distribution for assigning significance, falls between these extremes. Another reason for differences is the imposition of strict thresholds designed to control different error rates

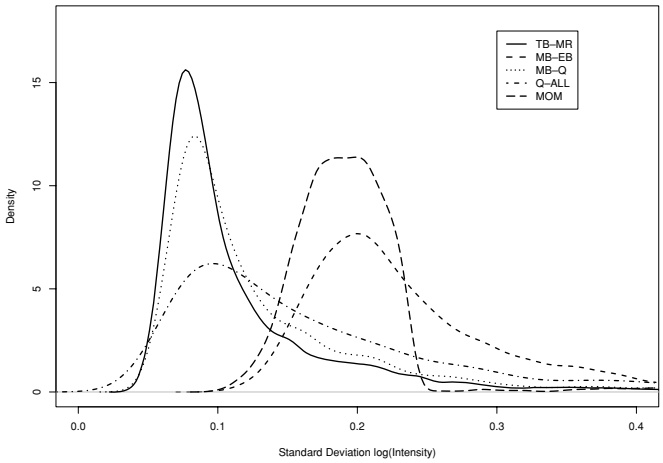

Figure 2. Sample standard deviations of transcripts identified as DE by each of the five methods. Sample means were very similar across methods (not shown).

across methods. When instead considering average evidence given by each approach, there is increased agreement among the methods in terms of genome regions identified.

Figure 3 identifies regions of enhanced linkage (hot spots) for each method, as measured by average evidence in favor of linkage (average is taken across all transcripts). The hot spot D2Mit241 is adjacent to D2Mit9, which has recently been

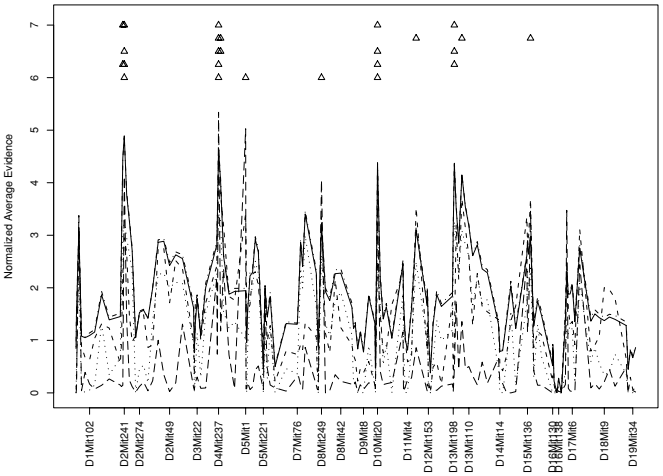

Figure 3. Evidence of linkage for each approach (LOD for TB-MR, posterior probability for MB-EB and MOM, and 1 − (q-value) for MB-Q and Q-ALL) averaged over transcripts and normalized by the sum of the evidence over all markers. The five markers with the strongest evidence of mapping transcripts are indicated by triangles for each method. Triangles represent (from top to bottom) TB-MR, MB-EB, MB-Q, Q-ALL, and MOM. D4Mit237 is among the top five markers for each method; D2Mit241 and D10Mit20 are identified by TB-MR, MB-Q, Q-ALL, and MOM. Note that although hot spot regions are identified in common across approaches, the lists of transcripts mapping to these regions are largely different.

identified as an obesity-modifier locus (Stoehr et al., 2004). Two additional regions identified by at least four of the five methods (on chromosomes 4 and 10) are not yet known to be involved in diabetes although we note that the region identified on chromosome 4 has been implicated in other analyses done in the Attie lab. The two regions identified by MOM alone on chromosomes 5 and 8 have been identified by other groups in earlier studies: D5Mit1 is a location known to affect triglyceride levels (Colinayo et al., 2003) and D8Mit249 is the marker on our map closest to the "fat" gene which is known to affect both diabetes and obesity (Naggert et al., 1995).

6. Discussion

With the advent of microarrays, it is now relevant to consider the QTL mapping problem with thousands of expression traits simultaneously. We have demonstrated that novel applications of existing methods for traditional QTL mapping or microarray studies do not perform well. In particular, a repeated application of standard QTL methods to each transcript results in inflated FDR. A similar inflation is observed if methods for identifying DE transcripts are repeatedly applied at every marker. Much of the inflated FDR results from not correcting for multiple tests across transcripts in the former case and across markers in the latter.

The Q-ALL approach, which combines tests across markers and transcripts simultaneously, is perhaps better justified. This approach is valid provided the p-values are weakly dependent (Storey et al., 2004); and Storey and Tibshirani (2003) hypothesize that weak dependence is the most likely form of dependence in genomewide studies such as the eQTL study of Brem et al. (2002). This hypothesis remains to be verified. We found that Q-ALL did not control the FDR at the target level in our simulations. This could be due to the fact that the simulation induces dependence that does not satisfy the assumption of weak dependence or that the p-values calculated from Student's t-test are not accurate. We find little evidence for the latter in our simulation set up. Further consideration of these issues is warranted.

To address the eQTL mapping problem, we propose a MOM model that shares information across markers and transcripts. The general method is flexible in specification of component densities and different forms will be appropriate for different data sets. Diagnostics such as those prescribed in Newton and Kendziorski (2003) should always be checked. To facilitate comparisons to Gaussian-based methods, we here considered component densities obtained from log-normal–normal hierarchical models. Simulations demonstrate that FDR is well controlled, without a sacrifice in power.

The conditions under which data are simulated are always questionable, and particularly so here as the methods compared vary considerably in underlying assumptions. To evaluate these approaches without biasing the results in favor of any one method, we have proposed a simulation framework that allows for evaluation of Bayesian-based methods that share information across units of interest (here, transcripts and markers) as well as those that do not. The framework is in no way designed to capture the many complexities of eQTL data, but it does provide some useful information regarding operating characteristics, and will serve as the basis for the development of more realistic simulation settings.

In addition to simulations, the methods were also compared based on results from a (B6 $\times$ BTBR)F_2 mouse cross in a study of diabetes. A number of differences were observed. Most notably, TB-MR and MB-Q identify traits with relatively small standard deviations. This type of behavior motivated the Bayes approach considered here, as information across transcripts can be shared to better estimate a transcript-specific variance and help prevent spurious identifications; other Bayesian approaches in the context of microarray studies are similarly motivated (Baldi and Long, 2001; Newton et al., 2001; Tusher, Tibshirani, and Chu, 2001; Lonnstedt and Speed, 2002; Kendziorski et al., 2003).

Figure 3 shows that in spite of these differences, there are regions of enhanced linkage identified in common among the approaches. These hot spot regions provide support for each approach to some extent and are of most interest to a biologist. The first region we considered is adjacent to one recently identified as an obesity-modifier locus (Stoehr et al., 2004). Two other identified regions are not yet known to be involved in diabetes, but are of particular interest considering they are identified by at least four of the five methods considered here. Of more interest to those evaluating these approaches are regions that are not identified in common across methods.

In particular, there are two regions identified by MOM alone. We have yet to confirm that these regions of enhanced linkage are real. However, we are encouraged by the results for two reasons. The first is that these regions have been identified by other groups in earlier studies: D5Mit1 is a marker linked to triglyceride levels (Colinayo et al., 2003) and D8Mit249 is the marker on our map closest to the "fat" gene, which is known to affect both diabetes and obesity (Naggert et al., 1995). The second reason is that there is good evidence that MOM may be better able to identify the types of transcripts mapping to hot spot regions (so called *trans* transcripts). Basically, transcripts can be labeled as *cis* or *trans*; *trans* transcripts are transcripts in which the expression is regulated by genes perhaps distant from the physical location of the transcript. They generally have higher variability compared to *cis* transcripts, transcripts that are self-regulated. The vast majority of traits mapping to a hot spot region are not physically located at the region; by definition, these traits are *trans* traits. As shown in Figure 2, MOM generally identifies transcripts with larger variability; most likely, these are *trans* traits. A close evaluation of these hot spot regions is underway.

In summary, eQTL mapping promises to be among the most statistically challenging problems involving microarray data; and the methods developed for the design and analysis of traditional QTL mapping or microarray studies will not directly apply. The question of selecting the most informative subjects to be phenotyped has been addressed (Jin et al., 2004), but most design and analysis questions for eQTL studies remain open. We have here considered a central problem in the analysis of eQTL data—that of identifying the collection of mapping transcripts and the genome locations to which they map. We have shown that novel applications of some existing methodologies do not fare well and have proposed an alternative approach, the MOM model. The MOM model should prove useful in improving the specificity of eQTL identifications. Specifically, by considering one full model for the

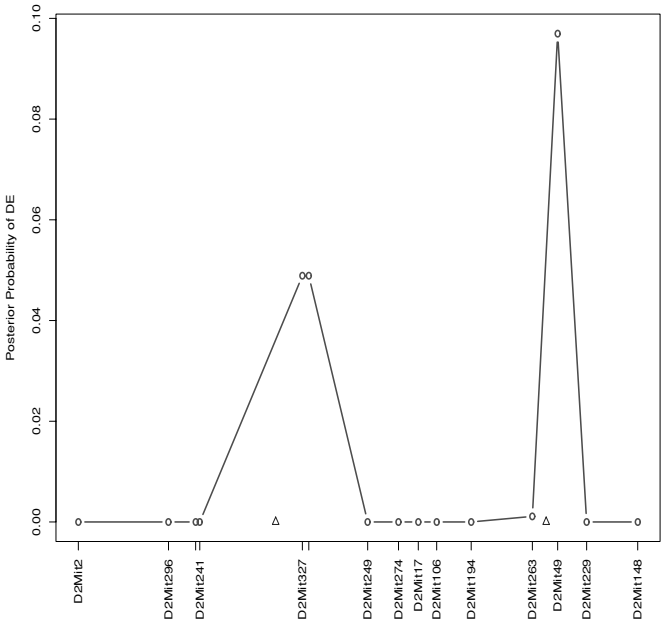

Figure 4. Simulation results from 5000 simulated transcripts; 1000 have expression levels determined by 2 QTL (triangles). QTL genotypes were defined by the marker genotypes at the QTL locations. These markers were removed from the analysis to simulate QTL in between markers; the QTLs are not interacting. Intermarker distance surrounding the first QTL is 22.3 cM with the QTL 16.5 cM from D2Mit241. There are 5.5 cM surrounding the second QTL, which is 3.0 cM from D2Mit263. The estimated proportion of DE transcripts is given on the y-axis. As shown, posterior probabilities of DE are highest at the markers nearest the QTL.

data, multiple tests across markers and transcripts are accounted for and FDR can be controlled without a sacrifice in power. Two regions identified by MOM alone are known to be involved in diabetes, providing further support for this approach. Additional validation studies are required.

The question of the best way to find multiple eQTL remains open. We here use HPD regions to identify the most likely locations to which mapping transcripts are linked. Figure 4 suggests that this approach may be useful for identifying multiple loci, even when the loci lie between markers. In some cases, markers closest to the loci will have the highest posterior probability of DE and, in this way, interesting regions will be identified using the MOM model. The precise conditions under which this is the case remain to be identified. Explicit consideration of a multiple loci model should certainly improve upon the MOM model, particularly when multiple eQTL are interacting. Interval mapping in the context of the MOM model should also prove useful, as identified genome regions are often large. Finally, a substantial benefit is expected by incorporation of sequence and other available information. In the context of the MOM model, information regarding the physical location of transcription factors could inform priors on the mixing proportions while functional categories could be used to more appropriately identify gene clusters, thereby improving model accuracy, power, and eQTL identification.

Acknowledgements

This work was supported in part by HHMI 133-ES29 to C.M.K. as well as NIDDK 58037 and NIDDK 66369 to A.D.A. The authors wish to thank Geoff MacLachlan, Michael Newton, Brian Yandell, and Ping Wang for useful discussions. We also thank Ping Wang for simulation studies of the MOM model, not shown here.

References

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17,** 509–519.

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

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Laboratory Animal* **30,** 44–52.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). *Journal of the Royal Statistical Society, Series B* **64,** 641–656 and 737–775 (discussion).

Carlin, B. and Louis, T. (1998). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall.

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

Coleman, D. L. and Hummel, K. P. (1973). The influence of genetic background on the expression of the obese (Ob) gene in the mouse. *Diabetologia* **9,** 287–293.

Colinayo, V. V., Qiao, J. H., Wang, X., Krass, K., Schadt, E., Lusis, A. J., and Drake, T. A. (2003). Genetic loci for diet-induced atherosclerotic lesions and plasma lipids in mice. *Mammalian Genome* **14,** 464–471.

Cox, N. J. (2004). An expression of interest. *Nature* **12,** 733–734.

Dudoit, S., Yang, Y. H., Speed, T. P., and Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12,** 111–139.

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151,** 373–386.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **1,** 70–86.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment.

Journal of the American Statistical Association **96,** 1151–1160.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264.

Jin, C., Lan, H., Attie, A. D., Bulutuglo, D., Churchill, G. A., and Yandell, B. S. (2004). Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics* **168,** 2285–2293.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22,** 3899–3914.

Lan, H., Rabaglia, M. E., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Zou, F., Yandell, B. S., and Attie, A. D. (2003a). Gene expression profiles of non-diabetic and diabetic obese mice suggest a role of hepatic lipogenic capacity in diabetes susceptibility. *Diabetes* **52,** 688–700.

Lan, H., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Yandell, B. S., and Attie, A. D. (2003b). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164,** 1607–1614.

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

Lonnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12,** 31–46.

Lund, M. S., Sorenson, P., Guldbrandtsen, B., and Sorensen, D. A. (2003). Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* **163,** 405–410.

Naggert, J. K., Fricker, L. D., Varlamov, O., Nishina, P. M., Rouille, Y., Steiner, D. F., Carroll, R. J., Paigen, B. J., and Leiter, E. H. (1995). Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nature Genetics* **10,** 135–142.

Newton, M. A. and Kendziorski, C. M. (2003). Parametric empirical Bayes methods for microarrays. In *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, E. S. Garrett, R. Irizarry, and S. L. Zeger (eds), 254–271. New York: Springer Verlag.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8,** 37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Parmigiani, G., Garrett, E. S., Irizarry, R., and Zeger, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer Verlag.

R Development Core Team. (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schadt, E., Monks, S., Drake, T. A., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302.

Stoehr, J. P., Nadler, S. T., Schueler, K. L., Rabaglia, M. E., Yandell, B. S., Metz, S. A., and Attie, A. D. (2000). Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. *Diabetes* **49,** 1946–1954.

Stoehr, J. P., Byers, J. E., Clee, S. M., Lan, H., Boronenkov, I. V., Schueler, K. L., Yandell, B. S., and Attie, A. D. (2004). Identification of major quantitative trait loci controlling body weight variation in ob/ob mice. *Diabetes* **53,** 245–249.

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

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA* **100,** 9440–9445.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* **66,** 187–205.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98,** 5116–5121.

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genetics* **35,** 57–64.

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L., and Friedman, J. M. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature* **372,** 425–431.

Received October 2004. *Revised June* 2005.
Accepted June 2005.